



Universidade do Minho
Escola de Engenharia

Rebeca Tatiana de Souto Santos

**Metabolomics-Based Approaches For
Food Authentication and Traceability**



Universidade do Minho

Escola de Engenharia

Rebeca Tatiana de Souto Santos

Metabolomics-Based Approaches For Food Authentication and Traceability

Doctoral Thesis

Doctoral Program in Chemical and Biological Engineering

Work developed under the supervision of:

Professor Miguel Rocha

Professor Marcelo Maraschin

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositoriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International

CC BY-NC-SA 4.0

<https://creativecommons.org/licenses/by-nc-sa/4.0/deed.en>

Acknowledgements

I would first like to thank European Social Fund under the scope of Norte2020 - Programa Operacional Regional do Norte for doctoral advanced training I was awarded (call NORTE-69-2015-15), without which this work would not be possible. I also thank my main host institution, the Centre of Biological Engineering at the University of Minho.

I also wish to express my gratitude to my supervisors, Professor Miguel Rocha and Professor Marcelo Maraschin. I thank them for offering me this opportunity and for all their support. I am immensely grateful for the guidance and feedback they provided throughout this project.

I would also like to thank Professor Eugénio Campos Ferreira for his availability, kindness, and opportunity to develop work in the BIOSYSTEMS group.

I also thank to the collaborations for the kindly supplied of the metabolomics data used in this dissertation, mainly the Enology Laboratory of INIAV - Polo de Inovação de Dois Portos, the University of Aveiro, Portugal, the federation “Federação das Associações de Apicultores e Meliponicultores de Santa Catarina” (FAASC) Brazil, the company “Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina”(EPAGRI) Brazil and the Federal University of Santa Catarina (UFSC) Brazil.

I would like to thank all of my colleagues from the BISBII research group for the friendly and collaborative working environment, and for their useful comments and suggestions, especially Sara Cardoso for her technical help with the R package specmine and Delora Baptista, Telma Afonso, Patrica Dias for their friendship till today.

A special thanks to Saulo Rodrigues for the words of support and friendship that started with this work and that will continue for a lifetime.

Finally, a special thanks to my close friends Cristina Costa, Manuela Tiago, Carlos Alberto, my beloved Parents and Brother and my dear boyfriend, Pedro Tiago Evangelista, without a doubt your support, encouragement, patience, kindness, friendship and love have been precious in last years. To you, I dedicate the commitment of these years.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the Universidade do Minho.

Abordagens Metabolómicas para Autenticação Alimentar e Rastreabilidade

Nos últimos anos, a procura do consumidor por produtos alimentares naturais autênticos aumentou consideravelmente, assim como a ocorrência de eventos de adulteração. Paralelamente, a conscientização sobre a responsabilidade de produtores e vendedores no combate à fraude alimentar tem vindo a crescer. O uso de ferramentas estatísticas (ST) e algoritmos de aprendizado de máquina (ML), em conjunto com abordagens metabolómicas, podem ajudar numa ampla gama de aplicações na área da autenticação alimentar. Os modelos de ML permitem identificar e entender quais os fatores com maior influência na autenticação alimentar, tais como as condições biológicas e ambientais. A Ressonância Magnética Nuclear (NMR) e metodologias metabolómicas espectrais são técnicas analíticas avançadas e abrangentes amplamente utilizadas na autenticação de alimentos. No entanto, existe a necessidade de construir modelos mais precisos de forma a prever a autenticidade dos produtos alimentares, garantindo a sua segurança e qualidade para o bem-estar dos consumidores e da economia.

Esta dissertação foca-se no uso da análise estatística multivariada (MSA) em dados metabolómicos de produtos alimentares naturais de alto valor, como o vinho e o mel, com recurso à aplicação de modelos de ML a vários problemas de classificação, incluindo a determinação de origens botânicas e geográficas. Modelos de ML foram construídos para discriminar e prever variáveis relacionadas com os problemas de autenticação dos alimentos estudados, incluindo a previsão de diferentes influências climáticas e a previsão da idade de armazenamento com base em análises de dados metabolómicos. Diferentes estratégias de pré-processamento, seleção de variáveis, técnicas de redução de dimensionalidade e métricas de avaliação foram testadas. Foi avaliado o impacto de diferentes métodos utilizados no desempenho desses modelos. Com base nessas análises, foram identificados os modelos com melhor desempenho preditivo. A interpretação dos modelos foi realizada comprovando que com o uso de modelos de ML é possível associar características químicas importantes com a classificação de alimentos naturais no processo de autenticação. ST e ML com tratamentos de pré-processamento adequados e métricas de avaliação são um pipeline adequado para diferentes problemas em autenticação de alimentos, como demonstrado neste trabalho. Apesar de algumas limitações, tais como, o tamanho dos conjuntos de dados dos nossos estudos, as nossas descobertas são úteis para orientar o desenvolvimento de novas abordagens usando a metabolómica para diversos problemas existentes em autenticação alimentar conjugando com ST e ML em amostras de vinho e mel. Para realizar análises de dados metabolómicos, diferentes abordagens MSA e de ML foram usadas, testadas e validadas usando o *package* da linguagem de programação *R specmine*, um software desenvolvido pelo grupo de investigação, e um conjunto de bibliotecas da linguagem de programação *Python*.

Palavras-chave: Autenticação de Vinho e Mel, Metabolómica, Ferramentas Estatísticas e Aprendizado Máquina.

Metabolomics-Based Approaches For Food Authentication and Traceability

In recent years, the consumer demand for authentic natural food products has increased concurrently with an expand of adulteration events. At the same time, the awareness regarding the responsibility of retailers and producers to combat food fraud has also grown.

The use of statistical tools (ST) and machine learning (ML) algorithms, in conjunction with high-throughput metabolomic-based approaches, can help in a wide range of applications in food authentication. ML models allow to identify and understand the most influential factors in food product authentication, such as biological and environmental conditions. Nuclear Magnetic Resonance (NMR) and other spectral metabolomic methodologies are advanced and comprehensive analytical techniques widely used in food authentication. Nonetheless, there is a need to construct more accurate models to predict the authenticity of food products to ensure their safety and quality for the well-being of the consumers and the economy.

This dissertation is focused on the use of multivariate statistical analysis (MSA) of metabolomics data from high-value natural food products, as wine and honey, addressing the application of ML models to several classification problems, including the botanical and geographical origins. ML models were also built to discriminate and predict variables related to food authentication problems studied, including the prediction of different climatic influences on production samples, and the prediction of the age of storage based on metabolomics data analyses.

Different preprocessing strategies, feature selection, dimensionality reduction techniques, and evaluation metrics were tested. The impact of different methods used on the performance of these models was evaluated. Based on these analyses, the models with better predictive performance were identified. Models interpretability was performed showing that by using ML models it is possible to associate important chemical features with food authentication classification. ST and ML with adequate preprocessing treatments and evaluation metrics are a suitable pipeline for different food authentication problems, as shown in this work.

Despite some limitations of the datasets sizes of our studies, our findings are useful to guide the development of new metabolomics based approaches for different food authentication issues using ST and ML algorithms in wine and honey samples. To perform metabolomic-data analyses, different MSA and ML approaches were used, tested, and validated using the open source *R* package *specmine*, a software developed by the host group, and a set of open source Python libraries.

Keywords: Wine and Honey Authentication, Metabolomics, Statistical and Machine learning tools.

Contents

List of Figures	xi
List of Tables	xvii
Acronyms	xxi
1 Introduction	1
1.1 Context and Motivation	1
1.2 Research Objectives	3
1.3 Outline	4
2 Metabolomics	6
2.1 Definitions and general workflow	6
2.2 Sample Preparation	9
2.3 Data Acquisition: analytical techniques	9
2.3.1 Mass Spectrometry	10
2.3.2 Nuclear Magnetic Resonance	10
2.3.3 Infrared Spectroscopy	10
2.3.4 Ultraviolet-visible Spectroscopy	11
2.3.5 Raman Spectroscopy	11
2.4 Data Preprocessing	11
2.4.1 Missing values and Outliers	12
2.4.2 Spectra Treatment	12
2.4.3 Mean-centering	14
2.4.4 Scaling methods	14
2.4.5 First and second-order derivatives	14
2.4.6 Data transformation	15
2.5 Data Analysis	15

2.5.1	Metabolite Annotation and Metabolite Identification	15
2.5.2	Exploratory Analysis	16
2.5.3	Univariate Analysis	16
2.5.4	Unsupervised Multivariate Analysis	18
2.5.5	Supervised Multivariate Analysis	19
2.5.6	Model Interpretability	22
2.5.7	Tools and databases	23
2.5.8	specmine and Webspecmine	24
2.5.9	Workflow and Data Analysis pipeline	25
3	Metabolomics for authentication in wine and honey food products	27
3.1	Food Authentication and Traceability	27
3.1.1	Food Control	28
3.1.2	Food Origin and Traceability	29
3.2	Wine	32
3.2.1	Metabolomics and Analytical techniques in Wine Authentication	33
3.2.2	Variety and Geographic Origin in Wine	33
3.2.3	Vintage and Age Determination	34
3.2.4	Adulteration	35
3.3	Honey	36
3.3.1	Honey Authenticity and its related issues	37
3.3.2	Classification for Floral Origin in Honey	37
3.3.3	Classification for Floral and Geographical Origin	38
3.4	Challenges, limitations and possible paths for further improvement.	39
3.5	Food Products studied in Data Analyses and their Authentications issues	39
4	Harvest, Age of bottle and Climatic factors prediction using statistical and machine learning approaches on metabolomics data of Brazilian <i>Cabernet Sauvignon</i> wines	41
4.1	Introduction	41
4.2	Results and Discussion	43
4.2.1	Metabolites annotation from ¹ H-NMR resonance peaks – qualitative and untargeted metabolomics analysis	43
4.2.2	Discrimination of the harvest year	48
4.2.3	Wine Age of Bottle Storage and Climatic factors influence	56
4.2.4	¹ H-NMR resonances with capability to discriminate wine samples by harvest, age of bottle storage and climatic factors	62
4.3	Conclusions and Future work	64
4.4	Materials and Methods	65

4.4.1	Samples	65
4.4.2	Age of Bottle Storage Data	65
4.4.3	Meteorological Data	66
4.4.4	Chemical Experimental Analysis	66
4.4.5	Statistical Analysis and Chemometrics	67
5	Multivariate statistical analysis and machine learning approaches to classify musts of Portuguese native grapevine varieties using FT-IR	70
5.1	Introduction	71
5.2	Results	73
5.2.1	Exploratory analysis	73
5.2.2	Statistical analysis	74
5.2.3	Supervised Data Analysis: Musts of grapevine variety classification	77
5.2.4	Variety classification	77
5.2.5	Features analysis	80
5.2.6	Features analysis of the best models performance for each Red Variety	81
5.2.7	Comparative analysis between Statistical, Machine Learning approach and Shap- ley values results	86
5.3	Discussion	88
5.4	Conclusions	90
5.5	Materials and Methods	90
5.5.1	Samples	90
5.5.2	Methods	91
5.5.3	Statistical Analysis and Chemometrics	91
6	Machine learning approaches to classify honeys based on flora and geographic origin from Santa Catarina, Brazil using ¹H-NMR	94
6.1	Introduction	95
6.2	Results	98
6.2.1	Exploratory analysis	100
6.2.2	Statistical analysis	101
6.2.3	Multivariate data analysis	101
6.2.4	Geographic Region and Flora classification	101
6.2.5	Features analysis of models classification	106
6.2.6	Comparative analysis between the results of Geographic and flora classifications	113
6.3	Discussion	115
6.4	Conclusions	116
6.5	Materials and Methods	117

6.5.1	Samples	117
6.5.2	Methods	118
6.5.3	Statistical Analysis and Chemometrics	118
7	Conclusions and Future Work	121
7.1	Main contributions	121
7.2	Future work	123
	Bibliography	125
	Appendices	
A	Supplementary Figures	144
A.1	Supplementary Figures for Chapter 5	144
A.2	Supplementary Figures for Chapter 6	148
B	Supplementary Tables	153
B.1	Supplementary Tables for Chapter 5	153
B.2	Supplementary Tables for Chapter 6	156

List of Figures

1	Percentage of publications distributed over different analytical techniques applied to food authentication until 2016. Adapted from [10].	6
2	More common Metabolomics Purposes and based Approaches. Adapted from [9].	8
3	Standard workflow on metabolomics food authentication. Adapted from [9].	9
4	Graphical interface of webspecime front-end.	25
5	General workflow used for the data analysis pipeline.	25
6	Used Workflow on this work and the correspondent Python libraries.	26
7	Top 10 countries on food authentication publications. Adapted from [10].	28
8	Official GIs logos of the European Union (PDO, PGI and TSG) according to Commission Implementing Regulation (EU) No. 668/2014. Adapted from [10].	30
9	Collective number of PDO, PGI and TSG products per EU country. Adapted from [10].	30
10	A: ¹H-NMR spectra plot for all wine samples with distinct harvest years. B: Zoom of ¹H-NMR spectra showing the amino acids region interval [2.25 - 2.00 ppm]. Wine samples were coded in color according to their harvests as follows: 1992 (black), 1994 (red), 1996 (green), 1999 (dark blue), 2002 (light blue), and 2005 (pink).	48
11	Total of amino acids concentration of <i>Cabernet Sauvignon</i> wines (BCS) according to their harvests.	49
12	Amino acids concentration (ml/L) of <i>Cabernet Sauvignon</i> wines according to their harvests.	50
13	A: PCA 2D scores plot (PC1, PC2) from ¹H-NMR data. B: PCA plot of the first five PCs from ¹H-NMR data. The wine samples are presented in different colors corresponding to the harvests groups: 1992 (dark-green), 1994 (orange), 1996 (purple), 1999 (pink), 2002 (light-green), and 2005 (yellow).	53

14	A: PCA 2D scores plot (PC1, PC2) from amino acids concentrations. B: PCA plot of the first five PCs from amino acids concentrations. The wine samples are presented in different colors corresponding to the harvests groups: 1992 (dark-green), 1994 (orange), 1996 (purple), 1999 (pink), 2002 (light-green), and 2005 (yellow).	54
15	A: Dendrogram plot of the hierarchical clustering, with Euclidean distance between wine samples from $^1\text{H-NMR}$ data. B: Dendrogram plot of the hierarchical clustering, with Euclidean distance between wine samples from amino acids concentrations. The wine samples are presented in different colors corresponding to the harvests groups: 1992 (black), 1994 (red), 1996 (green), 1999 (blue), 2002 (light-blue), and 2005 (pink).	54
16	Plots for the linear regression models showing the influence of age of bottle storage on $^1\text{H-NMR}$ resonances and one amino acid with highest adjusted R^2. A: 3.74 ppm resonance. B: 4.28 ppm resonance. C: Pro amino acid concentration.	57
17	Plot for the linear regression model showing climatic influences on 3.60 ppm resonance with a highest adjusted R^2. The wine samples are presented in different colors corresponding to the harvests groups: 1992 (green), 1994 (yellow), 1996 (brown), 1999 (light-blue), 2002 (red), and 2005 (dark-blue).	60
18	Plots for the linear regressions models showing climatic influences on $^1\text{H-NMR}$ resonances with highest adjusted R^2. A: 2.05 ppm resonance. B: 2.29 ppm resonance. The wine samples are presented in different colors corresponding to the harvests groups: 1992 (green), 1994 (yellow), 1996 (brown), 1999 (light-blue), 2002 (red), and 2005 (dark-blue).	61
19	$^1\text{H-NMR}$ resonances with best predictive capability to distinguish wine samples by harvest (colored green), age of bottle storage (colored blue), and climatic factors (colored yellow) influences.	64
20	FT-IR spectra plot for wine must samples grape varieties without water region ($1450\text{-}1800\text{cm}^{-1}$). A: Plot of FT-IR spectra of red musts of grapevine varieties samples. B: Plot of FT-IR spectra of white musts of grapevine varieties samples.	74
21	Top 20 most important features ranked by mean absolute SHAP values. The features are ordered according to their importance of the best model classification for each red wine must variety. The Mean Shap values and their average impact on model output are shown for each feature. A: SHAP values and the average impact of feature's on model of PLS-DA to classify Syrah. B: SHAP values and the average impact of feature's on model of LR to classify Touriga Nacional. C: SHAP values and the average impact of feature's on model of SVC to classify Aragonez.	84

22	The 20 most important features are shown and ordered according to their importance to the best model classification for each red wine must variety. The Shap values and their impact on model output are shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of PLS-DA to classify Syrah. B: SHAP values and impact of feature's on model of LR to classify Touriga Nacional. C: SHAP values and the impact of feature's on model of SVC to classify Aragonez.	84
23	Top 17 most important features ranked by mean absolute SHAP values. The features are ordered according to their importance of the best model classification for each white wine must variety. The Mean Shap values and their average impact on model output are shown for each feature. A. SHAP values and the average impact of feature's on model of PLS-DA to classify Arinto. B. SHAP values and the average impact of feature's on model of PLS-DA to classify Viosinho.	86
24	The 17 most important features are shown and ordered according to their importance to the best model classification for each white wine must variety. The Shap values and their impact on model output are shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A. SHAP values and impact of feature's on model of PLS-DA to classify Arinto. B. SHAP values and impact of feature's on model of PLS-DA to classify Viosinho.	86
25	Synopsis of the results with higher performance and the data analysis on FT-IR spectrum. The squares present the classification model and the metric score of best model performance to each wine must variety.	89
26	The eleven agroecological regions of Santa Catarina Estate and the most common flora honeys. Adapted from https://ciram.epagri.sc.gov.br/apicultura/monitor-colmeias.jsp	98
27	¹H-NMR spectra spectra plot for honey samples based on region production(A) and flora origin (B) at interval of 3.0ppm-5.5ppm. C: Zoom plot of ¹ H-NMR spectra of all honey samples at interval of 7.7ppm-8.6ppm. D: Zoom plot of ¹ H-NMR spectra of all honey samples at interval of 7.6ppm-7.1ppm. E: Zoom plot of ¹ H-NMR spectra of all honey samples at interval of 1.1ppm-2.3ppm.	100
28	Top 20 most important features ranked by mean absolute SHAP values. The features are ordered according to their importance of the best model classification for each production region. The Mean Shap values and their average impact on model output is shown for each feature. A: SHAP values and the average impact of feature's on model of LR to classify 1A. B: SHAP values and the average impact of feature's on model of LR to classify 1B.	109

29	The 20 most important features are shown and ordered according to their importance to the best model classification for each production region. The Shap values and their impact on model output is shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of LR to classify 1A. B: SHAP values and impact of feature's on model of LR to classify 1B.	110
30	Top 20 most important features ranked by mean absolute SHAP values. The features are ordered according to their importance of the best model classification for each each production region. The Mean Shap values and their average impact on model output is shown for each feature. A: SHAP values and the average impact of feature's on model of PLS-DA to classify 2C. B: SHAP values and the average impact of feature's on model of RF to classify 3B. C: SHAP values and the average impact of feature's on model of LR to classify 3C.	110
31	The 20 most important features are shown and ordered according to their importance to the best model classification for each production region. The Shap values and their impact on model output is shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of PLS-DA to classify 2C. B: SHAP values and impact of feature's on model of RF to classify 3B. C: SHAP values and the impact of feature's on model of LR to classify 3C.	111
32	Top 20 most important features ranked by mean absolute SHAP values. The features are ordered according to their importance of the best model classification for each each production region. The Mean Shap values and their average impact on model output is shown for each feature. A: SHAP values and the average impact of feature's on model of SVC to classify 4A. B: SHAP values and the average impact of feature's on model of PLS-DA to classify 4B. C: SHAP values and the average impact of feature's on model of SVC to classify 5.	111
33	The 20 most important features are shown and ordered according to their importance to the best model classification for each production region. The Shap values and their impact on model output is shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of SVC to classify 4A. B: SHAP values and impact of feature's on model of PLS-DA to classify 4B. C: SHAP values and the impact of feature's on model of SVC to classify 5.	112

34	Top 20 most important features ranked by mean absolute SHAP values. The features are ordered according to their importance of the best model classification for flora origin. The Mean Shap values and their average impact on model output is shown for each feature. A: SHAP values and the average impact of feature's on model of PLS-DA to classify "Uva do Japão". B: SHAP values and the average impact of feature's on model of SVC to classify Other. C: SHAP values and the average impact of feature's on model of SVC to classify "Silvestre".	113
35	The 20 most important features are shown and ordered according to their importance to the best model classification for flora origin. The Shap values and their impact on model output is shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of PLS-DA to classify "Uva do Japão". B: SHAP values and impact of feature's on model of SVC to classify Other. C: SHAP values and the impact of feature's on model of RF to classify "Silvestre".	113
36	Synopsis of the results whit higher performance and the data analysis on ¹H-NMR spectra for the regions honeys productions and flora. The squares present the classification model and the metric score of the best model performance to flora origins. And the colored lines represent the links between flora and production regions.	115
37	Principal Components Analysis (PCA) 2D scores plot (PC1, PC2) from FT-IR spectra of red wine musts samples of varieties Aragonez, Syrah and Touriga Nacional (Touriga Nacional). Explain variance of Principal Componets 1 and 2:PC1 = 69.8% PC2 = 21.6%.	144
38	t-distributed Stochastic Neighbor Embedding (t-SNE) from FT-IR spectra of red wine musts samples of varieties Aragonez, Syrah and Touriga Nacional (Touriga Nacional)	145
39	Dendrogram plot of the hierarchical clustering analysis (HCA), with Euclidean distance between samples, from FT-IR spectra of red wine musts samples of varieties Aragonez, Syrah and Touriga Nacional (Touriga Nacional).	145
40	Principal Components Analysis (PCA) 2D scores plot (PC1, PC2) from FT-IR spectra of white wine musts samples of varieties Arinto and Viosinho. Explain variance of Principal Components 1 and 2: PC1 = 43.4% PC2 = 30.4%.	146
41	t-distributed Stochastic Neighbor Embedding (t-SNE) from FT-IR spectra of white wine musts samples of varieties Arinto and Viosinho.	146
42	Dendrogram plot of the hierarchical clustering analysis (HCA), with Euclidean distance between samples, from FT-IR spectra of white wine musts samples of varieties Arinto and Viosinho.	147

43	Principal Components Analysis (PCA) 2D scores plot (PC1, PC2) from ¹H-NMR spectrum of honey samples of all SC Regions. Explain variance of Principal Components 1 and 2: PC1 = 52%; PC2 = 29%; PC3 = 0.5%.	148
44	t-distributed Stochastic Neighbor Embedding (t-SNE) from ¹H-NMR spectrum of honey samples of all SC Regions.	149
45	Dendrogram plot of the hierarchical clustering analysis (HCA), with Euclidean distance between samples, from ¹H-NMR spectrum of honey samples of all SC Regions.	149
46	Principal Components Analysis (PCA) 2D scores plot (PC1, PC2) from ¹H-NMR spectrum of honey samples of Flora origins. Explain variance of Principal Components 1 and 2: PC1 = 52%; PC2 = 29%; PC3 = 0.5%.	150
47	t-distributed Stochastic Neighbor Embedding (t-SNE) from ¹H-NMR spectrum of honey samples of Flora origins.	151
48	Dendrogram plot of the hierarchical clustering analysis (HCA), with Euclidean distance between samples, from ¹H-NMR spectrum of honey samples of Flora origins.	152

List of Tables

1	Metabolite Annotation and Metabolite Identifications tools.	16
2	Available free tools for metabolomics and spectral data.	23
3	Metabolomics Approaches on Food Authentication.	32
4	Metabolomics Approaches for Wine Authentication and Traceability.	36
5	Metabolomics Approaches for Honey Authentication and Traceability.	38
6	Metabolites and their ¹H chemical shifts annotated by 500 MHz ¹H-NMR with higher scores for frequency, organism and solvent.	44
7	ANOVA results - peaks with the lowest corrected p-values (< 0.01). The first column shows the peaks, the second one the respective p-value, the third column includes the corrected p-value with False Discovery Rate (FDR) method, and the final column the result of the Tukey's HSD test, which consists of the pairs of harvest years that were significantly different in terms of means for each peak.	51
8	ANOVA results - amino acids with the lowest corrected p-values (< 0.05). The first column contains the considered amino acids, the second one the respective p-value, the third column includes the corrected p-value with False Discovery Rate (FDR) method, and the final column the result of Tukey's HSD test, which consists of the pair of harvest years that were significantly different in terms of means for each amino acid concentration.	52
9	ML models results for the ¹H-NMR data and the amino acids concentration with cross-validation accuracy. The first column contains the considered input data as features of the ML model, the second one the applied ML model, the third one the result with the accuracy of the correspondent ML model computed using cross-validation, and the last column the features (¹ H-NMR resonances - ppm and amino acids) most relevant, with those marked with bold type indicating those shared with ANOVA results.	55
10	Climatic Data by each harvest year	58

11	Results for linear regression analysis of the climatic influence over $^1\text{H-NMR}$ data. The first column contains the considered peaks, the second column presents the influence type, followed by the respective corrected p-value ($\leq 0,0001$), adjusted R^2 ($> 50\%$), and coefficient of the LR model. Bold type indicates $^1\text{H-NMR}$ resonances with highest adjusted R^2	58
12	Regions of $^1\text{H-NMR}$ resonances (ppm) with predictive capability to distinguish samples selected by ANOVA and ML models.	63
13	ANOVA results. Wavenumber (cm^{-1}) with the lowest corrected p, values (< 0.01). The first column shows the t-test, which consists of the pairs of grapevine musts varieties that were significantly different in terms of means for each feature, the second one the respective spectroscopic bands, the third column includes the chemical groups related to the wavenumbers identified.	76
14	Performance scores of the best classification models for the grapevine musts varieties studied. All models had pre-process methods(SNV plus Scaling) with the application of Feature Selection (Drop of constant features with 90% of correlation). The MCC evaluation metric followed a threshold based on the calculated PR Threshold mean. Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC:Matthews Correlation Coefficient; BA: Balanced Accuracy; PR Threshold: Precision-Recall mean Threshold; SD: Standard Deviation.	80
15	Comparative analysis between Statistical, Machine Learning approach and Shapley values results. Note: Arg: Aragonez; TN: Touriga Nacional; SY: Syrah; LR: Logistic Regression; SVC: Support Vector Classification; PLS-DA: Partial Least Squares-Discriminant Analysis; (-): Negative impact on the model output; (+): Positive impact on the model output.	87
16	Hyper-parameters used by the distinct models.	93
17	Geographical information of the honey samples.	99
18	Flora information of the honey samples. Uva mix samples are multiflora honeys with presence of "Uva do Japão" flora. Other honey samples are multiflora honeys with different types of flora as cinnamon, fruit trees, native trees, "Canudo de Pito" tree, "Bracatinga" tree, and other plants from each region.	99
19	Feature with highest variance and the correspondent correlated features group. First column presents the feature with the higher variance correlated to a set of peaks given a predefined threshold. These features are the selected to the new sub-dataset used at all models classification. And these features were used by Shapley method to interpret their impact on each model classification. Second column are the correspondent correlated features to each feature. Note: * are features with high variance and with lower threshold correlation (0.90) with other features.	103

20	Performance scores for best models classification for all Regions. All models had pre-process methods with the application of Feature Selection (Drop of constant features with 90% of correlation). The MCC evaluation metric followed a threshold based on the calculated ROC Threshold mean. Rows with background color gray are the regions with higher classification models. Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; ROC Threshold: Receiver Operating Characteristic. .	104
21	Performance scores for best models classification for all Flora honeys. All models had pre-process methods with the application of Feature Selection (Drop of constant features with 90% of correlation). The MCC evaluation metric followed a threshold based on the calculated ROC Threshold mean. Rows with background color gray are the regions with higher classification models. Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; ROC Threshold: Receiver Operating Characteristic. .	105
22	Best seven Mean absolute SHAP values and their impact on model output of the best models classification for Geographic Regions. Note: (-): Negative impact on the model output; (+): Positive impact on the model output.	109
23	Best seven Mean absolute SHAP values and their impact on model output of the best models classification for Flora groups. Note: (-): Negative impact on the model output; (+): Positive impact on the model output.	112
24	Hyper-parameters used by the distinct models.	120
25	Feature with highest variance and the correspondent correlated features group for Red varieties. First column presents the feature with the higher variance correlated to a set of FT-IR bands given a predefined threshold. This features are the selected to the new sub-dataset used at all red models classification. And these features were used by Shapley method to interpret their impact on each wine grape must variety model classification. Second column are the correspondent correlated features to each feature. Note: 1015*, 1123*, 1127*, 1130* are the features computed by Shapley value with zero correlations to other features, in our work designated as unique features. The wasenumbers at 930**, 980**, 1080** and 2639** were features computed by Shapley values for red and white varieties. . . .	153

26	Feature with highest variance and the correspondent correlated features group for White varieties. First column presents the feature with the higher variance correlated to a set of FT-IR bands given a predefined threshold. This features are the selected to the new sub-dataset used at all white models classification. And these features were used by Shapley method to interpret their impact on each wine grape must variety model classification. Second column are the correspondent correlated features to each feature. Note: 2924* is the only feature computed by Shapley value with zero correlations to other features, in our work designated as unique features. The wasennumbers at 930**, 980**, 1080**, 2639** were features computed by Shapley values for red and white varieties.	154
27	Supplemental Table: Performance scores for all models classification for all wine musts varieties. All models had pre-process methods (SNV plus Scaling) with the application of Feature Selection (Drop of constant features with 90% of correlation). Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; PR Threshold: Precision-Recall Threshold; SD: Standard Deviation.	155
28	Supplemental Table: Performance scores for all models classification for all Regions. All models had pre-process methods with the application of Feature Selection (Drop of constant features with 90% of correlation). Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; PR Threshold: Precision-Recall Threshold; SD: Standard Deviation.	156
29	Supplemental Table: Performance scores for all models classification for all Flora honeys. All models had pre-process methods with the application of Feature Selection (Drop of constant features with 90% of correlation). Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; PR Threshold: Precision-Recall Threshold; SD: Standard Deviation.	157

Acronyms

¹H-NMR Hydrogen Proton Nuclear Magnetic Resonance

1D-NMR one-dimensional NMR spectrum

2D-NMR two-dimensional NMR spectrum

Ala Alanine

ANN Artificial Neural Network

ANOVA Analysis of Variance

BA Balanced Accuracy

CE Capillary Electrophoresis

CS Cabernet Sauvignon wines

Cys Cysteine

D₂O deuterium oxide

D-UPLS least-squares and discriminant unfolded partial least-squares

dGMP 2-deoxyguanosine monophosphate

EU European Union

FIR Far-Infrared

FSA Food Standards Agency

FT-ICR Fourier transform ion cyclotron resonance

FT-IR Fourier-transform infrared spectroscopy

GBC Gradient Boosting Classifier

GC Gas Chromatography

Glu Glutamic acid

Gly Glycine

HCA Hierarchical Clustering Analysis

Hpro Hydroxyproline

IM Ion Mobility

IR Infrared Spectroscopy

Iso Isoleucine

KNN K-Nearest Neighbors

LC Liquid Chromatography

LDA Linear Discriminant Analysis

Leu Leucine

LnR Linear Regression model

LR Logistic Regression model

Lys Lysine

MCC Matthews Correlation Coefficient

MIR Mid-Infrared

ML machine learning

MRL maximum residue limits

MS Mass Spectrometry

MSA Multivariate Statistical Analysis

MSC Multiplicative Scatter Correction

NIR Near-Infrared

NMR Nuclear Magnetic Resonance

OIV International Organisation of Vine and Wine

OPLS-DA Orthogonal Projections to Latent Structures Discriminant Analysis

PCA-DA Principal Component Analysis-Discriminant Analysis

PCA Principal Component Analysis

PDO Protected Designations of Origin

PGI Protected Geographical Indication

Phe Phenylalanine

PLS-DA Partial Squares-Discriminant Analysis

PLS-r Partial Least Squares Regression

PR Threshold Precision-Recall threshold

PR-AUC Area Under the Precision-Recall Curve

Pro Proline

RF Random Forest Classifier

RMSE Root Mean Square Error

ROC-AUC Area Under the Receiver Operating Characteristic curve

SD Standard Deviation

Ser Serine

SHAP Shapley Additive exPlanations

SIMCA Independent Modelling by Class Analogy

SNV Standard Normal Variate

SPME-GC-MS Solid-Phase Microextraction followed by Gas Chromatography-Mass Spectrometry

ST statistical tools

SVC Support Vector Classifier

SVM Support Vector Machines

t-SNE t-distributed Stochastic Neighbor Embedding

TGS Traditional Specialty Guaranteed

Thr Threonine

TSP sodium 3-(trimethylsilyl)-propionate-2,2,3,3- δ^4

Tyr Tyrosine

UK United Kingdom

UPLC Ultra-Performance Liquid Chromatography

UV-vis Ultraviolet–visible spectroscopy

Introduction

1.1 Context and Motivation

Nowadays, the agriculture industry is the second largest industrial sector in the European Union. Globalization has led to the demand for the acquisition of food products independently of the season, the location, and the availability. This context has been leading to a growth of production, but also creating opportunities for frauds, adulterations and contaminations events [1].

The concerns with agricultural authenticity and traceability are ancient and, since the control of food supplies on World War II, new efforts were made to ensure “food security” [2]. Proper regulation has been made and applied to food processes and distribution. However, several events of food contamination, adulteration and fraud continue to occur [1]. More recently, the horse meat scandal forced the UK government to create new guidelines and strategies applied to food systems to deal with these crimes [1].

The rapid development of metabolomics-based approaches combined with the multivariate-data analysis tools, have shown potential to face the challenges in food authentication and traceability [3].

Metabolites are a diverse group of small molecules, including lipids, amino acids, peptides, nucleic acids, organic acids, vitamins, thiols and carbohydrates, which represent the results and/or products of the cell processes [4]. These metabolites are modulated by genetic and environmental factors, which promote variations on the metabolite levels, as a result of biological systems responses [5]. The metabolome provides a comprehensive view of the cells function, and the identification and quantification of these compounds, allows the recognition of a correct metabolomics profile.

The emerging field of metabolomics, lead to robust, efficient, sensitive, and cost-effective analytical methodologies, and also resulted in a large amounts of data. To cover the needs of data processing, it is necessary to combine the metabolomics analytical techniques with the multivariate statistical analysis tools for data extraction and interpretation [6], [7].

Different computational tools regarding spectral data analysis are now available, for specific purposes or specific tasks, covering a wide range of experimental techniques. As examples, the platforms *ChemoSpec* and *hyperSpec* handle spectral data from Raman, Infrared Spectroscopy (IR) and Ultraviolet–visible (UV-vis) spectroscopy techniques. More recently, another free R package was developed in the host group,

named *specmine* [8]. This framework allows metabolomics and spectral data analysis, addressing the development of customizable data analysis pipelines, covering different types of metabolomics and spectral data, such as IR, UV-vis, Nuclear Magnetic Resonance (NMR), Mass Spectrometry (MS) and Raman spectroscopies. All of these tools were developed to run over the R scientific computing platform.

Metabolomics-based approaches for food authentication and traceability are currently hindered by several limitations. These include the limited availability of freely accessible repositories, the generation of large amounts of data that require appropriate pre-processing and analysis methods, and the lack of standardization in analytical techniques and chemometric approaches [9].

In the context of this dissertation, related studies presented reduced dataset sizes and limited availability of statistical and machine learning models, which further compromises the reliability and reproducibility of the performed tests and studies.

Therefore, it is crucial to address these limitations to enhance the validity and generalizability of metabolomics-based approaches for food authentication and traceability. Specifically, efforts should be focused on increasing the availability of high-quality data repositories, developing more standardized workflows for data analysis and chemometric approaches, and increasing the number and diversity of available statistical and machine learning models.

This dissertation stands out by employing a model interpretability approach, which is not commonly used in other food authentication works. This approach enabled a better understanding of the models used in this study and helped identify the most influential features.

In addition to model interpretability, this study utilized chemometrics tools such as preprocessing methods, multivariate statistical analysis (MSA), and several machine learning (ML) approaches to analyze metabolomics data from two high-value food products (honey and wine). These tools aided in chemical identification, standardization, classification, and prediction of food product variables for different authenticity issues.

Furthermore, appropriate feature selection methods and metrics tools were employed based on the studied datasets, ensuring the accuracy of the models. Finally, state-of-the-art methods for food authentication were explored using metabolomics-based approaches combined with statistical and machine learning models.

The metabolomics data utilized in this work are from the following collaborations: (i) primarily from the Federal University of Santa Catarina (UFSC), Santa Catarina Brazil, concerning the Wine and Honey samples, and (ii) from the Instituto Nacional de Investigação Agrária e Veterinária (INIAV), Dois Polos, Portugal, regarding the wine musts samples. The limited sample size of the metabolomics datasets poses certain difficulties; however, it is noteworthy that these samples were obtained as part of an initial collaboration between the University and producers, considering the quality of products offered by beekeepers and their representation in the market. These works represent pilot studies that will be expanded in subsequent years. Consequently, the investigations presented in this dissertation are the first to focus on honey samples from Santa Catarina Estate and also the first to employ wine musts samples from Portuguese grape varieties grown in PDO regions of Portugal.

Overall, this study provides a robust framework for analyzing metabolomics data with applications to food authentication studies, and highlights the importance of model interpretability in such analyses.

1.2 Research Objectives

In light of the context described above, the primary objective of this research was to contribute to the development of more effective methods for data preprocessing and analysis in cases related to authentication issues, particularly in classification problems concerning food origin. To achieve this objective, the R package *specmine* and Python libraries were employed to analyze untargeted metabolomics-data from two high-value food products, namely Wine, Wine musts, and Honeys.

The research focused on the exploration, integration, and improvement of data analysis from different experimental techniques, with an emphasis on IR and $^1\text{H-NMR}$ spectroscopies. However, it should be noted that the research did not specifically address traceability issues.

Overall, this study sought to advance the field of data analysis for food authentication by enhancing the methods of data preprocessing and analysis, leveraging the power of multiple experimental techniques to achieve more accurate classification of food origin.

This work encompasses the following specific objectives:

- To review the state-of-the-art in metabolomics-based data approaches for authenticity and traceability, focusing on the food products of interest;
- Use experimental untargeted metabolomics data for the selected food products (wine, wine must and honey) using spectroscopic analytical techniques ($^1\text{H-NMR}$ and FT-IR);
- Implement pipelines for metabolomics data analysis using the functionalities of the R package *specmine* and Python libraries, namely regarding preprocessing, multivariate statistical analysis tools, machine learning, feature selection methods and model interpretability;
- Each metabolomics data set utilized in this work, has an emphasis on specific food authentication issues, namely:
 - To understand the effect of harvest, the age of bottle storage, and climatic influences on Brazilian *Cabernet Sauvignon* wines.
 - To classify Portuguese grape must varieties using FT-IR using multivariate statistical analysis and machine learning approaches.
 - To classify honeys based on botanical and geographic origin from Santa Catarina, Brazil using $^1\text{H-NMR}$ combine with machine learning approaches and model interpretability.
 - The main work is briefly explained in the outline.

- To present the results on scientific publications and to standardize the selected approaches for authentication processes on the food products selected.

1.3 Outline

This manuscript is structured as follows:

In Chapter 1 (the present chapter), the research problem is identified and the main research objectives are set forth. At the end of the chapter, a brief overview of the structure of the manuscript is provided.

Chapter 2 provides the theoretical background for this work. The first part of the chapter introduces some brief definitions of Metabolomics and the general workflow needed to develop the most appropriate metabolomics approaches on food authenticity. In the second part, presents ST and ML, from the basic concepts underlying these algorithms to common problems and the strategies, including the preprocessing data, the data analysis with the statistical tools, and the most common models used in unsupervised and supervised analyzes of food authentication. In addition, model interpretability concept is explained, and the methodology behind is described, a new approach that it will be used on this dissertation to food authentication problems.

Chapter 3 Provides the context of the Food authentication and traceability origin-specific. A brief explanation of some concepts and definitions concerning food Authentication, Traceability, Food Control, Food Origin (biological and geographical) is provided. Additionally, is presented the food products honey and wine, and the most commons approaches of metabolomics and analytical techniques applied for their authentication. Finally, a critical overview of recently published studies that have employed Statistical analyses and Machine Learning methods to perform food authentication in the food products, honey and wine. It was also address the limitations of these approaches and provide a discussion on possible paths for further improvement.

Chapter 4 presents a study that contributes to understand the influences of the age of bottle storage and of climatic factors on the wines' chemical composition. The study uses a multidimensional metabolomics data set of ^1H -NMR spectra and metabolites concentrations from 30 Brazilian *Cabernet Sauvignon* wine samples, with distinct harvest years. Metabolomics based-approaches were combined with ST and ML to discriminate wine samples according to their harvest years. Qualitative analysis revealed that wines with different harvest years presented similar composition of metabolites, using metabolites annotation software developed by the host-group. In the end of the work, Linear Regression (LnR) models, allowed to discriminate ^1H -NMR resonances that were most influenced by the age of bottle storage and the climatic factors, being the temperature the factor with most influence on this case study.

Chapter 5 contemplates the study of Portuguese native wine grapes discrimination based on varieties. It is an exploratory study aiming to assess the capability of common high-speed analytical devices present in the wine industry, such as Fourier-transform infrared spectroscopy (FT-IR), used to control wine must maturation (grape juices before fermentation), to classify different Portuguese native grape musts varieties.

Another aim is to describe which specific spectral bands allow discrimination of a given variety and how they influence on the model output. Due to the complexity of the generated signals, specific spectral and data preprocessing treatments were needed and were tested different methods. It was also performed five supervised models (Random Forest (RF), Gradient Boosting Classifiers (GBC), and Partial Squares-Discriminant Analysis (PLS-DA), Support Vector Classifiers (SVC), Logistic Regression (LR)) to each variety classification and different metrics evaluation were tested and analyzed. Finally, features influencing the models output were identified using Shapley values, and these features were linked to the correspondent functional chemical groups. The work demonstrated that conjugating statistical tools, machine learning approaches and using model interpretability was possible to maximize the mean Area Under the Precision-Recall Curve (PR-AUC) of wine must variety classification and to better understand the biological features with importance for the variety classification. This chapter provides a description of the different ML models and evaluation metrics that were tested, and the correspondent analysis.

Chapter 6 presents the exploratory study to classify Santa Catarina (SC) honeys according to their region of production (from the eleven agroecological regions) and flora group origin (five flora honeys, “Uva do Japão”, “Silvestre”, Eucalyptus, “Uva mix” and Other (representing several multifloras)). Another aim is to describe which specific features allow each classification of a given region and flora group and how they influence the model output. The approach encompasses the following stages: (i) exploratory and statistical analysis; (ii) origin classification (region and flora), feature analysis, and (iii) comparative analysis between the results of geographic and botanical classifications. RF, GBC, PLS-DA, SVC, LR and different metric evaluation were tested. LR, SVC, RF, and PLS-DA, were the supervised models with the highest performances of mean Area Under the Curve of Receiver Operating Characteristic (ROC-AUC) score for the different geographic origins. PLS-DA and SVC were the supervised models with the highest performances of ROC-AUC score for the different flora origins. The features influencing the model output identified by Shapley values were linked to the correspondent functional chemical groups and a comparative analysis based on the feature importance results for region of production and flora origin was conducted. It is observed that there are features with importance in some regions which are also considered features importance in some flora groups. These results highlight the possibility to exist some common features (resonances) with influence on flora and the correspondent producer region at the same time. This exploratory work provide useful information for honeys produced in Santa Catarina, Brazil regarding botanical and geographic origin authentication.

The main conclusions of this thesis are summarized in Chapter 7. The work undertaken in this thesis is briefly analyzed and future perspectives and some ideas for improvement are also proposed in this chapter.

Metabolomics

2.1 Definitions and general workflow

Different analytical techniques for the suitability of food authentication processes have been assessed throughout the years. Chromatographic and molecular techniques are the most common techniques used and presented in studies of food authentication.

Nonetheless, there are non-invasive techniques able to achieve good performances which are being used for the food authentication process, like: isotopic, vibrational, UV-vis, fluorescence spectroscopy, elemental techniques, and nuclear magnetic resonance presented in Figure 1 [10].

This chapter describes briefly main analytical techniques employed for food authentication with emphasis in the methods used in this work, regarding wine and honey authentication.

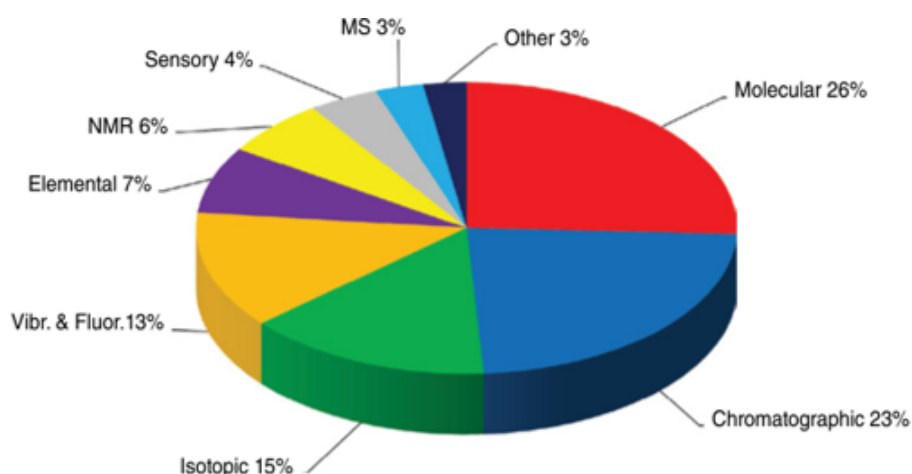


Figure 1: **Percentage of publications distributed over different analytical techniques applied to food authentication until 2016.** Adapted from [10].

The field of Metabolomics is one of the main omics areas, and consists of the exhaustive study of the whole metabolome composition of a particular system or organism and their environmental interactions, focusing on small chemical molecules involved in cell processes called metabolites [9], [11].

There are a considerable variety of metabolites, such as carbohydrates, lipids, proteins, amino acids, amines, steroids, phenolic compounds, carotenoids, alkaloids, or volatile compounds, among others.

Determining the authenticity of food products could involve different methodologies, depending on the purpose and the extension of the food analysis [11], [9], [10].

Therefore by choosing the most suitable metabolomics approach(s) and method(s), it is possible to clear up different issues and unsolved problems, such as the ones in food authentication [9].

The constant evolution and the new developments in analytical techniques, instrumentation, analytical software, statistical methods, and computational techniques allow to accelerate or improve data collection, data analysis, and data interpretation [12], [13].

Nowadays, the analytical techniques used in metabolomics approaches are robust, sensitive, reproducible, and more economical, meaning that it is possible to validate the results of more studies and also to standardize methodologies [12], [13].

Metabolomics analyses have been generally classified based on the selected methods used to study the metabolites. The targeted analysis focus on a specific group of metabolites with identification and quantification. Untargeted metabolomics studies follow a detection approach to a non-specific group of metabolites, to obtain patterns or fingerprints without the need for identification or quantification, as shown in Figure 2 [14], [9].

Based on the data manipulation, metabolomics studies can be categorized into four to four conceptual approaches:

- Targeted analysis: follow the identification and quantification of a small set of known metabolites related to a specific metabolism interaction, also called targets. This approach uses analytical technique(s) detection of interest compounds with the best performance.
- Metabolite profiling: the analysis focus on a large set of compounds related to a group of specific metabolites. The metabolites of this analysis do not require quantification and do not need to be identified. Semi-quantification is the procedure commonly performed.
- Metabolic footprinting: an approach with a study focus on the external and/or secreted metabolites by the cells. This analysis can follow a targeted procedure at a specific group of metabolites or through the spectra not directed to any selective metabolites group.
- Metabolic fingerprinting: This approach delivers a metabolites "signature" or fingerprint spectra of a sample of interest. The generated data is then compared to a large set of other samples to screen for differences. This approach does not focus on providing information about specific metabolites and is common used in classification problems based on its spectrum, as in food authentication studies.

Considering the specific objective of the analysis some metabolomics studies can be classified as

discriminative, informative, and/or predictive as shown in Figure 2[15], which can be achieved by two most common approaches: the *Target* and the *Untargeted* as presented in Figure 2[15].

In Food authentication studies when spectral signals are able to discriminate between samples, then the metabolites can be identified and the biological relevance of that compound elucidated, saving valuable analysis time. It is commonly used in forensics, among other fields, having been applied for instance in the discrimination of food quality and in food authentication. This approach will be emphasized throughout this dissertation.

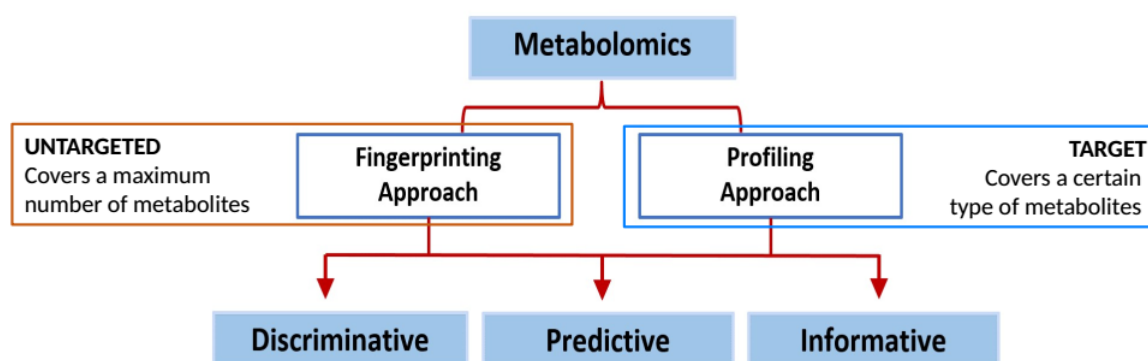


Figure 2: **More common Metabolomics Purposes and based Approaches.** Adapted from [9].

An example of a discriminative purpose using metabolomics approaches is assigning specific labels and certifications to food products, confirming their authenticity [10]. So depending on the study purpose (Predictive, Discriminative or Informative) a different metabolic-based approach is followed [16].

In summary, metabolomics approaches have the potential for solving many problems in food authentication, namely in the issues regarding food origin determination (botanical and geographical), traceability, and testing of possible adulterations.

Although the existence of different metabolomics approaches, the methodology steps are similar between studies. Often the metabolomics workflow follows a sequence of steps. The process starts with the Experimental design with the main question of the study, followed by Sample Preparation, then the selection of Data Acquisition, Metabolite extraction and Detection (when applicable) followed by Data treatment (data preprocessing and data processing) preceded the ending of Data Analysis, as presented in Figure 3. However, several studies choose more than one analytical technique for the metabolomics study, different data analysis methods can be also combined according to the investigation issue to interpret the results and achieve the most accurate analysis [3],[9]. This phenomenon contributes to the lack of standardized guidelines for every step of the workflow in metabolomics studies. Therefore, to ease the data integration and avoid duplication efforts, the Metabolomics Society organized an initiative (the Standards Metabolome Initiative) to formulate a minimum set of reporting standards describing the experiments associated with the statistical and chemometric analysis[16].

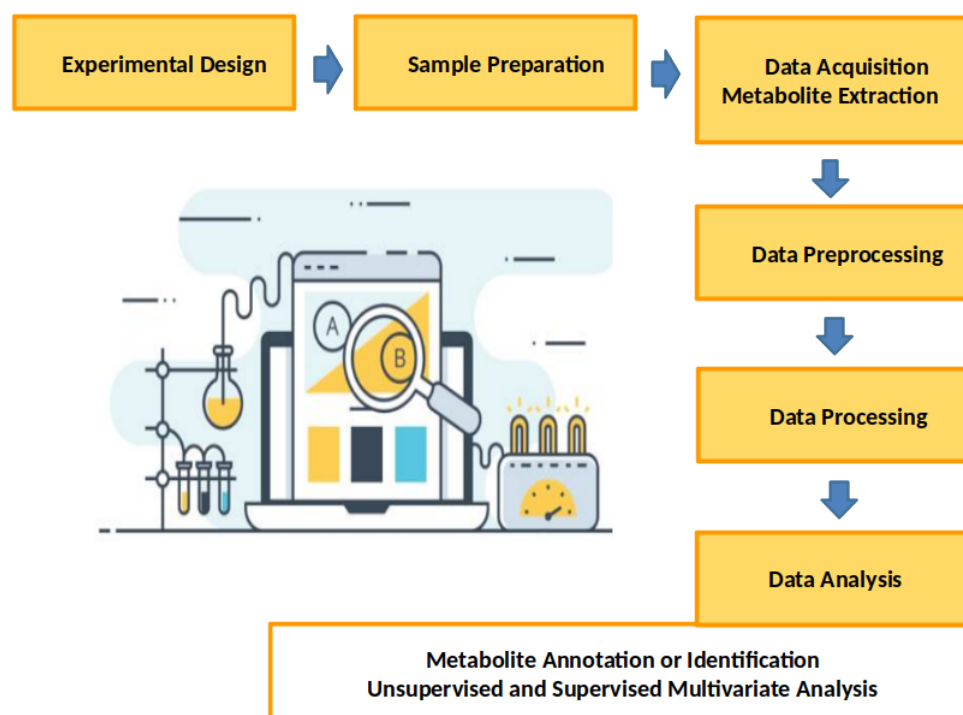


Figure 3: **Standard workflow on metabolomics food authentication.** Adapted from [9].

2.2 Sample Preparation

Preserving the quality and the diversity of the metabolites at the moment of the acquisition is a requirement in any metabolomics approach to achieve the most accurate analysis.

For this reason in any metabolomics pipeline, Sample Preparation is a step with the highest relevance to the whole metabolomics approach.

For that, it is necessary to avoid the cell's natural degradation process, to ensure the physics-chemical features of the metabolites according to their natural properties and location. This method is called the quenching method. Also, in this step, it is imperative to ensure the sample pureness, by removing totally or partially the solvents of the samples. Removing water from aqueous samples through this method can also prevent heat degradation. It is important to note that this method is related to the analytical technique selected[17].

2.3 Data Acquisition: analytical techniques

This section presents the most common analytical techniques used in metabolomics, GC/LC-MS, NMR, IR, UV-vis, and Raman spectroscopies.

2.3.1 Mass Spectrometry

Mass Spectrometry (MS) is one of the analytical techniques which allows the identification and quantification of metabolites by the determination of the mass charge ratio (m/z) of charges on compounds. This technique can be applied on single forms, as molecules, or on large complex mixtures of molecules or fragments, when the metabolites must be separated. This separation is performed before the MS and is achieved by combining other techniques such as Gas Chromatography (GC), Liquid Chromatography (LC), Capillary Electrophoresis (CE), or Ion Mobility (IM), according to the chemical properties and state of the samples (gas or liquid). After the separation, the different metabolites are then ionized in MS by adding a charge to determine their mass/charge ratio. The results are presented as spectral data, where the x-axis is the mass/charge ratio and the y-axis the intensity/quantity of the different ions in the metabolite [18].

2.3.2 Nuclear Magnetic Resonance

Another analytical technique is Nuclear Magnetic Resonance (NMR) spectroscopy, a non-invasive technique that identifies and quantifies even the metabolites with the same mass. NMR uses the magnetic properties of a selected atomic nucleus to determine the chemical and physical properties and location of those atoms or molecules. The variation of the external magnetic field causes the absorption and reemission of energy by the atomic nuclei to vary. This shift is calculated as the difference between the resonance and the reference substance frequencies, divided by the operating frequency of the spectrometer.

Regarding the atom selection, different metabolomics data can be generated based on the atom selected. The most common atom chosen is hydrogen since it is an abundant atom in biological samples (^1H -NMR). The results are shown as spectral peaks pattern signal where the x-axis represents the chemical shifts while the y-axis as the signal intensity, revealing a pattern [19].

In metabolomics approaches usually is selected the one-dimensional NMR spectrum (1D-NMR), and the two-dimensional NMR spectrum (2D-NMR) is chosen only when the metabolites compounds are impossible to identify through the 1D-NMR spectrum because the 2D-NMR allows separating overlapping peaks [18].

2.3.3 Infrared Spectroscopy

Infrared (IR) spectroscopy is another non-invasive analytical technique, which reveals the chemical composition of the metabolites present in the samples. This spectroscopy technique uses the resulting vibration from the absorption of specific frequencies according to the features of the structure of the metabolite.

A great advance in IR technique was the introduction of Fourier-Transform spectrometers, because this evolution improved the quality of infrared spectra by reducing the acquisition time, and also allowed the analysis of samples of small size without a sample preparation procedure. According with the analysis length, the results presented as spectrum can be divided in three main regions: the Far-Infrared (FIR),

400cm⁻¹ - 100cm⁻¹; the Mid-Infrared (MIR), 4000cm⁻¹ - 400cm⁻¹; and Near-Infrared (NIR), 13000cm⁻¹ - 4000cm⁻¹ [18].

2.3.4 Ultraviolet-visible Spectroscopy

Ultraviolet-visible Spectroscopy (UV-vis) is another non-invasive technique that reveals qualitative information about the metabolites samples by the emission of high energy radiation in the UV(200-400nm) and the visible (400-700nm) range, causing an electromagnetic transition in the molecules at the electron's level. The UV-vis spectrum results show absorbency values on the y-axis as the wavelength values (nm) on the x-axis [20].

2.3.5 Raman Spectroscopy

Raman spectroscopy is a quantitative or semi-quantitative analytical technique that uses vibrational, rotational, and other low-frequency modes to provide information about the chemical and physical forms of the metabolites. Essentially, this technique uses a tiny portion of light resulting from the interaction between the photon's light with the matter.

The technique uses a single frequency of radiation to irradiate the sample, and it is the radiation scattered from the molecule, one vibrational unit of energy different from the incident beam, which is detected. Most of the scattered light does not change its wavelength in the process, but part of it does, and such scattering is known as Raman scattering. The intensity of the scattered light is plotted against its frequency (cm⁻¹) and the result is a Raman spectrum of the sample.

This technique is less used due to problems with sample degradation and fluorescence. Therefore, recent advances in instrument technology have simplified the equipment and reduced the problems substantially. These advances, together with the ability of Raman spectroscopy to examine samples in a wide range of states and with minimal spectrum manipulation needed, have led to rapid growth in the application of the technique [21].

In this dissertation the focus is in ¹NMR and FT-IR spectroscopy techniques used to performed metabolomics based approaches to authentication of wine and honey study cases.

2.4 Data Preprocessing

Data pre processing is a important step in any metabolomics approach because it addresses the necessary changes to conduct over the data to make the samples analyzable and comparable.

Raw data is usually complex and needs to be clean and treated for subsequent analysis. Incomplete records, atypical and inconsistent data can affect the classifier model's performance and lead to misleading data interpretations [22],[23]. Also it is important to consider that data collection can be an expensive, and time-consuming. Thus to avoid missing data there are exist distinct preprocessing methods.

In metabolic approaches, the most common pre processing methods include missing values treatment, outliers removal, normalization or scaling, and spectral processing strategies.

With these considerations, data pre processing is a crucial step in any metabolomics approach because raw data has the information to answer the food authentication issues or questions.

This section covers the principal methods employed in the metabolomics-based approaches to food authentication cases supported by recent studies regarding FT-IR and ^1NMR spectroscopies, the techniques used in this dissertation.

2.4.1 Missing values and Outliers

A missing value consists of a variable with no assigned value caused by three common mechanisms.

The first mechanism is randomly missing values distributed in the data matrix with no dependency on variables. The second mechanism is randomly missing data associated with possible variable dependency to other known variables (X) but not to response variables (Y) [24]. The third mechanism originates non-randomly missing data in which the missing values have a pattern and dependency within missing values in a variable [25].

There are two methodologies for handling missing data: the removal or replacement. Removing missing values means the elimination of one or multiple variable(s) or a sample(s) that contains the missing data. It is also possible to remove missing values by eliminating variable(s) based on the high percentage of missing values (e.g. more than 5% of missing data per variable(s) or sample)[25].

Replacement is a method in which the missing values are replaced by a numerical value from mean, median, or using other more sophisticated methods (e.g. K-Nearest Neighbor, linear approximation, or other substitute methods) [25].

Outliers are data values that are distant from the other observations and can occur from variations in data measurement or caused by experimental errors. These values are usually excluded from the dataset to avoid interference with the analysis results [26], [23].

2.4.2 Spectra Treatment

IR and ^1NMR data processing methods consist of data treatment methods as baseline correction, and noise filtering processes as smoothing.

In ^1NMR data, peak detection and peak alignment, and data correction (which includes baseline, offset and background corrections) are also applied. Notwithstanding, the spectrum needs to be transformed from FID into frequency spectra using the preprocessing steps of zero-filling, apodization, Fourier transformation, and phase correction before the previously performing the aforesaid preprocessing methods [27].

Relative differences observed between samples are corrected using the methods of Normalization and Scaling. In IR data, mean centering and Standard Normal Variate (SNV) are often used, as well as first

and second-order derivatives calculations. In some metabolomics studies with ^1NMR data, logarithmic transformations are applied. However, this method does not deal with missing values, requiring a previous treatment of null data [28],[29].

Spectral peaks can suffer dislocations through the x-axis, which can occur due to changes in the chemical environment of the sample like solvents or pH [18]. Spectral alignment procedures are categorized into warping and segmentation. The warping method applied a non-linear transformation to the respective axis. The segmentation method applied a constant shift to all spectral points [18]. Another approach using fast Fourier transformation has been used to perform data alignment [30], [18], [26]. The process of spectral alignment allows to correct the position of matched peaks, that belong to the same metabolic feature(s), in a multiple spectra study [18].

Data correction, which can either be baseline, offset, or background, aims to eliminate the effect of certain signal variations through the spectra and background noise during the sample analysis. Distinct factors can produce these signal variations, as absorption associated with the sample holder and/or solvent used or by either experimental or instrumental variations [18], [11].

These correction methods are very important to avoid false results in data analysis and to not reveal metabolites that are not present in the samples [30].

The smoothing methods filter the spectra noise with significant impact in cases with low noise ratio signals or when the followed data analyses approaches are more sensitive to noise signals.

These methods help with the visual interpretation and robustness of the analysis, therefore they require a balance between the noise reduction and the peak retention, especially in the peaks with reduced absorbancy [30]. The Savitzky-Golay filter is one of the most popular smoothing methods. This procedure fits successive sub-sets of adjacent data points with a low-degree polynomial by the method of linear least squares, using a process known as convolution [31].

Another approach is to treat relative differences between samples instead of the absolute values allowing different data to be comparable and consistent. In this approach Normalization, Data mean centering, Scaling, Standard Normal Variate (SNV), Multiplicative Scatter Correction (MSC), and Derivative spectroscopy are the most common metabolomics preprocessing methods used.

Normalization methods contribute to a correct measurement of the features in the metabolomics analysis [11], [26], [18], [32]. The subtraction of the row mean to each data value and dividing by the standard deviation is the most common normalization method utilized in ^1NMR data [11], [26], [18], [32].

MSC method is a relatively simple processing step that attempts to account for additive and/or multiplicative effects in spectral data. It does so by estimating light scattering or change in path length for each sample relative to that of an ideal sample. MSC is probably the most widely used pre-processing technique for NIR [33].

SNV method, a weighted normalization is performed (not all points contribute to the normalization equally). Therefore, the values are subtracted from their mean, and then the result is divided by the standard deviation. Spectra treated in this manner have always zero mean value and a variance equal to

one and are thus independent of original absorbance values. SNV is the second most applied method for scatter correction of NIR data. [33].

2.4.3 Mean-centering

Data mean-centering is a column-wise transformation and consists of subtracting the mean spectrum from each sample, resulting in all columns with a mean of zero. This method accentuates the differences and not the similarities of the data samples, so it requires the combination with scaling methods [29], [26],[32].

2.4.4 Scaling methods

Scaling methods divide each variable by a factor, the scaling factor, which is different for each variable. Scaling methods allows samples with variables measured on different scales comparable [29]. Scaling has two subclasses considering the type of measure. Scaling data with dispersion such as the standard deviation as a scaling factor, this measure includes Auto-scaling, Range scaling, Pareto scaling, and Vast scaling. Scaling data using size as a measure, like the mean, includes the Level scaling.

- Auto-scaling: Compare metabolites based on correlations. All metabolites become equally important. May inflate the measurement errors.
- Range scaling: Compare metabolites relative to the biological response range.

Both Auto and Range scaling methods treat metabolites with equal importance but also give a gain effect in measurement errors. Range scaling also inflames outliers values. To deal with this inflation standardization methods are applied as SNV.

- Pareto scaling: Reduce the relative importance of large values, but keep data structure partially intact. This method keeps closer to the original measurement however is sensitive to large fold changes.
- Vast scaling: Focus on the metabolites with minor fluctuations. Aim for robustness when using prior group knowledge. Therefore is sensitive to higher variations without group structure.
- Level scaling: Focus on relative response. Indicate for identification of biomarkers. As auto and range, Level scaling can increase the measurement errors [29], [30].

2.4.5 First and second-order derivatives

First and second-order derivatives are the derivative spectroscopy most commonly calculated for metabolomics analyses. These methods allow to identify significant differences in the derivative mode present in spectra

samples that are very similar. The separation of overlapping peaks is achieved by differentiation of a zero-order spectrum and obtaining consecutive derivative spectra. These methods increase the discrimination without the separation of the analytes.

The first-order derivative consists of the rate of change of absorbance concerning wavelength. Using the first-order derivative the baseline shifts are eliminated because the resulted constant absorbance offset is zero. The second-order derivative has a very characteristic feature consisting of a negative band with a minimum at the same wavelength as the maximum on the original spectrum [34].

2.4.6 Data transformation

Data transformations, such as logarithmic and power, are non-linear conversions methods usually applied to correct data, as in the case of heteroscedasticity, which is the existence of absolute noise that increases with the rising of signal intensity. Despite both methods correcting the heteroscedasticity, logarithmic transformation does not deal with null values in the opposite of Power transformation that does not require this previous data treatment [28],[29].

2.5 Data Analysis

After the preprocessing methods, the metabolomics data is ready to be analyzed, interpreted, and compared. Metabolomics data reunite vast information and allow to perform different analyses according to the food authentication issue using chemometrics tools, such as Metabolites identification, Exploratory analyses, Univariate and Multivariate Statistical analyses, Unsupervised and Supervised analyses (machine learning) including Feature importance and interpretation.

2.5.1 Metabolite Annotation and Metabolite Identification

The Annotation and the Identification of metabolites are two distinct methods used in metabolomics approaches achieved by using NMR and GC/LC-MS techniques.

The Metabolite annotation is the assignment of metabolite candidates to the signal based on matching spectra of a studied sample with reference database or library entries spectra.

Metabolite identification entails the comparison of the metabolite's properties of an authentic standard obtained under identical analytical conditions to those of the identified metabolites. [35].

Here it discusses the metabolites annotation by ^1NMR , a method utilized in this dissertation.

The annotation of metabolites from ^1NMR data is possible by matching the spectra of measured ^1NMR peaks against the ones from each reference metabolite, acquired under similar conditions to those of the samples analyzed.

Each metabolite has a unique and specific spectrum with spectral shifts or differences resulting from

the acquisition conditions. Annotated metabolites are those candidate metabolites from the studied samples that reveal to have more peaks or peak groups matched to the ones present in reference databases.

To achieve a metabolite annotation with a higher score of matching needed reference spectra with quality. Meaning to have reference spectra with the different frequencies of spectra acquisition, different types of solvents, also distinct biological origin sources, and other characteristics considered to influence the performance of metabolites annotation [18], [35].

There are different tools to perform the metabolite annotation and identification like *R packages*, *Web tools* and *Software* as presented below in 6.

Table 1: **Metabolite Annotation and Metabolite Identifications tools.**

Tool	Type	Sepctral Data	URL
MAIT	R package	GC-MS	https://bioconductor.org/packages/release/bioc/html/MAIT.html
BATMAN	R package	1D-NMR	https://batman.r-forge.r-project.org/
rNMR	R package	NMR	https://nmrfam.wisc.edu/software
MetaboHunter	Web Tool	1D-NMR	http://www.nrcbioinformatics.ca/metabohunter/
MetaboHunter	Python library	1D-NMR	https://github.com/mfitzp/metabohunter
Bayesil	Web Tool	NMR	https://http://bayesil.ca/
MetaboMiner	Software	2D-NMR	http://wishart.biology.ualberta.ca/metabominer/
ACD	Software	UV, IR, Raman and more	https://www.acdlabs.com/products/spectrus-platform/

2.5.2 Exploratory Analysis

Exploratory analysis performs the initial investigations on data to discover patterns, and anomalies, test hypotheses, and check assumptions with the help of statistical methods and graphical representations. This first approach to the data usually starts with the spectral visualization of the sample's metabolomics data [36].

2.5.3 Univariate Analysis

The univariate analysis focuses the data study on one variable at a time, which is easy to perform and interpret [18]. Based on the data distribution two types of tests can be used: Parametric tests, the data assume a normal distribution of values, also called Gaussian distribution. The parametric test can use a reduced sample size. Non-parametric tests are also powerful statistical tools because use few assumptions

about the underlying distribution of the data. These tests are applied to data obtained from a non-normal distribution or one containing outliers [37].

Among the most popular univariate analysis techniques used in metabolic approaches are Parametric tests such as t-tests and Analysis of Variance (ANOVA). Nonparametric tests often used include the Kruskal-Wallis, Kolmogorov-Smirnov, and Wilcoxon signed-rank tests. Regression analysis is also employed.

Parametric tests

- A t-test is an example of a parametric test, and it is a statistical hypothesis test following a Student's t distribution under the null hypothesis. This statistical test compares the data by determining if two datasets are significantly different from each other in terms of their means[18].
- Analysis of Variance (ANOVA) is a selection of statistical models used to get a comparison between more than two groups. The one-way ANOVA method is an upgrade of the t-test similar to multiple two-sample t-tests. The two-way ANOVA is an extension of the one-way ANOVA and examines the influence of two different categorical independent variables on one continuous dependent variable. Besides assessing the main effect of each variable, it also estimates if there is any interaction between them [18].

Non-Parametric tests

- Kruskal-Wallis test is frequently used when it is necessary to determine if three or more independent samples originated from the same population. A significant Kruskal-Wallis test indicates that at least one sample stochastically dominates one other sample.
- Kolmogorov-Smirnov Two-Sample test is used to determine if two independent samples are from the same population or two populations with the same distribution pattern. It is a test sensitive to distribution differences, and useful with small samples.
- Wilcoxon signed-rank test is a paired difference test. It is used when comparing two related samples or repeated measurements on a single sample to assess if their population mean ranks differ. The test focuses on the magnitude and direction of the differences for matched pairs [18] [38].

Regression analysis: the linear regression test analysis is one of the most widely used statistical tools. The regression analysis allows the investigation and modeling of the relationship between a scalar dependent variable (y) with one or more independent variables (regressors) (x), assuming a linear relationship between them. When there is only one independent variable is called simple linear regression, and for more than one independent variable the process is called multiple linear regression. The regression coefficients are often estimated using the least squares method, which attempts to minimize the sum of the squares of the errors made in the results of every single equation [39].

However, in many results from analytical experiments, there are a large number of variables to analyze, so it is performed a multivariate analysis could mean a better characterization of the data [18].

2.5.4 Unsupervised Multivariate Analysis

When there are a large number of variables to analyze, it is performed a multivariate analysis could mean a better characterization of the data [18]. The unsupervised multivariate analysis considers the use of all or many metabolomics features (variables) simultaneously to identify relations between them, which may this increase the quality of the data evaluation [18].

The unsupervised multivariate analysis uses unlabeled data meaning no information about natural groups in the data. This analysis is normally used to visualize and summarize data and, thus, detect patterns that can be related to biological variables [18]. Principal Component Analysis (PCA), t- Stochastic Neighbour Embedding (t-SNE), and the Clustering methods such as Hierarchical Clustering Analysis (HCA) are the most common methods used.

PCA is a linear dimension reduction technique with the main goal to maximize variance and preserve large pairwise distances, by converting a set of observations of possibly correlated variables into a set of uncorrelated variables (principal components). PCA aims to explain the maximum data variability possible with few principal components as possible. In general, principal components can compute up to the total number of variables. It can be seen as a method to compute a new coordinate system formed by the latent variables, which are one orthogonal, and where only the most informative dimensions are used. Latent variables from PCA optimally represent the distances between the objects in the high-dimensional variable space [32].

t-SNE is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data used for visualization in a low-dimensional space of two or three dimensions. This technique gives visual information about how data is arranged in a high-dimensional space. t- SNE operates in two stages, in the first stage it generates probability distribution of high dimensional data and in the second stage it generates similar probability distribution of lower dimensional representation of the data points. It then minimizes the distance between the two-probability distribution using Kullback–Leibler divergence. t-SNE overcomes the shortcoming of the crowding problem suffered by other manifold learning methods [40].

Cluster analysis is a statistical procedure used to find groups of similar objects. There are two popular types of clustering methods, HCA and Non-Hierarchical clustering. HCA used one of two methods (agglomerative or divisive clustering) to rearrange objects and partitions in a hierarchy tree-like dendrogram.

- Agglomerative: each of the objects forms a separate cluster, resulting in n clusters in the first level. In the next level, the two closest clusters are merged, and so on, until finally, all objects are in a single cluster.
- Divisive: groups all objects in one single cluster, in the first level of the hierarchy. In the next level, this cluster is split into two smaller clusters, and so on, until finally, each object forms a separate cluster. The result greatly depends on the distance between samples [32].

Non-hierarchical clustering methods aim to divide a dataset into a predefined number of clusters, organizing data objects into a set of non-overlapping flat groups, by commonly using iterative algorithms

which optimize the chosen criterion, which means clustering is one of the formulations for the objective function of these processes [18]. The widely used K-means method is an iterative procedure used to find the best partition of k classes. This non-hierarchical cluster model uses the initial partition of k groups usually randomly generated. In addition, K-means results are highly dependent on the initial partition and the number of classes k assigned. For each iteration, the mean vector of the class is recalculated. And the samples are disposed of at the nearest center. This procedure continues until the final criterion is reached as the minimum or the maximum number of iterations reached [18].

2.5.5 Supervised Multivariate Analysis

Supervised multivariate analysis, consists of predictive models used to determine output from certain data input. The predictive models require a dataset for training and learner algorithms able to generalize from data, with the aim of performing accurate predictions on new, unseen examples [32], [26], [41]. There are two main approaches to reach a predictive model: classification when the response variable is categorical, or regression when the output variable is continuous.

2.5.5.1 Model Evaluation

Model evaluation uses error metrics to understand if the model is able to be performed with unseen data. The simplest error metric is accuracy, which is the proportion of samples correctly classified, while the Kappa Statistic compares the observed accuracy with the expected, determining how close the classification from the model is to real classification. Another error metric is the mean Area Under the ROC Curve (AUC), this metric measures the performance of a two classes classifier, a common metric used in a classification problem. In cases with imbalanced data there are also different metrics that can be used as Balanced accuracy or Matthew's correlation coefficient. Other metrics can be selected as Precision, Recall, f1-score, Log Loss and others depending on the model's problem. To evaluate regression problems, there are essentially two error metrics: Root Mean Square Error (RMSE) and coefficient of determination [41].

In cases when the complexity of the model is high and there is insufficient training data, an event called overfitting occurs. This results in a model well suited for the training data, but the model is unable to generalize to unseen data. To deal with this problem it is recommended to increase the size of the training set, but in many cases that approach is not always possible. Thus, other statistical methods have to be utilized to assess the generalization capability of the model, namely, cross-validation schemes, or bootstrapping [41].

2.5.5.2 Feature Selection

Feature selection techniques are a procedure to reduce model complexity with different methods such as filters, wrappers, and embedded. The method considers subsets of the original variables with relevance by removing irrelevant, redundant, or noisy data. Feature selection also allows the model to learn

faster, improves the predictive accuracy, and simplifies the interpretation of both the model and the data [42]. Another approach to reducing model complexity is the use of unsupervised dimensionality reduction techniques, such as PCA [43].

2.5.5.3 Model Validation

After the model construction or selection, the model need to be validated to ensure the quality of a given predictive model. Model validation is crucial to choose the best model to utilize, out of a set of selected models. For that the model need to be tested on a set of examples that were not used to train the model. The available data is divided into the train set and the test set. There are three of most popular procedures to validate the models performance:

- K-fold cross-validation divide randomly the input dataset into k subsets that are equal in size and mutually exclusive. The estimator is evaluated k times, using one of the folds as the test set and the remaining folds as the training set. The cross-validated model performance is the average of the performance score calculated for the iterations [44].
- Leave-one-out cross-validation is the case where k is equal to the number of samples in the training set.
- Bootstrapping creates a validation set by sampling from the original dataset with replacement.

2.5.5.4 Machine Learning Models and Algorithms

Among the most popular supervised methods for classification are: Logistic Regression (LR), Partial Least Squares Discriminant Analysis (PLS-DA), Linear Discriminant Analysis (LDA), Soft Independent Modelling by Class Analogy (SIMCA) and Decision trees.

Methods for regression tasks include Linear Regression, Partial Least Squares Regression (PLS-r) and Regression trees.

Methods such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Artificial Neural Network (ANN) and Random Forests can be used for both classification and regression tasks

Logistic regression is a linear model used when the dependent variable is binary, a common classification problem. This model uses a sigmoid (logistic) function to model the probability of a given example belonging to a given class. Is a model similar to linear regression in which the estimation of model parameters is an optimization problem [45].

PLS-DA is a multivariate dimensionality-reduction, a variant of PLS-r model that allows qualitative analysis or discriminant analysis [46]. As in PLS-r model the construction of the PLS-DA is based on the independent and dependent variables covariance being the main difference that PLS-DA works with categorical dependent variables. Each column matrix of y corresponds to a class and contains either 1 if the sample belongs to the class, or 0 otherwise (a complete disjunctive coding). PLS-DA in advantage

of SIMCA focuses on separating the classes. PLS-DA is a common model used in metabolomics [47],[48], [49] and with an increasing application in food authentication as can be seen in chapter 3.

LDA is a discriminant model that attempts to find a linear combination of features (latent variable) that characterizes or separates the classes of objects or events. For each class, the statistical properties of the data are calculated and then used to make predictions. This method assumes a Gaussian distribution of the data, meaning that each attribute has the same variance. It is a model related to PCA, although PCA does not take into consideration the underlying class structure, and in some cases is used for data dimensionality reduction [50].

SIMCA is a discrimination model used to solve multi-class problems based on PCA and is thus suitable for high-dimensional data. The idea is to describe the multivariate data structure of each group separately in a reduced space using PCA. The special feature of this method is that PCA is applied to each group separately and also the number of principal components is selected individually and not jointly for all groups. Due to the use of PCA, this approach works even for high-dimensional data with rather a small number of samples, whereas methods like LDA can become unstable under such conditions. In addition, SIMCA also provides information about the relevance of different variables to the classification, or measures of separation [32].

Linear regression methods are one of the simplest machine learning regression algorithms. Linear regression assumes that the relationship between the input features and the output variable is linear, and attempts to find the linear function that best fits the data.

Partial Least Squares regression (PLS-r) is the most used model in regression problems and requires a dependent and independent variable for generating a regression line.

The PLS algorithm takes into account the covariance between the independent and dependent variables, meaning searching for the variances of the variables and the correlation between them. The model structure transforms the independent variables into a set of intermediate linear latent variables called components. During model development, a relatively small number of components are calculated, and it is the number that determines the complexity of the model. The fewer the number of latent variables, the more robust the model is. This means that the model is stable concerning external disturbances, however, it increases the risk of being under-fitted. It is therefore necessary to choose the number of latent variables carefully to create an optimized model with a high prediction performance [32].

KNN is able to be used in classification and regression problems. This model does not require to be fitted. It is a type of model that works by using distance measures as Euclidean distance measure and the k -nearest neighbors. This model does not require linearly separable groups nor compact clusters for the groups of classes, being easily applied to multi-class problems [32].

SVMs were originally created to be used on classification problems. However, nowadays, it has been extended to regression problems with SVM-R and LS-SVM. In their structure, SVM models use kernel-based machine learning methods able to model non-linear problems. They start by mapping the input features to a new high-dimensional feature space using a kernel function. In this new feature space, the model searches for the optimal linear hyperplane that maximizes the separation between classes, called margin. The input examples that define the margin constitute the support vectors. Therefore, the SVM model is only

dependent upon a subset of the training data [51], [32]. To achieve a efficient and robust model, SVMs requires the optimization of parameters which also allow to avoid overfitting problem [32].

ANNs extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of these features. These models can predict quantitative or classes according to the type of network developed. Generally, the model structure used three layers of interconnected neurons. The first layer consists of an input layer with the correspondent independent variables with one neuron per column, the second (or more) layer(s) is hidden with k neurons that correspond to the weights which will have to be trained to build the model and the final layer correspondent to the output with the dependent variables with one neuron per column. The connections between different neural units can be enforcing or inhibitory in their effect on the activation state of connected neural units, through a limiting threshold function [52]. To use ANN models it is recommended to conduct several iterations because these are non-linear stochastic methods meaning that in each model performance different results can be produced. These models should be used with caution however when applying modeling parameters robust models can be generated [52].

Decision trees are recursive data structures including decision rules, which may be inferred from the input data through appropriate algorithms [53].

A set of weak classifiers can be combined by using two major classes ensemble algorithms, to create a strong learner with improved performance. In this context ensemble corresponds to multiple models of the same type trained together. The ensembles help to minimize the effects of noise, bias and variance.

In Bagging the training set is sampled with replacement, and each learner can be trained in parallel, while in Boosting a serial process is followed where the training data is sampled with repetition, weighting more heavily the samples the were previously misclassified by the previous models in the sequence

Random Forest classifiers utilize the Bagging algorithm, and build a large collection of trees, and then averages their output to improve the classification performance [52].

The Gradient boosting (GB), utilizes a gradient descent based method is used to minimize an arbitrary differentiable loss function, while employing a decision tree as base learner. As the name implies this trees are trained using a boosting algorithm, guided by the negative gradient of the loss function being minimized [54].

Some of these and other machine learning models used in metabolomics data analysis applied to food authentication studies available in the literature are listed on the fifth column of Table 1 in chapter of food authentication.

2.5.6 Model Interpretability

Machine learning models are generally considered 'black boxes', but the ability to easily interpret the results of a predictive model would improve our understanding of the factors underlying the differences between variables, rather than merely predict them. This information can be useful in metabolomics to improve the authentication process and avoid events of fraud. To assess the feature importance, the Shapley Additive

exPlanations (SHAP) value estimation method was adopted from the field of game theory. This method can be applied for classification problems, as in this dissertation, by computing the average marginal contribution of a feature value across all possible coalitions. [55], [56], [57],[58].

By computing the Shapley values of the best classification models, the selected features with the higher impact on the model performance and their contributions to the model performance are analyzed. An important note to consider is that to compute Shapley values, KernelShap needs to use a reduced number of features due to the time required to work, and also correlated features should be previously eliminated to avoid weighting the features wrongly, leading to an incorrect interpretation of the model performance.

2.5.7 Tools and databases

There are several tools and web tools which allow the processing and analysis of metabolomics data. Specifically, *webtools* have the advantage to help on the comprehension and extraction of knowledge from metabolomics data, without the need of programming skills.

To perform metabolite annotation or identification, pathway analysis, enrichment analysis and biomarker identification, the researcher use metabolomics databases such as the presented bellow on 2.

Nowadays, with the increasing number of open and web accessible tools, the performance of the analysis and interpretation have also become more precise, robust and also reproducible.

Of the web-based tools for metabolomics and spectral data analysis, MetaboAnalyst is the most comprehensive available tool.

The *specmine* R package, developed by the host group is another freely available tool that provides a set of methods for metabolomics data analysis, including data loading in different formats, preprocessing, metabolite annotation, univariate and multivariate data analysis, machine learning, and feature selection.

Table 2: **Available free tools for metabolomics and spectral data.**

Tool	Type	Sepctral Data	URL
chemometrics	R package	Chemical data	https://CRAN.R-project.org/package=chemometrics
ChemoSpec	R package	NMR, IR and Raman	https://CRAN.R-project.org/package=ChemoSpec
COLMAR	Web tool	NMR	https://spin.ccic.ohio-state.edu/index.php/colmar
hyperSpec	R package	UV-vis, IR, NMR, MS, Raman, ...	https://CRAN.R-project.org/package=hyperSpec
MeltDB	Web tool	GC-MS and LC-MS	https://meltdb.cebitec.uni-bielefeld.de/cgi-bin/login.cgi
MetaboAnalyst	Web tool	NMR, LC-MS and GC-MS	http://www.metaboanalyst.ca

metabolomics	R package	NMR, GC-MS, LC-MS and MS	https://CRAN.R-project.org/package=metabolomics
MetaboMiner	Java software	NMR	https://wishart.biology.ualberta.ca/metabominer
metaP-Server	Web tool	NMR, GC-MS, LC-MS and MS	http://metap.helmholtz-muenchen.de/metap2/
muma	R package	NMR, GC-MS, LC-MS and MS	https://CRAN.R-project.org/package=muma
MVAPACK	Toolkit	NMR and MS	https://bionmr.unl.edu/mva-pack.php
OpenMS	C++ library	LC-MS	https://www.openms.de
specmine	R package	NMR, GC-MS, LCMS, UV-vis, IR, Raman, ...	https://CRAN.R-project.org/package=specmine

2.5.8 specmine and Webspecmine

The *specmine* is an R package created by the host research group for the integrated analysis of metabolomics and spectral data which fills an unmet need of metabolomics. The need for an easy way to process disparate data sources in the open source statistical computing program R. The package handles datasets in different formats (GC/LC-MS, NMR, IR and UV-Vis), helps with the pre-processing data, metabolite annotation, and provides machine learning support [26].

The WebSpecmine (Figure 4) is a website for metabolomics data analysis and mining, which provide simple, interactive front-end and easy-to use tools to cover the main steps of the metabolomics data analysis data, such as NMR, MS, IR, Raman, UV-Vis and concentrations data. It contains modules for data reading and dataset creation, data pre-processing and analysis, as well as metabolite identification, by using implemented functions from the R package *specmine*. The webspecimine is easy to use, the user does not need to know any kind of programming language, and the results are presented using graphical visualization, so that they can be easily interpretable. Finally, the platform allows the download of any format text, table or graphical.

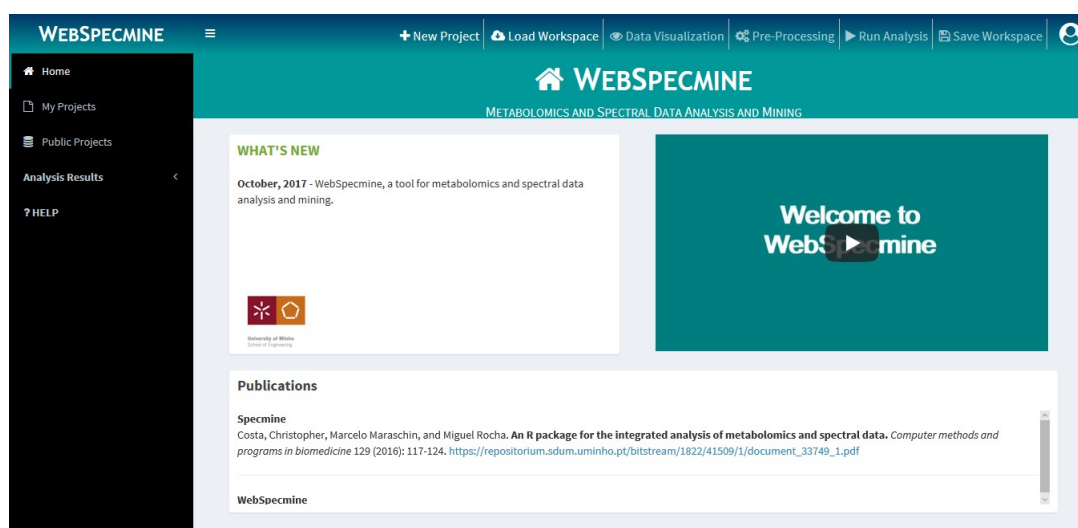


Figure 4: **Graphical interface of webspecmine front-end.**

2.5.9 Workflow and Data Analysis pipeline

The data analysis workflow in this dissertation encompassed the data preprocessing, statistical and machine learning modeling, and model interpretability. In Figure 5 is presented the general workflow used in the three works presented in this dissertation.

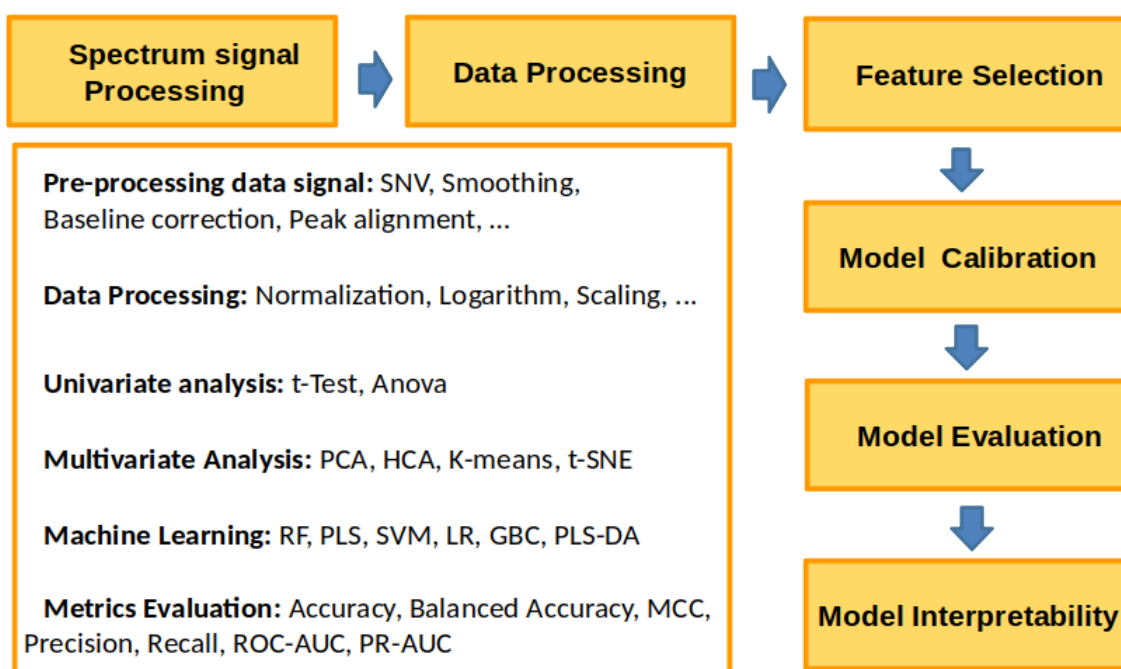


Figure 5: **General workflow used for the data analysis pipeline.**

Two distinct pipelines were employed: The first pipeline utilized the *specmine* R package, developed by the host research group, for spectral data treatment of untargeted metabolomics data of wine samples

from $^1\text{H-NMR}$ and amino acid concentration present in Chapter 4. This approach also included exploratory, statistical, unsupervised, supervised, feature importance and metabolite annotation analysis.

The second pipeline was used for the data analysis of untargeted metabolomics data from wine musts samples using FT-IR (Chapter 5) and honey samples using $^1\text{H-NMR}$ (Chapter 6), employing freely available Python libraries. Notably, the spectral signal preprocessing of honey samples using $^1\text{H-NMR}$ was performed using the *specmine* R package. The data analysis in these pipelines included exploratory, statistical, unsupervised, supervised, and model interpretability analyses, performed using Python scripts and libraries such as Pandas, Numpy, ResearchPy, Scipy stats, Scikit-learn and SHAP.

The comprehensive utilization of these pipelines in the present study filled a critical need in the field of metabolomics. More details are presented in methods section in each Chapter of the presented works. In Figure 6 a schematic link of the workflow steps and the respective Python libraries is shown.

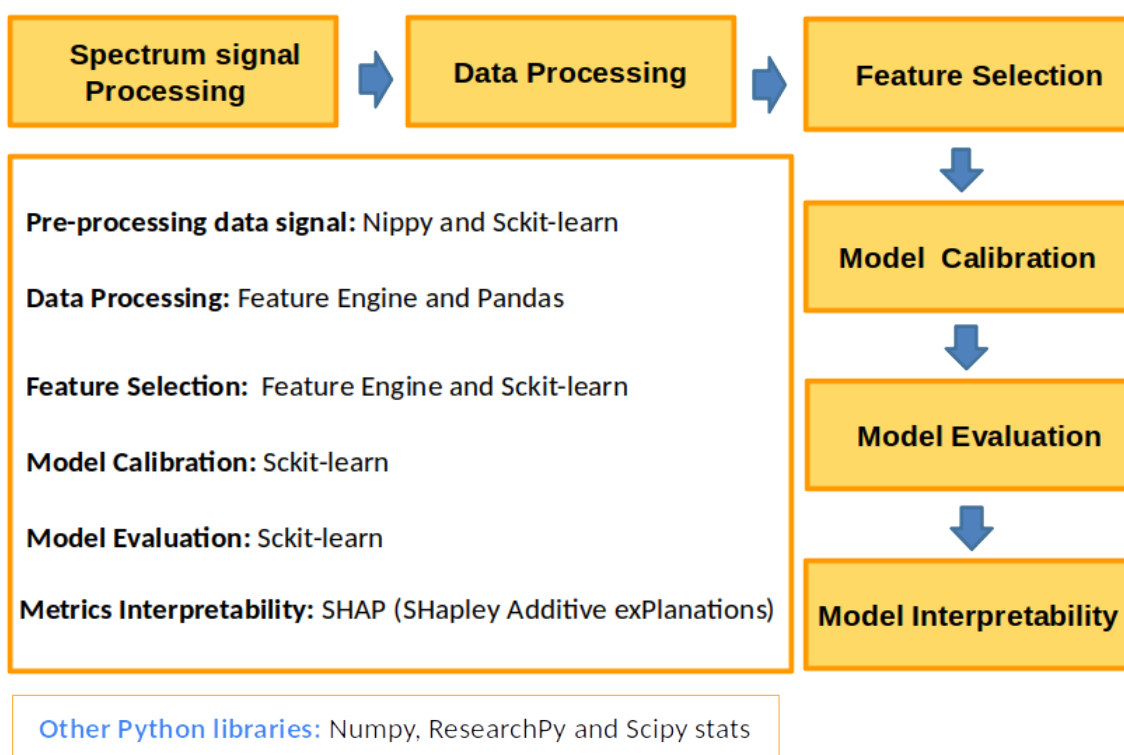


Figure 6: **Used Workflow on this work and the correspondent Python libraries.**

Metabolomics for authentication in wine and honey food products

3.1 Food Authentication and Traceability

The concerns with food authenticity started after the end of World War I, with the term “food security” associated with the traumatic experience lived by European countries who suffered from control and domination of food supplies [2].

In the present days, the Agricultural industry in the European Union (EU) is the second largest industrial sector, and with the globalization of the food market, the demand for the acquisition of certain food products despite the season, location, and availability leads to an increase of the production and distribution guided according with restricted regulation.

Economic concerns, and more importantly, safety issues are the main reasons for the regulation of the food systems applied in the EU, and integrated within those guidelines is the directive of the consumer’s rights to receive truthful information about the food they buy (EU regulation No. 178/2002).

The main aim of these guidelines is the prevention of fraudulent or deceptive practices, the adulteration of food, and the exercise of other practices which may deceive the consumers [59].

Currently, in Europe, there is also specific legislation for food products according to geographic origin - one of the main issues of authentication. This regulation protects high-quality food products with specific features derived from special geographical conditions or with traditional methods of production [60]. This special certification tries to ensure the safety of the food production system, and the commercial valuation for consumers and producers; aiming to reduce the cases of counterfeit food production and commercialization.

To better understand the present work, it is important to present a brief explanation of some concepts and definitions.

Food authenticity is an ancient field that tries to validate the veracity of the description of food products and their ingredients, according to the label description [61]. The food product needs to follow the regulation and the standard procedures applied to the food system (production, distribution, and commercialization), adopted by the country to ensure quality and safety. These concerns are integrated with the

Food authentication process [62].

This process may differ according to the regulation adopted by the country, which may lead to a panoply of analysis tests to apply. The aforementioned process may also include distinct research topics, such as food origin or traceability, food control, or adulteration. Thus, having these concerns in mind, there is a need to select the proper analytical test and the best chemometric data analysis approach. While the increase of tests also increases the number of studies, in Figure 7 is showing the countries with the most publications related to the authentication process.

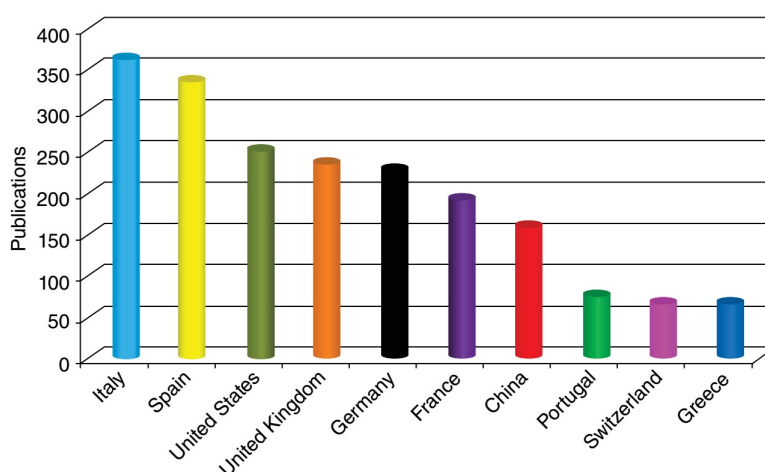


Figure 7: **Top 10 countries on food authentication publications.** Adapted from [10].

3.1.1 Food Control

Food Control is one of the most important topics within the food analysis field, and it has to be applied in most stages of production to ensure the quality and the safety of the food products. These concerns encompass the verification of the quality and quantity of certain ingredients, and other elements connected with the legislation applied to the last stage of the production before reaching the final consumers.

Notwithstanding, to ensure the safety of the product the analytical exams are performed with the purpose to assess: (1) the presence of selected compounds in foods that may be present below certain thresholds (maximum residue limits - MRL); (2) the detection of microbial-related spoilage; (3) the determination of allergens; (4) the detection of environmental contaminants, as well as banned external compounds; and (5) the assessment of the occurrence of natural toxins [3].

To certify the quality of the product – which is the verification of quality and quantity of certain ingredients and other elements, - analytical tests are conducted to check the presence of food adulterants. Adulteration is described as the deliberate alteration of the product by adding or removing ingredients from the food product, aiming to get an increase in profit. According to Food Standards Agency (FSA) (United Kingdom Agency), there are two main types of adulteration: (1) the transaction of food which is unfit and potentially harmful; and (2) the deliberate misdescription of food [63] [64]. Fortunately, most of the food

adulteration cases do not have the intention to harm the consumers wellbeing [65] [66], but have as main consequence to reduction of consumer confidence leading to loss of product value [66] [10].

Adulteration is not a new practice, in 1820, German chemist Fredrick Accum published his book “A Treatise on Adulterations of Food and Culinary Poisons” in which he described methods of adulteration and analytical techniques to detect them, applied to tea, coffee, bread, beer, and pepper. He also highlighted the quality of the water used for food and drug production. His work still serves as a key element in the quality and safety schemes, at present days [67] [68][10].

3.1.2 Food Origin and Traceability

In Europe, the main concerns related to food authenticity are food origin and traceability. Food origin is concerned with the geographical place or region where the food or its ingredients were produced. This also includes the floral species, the traditional methods of production, and the animal diet and/or activity [60]. To protect the food product according to their geographical origin, EU is applied specific legislation for these authentication issues: (1) the protected designation of origin (PDO); (2) the protected geographical indication (PGI) labels for “mountain products” and “island farming products” (EU regulation No.1151/2012); and (3) “traditional specialty guaranteed” (TSG), a previously introduced term (EU regulation No. 509/2006) [10].

This special certification has the purpose to protect the reputation of the regional foods and promoting good practices in rural and agricultural activity. Those practices help the producers to obtain the best prices for their authentic products, reducing the unfair and misleading competition from non-genuine products, usually with inferior quality. With this regulation, there is a reinforcement on the consumer perception about the special quality of the product [69].

Another important definition is the term traceability, meaning the path covered from its first production, the geographical origin, to the final consumer. In developed countries, traceability has the legal obligation to follow the regulation (EU regulation No.178/2002) directed to ensure the quality management schemes and standards (i.e., HACCP, ISO 22000:2005, FSSC 22000, EurepGAP, BRC, etc.).

The large applications of the Food Authentication process with their economical profit managed to gain exponential attention in the field, especially for producers, manufacturers, sellers, consumers, agencies, and governments.

The food authentication processes managed to lead the attention of producers, manufacturers, sellers, consumers, agencies, and governments due to the large sums of money involved at the industrial level. In the last decades, there has been an increasing interest in scientific research supported by the coordination actions of the European Union, including “Food Integrity”, “MoniQa” and “TRACE” within HORIZON 2020.

The process of food authenticity reveals the unique features or the several attributes of food products in evaluation. These specific features add high value to the place and/or the production method, recognized by proper certifications granted by the aforementioned labels PDO, PGI and TSG as presented in Figure

8 and nowadays there are a growing number of products with the authenticity certification as shown in Figure 9.



Figure 8: **Official GIs logos of the European Union (PDO, PGI and TSG) according to Commission Implementing Regulation (EU) No. 668/2014.** Adapted from [10].

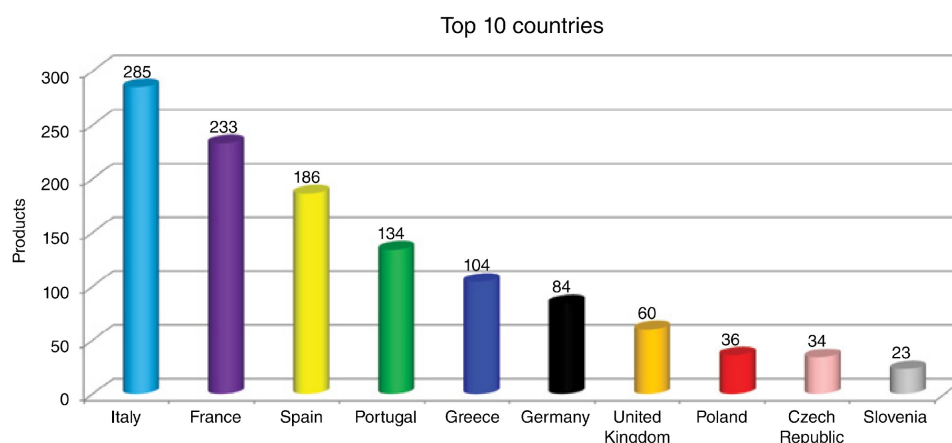


Figure 9: **Collective number of PDO, PGI and TSG products per EU country.** Adapted from [10].

The creation and implementation of specific laws, regulations, and guidelines regarding food authentication in conjunction with the exponential growth of the food industry to profits, did not prevent the occurrence of events concerning food contamination, adulteration, and fraud.

In the last years, the horse meat scandal, the melamine adulterated baby milk formula, and more recently the eggs contaminated with pesticides, lead to the alertness of the consumers concerning the quality and authenticity of the food [1]. These events are forcing governments, food agencies, and producers to inquire about new authentication approaches to apply to the food system to deal with food crime [1].

Due to these types of criminal events, the fields of authentication and traceability are under constant pressure regarding the continuous improvement and development of robust, efficient, sensitive, and cost-effective methodologies [3].

In the last decades, there has been continuous work on developing, improving, and standardizing methods and practices to assure food authenticity. Nowadays, new studies have proved the potential of metabolomics based approaches in the field of food science and nutrition [3]. Some of these examples are presented in Table 3.

The combination of analytical techniques in conjunction with bioinformatic tools for data analysis based on Metabolomics gives the capacity to face some of the important challenges in the field [3].

As presented in the previous chapter, metabolomics, is one of the main areas in the field of -omics techniques, and consists of the exhaustive study of the whole small metabolite composition of a particular system or organism. In practice, food metabolomics aims to analyze the chemical composition of the food metabolome, meaning some quite diverse compounds. For this, several studies choose more than one analytical approach and combine the interpretation of several results. This is the main reason for the lack of standard guidelines described for every step of the workflow in metabolomics studies as mentioned and explained in section 2.1 of chapter metabolomics.

Despite, there is no standard workflow due to the non-applicability of the aforementioned steps in all studies [9], in the food authentication process metabolomics follow a common number of steps where the analytical techniques and chemometric approach are selected according to the purpose of the study, as presented in Figure 3 (Chapter 2). [9].

Most food authenticity studies are based on Metabolomics approaches and are applied to food products such as alcohol drinks (beer, wine, spirits), honey, olive oil, milk and cheese products, meat, spices, and juices [70]. The studies reviewed in this work have focused on particular techniques, confirming the proof-of-concept, and also presented robust statistical results based on a large number of samples, contributing to creating predictive models. These studies are presented in this subsection along with the analytical techniques, data analysis tasks, the authentication purpose, the food product tested, and their references in Table3.

Table 3: **Metabolomics Approaches on Food Authentication.**

Food	Purpose	Sepctral Data	Data Analysis	Reference
Beer	Quality and Traceability	FT-IR	PCA, PLS-r	[71]
Juice	Adulteration	FT-IR	PCA, SIMCA	[72]
Milk	Adulteration	FT-IR	SIMCA, PLS-r	[73]
Indian Tea	Biological Origin	NIR, UV-VIS	PCA, K-means, ANN	[74]
Coffee	Quality, Origin, Traceability	UV-VIS	PCA, SIMCA, LDA	[75]
Tequila	Quality and Origin	UV-VIS	PCA, PLS-LDA	[76]
Salmon	Biological and Geographic Origin	NMR	SVC	[77]
Propolis	Harvest Origin	NMR	DA	[78]
Apples	Quality and Traceability	UV-vis, GC-MS, LC-MS	PCA	[79]
Berries	Biological Origin	LC-MS	Metabolite ID	[80]
Japanese Tea	Quality	UPLC/TOF-MS	PCA	[81]
Wine	Biological Origin	NMR	PCA, PLS	[82]
Honey	Biological and Geographic Origin	NMR	PCA, OPLS-DA	[83]

3.2 Wine

Wine is one of the food products with higher market value, and a recent census shows that is a widely consumed drink, being the second alcoholic drink most consumed after beer. According to the International Organisation of Vine and Wine (OIV) in 2017 USA, France, Italy, Germany, and China, were the five countries with half of the world's wine consumption (IOV, 2018). However, per capita, Portugal is the country with higher wine consumption (2016, OIV), mainly due to its consumer culture and traditional production methods. Nowadays, there is an increasing number of country producers, such as China, Peru, and Australia, owing to their high market value. Issues concerning authenticity, safety, and quality are relevant to consumers and producers [10]. Like other food products, such as honey or olive oil, wine is a product with authenticity certifications associated (Protected Designation of Origin - PDO, and Protected Geographical Indication - PGI). These certifications increase the commercial value of wine and ascribe the specific features related to its floral and geographical origins or attributes derived from specific production methods [10], [69]. These labels ensure the perception of the wine's authenticity, quality, and safety leading to the empowerment of the wine market value.

Wine and other food products (such as olive oil and honey) possess complex matrices, resulting from the combination of several modulator features like floral and geographical origin, environmental conditions, fermentation, age, and other modulator agents. Regardless of its complex matrices and its high market value, several adulterations events occur, making the insurance of wine authenticity a demanding work [1]. External factors (due to environmental constraints) can also compromise the food product characteristics.

Notwithstanding, these modifications do not compromise the authentication process [84].

Wine authentication is a process where validation methods are used to verify the label description of the wine bottle and also the verification of the followed standard guidelines on production, distribution, and commercialization. These authenticity validation processes entail the verification of stamps in the wine label bottles and/or in transportation boxes. Other authenticity validation methods are analytical approaches applied to verify the wine composition.

Safety and quality are related to the presence, absence, and quantification of specific elements related to the type of the product and safety is related to the control of limits of selected compounds, the determination of allergens; the detection of environmental contaminants and the assessment of the occurrence of toxins [3]. Quality means the certification of the quality and quantity of certain ingredients and other elements verified in the food product.

Adulteration is described as the deliberate alteration of the product by adding or removing ingredients from the food product, aiming to get an increase in profit. According to Food Standards Agency (FSA) (UK's Agency) there are two main types of adulteration: (1) the transaction of food which is unfit and potentially harmful; and (2) the deliberate misdescription of food [63], [64].

3.2.1 Metabolomics and Analytical techniques in Wine Authentication

In wine composition, there are a large variety of metabolites, organic acids and diols (tartaric acid, succinic acid, acetic acid, glycerol, 2,3-butanediol, lactic acid), amino acids (proline, alanine, arginine), carbohydrates (glucose, fructose), phenolic compounds, polyphenols (gallic acid, 2-phenyl ethanol, tyrosol) [85], and polyphenolic secondary metabolites (tannins, flavonoids, and anthocyanins) among other metabolites [86]. Through the analysis of these small molecules combined with chemometric tools for data analysis, it is possible to achieve different purposes like the discrimination, and prediction of specific features related to authentication [9] achieved by two different approaches: the fingerprinting and the profiling [15].

In wine authentication the use of metabolomics has the potential for solving two distinct hurdles, namely: (i) the determination of the origin (variety and geographical); (ii) testing for possible adulteration [10]. So depending on the study purpose, a different metabolomics-based approach may be followed [16].

The analytical techniques used in metabolomics approaches are robust, sensitive, and reproducible, leading to standardization processes aiming to validate more studies and the standardization of methodologies. NMR and MS coupled with others separation techniques (such as GC or LC) are the metabolomics methodologies more commonly used in wine authentication.

3.2.2 Variety and Geographic Origin in Wine

Due to its economic market value and consumer's importance/attention/interest, as mentioned in the above section, several different metabolomics approaches have been used to ensure the variety and geographic origin in wine authentication, over the last decade.

In 2009, two investigation groups presented the first applications to discriminate wine samples from different varieties. Both studies used ^1H -NMR and non-supervised models (PCA) and the results revealed the high potential and feasibility of fingerprinting in wine variety authentication[87][82].

To distinguish three different grapes varieties that have to grow under the same conditions, Portuguese cultivars ‘Trincadeira’, ‘Aragonês’, and ‘Touriga Nacional’, H-NMR was used combined with multivariate statistical analysis. In this approach non-supervised and supervised methods, respectively, principal component analysis (PCA) and partial least-squares-discriminant analysis (PLS-DA) reveal a highly reproducible and effective analysis in a wide range of compounds from the grapes metabolomes [88].

A recent study with a large dataset (272 samples) used GC coupled with MS to discriminate white wine German varieties. The variety origin was achieved using PLS-DA and Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) models where samples were correctly classified with high scores. External samples were also used and the models reveal an accuracy of 97% for Silvaner, 93% for Riesling, 91% for Pinot Gris/Blanc, and 80% for Muller-Thurgau for classification. Other results are present in the study and the major conclusions reveal indicate the selected strategy as reliable and relevant for white wine varietal classification [89].

The combination of different techniques allied with multivariate analysis is also used in some studies to determine geographical origins. In this case, the group used the combination analytical techniques MS and NMR, to discriminate between commercial wines of Greco Bianco grape variety, produced in two different Italian regions (Calabria and Campania). The metabolomics data analysis, under PLS-DA, supervised models reveal a good differentiation of the wine samples according to their geographical origin [90].

Another metabolomics combination approach using NMR and solid-phase microextraction followed by gas chromatography-mass spectrometry (SPME-GC-MS), was applied to classify white wine samples from different PDOs in Galicia regions Spain, with multivariate analyses. The results of the Principal Component Analysis-Discriminant Analysis (PCA-DA) method, reveals an improvement in metabolomics discrimination of the geographical regions resulting from the combination of the analytical methodologies [91].

The use of MS wine authentication is another selected metabolomics wine approach. A study from 2015, used as metabolomics markers, anthocyanins from LC-MS data, to classify Argentinian red wines according to variety and geographical origin. The alternating least-squares and discriminant unfolded partial least-squares (D-UPLS) models allowed the identification of metabolomics compounds responsible for both types of wine varieties and geographic origins [92].

3.2.3 Vintage and Age Determination

Wine vintage and wine aging are frequently associated with wines possessing a high-quality standard classification. A vintage wine has a set of unique features. These characteristics are the product of environmental factors (e.g. grapes harvested in a specific year and/or location) and production techniques.

After the harvest, wine may be put through an aging process. For example, vintage Port wine is considered by many as the jewel of Port Wine and is the only Port wine aged in bottle IVDP, 2018.

The capability to improve the quality of wine through the aging process requires time and financial investment (usually five years) and is influenced by biological factors (such as grape variety, vintage, wine region) and human modulator agents (viticultural practices and winemaking process)[93]. A wine denoted with PDO certifications, vintaged, aged, and presenting high-quality features, commonly has a high market value and its price can increase with time, so the identification of metabolomics markers for their characterization and authentication has a significant economic impact.

A study where these issues were evaluated was conducted by Cuadros-Iñostroza et al., in 2010, using a large dataset with 400 samples from different red monovarietal wines. This work followed a metabolomics untargeted and target using approaches using UPLC-FT-ICR-MS data. The results of PCA, HCA, and LDA models allow for discrimination of the wines, according to the type of cultivar, vintage, and region [94].

Using ^1H -NMR and FT-IR data, collected from different varieties of german white wines, the models (PCA, PLS-DA) reveal significant results to identify different metabolomics compounds and also highlighted differences in different vintages wines by combining [88].

Another study focuses on a metabolomics targeted approach to distinguish vintage wines and the aging process by using PCA, and PLS-DA models on H-NMR collected data [95].

3.2.4 Adulteration

The non-authorized addition of sugar or concentrated grape during or before the fermentation phase, to increase the natural ethanol, is one of the adulteration events that can occur to achieve a higher market price. To test this particular case, Imparo and their colleagues developed a profiling approach using ^1H -NMR to identify adulterated red wine with glucose syrup. The group used a multilinear regression method on a selected set of NMR variables and used the LDA model to get a classification of good separation of different mixture classes [96].

Another case, considered by many countries as wine adulteration practice is the addiction to red color in wines, specifically the enrichment of wines with anthocyanins. Both samples (originals and adulterated) were analyzed by FT-NIR and ^1H -NMR metabolomics approaches. The results from the multivariate analysis, PCA and PLS-DA, reveal a better classification model on NMR data compared to NIR data with an efficacy validation of 95% focusing on the 6.5 and 9.5 ppm region for aromatic compounds [97].

In Table 4 is presented a selection of the most relevant studies in wine authentication metabolomics-based approaches using NMR, MS, and IR techniques followed by data analysis.

Table 4: **Metabolomics Approaches for Wine Authentication and Traceability.**

Purpose of analysis	Techniques	Preprocessing	Data Analysis	References
Geographical origin	UV, Vis, NIR, MIR	-	PCA, PLS-DA, SIMCA	[98]
	UV/Vis, NIR, MIR	SNV, 2nd Derivative smoothing	PCA, LDA, SIMCA	[99]
	¹ H-NMR	Bucketing, normalization, centering	PCA, DA, HCA	[100]
	¹ H-NMR	Bucketing, normalization, centering	PCA, PLS-DA	[82]
	NIR-vis	SNV, 2nd Derivative transformation, correction	PCA, PLS-DA, SLDA	[101]
	¹ H-NMR	Mean-centering, scaling	PCA, PLS, OPLS	[102]
	FT-IR	Transformation to absorbance unit	PCA, CA, LDA, CART	[103]
	FT-IR	Normalization, 2nd derivative	PCA	[104]
	NIR, Vis	SNV, 2nd Derivative transformation, smoothing	PLS	[105]
	UV, IR	SNV, Smoothing, 2nd derivative	PCA, LDA, SIMCA	[106]
	UV	Autoscaling	PCA, HCA, PLS-DA	[107]
	¹ H-NMR	Baseline correction, bucketing	PCA	[108]
	¹ H-NMR	Baseline correction, normalization, mean-centering, scaling	PCA, PLS-DA	[109]
	FT-NIR	SNV, MSC, 2nd Derivative	PCA, DA, PLS-DA	[110]
Geo. origin, vintage	¹ H-NMR	Normalization, correction of chemical shift, scaling	PCA, ECVA, iECVA	[111]
Bot. and geo. origin	¹ H-NMR	Bucketing	PCA, LDA, MANOVA, NCM	[112]
	¹ H-NMR	Baseline correction, bucketing, centering, scaling		[87]
Variety origin, vintage	UPLC-FT-ICR-MS	Normalization, log-transformation, scaling, centering	HCA, PCA, LDA	[94]
Variety origin	ESI-FT-MS	Normalization	PCA, LDA	[113]
	HPLC-TOF-MS	Mean-centering, log-transformation	PCA, PLS-DA	[114]
	CG-MS	Autoscaling	PCA, SLDA	[115]
	FT-MIR	Centering, scaling	PCA, LDA	[116]
	¹ H-NMR	Bucketing, normalization, COW, mean-centering	PCA, OPLS	[82]
Age Origin	MIR	-	PCA, LDA	[117]
	UV-vis	Mean centering, Smoothing, 1st and 2nd derivatives, SNV, Orthogonal Signal Correction	PLS-r	[118]
	¹ H-NMR	-	PCA, PLS	[119]
Adulteration	¹ H-NMR	Bucketing	LDA, ANN	[96]
Origin	MIR	SNV	PCA, PLS	[120]
Origin	UV-vis	1st derivate	PCA, SIMCA.	[121]

3.3 Honey

Honey is an extensively investigated food matrix [122], a naturally sweet food product produced by honey bees (*Apis mellifera* L.) synthesized from various flower nectars and it is known for its therapeutic properties and antibacterial action since ancient times [123]. The composition of honey consists of an aqueous solution (20%); two predominant sugars, fructose, and glucose in approximately equal amounts (about 31–38%); and also small quantities of organic acids, amino acids, minerals, enzymes, and pollen grains derived from honey collection Codex Alimentarius, 2001.

Many factors can influence the basic properties of honey including the nectar provided by different

plants and the harvesting conditions compromising the authentication process [84]. Human handling can also affect the quality and composition of the honey, leading to adulterations events, such as, the overfeeding of bees with sucrose, harvesting before bee's maturity and the addition of sugar syrups (authentication process issues) [84] [124].

3.3.1 Honey Authenticity and its related issues

Unifloral kinds of honey are the most expensive and a higher price is also commanded by honey produced in a specific geographical location. So the floral and/or geographical origin predetermines the price of honey. Due to the growing global trade and consequent increase in market value, the production of honey is susceptible to fraud. Therefore authorities, such as the E.U. created the Honey Directive and other legislation and started to implement international standards for honey authenticity.

The two major concerns in honey authenticity are concerned with: (i) The origin of the honey (floral and/or geographical origin), and (ii) the honey production process. These concerns are similar to other food products that evolve with food authentication. Like other food products authentication, profiling, and fingerprint approaches based on metabolomics data, have shown great results in the field of floral and geographical origin, adulteration detection for honey, and other events in table II.

3.3.2 Classification for Floral Origin in Honey

In 2010, Schievano and his research team, have used the NMR spectroscopy for the floral origin differentiation of honeys types, followed by other research studies years later.

More recently, a high-resolution H-NMR study used a larger number of samples ($n=328$) combined with chemometric data analysis to also identify different the floral origins of honey types and to quantify 13 of the selected components of honey [125].

In this particular study, the major goal was to differentiate 118 monofloral Italian honey types (acacia, orange, eucalyptus, linden, chestnut, ...), followed by the identification of specific key components for each monofloral origin allowing the proper classification of each honey floral type origin. The statistical data analysis was performed using both unsupervised (PCA) and supervised models (PLS-DA and OPLS-DA), whereby the supervised techniques demonstrated a high power efficiency for the prediction of floral origin.

Schievano et al. obtained a reasonable discrimination. This might be due to different sets of samples (with distinct intrinsic variations) and/or different pre-processing techniques as well as the potentially limited meaning of PCA in this case[126].

More recently, a high-resolution ^1H -NMR study use a larger number of samples ($n=328$) combined with chemometric data analysis to also identify different the floral origins of honey types and to quantify 13 of the selected components of honey [125].

3.3.3 Classification for Floral and Geographical Origin

Another recent study used H-NMR metabolomics approaches and chemometric tools to determine the floral origin of Finland honey and also their geographical origin, presented by Kortensniemi and their colleagues in 2016. In this investigation, key markers of authentication for different honey types were identified (buckwheat, clover, dandelion, heather, Himalayan balsam, honeydew, linden, lingonberry, multiflora) and the fingerprints of some samples with full discrimination were accomplished with unsupervised (PCA) and supervised (OPLS-DA) models. In the PCA, melezitose, glucose, and fructose explained 83% of the variation, while the origin-specific non-saccharide metabolites were the discriminatory compounds in OPLS-DA loadings [127].

To classify floral and geographical origin and to identify fraudulent honey productions, there are other studies with relevant results using metabolomics data approaches combining chemometric tools for data analysis where the data came from other analytical techniques. Gök in 2015, differentiated and classified 120 honey Turkish types by using FTIR spectroscopy techniques combine with unsupervised models (PCA) and hierarchical cluster analysis [128]. Woodcock in 2007 used NIR spectroscopy to evaluate the potential of the technique to confirm the geographical origin of several samples of honey (Irish, Mexican, Irish, Argentinean, Czech, Hungarian, and Spanish honey types) [129].

In Table 3.4 is presented a selection of the most relevant studies in wine authentication metabolomics-based approaches using NMR, MS, and IR techniques followed by data analysis.

Table 5: **Metabolomics Approaches for Honey Authentication and Traceability.**

Purpose of analysis	Techniques	Preprocessing	Data Analysis	Reference
floral origin	¹ H-NMR	Shift correction normalization, 1st derivative, mean-centering, scaling	PCA, HCA, KNN, SIMCA, PLS-DA	[130]
	HPLC-UV	Integration (s/n <3), scaling to zero mean and unit variance	PCA, LDA	[131]
	NIR	1st and 2nd derivative, Savitzky-Golay smoothing bucketing, normalization	PCA, MD-DA, BP-ANN	[132]
	¹ H-NMR	bucketing, normalization	PCA, O2PLS-DA, SVM	[126]
	¹ H-NMR	bucketing normalization, DmodX	PCA, PLS-DA	[133]
	FT-IR	-	PCA	[134]
	2D H-NMR	Normalization	PCA, GDA	[135]
	NIR	-	PLS, BONN	[136]
	SIFT-MS	-	SIMCA	[137]
	GC-MS/sensory analysis	-	PCA, CA	[138]
	¹ H-NMR	-	PCA	[139]
	¹ H-NMR	-	GP	[140]
	¹ H-NMR	-	PCA	[141]
	¹ H-NMR and	-	PLS	[142]
	and voltammetric-electronic-tongue (VET)	-		
Bot. and Geo. origin Floral origin	GC-MS	Normalization	OPLS-DA, OPLS-HCA, SIMCA	[143]
	Raman	MSC	PCA, PLS-DA, SVM	[144]
	NIR	1st and 2nd derivative, SNV	SIMCA, UNEQ, POTFUN	[145]
	FT-IR	SNV, 1st and 2nd derivative	FADA, PLS, PCA	[146]
	NIR	None, different derivatives, SNV	PLS	[147]
	¹ H-NMR	Mean centering, unit variance	PCA, PLS-DA	[148]
	H-NMR	None, different derivatives, SNV	PLS-DA, GP, PLS-GP	[149]
	FT-IR	SNV, 1st and 2nd derivative	PLS, FDA, SIMCA	[150]

Adulteration	GC-GC-TOFM	-	LDA, DPLS, SVM	[151]
	Raman	AirPLS	PLS-DA	[152]
	¹ H-NMR	None	FA, GDA	[153]
	NIR	Smoothing, SNV	PCA, LS-SVM, SVM, BP-ANN, LDA, KNN	[154]
	FT-IR	-	PCR, PLS	[155]
	FT-IR	Baseline correction, normalization, peak correction, outlier removal	PCA, Linear Discriminant Analysis (LDA)	[156]

3.4 Challenges, limitations and possible paths for further improvement.

The rapid growth of metabolomics-based approaches in food authentication in conjunction with advances of the analytical techniques and data analysis, permits to inform, discriminate and predict several characteristics of food products.

These progresses can help to solve problems in safety and quality arenas and at the same time provide important information to the food industry.

Despite a distinct range of metabolomics analyses performed a diversified set of food items, most of the studies can be only considered as discriminative (Table 4 and Table) with very few compounds identified.

Nowadays, there are more sensitive, precise and robust techniques which can analyse several features, however like mentioned before the large amount of data can make difficult the analysis.

It is necessary not only to implement new data analysis tools but also to validate and to reproduce and standardize the procedures. The major hurdles in metabolomics approaches concerning profiling and fingerprint in food authentication are: the reduced number of public databases with the information concerning the distinct food metabolites composition, and the respective steps of data analysis.

Another important challenge is the lack of data regarding metabolites produced based on external environmental factors, such as the weather changes and the different seasons conditions, which may affect the selection of metabolites used as biomarkers for the food authentication.

3.5 Food Products studied in Data Analyses and their Authentications issues

In the context of this dissertation, chemometrics tools such as preprocessing methods, multivariate statistical analyses (MSA), and several machine learning (ML) approaches were utilized on metabolomics data from two high-value food products (Wine, Wine musts and Honey samples), and with PDO importance. These tools aided in chemical identification, standardization, classification, and prediction of these food product variables for authenticity issues, mainly related to origin classification problems.

In the first study the purpose of the data analyses was to classify *Cabernet Sauvignon* wine samples according to their harvest year, to determine which and how the climatic influence factors influence the ^1H -NMR data, and to predict the age of bottle storage of the samples.

In the second study the focus was on the variety classification of wine musts samples of four native Portuguese grape varieties and one international red grape variety, all produced in two Portuguese PDO regions. This study also aimed to interpret how the features with higher importance modulate the classification model.

The last study classified geographical and flora Brazilian honey samples from eleven regions with different flora origins. Model interpretability for the model's whit higher performance and the relation between floral and geographical origins were explored.

The main studies in literature related to these analyses are presented and discussed in each of the presented study in their correspondent chapter.

Harvest, Age of bottle and Climatic factors prediction using statistical and machine learning approaches on metabolomics data of Brazilian Cabernet Sauvignon wines

Abstract

Wine possesses a complex matrix resulting from the combination of several factors such as grape variety(ies) and terroir, among others. This study contributes to understand the influences of the age of bottle storage and of climatic factors on the wines' chemical composition. A multidimensional metabolomics data set of ^1H -NMR spectra and metabolites concentrations from 30 Brazilian Cabernet Sauvignon wine samples, with distinct harvest years, were collected and analyzed. Metabolomics based-approaches were combined with quantitative statistical and machine learning tools. Qualitative analysis revealed that wines with different harvest years presented similar composition of metabolites. Univariate ANOVA shows that 30 relevant ^1H -NMR peaks are distributed in three regions, e.g., phenolic acids, sugars, and amino acids and 5 amino acids allowed to discriminate wine samples according to their harvest years. Supervised models were used to predict the harvest years from ^1H -NMR data, with an accuracy of 92%. Using linear regression models, the data allowed to discriminate ^1H -NMR resonances that were most influenced by the age of bottle storage and the climatic factors, being the temperature the factor with most influence on this case study.

4.1 Introduction

Winemaking has a high cost and is a complex process based on the combination of grape varieties, terroir effects, fermentation, harvest timing/ methods, storage conditions and age, as well as many other conditions capable of influencing the wine quality and authenticity [1], [157], [158]. Complementary to the challenges of the winemaking process and wine metabolome analysis, the map of wine production has expanded to regions, such as Argentina, Brazil, Chile, China, United States, South Africa, New Zealand

and others, considered as the New Worlds for wine production [159].

Cabernet Sauvignon (CS) is one of the best-known red wine varieties and it is produced in many regions of the world. This cultivar is known for having a long life-cycle (around 214 days), requiring higher temperatures in all phenological stages. Following climatic conditions similar to the ones in Europe, the produced wines have a good quality [93]. Based on different climatic conditions, such as high insulation, humidity, rain, and temperature, maturation parameters can be changed, influencing the times of harvest and storage and ultimately the wine quality [93].

Complementary to the control of the wines' quality and their authentication, studies focusing on the factors influencing wine's chemical composition have been carried out, following the global climatic changes in the wine production regions [157]. Due to the complexity of winemaking, several chemical and omics approaches have been successfully applied to the wine science to better understand the mutual wine composition interferences [85], [157], [160]–[163]. Regardless of these advances on wine sciences, the influence of each individual (a)biotic factor in the wine chemical profiles remains, in most cases, unclear and not well established [85], [157], [160], [161].

The combination of different metabolomics tools, such as Nuclear Magnetic Resonance (NMR) and other spectral techniques, leads to robust, efficient, sensitive, and cost-effective analytical methodologies, resulting in large amounts of data on wine metabolomics profiles. ^1H -NMR has the sensitivity to identify chemical shifts in each metabolite per sample, as a result of structural and environmental changes. Even small environmental changes can lead to a significant effect on the spectrum. This means that this technique has the capability to distinguish two similar molecules with identical spectra [164]. The resulting data provide an attractive resource for studying the coordinated behavior of wine chemical profiles. However, to cover the needs of data processing and data mining it is necessary to combine metabolomics based approaches with powerful computational tools, such as multivariate data analysis or machine learning tools. These techniques have shown the potential to face challenges in food science, and, more specifically, in wine science. Statistical multivariate analysis and machine learning employ several methods that allow simultaneously the study of many metabolomics features to identify relations between them, which may increase the quality of the data evaluation [18].

In our study, an analysis was performed considering the influence of combined factors such as the harvest years, the age of bottle storage, and a selected set of climatic factors on metabolomics datasets of CS wines. For that, a comprehensive analysis was conducted using bioinformatics tools, including multivariate statistical analysis and machine learning, over untargeted metabolomics data from ^1H -NMR spectra and targeted metabolomics data from Gas Chromatography (GC) (amino acids) for metabolite quantification. The wine samples were produced in the same location in southern Brazil, in distinct years, with similar climatic conditions, and different ages regarding the bottle storage. Based on the combination of the metadata available and metabolomics datasets, the study intends to discriminate and predict wine samples with distinct harvest years and also to identify the influence of the age of bottle storage and of a set of selected climatic factors.

4.2 Results and Discussion

This study contemplates the analysis of a multidimensional metabolomics data set of 30 Brazilian CS wine samples with distinct harvest years (i.e., 1992, 1994, 1996, 1999, 2002, and 2005) and also distinct ages of storage (i.e., 1, 4, 7, 10, 12, and 14 years). Firstly, an untargeted and qualitative metabolomics analysis was applied to the ^1H -NMR spectra data set, followed by a target and quantitative analysis of 14 specific metabolites by GC-FID (amino acids). As mentioned in the methods section below, each group of wine samples was produced in distinct years, in the same geographic origin and kept under similar conditions of light, humidity, and temperature over the time of storage. Metabolomics based-approaches were used in all samples at the same time and the resulting data were combined with univariate and multivariate statistical analyzes, and machine learning tools to answer to the aims of the study.

4.2.1 Metabolites annotation from ^1H -NMR resonance peaks – qualitative and untargeted metabolomics analysis

To understand the effect of the harvest year on the wine's chemical composition two qualitative analyses were carried out: one using the ^1H -NMR data from all wine samples (considering all the harvest years) and the other focused on each wine sample group (wine samples with the same harvest year).

The qualitative analysis of the ^1H -NMR profiles of the CS wine study model revealed that samples produced over the years investigated showed the same set of annotated metabolites. A total of 37 metabolites with a high score for annotation were detected in the wine samples, being all of them important for wine composition. The annotation of these metabolites was based on a new algorithm developed by the host group [165], by comparing the wine samples spectra with the spectral references available in ^1H -NMR libraries from free databases. The algorithm takes into account a set of parameters for ^1H -NMR spectra acquisition, as the type of solvent, the frequency of acquisition, and the organism. Additionally, glutathione, carnitine, 2'-deoxyguanosine monophosphate (dGMP), isoleucine, thiamine monophosphate, deoxycarnitine, and some biogenic amines (e.g., putrescine, cadaverine, spermine spermidine, histamine, tyramine, and tryptamine) were tentatively annotated according the algorithm used, but require further investigations by 2D-NMR experiments and other analytical techniques to prove its occurrence in the wine samples studied.

The main metabolites found in this work have been previously annotated in CS wines, such as alcohols, sugars, acids, phenolic compounds, esters, aldehydes, and some amino acids [85], [166], [167]. From the annotated metabolites, several of them are relevant for wine composition and they have been previously reported in wine samples using other analytical techniques. Nonetheless, they are not so commonly found by using ^1H -NMR spectroscopy. For instance, sorbitol and myo-inositol, naturally present in grapes, are polyhydric or sugar alcohols which may play a sensory and potential balance role in winemaking [93] particular in dessert wines [168] and are frequently found in wine by other analytical techniques, but only in minor amounts [93].

Several free amino acids were annotated in our study, such as lysine, alanine, leucine, serine, glutamic acid, tyrosine, cysteine, phenylalanine, aspartic acid, threonine, histidine, methionine, glycine, aminobutyric acid, proline, and others. These metabolites are important for the fermentation process allowing the normal activity and grow of yeast, also to originate vitamins and to integrate the flavor and aroma of wine [169].

With these results, our study brings a new highlight to the use of the ^1H -NMR technique for the annotation of relevant metabolites with lower concentrations to improve the wine chemical profile characterization, classification and the winemaking process as well. In Table 6, the annotated metabolites with higher score values concerning importance and relevance in wine composition are presented. Bold type indicates amino acids quantified in this study.

Table 6: **Metabolites and their ^1H chemical shifts annotated by 500 MHz ^1H -NMR with higher scores for frequency, organism and solvent.**

Compound	Molecular formula	δH , ppm (multiplicity, J, Hz, assignment))
Lysine	$\text{C}_6\text{H}_{14}\text{N}_2\text{O}_2$	1.37; 1.40; 1.43; 1.46 (m, γ -CH ₂); 1.49; 1.52; 1.55; 1.65; 1.68; 1.71 (m, δ -CH ₂); 1.74; 1.77; 1.80; 1.83; 1.86 (m, β -CH ₂); 1.90; 2.98; 3.03(t, ϵ -CH ₂); 3.64; 3.67; 3.71 (t, 6.07, α -CH)
Serine	$\text{C}_3\text{H}_7\text{NO}_3$	3.80; 3.83 (dd, 5.51/3.79, α -CH); 3.86; 3.89; 3.92; 3.95; 3.98; 4.01 (m, β -CH ₂)
Methionine	$\text{C}_5\text{H}_{11}\text{NO}_2\text{S}$	2.05; 2.08; 2.11 (m, β -CH ₂); 2.14 (s, SCH ₃); 2.17; 2.60; 2.63 (t, 6.2, γ -CH ₂); 2.66; 3.77 (t, 7.57, α -CH); 3.80; 3.83
Tryptophan	$\text{C}_{11}\text{H}_{12}\text{N}_2\text{O}_2$	3.25; 3.28 (dd, 15.31/8.08, β -CH); 3.34; 3.43; 3.46 (dd, 15.33/4.77, β' -CH); 3.49; 3.52; 4.01; 4.04 (dd, 8.12/4.02, α -CH); 7.17 (m, C ₅ H-ring); 7.28 (m, C ₆ H-ring); 7.30 (s, C ₂ H-ring)
Phenylalanine	$\text{C}_9\text{H}_{11}\text{NO}_2$	3.06 (d, 7.88, β' -CH ₂); 3.22 (m, β' -CH); 3.25; 3.28; 3.95; 3.98 (dd, 7.86/5.31, α -CH); 4.01; 7.28; 7.30 (d, 6.96, C ₂ H, 6H-ring); 7.35 (m, C ₄ H-ring); 7.37 (m, C ₃ H, 5H-ring)
Tyrosine	$\text{C}_9\text{H}_{11}\text{NO}_3$	3.00 (dd, 14.6, β' -CH); 3.03; 3.06; 3.19 (dd, 7.88, β' -CH); 3.22; 3.89 (dd, 5.16, α -CH); 3.92; 3.95; 7.17 (d, 7.97, C ₂ H-ring)

Sucrose	$C_{12}H_{22}O_{11}$	3.43; 3.46 (t, 9,29, G4H); 3.49; 3.52; 3.55 (dd, 9,89/3.87, G2H); 3.57; 3.64 (s, F1H); 3.71; 3.74 (t, 9.55, G3H); 3.77; 3.80; 3.83; 3.86 (dd, 6.46/3.25, F6H); 3.89 (dd, 6.51/3.65, F5H); 3.92; 4.01; 4.04 (t, 8.57, F4H); 4.08; 4.21 (d, 3,88, F3H)
Fructose	$C_6H_{12}O_6$	3.52; 3.55; 3.57 (m, C1H); 3.60; 3.64; 3.67; 3.71; 3.74; 3.77 (dd, 12.72/1.05, C6H); 3.80; 3.83; 3.86; 3.89 (dd, 9.98/3.45, C4H); 3.92; 3.98; 4.01; 4.04 (m, C5H); 4.08 (m, C5H); 4.12 (m, C3H)
Glycerol	$C_3H_8O_3$	3.52; 3.55 (m, C1H2); 3.57; 3.60; 3.64 (m, C3H2); 3.67; 3.74; 3.77 (t, 6.51, C2H); 3.80
Leucine**	$C_6H_{13}NO_2$	0.91; 0.94 (d, 6.1, δ -CH); 0.98 (d, 6.1, δ -CH); 1.62; 1.65; 1.68 (m, γ -CH); 1.71 (m, β -CH2); 1.74; 3.67; 3.71 (t, α -CH); 3.74
Myo-inositol	$C_6H_{12}O_6$	3.22 (t, 9,49, C5H); 3.25; 3.28; 3.49; 3.52; 3.55; 3.57 (dd, 9.97/2.88, C1H, C3H); 3.60 (t, 9,98, C4H, C6H); 3.64; 4.01; 4.04 (t, 2,85, C2H); 4.08
Proline	$C_5H_9NO_2$	1.94; 1.96; 1.99; 2.02 (m, γ -CH); 2.05; 2.08 (m, β -CH); 2.11; 2.29; 2.33; 2.36; 2.35 (m, β' -CH); 2.39; 3.28; 3.31; 3.34 (m, δ -CH); 3.37; 3.40 (m, δ' -CH); 3.43; 4.08; 4.10 (m, α -CH)
Mannose	$C_6H_{12}O_6$	.34; 3.37 (dd, 9.61/6.36, CH5) 3.40; 3.52; 3.55; 3.57 (t, 9.71, CH4); 3.60; 3.64 (m, CH3); 3.67; 3.71; 3.74; 3.77; 3.80; 3.83; 3.86; 3.89 (dd, 4.15/2.13, CH6); 3.92; 3.95
Maltose	$C_{12}H_{22}O_{11}$	3.22; 3.25; 3.28 (dd, 9.45/8.06, CH3); 3.37; 3.40 (t, 9.57, CH12); 3.43; 3.52; 3.55; 3.57 (m, CH6/CH10); 3.60; 3.64 (m, CH1); 3.67 (m, CH11); 3.71 (m, CH13); 3.74; 3.77 (m, CH2); 3.80; 3.83 (m, CH16); 3.86; 3.89 (dd, 12.30/1.95, CH7); 3.92; 3.95 (m, CH2); 3.98 (m, CH2); 4.01; 4.62; 5.21 (d, 3.81, CH4); 5.25

Adenosine	$C_{10}H_{13}N_5O_4$	3.80; 3.83; 3.86 (dd, 12.89/3.53, CH15); 3.89; 3.92 (dd, 12.80/2.77, CH15); 3.95; 4.28; 4.31 (q, 3.32, CH5); 4.41 (dd, 5.01/3.46, CH4); 4.46
Uridine	$C_9H_{12}N_2O_6$	3.77; 3.80 (dd, 12.73/4.41, CH5); 3.83; 3.89; 3.92; 3.95; 4.12; 4.21 (dd, CH3); 4.25; 4.31; 4.34 (dd, 12.02/3.99, CH4); 4.38
Guanosine	$C_{10}H_{13}N_5O_5$	3.80; 3.83 (d, 4.07, CH1); 3.86; 3.89; 3.92; 3.95; 4.25 (q, 3.75, CH2); 4.28; 4.41 (dd, 5.11/3.88, CH3); 4.46
Cytidine	$C_9H_{13}N_3O_5$	3.77; 3.80 (dd, 12.90/4.53, CH12); 3.83; 3.89; 3.92 (m, CH12); 3.95; 4.08; 4.12 (m, CH9); 4.17; 4.21; 4.28; 4.31 (t, CH8)
2'-deoxyuridine	$C_9H_{12}N_2O_5$	2.36; 2.39 (m, CH10); 2.42; 2.44; 3.71; 3.74; 3.77 (dd, CH12); 3.80; 3.83 (dd, CH12); 3.86; 4.01; 4.04 (m, CH8); 4.08; 4.41; 4.46 (dt, 6.45, 4.03, CH9); 4.49
Arabitol	$C_5H_{12}O_5$	3.55; 3.57 (dd, 8.29/1.55, CH1); 3.60; 3.64; 3.67 (m, CH6); 3.71; 3.74; 3.77; 3.80; 3.83 (dd, 11.74/2.63, CH3); 3.86; 3.89; 3.92 (m, CH2); 3.95
Sorbitol	$C_6H_{14}O_6$	3.57; 3.60; 3.64 (m, CH4); 3.67; 3.71 (d, 286, CH6); 3.74 (m, CH3); 3.77; 3.80 (d, 2.91, CH1); 3.83 (m, CH2/CH5); 3.86
Trehalose	$C_{12}H_{26}O_{13}$	3.40; 3.43 (t, 9.45, CH4/CH10); 3.46; 3.60; 3.64 (dd, 9.91/3.83, CH2/CH8); 3.67; 3.71; 3.74; 3.77 (m, CH2/CH16); 3.80; 3.83; 3.86; 3.89; 5.21 (d, 3.79, CH1/CH7)
Salicin	$C_{13}H_{18}O_7$	3.46; 3.49 (m, CH2); 3.52; 3.57; 3.60; 3.64; 3.67; 3.71 (dd, 12.00/5.40, CH11/CH12); 3.74; 3.77; 3.80; 3.89 (dd, 12.01/2.10, CH11/CH12); 3.92; 3.95; 5.09; 5.12; 7.13; 7.17 (dd, CH14); 7.35 (dd, 7.43/1.33, CH20); 7.37
Methionine	$C_5H_{11}NO_2S$	2.05; 2.08; 2.11 (m, β -CH2); 2.14 (s, SCH3); 2.17; 2.60; 2.63 (t, 6.2, γ -CH2); 2.66; 3.77 (t, 7.57, α -CH); 3.80; 3.836

Argininosuccinic acid	C ₁₀ H ₁₈ N ₄ O ₆	1.62; 1.65; 1.68 (m, γ -CH ₂); 1.71; 1.74; 1.77; 1.96; 2.36; 2.39; 2.48; 2.51(dd, 16.13/9.61, CH14); 2.54; 2.57; 2.63; 2.66; 2.69; 2.78; 2.81 (dd, 16.10/3.43, CH14); 2.84; 2.87; 3.25; 3.28 (t, 6.87, δ -CH ₂); 3.34; 3.60; 3.64; 3.67; 3.74; 3.77 (t, 6.11, α -CH); 3.80; 4.21; 4.25 (dd, 9.58/3.30, CH4); 4.28; 4.31
2'-deoxyinosine	C ₁₀ H ₁₂ N ₄ O ₄	2.78; 2.81; 2.84 (m, CH ₃); 2.87; 3.74; 3.77; 3.80; 3.83; 3.86 (m, CH16); 4.17 (m, CH5); 4.21; 4.62 (m, CH4)
Valine	C ₅ H ₁₁ NO ₂	0.98 (d, 7.0, γ -CH ₃); 1.04 (d, 7.0, γ' -CH ₃); 1.07; 2.21; 2.27 (m, β -CH); 2.29; 3.57 (d, α -CH); 3.60
Cysteine	C ₃ H ₇ NO ₂ S	2.98; 3.00; 3.03; 3.06 (m, β -CH ₂); 3.92; 3.95; 3.98 (dd, 7.88/3.11, α -CH)
Alanine	C ₃ H ₇ NO ₂	1.46 (d, 7.26, β -CH ₃); 1.49; 3.74; 3.77 (q, 7.22, α -CH)
Choline	C ₅ H ₁₄ NO	3.19 (s, N(CH ₃) ₃); 3.49 (dd, 5.79/4.14, CH ₃); 3.52; 4.01; 4.04 (ddd, CH ₂); 4.08
Threonine	C ₄ H ₉ NO ₃	1.30 (d, γ -CH ₃); 3.55; 3.57 (d, α -CH); 4.21; 4.25 (m, β -CH ₂); 4.28
Glutamate	C ₅ H ₉ NO ₄	1.99; 2.02 (m, β -CH ₂); 2.05; 2.08; 2.11; 2.14 (m, β' -CH ₂); 2.17; 2.27; 2.29 (m, γ' -CH ₂); 2.36 (m, γ -CH ₂); 2.39; 2.42; 3.71 (t, 7.16, α -CH); 3.74; 3.77
N-Acetyl-L-aspartic	C ₄ H ₇ NO ₄	2.02 (s, CH ₃); 2.48; 2.51 (dd, 15.64/9.98, β -CH ₂); 2.54; 2.66; 2.69 (dd, 15.64/9.98, β' -CH ₂); 2.72; 2.75; 4.38; 4.41 (d, 6.75, N(CH ₂))
Histidine**	C ₆ H ₉ N ₃ O ₂	3.19 (dd, 15.53/7.71, β -CH ₂); 3.22 (dd, 16.08/4.89, β -CH ₂); 3.25; 3.28; 3.95; 3.98 (dd, 7.71/4.97, α -CH) ; 4.01
Ethanolamine	C ₂ H ₇ NO	3.80 (m, OCH ₂); 3.83
Alpha-aminobutyric	C ₄ H ₉ NO ₂	0.94; 0.98 (t, 7.56, γ -CH ₃); 1.90 (m, β -CH ₂); 3.71 (dd, α -CH); 3.74
Asparagine	C ₄ H ₈ N ₂ O ₃	2.81 (m, β' -CH ₂); 2.84 (m, β -CH ₂); 2.87; 2.98; 3.95; 3.98; 4.01 (dd, 7.67/4.25, α -CH)
Glycine	C ₂ H ₅ NO ₂	3.52 (s, α -CH)

4.2.2 Discrimination of the harvest year

4.2.2.1 Amino acids quantification – targeted metabolomics analyses

Based solely on the annotated metabolites by ^1H -NMR data, it was not possible to discriminate wine samples according to their harvest year, because the wine samples presented the same set of identified metabolites. This result was expected since the grape variety, the region of production, and all the wine-making processes (including fermentation, storage, and preservation processes) were the same over the harvests. Still, despite the identified metabolites being the same, the concentrations thereof and the wine ^1H -NMR profiles have differences between the samples from distinct harvest years. These discrepancies among the wine samples were due to the different ages of the wines stored in bottles and micro climatic changes considering the different harvest years.

Looking at the ^1H -NMR spectra plot of all CS wine samples from different harvest years shown in Figure 10A, it can be observed that ^1H -NMR spectra profiles were very similar among all samples. However, looking at the amino acids, sugars and phenolics regions, there are some resonances with different intensity levels. As an example, Figure 10B shows the amino acids region interval [2.25 -2.00 ppm], where it is possible to observe in detail the aforementioned differences.

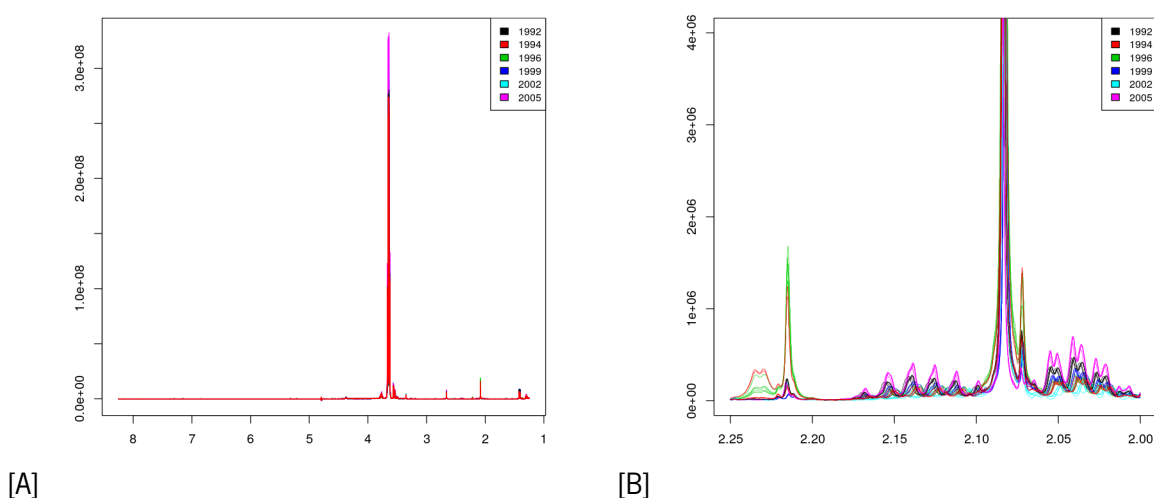


Figure 10: **A: ^1H -NMR spectra plot for all wine samples with distinct harvest years. B: Zoom of ^1H -NMR spectra showing the amino acids region interval [2.25 - 2.00 ppm].** Wine samples were coded in color according to their harvests as follows: 1992 (black), 1994 (red), 1996 (green), 1999 (dark blue), 2002 (light blue), and 2005 (pink).

Considering the three regions in the ^1H -NMR spectra identified with differences in the wine samples from distinct harvests, some amino acids were selected for further quantitative analysis because of their relevance for wine composition. Quantitative analyses were performed on amino acids by using GC.

Looking to the total amount of amino acids for the distinct harvest years, differences were observed among the wine samples, as presented in Figure 11. Also, we report individual amino acids concentrations,

as shown in Figure 12) where it is also observed that wine samples with distinct harvest years revealed the same amino acids constituents, but with different concentrations. It is observed that proline (Pro) has a higher concentration in wine samples from the harvest years 2005 and 2002, than the remaining harvests. It seems that Pro concentration decreases in wine samples with a higher age of bottle storage, like the wine samples from the 1994, 1996 and 1999 harvests. Glutamic acid (Glu) is another amino acid whose concentration is clearly higher in the 2005 wine samples, comparing with the remaining ones. Indeed, it seems that Pro and Glu suffer changes in their concentrations in wine samples from distinct harvest years.

Regarding the contents of amino acids herein described, similar values have been detected in CS wines and other red wines produced in Brazil and other countries [170]–[179], [167], [180], [181] despite the fact the ecological conditions in the sites of production differ from those found in southern Brazil.

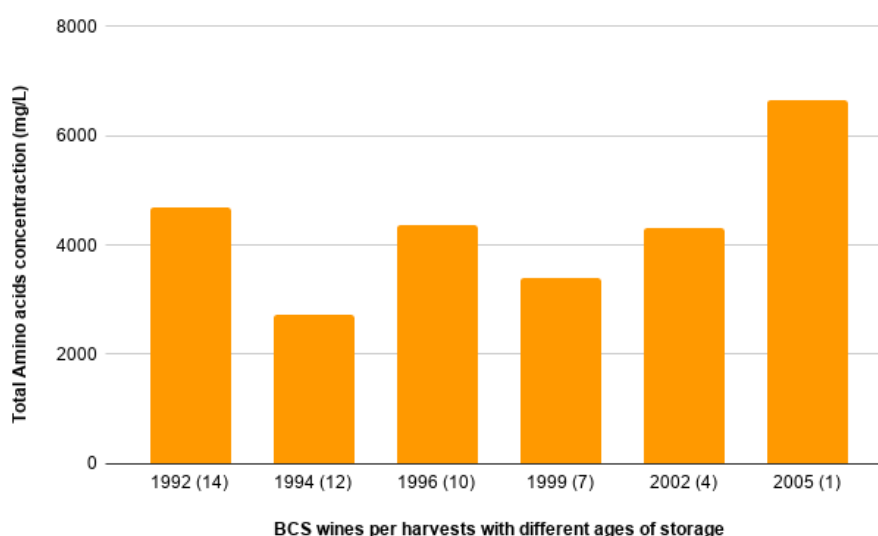


Figure 11: **Total of amino acids concentration of *Cabernet Sauvignon* wines (BCS) according to their harvests.**

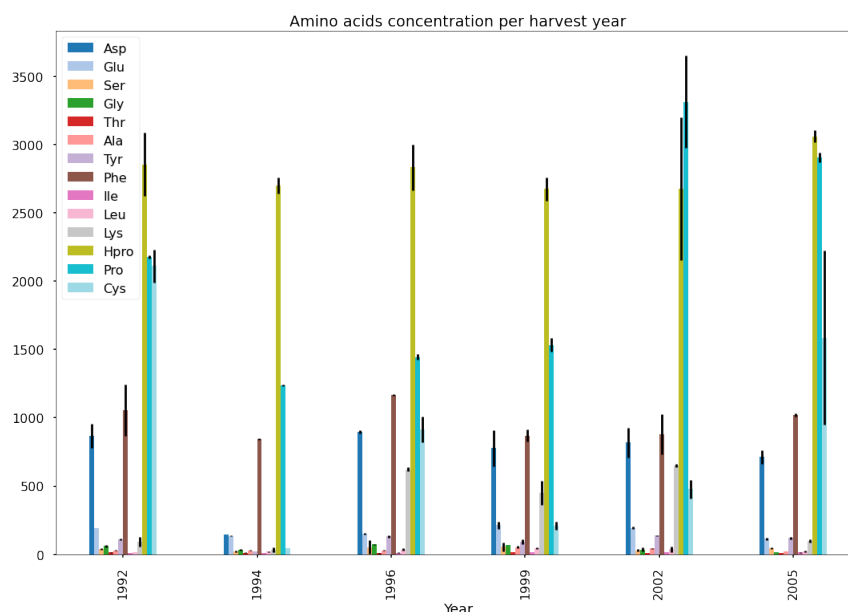


Figure 12: **Amino acids concentration (ml/L) of *Cabernet Sauvignon* wines according to their harvests.**

4.2.2.2 Univariate Statistical Data Analysis

Based on the previous results, to extract all the relevant information from the metabolomics data to discriminate the wine samples according to the harvest years, statistical analysis and machine learning were applied to the $^1\text{H-NMR}$ data and amino acid concentrations. Univariate analysis methods are the most commonly used for exploratory analysis, providing a preliminary overview about features potentially significant to discriminate the samples.

Our models from ANOVA, with post-hoc Tukey test, show that 30 $^1\text{H-NMR}$ peaks and the variables measuring concentrations of 5 amino acids (Ala, Tyr, Asp, Pro, Lys) may have the capability to linearly discriminate wine samples according to the harvests, as presented in Tables 7 and 8 respectively.

Looking in more detail, the 30 selected peaks from ANOVA are distributed in three important regions of the $^1\text{H-NMR}$ spectra, i.e., polyphenols, sugars alcohols, and amino acids. The regions of polyphenols includes resonances from 5.21 to 7.13 ppm, the one related to sugars includes from 3.46 to 4.28 ppm, while the amino acids region includes resonances from 0.85 to 3.40 ppm. These findings from ANOVA matched with the previous results observed in $^1\text{H-NMR}$ plot spectra (Figure 10A, 10B). Changes during the aging process and due to the climatic conditions can be the cause of the differences found in the aromatic $^1\text{H-NMR}$ regions, as suggested by [93], [169].

Concerning the amino acids concentrations, 4 out of the 5 amino acids (Ala, Tyr, Asp, and Lys) identified by ANOVA are precursors of biogenic amines. In wines, biogenic amines increase their concentrations by microorganism's activity, naturally happening during the vinification and aging processes and they can be affected by temperature and humidity on those processes or during storage. Consequently, amino acids as precursors of biogenic amines change their concentrations as they are used in the biosynthesis

of biogenic amines by yeasts. Besides, the contents of amino acids can also suffer changes as they are turned into other metabolites such vitamins during the grape growth and ripening stages and over the storage of wines[93], [169]. The results from ANOVA matched with those previously observed regarding differences in the $^1\text{H-NMR}$ plot spectra region of amino acids (Figure 10(A) and (B)). However, it is important to keep in mind that the ANOVA results point to a possible linear relationship without accounting for its strength.

Table 7: **ANOVA results - peaks with the lowest corrected p-values (< 0.01)**. The first column shows the peaks, the second one the respective p-value, the third column includes the corrected p-value with False Discovery Rate (FDR) method, and the final column the result of the Tukey's HSD test, which consists of the pairs of harvest years that were significantly different in terms of means for each peak.

Peak (ppm)	p-value	FDR	Tukey test (pairs of harvest years)
3.60	1.64E^{-13}	1.87E^{-11}	1999-1992; 1999-1994; 1999-1996; 2002-1999; 2005-1999
2.05	1.33E^{-12}	7.60E^{-11}	1994-1992; 1996-1992; 1999-1992; 2002-1992; 2005-1992; 1999-1994; 2002-1994; 2005-1994; 2002-1996; 2005-1996; 2002-1999; 2005-1999; 2005-2002
2.14	6.44E^{-12}	2.45E^{-10}	1994-1992; 1996-1992; 1999-1992; 2002-1992; 2005-1994; 2005-1996; 2002-1999; 2005-1999; 2005-2002
2.29	9.81E^{-10}	2.80E^{-08}	1994-1992; 1996-1992; 2002-1992; 1999-1994; 2002-1994; 1999-1996; 2002-1996; 2002-1999; 2005-2002
2.39	3.48E^{-09}	7.92E^{-08}	1994-1992; 1996-1992; 2002-1992; 2005-1992; 2005-1994; 1999-1996; 2005-1996; 2002-1999; 2005-1999; 2005-2002
3.89	5.29E^{-08}	1.01E^{-06}	2002-1992; 2005-1992; 1999-1994; 2005-1994; 1999-1996; 2005-1996; 2002-1999; 2005-2002
3.40	2.88E^{-07}	4.69E^{-06}	1994-1992; 2002-1992; 2005-1992; 2005-1994; 2005-1996; 2005-1999; 2005-2002
1.37	1.52E^{-06}	2.16E^{-05}	1996-1992; 1996-1994; 1999-1994; 1999-1996; 2002-1996; 2005-1996
2.21	1.77E^{-06}	2.24E^{-05}	1996-1992; 1996-1994; 1999-1994; 1999-1996; 2002-1996; 2005-1996
7.13	1.99E^{-06}	2.27E^{-05}	1994-1992; 1996-1992; 1999-1994; 1999-1996; 2002-1996; 2005-1996
3.74	3.58E^{-06}	3.69E^{-05}	2002-1992; 2005-1992; 2002-1994; 2005-1994; 1999-1996; 2002-1996; 2005-1996
4.12	6.25E^{-06}	5.77E^{-05}	1994-1992; 1999-1994; 2002-1994; 2005-1994; 1999-1996; 2002-1996; 2005-1996
4.28	6.58E^{-06}	5.77E^{-05}	1999-1992; 2005-1992; 2005-1994; 2005-1996; 2005-1999; 2005-2002
3.37	8.34E^{-06}	6.79E^{-05}	1996-1992; 2005-1994; 2005-1996; 2005-1999; 2005-2002

2.17	1.10E ⁻⁰⁵	8.35E ⁻⁰⁵	2002-1992; 2002-1994; 2002-1996; 2002-1999; 2005-2002
2.27	1.23E ⁻⁰⁵	8.76E ⁻⁰⁵	2002-1992; 2002-1994; 2002-1996; 2002-1999; 2005-2002
3.00	1.56E ⁻⁰⁵	0.0001	2002-1992; 2002-1994; 2002-1996; 2005-1996; 2002-1999; 2005-2002
5.32	3.39E ⁻⁰⁵	0.0002	2005-1992; 2005-1994; 2005-1996; 2002-1999; 2005-2002
2.57	4.32E ⁻⁰⁵	0.0003	2002-1992; 2002-1994; 2002-1996; 2002-1999; 2005-2002
1.52	0.00012	0.0007	2002-1992; 2002-1994; 2002-1996; 2005-2002
5.25	0.00014	0.0008	1994-1992; 1996-1992; 2002-1992; 2005-1994; 2005-1996; 2005-2002
0.85	0.00017	0.0009	1994-1992; 1996-1992; 1999-1992; 2002-1999; 2005-1999
1.77	0.00024	0.0012	1996-1992; 1999-1992; 2002-1992; 2005-1999
4.17	0.00030	0.0014	1994-1992; 1999-1994; 2002-1994; 2005-1994
1.40	0.00041	0.0019	1994-1992; 1996-1992; 2002-1992; 1999-1994; 1999-1996; 2002-1999
3.46	0.00045	0.0020	1994-1992; 1999-1994; 2005-1994
3.86	0.00087	0.0037	1996-1994; 2002-1996; 2005-1996
1.27	0.00099	0.0040	1996-1992; 2005-1996; 2005-1999
5.21	0.00113	0.0044	2005-1994; 2005-1999; 2005-2002
0.88	0.00116	0.0044	2002-1992; 2002-1996; 2005-1996; 2002-1999

Table 8: **ANOVA results - amino acids with the lowest corrected p-values (< 0.05)**. The first column contains the considered amino acids, the second one the respective p-value, the third column includes the corrected p-value with False Discovery Rate (FDR) method, and the final column the result of Tukey's HSD test, which consists of the pair of harvest years that were significantly different in terms of means for each amino acid concentration.

Amino acid	p-values	FDR	Tukey test (pairs of harvest years)
Alanine	1.85E ⁻⁰⁵	0.00026	1999-1992; 2005-1992; 1999-1994; 2002-1994; 1999-1996; 2005-1996; 2005-1999; 2005-2002
Tyrosine	4.51E ⁻⁰⁴	0.00316	1994-1992; 1996-1994; 2002-1994; 2005-1994
Aspartate	8.88E ⁻⁰⁴	0.00414	1994-1992; 1996-1994; 1999-1994; 2002-1994; 2005-1994
Proline	1.23E ⁻⁰³	0.00430	2002-1994; 2005-1994; 2005-1996; 2005-1999
Lysine	8.44E ⁻⁰³	0.02362	2002-1992; 2002-1994

4.2.2.3 Multivariate Data Analysis

The aforementioned results by ANOVA indicate the effect of the harvest year on the wine's chemical composition. Based on these findings, unsupervised and supervised multivariate statistical methods were applied to the metabolomics dataset to build descriptive and classification models to extract more relevant patterns. We started with Principal Components Analysis (PCA) and clustering algorithms to group relevant features and samples from the metabolomics data, and pinpointing to possible relevant information that may help to characterize the studied variables.

PCA is an unsupervised algorithm that projects the samples in a latent space using a linear transformation. This transformation may allow to separate the data and to discriminate the samples. This method can also be utilized to reduce the dimensionality of the data set, routinely applied on $^1\text{H-NMR}$ metabolomics studies by using a smaller number of dimensions than the original data set [182].

In this study, we plotted the first two dimensions (principal components 1 and 2: PC1, PC2) and the first five dimensions (principal components of the 5 PCs) as presented in Figure 13 A and B. These results represent the latent dimensions that explain the most variability in the data, coloring the sample points with their harvest year. It is clear that it is possible to discriminate well three of the six harvest years based on the $^1\text{H-NMR}$ data, specifically the harvest years 1992, 2002, and 2005. Still, this unsupervised model was not effective in providing a clear separation for the samples of the other harvest years, although some degree of separability is visible.

Concerning the metabolites concentration data, the model discriminated well only one of the six harvests, the one from 1994 (Figure 14 A and B). As in the results from $^1\text{H-NMR}$ data, this method was not effective in providing a clear separation for the samples into the different harvest years in a global way with even less discriminate power.

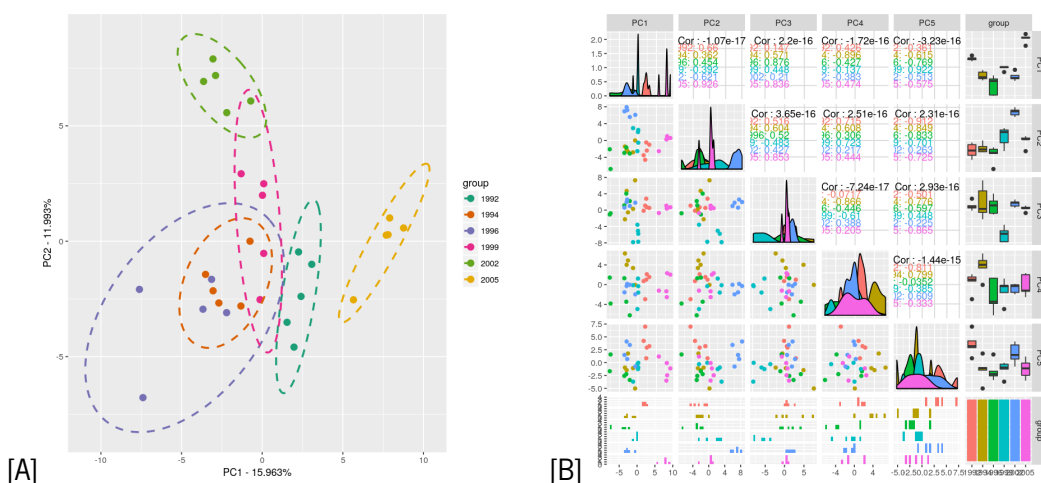


Figure 13: **A: PCA 2D scores plot (PC1, PC2) from $^1\text{H-NMR}$ data. B: PCA plot of the first five PCs from $^1\text{H-NMR}$ data.** The wine samples are presented in different colors corresponding to the harvests groups: 1992 (dark-green), 1994 (orange), 1996 (purple), 1999 (pink), 2002 (light-green), and 2005 (yellow).

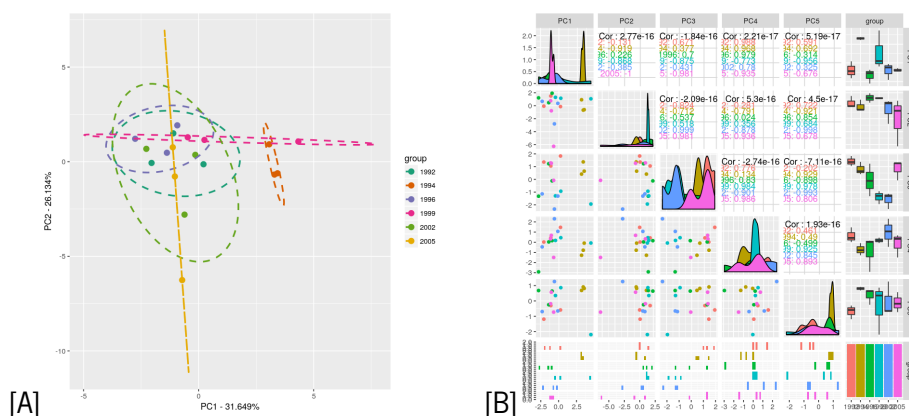


Figure 14: **A: PCA 2D scores plot (PC1, PC2) from amino acids concentrations. B: PCA plot of the first five PCs from amino acids concentrations.** The wine samples are presented in different colors corresponding to the harvests groups: 1992 (dark-green), 1994 (orange), 1996 (purple), 1999 (pink), 2002 (light-green), and 2005 (yellow).

Clustering gathers samples into a defined number of groups based on the used variables by utilizing a distance measure. Unlike discriminant analysis, the number and composition of such groups is often unknown, while techniques such as using elbow graphs can be utilized to estimate a possible number of clusters. Clustering approaches, such as Hierarchical Clustering Analyses (HCA) involve grouping biological samples using all peaks from the spectra or metabolite concentrations.

From the results of HCA, it is possible to observe that the samples from $^1\text{H-NMR}$ spectra were grouped in the considered harvest years quite well, with a small overlap between the years of 1994 and 1996. The samples from the 1992, 1999, 2002, and 2005 harvests were well discriminated (Figure 15 A). When performing the same method over the amino acids contents, the results revealed that only the 2002 harvest was well discriminated, while the remaining ones had significant overlap, as presented in Figure 15 B.

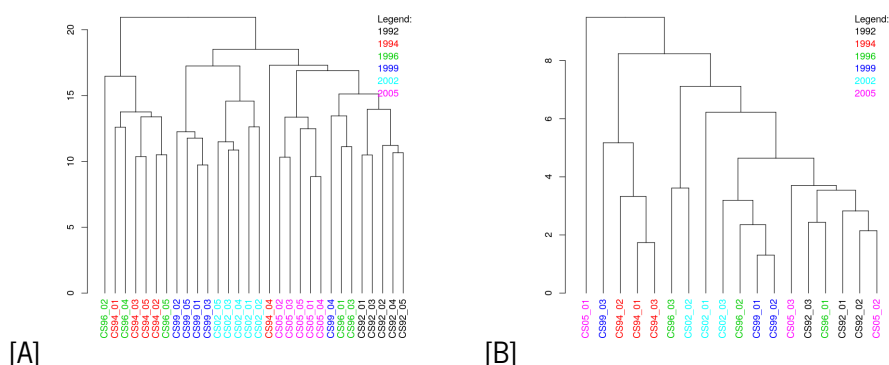


Figure 15: **A: Dendrogram plot of the hierarchical clustering, with Euclidean distance between wine samples from $^1\text{H-NMR}$ data. B: Dendrogram plot of the hierarchical clustering, with Euclidean distance between wine samples from amino acids concentrations.** The wine samples are presented in different colors corresponding to the harvests groups: 1992 (black), 1994 (red), 1996 (green), 1999 (blue), 2002 (light-blue), and 2005 (pink).

The overlap and the proximity of the samples from the harvests 1994 and 1996 can be related to the influences regarding the proximity of the age of storage (the wines have 12 and 10 years of age of bottle storage, respectively), as their harvest years had similar influences from three climatic factors (temperature, humidity and insulation, present in supplemental material). To better understand those influences, other analyses were made using an univariate linear regression model and are presented in section: Wine age of storage and climatic factors influence.

Afterwards, three distinct machine learning (ML) classification models were built, to identify which features were more relevant to discriminate the wine samples according to the harvest year including: Random Forests (RF), Partial Least Squares regression (PLS), and K-Nearest Neighbors algorithm (KNN). The results, using $^1\text{H-NMR}$ data as input features, show that these models have a very good performance, reaching 86-92% of accuracy. Utilizing only the data for amino acid concentrations as input for classifying the wines according to their harvest, the best model (RF) had a performance of 66%. Table 9 presents the results from all models.

For the $^1\text{H-NMR}$ data, the variables, i.e. resonances, with more importance in the classification by the RF model were similar to the ones returned by ANOVA (1.37, 1.40, 1.77, 2.05, 2.14, 2.21, 2.29, 2.39, 2.57, 3.37, 3.4, 3.60, 4.12, 4.28, and 7.13 ppm), while PLS and KNN models considered four other important resonances (2.11, 3.43, 3.95, and 3.98 ppm), still in the same spectral regions. Indeed, as in the previous results from ANOVA, the main features accounting for the classification models were resonances occurring at the amino acids, sugars, and phenolic regions of the $^1\text{H-NMR}$ spectra.

For the amino acid concentrations as input data, the variables with more importance in the classification by RF were again similar to the ones returning from ANOVA and Tukey test analyses, i.e., Ala, followed by Pro, Lys and Tyr. Only Ile was considered by the model as important features not previously selected by ANOVA and Tukey test analyses, what may point to non-linear relationships or relevant interactions among different variables regarding these amino acids.

Table 9: **ML models results for the $^1\text{H-NMR}$ data and the amino acids concentration with cross-validation accuracy.** The first column contains the considered input data as features of the ML model, the second one the applied ML model, the third one the result with the accuracy of the correspondent ML model computed using cross-validation, and the last column the features ($^1\text{H-NMR}$ resonances - ppm and amino acids) most relevant, with those marked with bold type indicating those shared with ANOVA results.

Input data (features)	Model	Accuracy	Features
Amino acids concentration	RF	66%	Alanine, Proline , Isoleucine, Lysine, Tyrosine
$^1\text{H-NMR}$ spectra	PLS	92%	0.85; 1.37; 1.40; 1.77; 2.21; 3.60; 3.86; 3.95; 3.98; 4.01; 4.28; 4.41; 5.21; 5.32; 7.13
$^1\text{H-NMR}$ spectra	RF	90%	1.37; 1.40; 1.77; 2.05; 2.14; 2.21; 2.29; 2.39; 2.57; 3.37; 3.40; 3.60; 4.12; 4.28; 7.13
$^1\text{H-NMR}$ spectra	KNN	86%	0.85; 1.40; 1.77; 2.05; 2.11; 2.14; 2.29; 2.39; 3.40; 3.43; 3.95; 4.01; 4.12; 5.25; 7.13

4.2.3 Wine Age of Bottle Storage and Climatic factors influence

In our study, regression analysis is used to identify which peaks and amino acid amounts are correlated with the age of bottle storage and climatic factors individually, considered as continuous (numerical) variables. Linear Regression (LR) models were also used to infer correlation between the independent (age of bottle storage and climatic factors) and dependent variables ($^1\text{H-NMR}$ resonances and amino acids concentrations). Note that regression analyses by themselves only reveal relationships between a dependent variable and a collection of independent variables in a fixed dataset [183]. With these considerations, other influences not considered in study may affect the metabolomics wine profiles.

4.2.3.1 Assessing the influence of Age of Bottle Storage

Bottle storage of wine is a long term process with an important influence in wine preservation, being also used to improve the quality and value of certain wines for a long term aging process, under adequate conditions [93], [169]. Some studies provide important analyses to elucidate the impact of the storage conditions influences on wine composition and quality, such as humidity, temperature, light, and packaging [184]–[188]. Other ones also reported the type of storage influence such the use of bottles or oaks on wine composition and quality [93], [169], [174], [189], [190]. However, the majority of the reported studies only comprise some months or a few years, such as two to five years, and are also focused in a certain type of compounds, such as phenolics and mainly anthocyanins [191], [192].

Due to the high storage cost, only few wine varieties with certain features benefit from the aging process [93], [193]. CS is one variety with good features for the aging process, such as the high tannin content, a certain range of pH values and other ideal features for storage conditions [174], [177], [191], [192], [194]. The pH values of the studied wine samples, before the experiment analyses, were between 3.64 and 3.90, as others also reported similar values of pH for the same wine variety produced in similar climatic conditions [180]. In our study, as mentioned before, the wine samples have distinct years of bottle storage, between 1 to 14 years.

To understand the effect of the aging on the wine's chemical profiles, a supervised approach was applied to the $^1\text{H-NMR}$ data set and the metabolite concentrations. In this case, a LR analysis was performed to better understand which peaks and metabolites were influenced directly by the age of bottle storage.

The results of the LR model over all $^1\text{H-NMR}$ spectra for the wine's age of bottle storage influence revealed that two $^1\text{H-NMR}$ resonance peaks at 3.74 ppm (coefficient: -0.17) and 4.28 ppm (coefficient: -0.16) presented significant p-values (<0.01) and high R^2 ($>55\%$). Both resonances decrease their intensities levels with the increase of the age of bottle storage, as represented in figures (Figure 16A, 16B).

Looking to amino acids concentration influenced by age of bottle storage, the LR models reveal one affected metabolite, Pro (with significant p-value <0.01 , $R^2 = 0.32$ and coefficient: -65.35), whose concentration decreases with the age of bottle storage, as represented in Figure 16(c). These results should be taken with a grain of salt due to the adjusted R^2 explaining only about one third of the variability in the data.

In wines, the concentration of different amino acids depends on the methods used for their determination. Usually, Pro, Ala, Arg, Glu, Ser, and Thr are the amino acids with highest concentration in wine samples [195]. Indeed, because Pro is not consumed in fermentation and maturation processes [196] its concentration is the highest among the remain amino acids and is also higher in some varieties like CS [197], being this confirmed in our data.

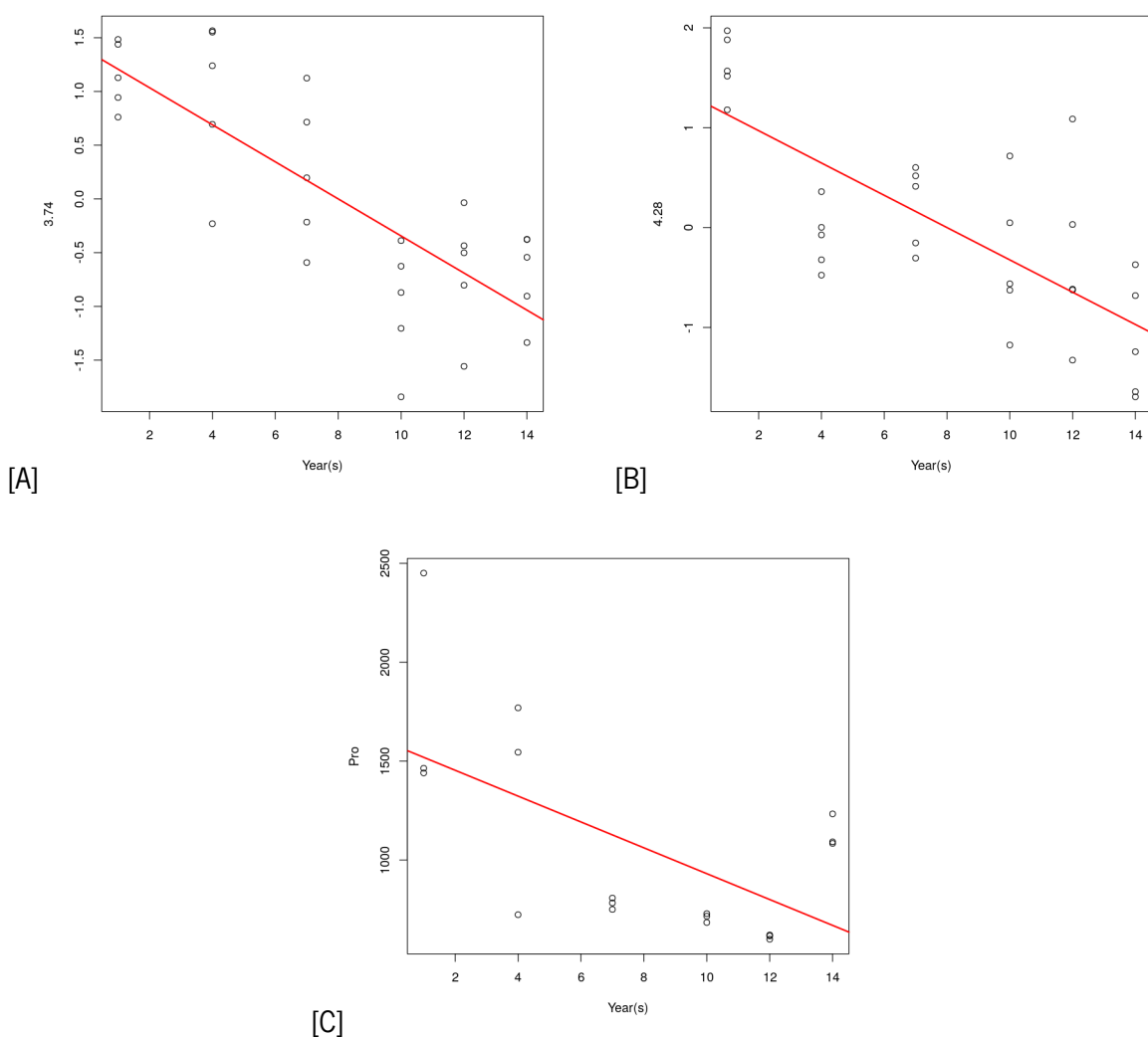


Figure 16: **Plots for the linear regression models showing the influence of age of bottle storage on ^1H -NMR resonances and one amino acid with highest adjusted R^2 .** A: 3.74 ppm resonance. B: 4.28 ppm resonance. C: Pro amino acid concentration.

4.2.3.2 Assessing the influence of Climatic conditions

Vineyard climatic conditions are some of the environmental factors which modulate the wine chemical composition and, consequently, give a specific fingerprint in the observed metabolomics profile. Furthermore, grape cultivars may respond differently to yearly climatic changes, so it is important to assess which are the climatic factors with highest impact on modulation of the wine metabolomics profile. Usually, CS

wines are produced in European regions with maturation average temperature of 17-20°C [198], [199]. In our study, the CS wines were produced in southern Brazil, under the same winemaking conditions over the harvests, with maturation average temperature of 15-26.4°C, having microclimatic variations among different years of production.

In our study, we considered temperature variations (°C) (minimum temperature (TempMin), median temperature(TempMean), and maximum temperature(Tempax)), including the minimum mean night temperatures (°C) (Cool Night Index (Cool Index)), and variations for Precipitation (mm) (PP), Humidity, (%) and Insulation levels ((kWh/m²) per day). Nonetheless, the analysis was conducted using the aforementioned climatic factors to answer the main question of the individual climatic influence. The climatic data used in this study is shown in Table 10.

Table 10: **Climatic Data by each harvest year**

Year	CoolIndex	Ins	Hum	PP	TempMean	TempMin	Tempax
1992	15.9	206.7	76.5	152.6	20.6	16.1	26
1994	15.7	210.7	75.5	134.0	20.6	16.1	25.5
1996	16.5	225.2	73.3	168.0	20.6	16.1	25.8
1999	18.7	223.9	72.3	83.9	20.5	16.0	26.0
2002	17.0	216.7	80.0	140.1	19.1	15.0	24.7
2005	17.0	240.3	70.3	99.1	20.4	15.6	26.4

The previous results showed differences in the ¹H-NMR profiles and amino acids concentrations of wine samples according to the different harvests, so it is important to know which are the climatic factors that influence the wine metabolomics profile, considering the case study.

The results of the LR model to study the influence over all ¹H-NMR peaks and amino acid concentrations reveal some combinations with a significant p-value ($\leq 0,0001$) and an adjusted R^2 above 50%. The LR models reveal 12 peaks (Table 11) showing a significant influence by climatic variables.

Table 11: **Results for linear regression analysis of the climatic influence over ¹H-NMR data.**

The first column contains the considered peaks, the second column presents the influence type, followed by the respective corrected p-value($\leq 0,0001$), adjusted R^2 ($> 50\%$), and coefficient of the LR model. Bold type indicates ¹H-NMR resonances with highest adjusted R^2 .

Peak (resonance, ppm)	Climatic influence	p-value	Adjusted R^2	Coefficient
2.05	T Max	9.69E⁻¹¹	77%	1.63
2.14	T Max	1.29E ⁻⁰⁶	56%	1.39
2.17	T Max	1.06E ⁻⁰⁶	56%	1.40
2.17	T mean	4.82E ⁻⁰⁶	51%	1.32
2.27	T mean	1.07E ⁻⁰⁷	63%	-1.45
2.27	T min	3.04E ⁻⁰⁶	53%	-1.79

2.29	T mean	2.69E⁻⁰⁹	71%	-1.54
2.29	T min	5.44E ⁻⁰⁷	58%	-1.87
2.39	T Max	1.37E ⁻⁰⁷	62%	1.47
2.57	T mean	5.42E ⁻⁰⁷	58%	-1.40
3.00	T Max	1.73E ⁻⁰⁶	55%	1.38
3.60	Cool Index	8.44E⁻¹¹	78%	0.89
3.60	Precipitation	4.07E ⁻⁰⁶	52%	-0.02
3.89	Humidity	4.06E ⁻⁰⁷	59%	0.24
3.89	T Max	9.23E ⁻⁰⁸	63%	-1.48
4.28	Insulation	5.65E ⁻⁰⁷	58%	0.07
5.32	Humidity	1.76E ⁻⁰⁶	54%	0.24

The results indicate that temperatures (maximum, mean, minimum, and Cool Night Index) are independent variables significant for 10 of the 12 selected peaks, while insulation, humidity, and precipitation influenced only four peaks. A peak at 3.60 ppm is the ¹H-NMR resonance most influenced by the Cool Night Index climatic factor (R^2 of 78%). This peak increases its intensity level with the increase of the minimum mean night temperature as presented in Figure 17, but also decreases its intensity with the increase of precipitation levels (Table11,line:13). The resonance at 2.05 ppm is influenced by maximum temperature (explained variance of 77%), meaning that the peak increases its intensity with the higher maximum temperatures (Figure18A). The signal at 2.29 ppm is influenced by both mean temperature (Figure 18B) and minimum temperature (Table11,line:8). Figures 17 and 18 presents the plots for the linear regression models showing the influence of the climatic factors on the three mentioned ¹H-NMR peaks with the highest adjusted R^2 . To note that, the 1999 harvest year is depicted as the isolated group of peaks with the highest value of Cool Night Index (Figure 17). Analogously, the isolated group of peaks regarding the mean temperature (T mean) concerns the 2002 harvest year (Figure 18B). These results show the influences of the studied climatic factors set on NMR resonances.

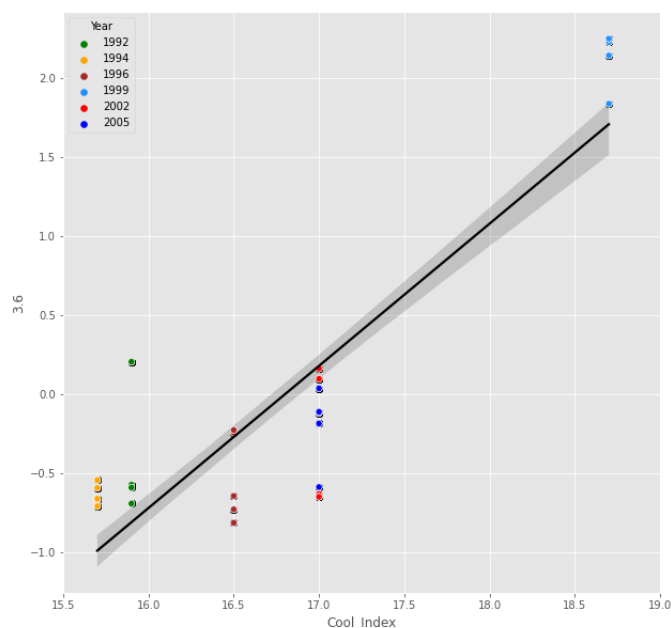
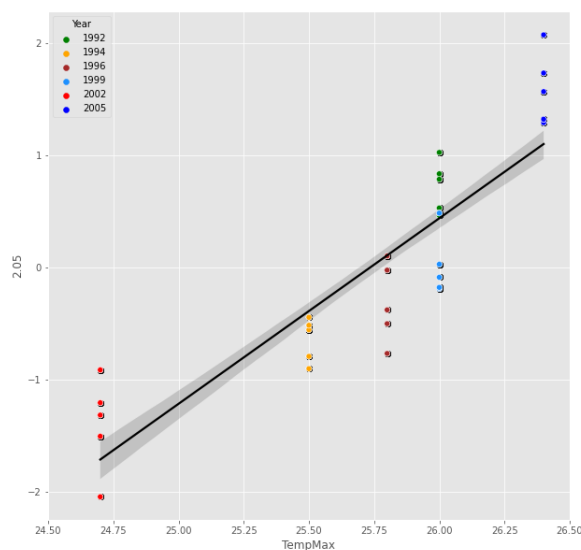
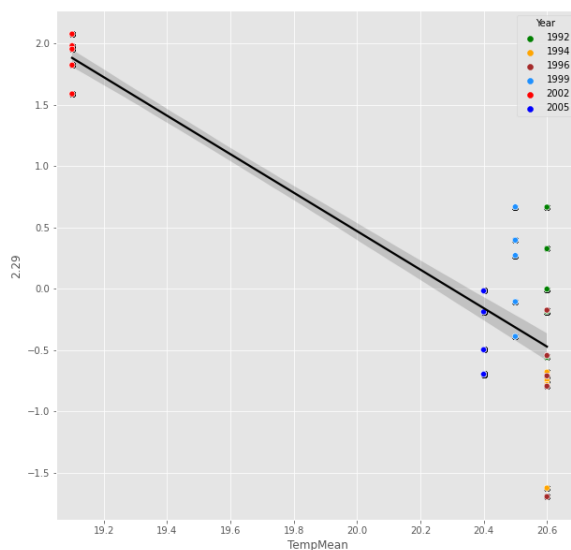


Figure 17: **Plot for the linear regression model showing climatic influences on 3.60 ppm resonance with a highest adjusted R^2 .** The wine samples are presented in different colors corresponding to the harvests groups: 1992 (green), 1994 (yellow), 1996 (brown), 1999 (light-blue), 2002 (red), and 2005 (dark-blue).



[A]



[B]

Figure 18: **Plots for the linear regressions models showing climatic influences on ^1H -NMR resonances with highest adjusted R^2 .** A: 2.05 ppm resonance. B: 2.29 ppm resonance. The wine samples are presented in different colors corresponding to the harvests groups: 1992 (green), 1994 (yellow), 1996 (brown), 1999 (light-blue), 2002 (red), and 2005 (dark-blue).

The majority of the remaining peaks influenced by the climatic factors are related to temperature. The results show that with higher temperatures most peaks have their intensities increased, while the resonances at 2.27, 2.29, 2.57 and 3.89 ppm decrease their intensity with higher temperatures. Considering the other climatic factors, with the increase of insulation, the signal at 4.28 ppm increases its level (this

peak is also known for decreasing its intensity with the age of storage), and with the increase of humidity the resonances at 3.89 ppm and 5.32 ppm increase their intensity levels.

Based on LR models for CS wines produced in the region under study, temperature seems to be the climatic factor with higher influence on ^1H -NMR profiles, specifically the minimum mean night temperature. The ^1H -NMR resonances concentrations that did not show a direct linear relationship with climate metadata were not selected. However, this does not mean that other non-linear relations cannot happen between these data.

An important note is that the identified resonances influenced by wine age of bottle storage are not the same as the peaks influenced by climatic factors, except for the 4.28 ppm one. This means that the two influential conditions have two distinct groups of peaks. With these results, these two groups of peaks can be used as metabolomics ^1H -NMR wine markers to assess the age of storage and climatic influence on wine CS produced in southern Brazil.

The main studies reported to understand the climatic influences on red wine quality concern to the influence of one factor or some related environmental factors on some wine metabolites, such as phenolics or other aromatic compounds, making difficult to understand the final metabolomics wine profile as a result from the climatic factors combinations.

Another study revealed that light exposure of Merlot berries increased the skin and pulp flavonols, histidine, and valine contents, and reduced the organic acids, GABA, and alanine contents [200]. A larger study used three varieties, *Merlot*, *Cabernet Franc* and *Cabernet Sauvignon* to study the influence of climate factors for a period of six years, reporting the effects of climate as highly significant regarding to vine behavior and berry composition based on the anthocyanin concentration [201]. Our study used only one wine variety and the corresponding amino acids concentrations, thus it is not possible to compare our results with other investigations. To overcome this limitation, the pipeline of our study should be applied on more wine varieties in conjunction with extra metabolomics data. This allows getting a more detailed characterization of wine profiles under the influence of the factors, such as age of storage and climatic factors. Considering the actual climatic changes on the world, vine management practices should be adjusted appropriately to regional growing conditions and grape cultivars. The results allowed us to putatively identify the relevant ^1H -NMR peaks in CS wines samples produced in different years under similar climatic influences.

4.2.4 ^1H -NMR resonances with capability to discriminate wine samples by harvest, age of bottle storage and climatic factors

Qualitative analyses of the identified ^1H -NMR resonances were performed after the statistical and chemometrics analysis, based on ANOVA, LR models and machine learning results. The main selected ^1H -NMR resonances with best predictive capability to distinguish wine samples based on harvest corresponding to the regions of amino acids, phenolic acids, and sugars (Table 12).

Table 12: **Regions of ^1H -NMR resonances (ppm) with predictive capability to distinguish samples selected by ANOVA and ML models.**

Phenolic acids	Sugars	Amino acids
5.32, 7.13	3.60, 3.74, 3.89, 4.12, 4.28	1.37, 1.40, 1.77, 2.05, 2.14, 2.17, 2.21, 2.29, 2.39, 2.57, 3.00, 3.37, 3.40

Based on the resonances obtained from the previous results, it is possible to observe that there are 22 peaks utilized in the discrimination of the harvest year. However, there are eight resonance peaks with the capability to discriminate harvests which are not influenced linearly by age of bottle storage, or the studied climatic factors set (in green in Figure 19). The resonance peak of 4.28 ppm has capability to discriminate wine samples by three factors under study (harvest, age of bottle storage, and the insulation climatic factor). There are twelve resonance peaks that allow to predict only the studied climatic factors and not age of bottle storage (in yellow in Figure 19), using a linear model. There is one resonance peak which can predict the age of bottle storage (3.74 ppm) (in blue in Figure 19).

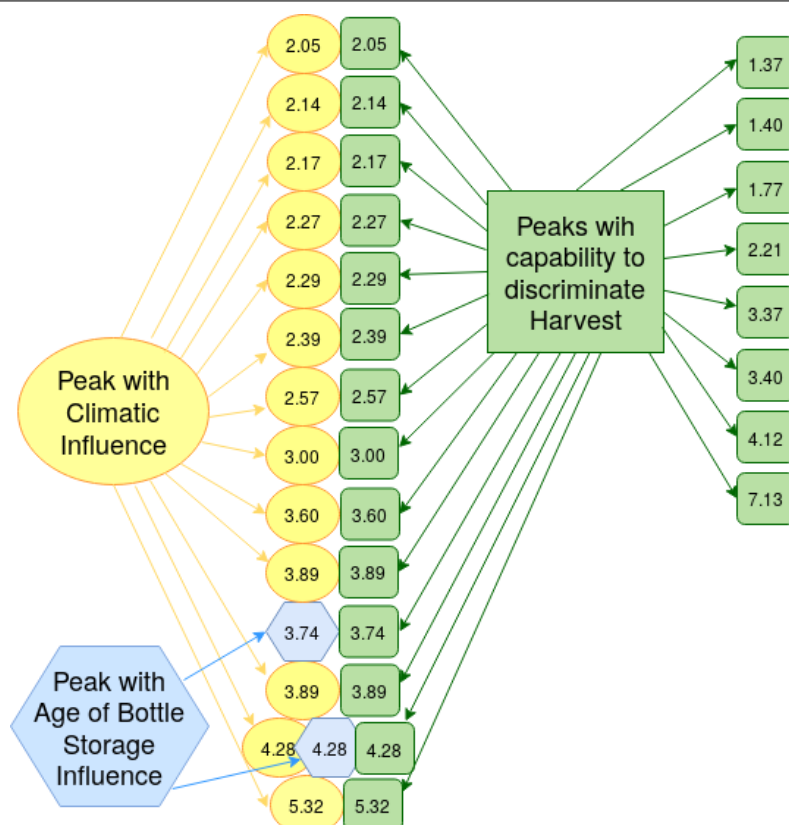


Figure 19: ^1H -NMR resonances with best predictive capability to distinguish wine samples by harvest (colored green), age of bottle storage (colored blue), and climatic factors (colored yellow) influences.

4.3 Conclusions and Future work

Our study addresses metabolomics profiles analyses on wine samples from six different harvest years, produced in the same place, considering metadata of different times of bottle storage (comprehending a larger period of time, one to fourteen years of bottle storage), and a set of selected climatic conditions.

In summary, the exploratory analysis has demonstrated that there is a close relationship between the harvests, the influences of the age of bottle storage and the selected climatic factors, in the wine's metabolomics profiles and in some metabolites. The study allowed to:

- Discriminate the wine samples according to the harvest year using supervised models on this dataset;
- Identify the climatic factors with higher influence on wine's metabolome;
- Identify the resonances and the metabolite concentrations influenced by each of the climatic factors investigated and age of bottle storage.
- Predict the age of bottle storage and climatic factors influence based on ^1H -NMR data.

Our study brings a new highlight to the use of metabolomics data combined with multivariate analysis and machine learning tools, for the use of identified resonances to study the influence of factors of interest on wine metabolomics profiles, e.g., a set of climatic factors and age of bottle storage, or other influences. To use the identified resonances as biological markers of the influence factors on wine metabolomics profiles, further studies must be conducted using more wine varieties and samples from other wine producing regions to know if the same resonances and metabolites are identified. Notwithstanding, this study also provides the generation of new hypothesis that may be explored in the future (i.e., if the same harvest, aging or climatic markers are found using distinct production processes or in distinct wine varieties). Further studies may also contemplate other factors not analyzed in this work, for example, soil types, grapevine nutritional management, and other influences.

4.4 Materials and Methods

4.4.1 Samples

Thirty *Cabernet Sauvignon* wine samples with different harvests years: 1992, 1994, 1996, 1999, 2002, and 2005 (five bottles from each harvest), produced in the Garibaldi county (29.2576° S, 51.5231° W), southern Brazil, were used in this study. Samples were kindly furnished by De Lantier Vinhos Finos Ltda in 2005 and were produced under the same winemaking conditions over the vintages, following regular practices adopted by the sample provider. A standardized set of barrel and bottle aging practices was adopted by the company over the vintages. At the end of the storage time in the barrels (twelve months) the wines were bottled, following their storage also under standardized conditions until commercialization. The bottle storage conditions were: temperature: 15°C to 18°C, with 80-85% air humidity in the dark, stopper: cork stopper. According to the technical services available at that moment, the dataset collected was representative of the climatic conditions registered over the years and harvests. After collection, the wine samples were transferred to the University of Aveiro, Portugal, where the experimental analyses performed. The information relative to the harvest year was considered as discrete categorical variables. The information relative to the age of bottle storage and the climatic conditions were considered as numerical continuous variables.

4.4.2 Age of Bottle Storage Data

Each sample has a respective age of bottle storage as follows: harvest 1992 (around 14 years of bottle storage), harvest 1994 (around 12 years of bottle storage), harvest 1996 (around 10 years of bottle storage), harvest 1999 (around 7 years of bottle storage), harvest 2002 (around 4 years of bottle storage) and wine samples with harvest 2005 approximately 1 year of bottle storage.

4.4.3 Meteorological Data

The climate data used in this study are from the meteorological stations owing to the winemaker (De Lantier Vinhos Finos Ltda) located in the region of the vineyards (i.e., Garibaldi county, Serra Gaúcha). The data contain as attributes the mean values for a set of meteorological variables, precipitation, insulation, humidity, and temperature (minimum, maximum, median, and minimum of night) taken from monthly observations during the vineyard cycles produced in the years of 1992, 1994, 1996, 1999, 2002, and 2005. The Cool Night Index (refer as Cool Index in the data analysis) was considered as the minimum average air temperature of the last month of the ripening stage of the grapes. The Cool Index is a night coolness variable which takes into account the minimum mean night temperatures during the months when grape berries ripening occurs. This index helps improving the assessment of the qualitative potentials of wine-growing regions, notably in relation to secondary metabolites in grapes [202].

4.4.4 Chemical Experimental Analysis

4.4.4.1 1D Nuclear Magnetic Resonance spectroscopy (^1H -NMR)

Samples (540 μL) of CS red wines were transferred to 5-mm NMR tubes and added of 60 μL of D_2O containing 0.024 g% of sodium 3-(trimethylsilyl)-propionate-2,2,3,3- δ^4 (TSP) as internal chemical shift standard. The pH of the samples was adjusted to 3.85 with 0.2N HCl and the internal atmosphere of the 5-mm NMR tubes was removed by injection of N_2 . High-resolution ^1H -NMR spectra were recorded on a Bruker Avance DRX-500 spectrometer, equipped with an autosampler device (NMRcase – Bruker Biospin), operating at 500.13MHz for proton and 125.77MHz for carbon, at 298K and nonspinning. For one dimensional ^1H -NMR spectra ($n=3$) were acquired using an modulate shape pulse (NOESYPR1DSP pulse sequence – Bruker library) with triple offset and amplitude scaling applied during the relaxation delay (15s) and the mixing time (100 ms) for saturation of water (4.87ppm) and ethanol (CH_3 at 1.17ppm and CH_2 at 3.66 ppm) resonances. A total of 576 transients were collected with 32768 data points over a 5482.45Hz bandwidth in the absolute intensity mode, with an acquisition time of 2.98s. The measurement time was 3 h and the signal-to-noise ratio of the spectra was improved by multiplying each free induction decay with an additional exponential factor corresponding to 0.3Hz. The digital resolution was $\pm 0.083656\text{Hz}$. In accordance with the need for standardized reporting of metadata or information describing large-scale metabolomics (Sumner et al., 2007)[203]), previously to the ^1H -NMR experiments of the wine samples, the minimum reporting standards related to the chemical analysis aspects of metabolomics studies were taken in consideration, in order to obtain accurate, reproducible, and comparable ^1H -NMR data sets.

4.4.4.2 Amino acid quantification using Gas Chromatography (GC)

A GC method was used to quantify 14 selected amino acids from the 30 wine samples. GC analyses were performed as previously described [204], with modifications. For the amino acids derivatization the lyophilized samples of wines were added of 30 μL acetonitrile and 30 μL N-methyl-N-(trimethylsilyl)

trifluoroacetamide (MSTFA) and kept at 60°C for 20 min, with occasional stirring in vortex. Derivatization was carried out in a 1-ml glass vial sealed with Teflon-coated rubber cap in order to avoid moisture. Each sample was individually prepared and analyzed by GC-FID in triplicate by injecting 1 μ L of the silylated derivatives into a gas chromatograph Hewlett-Packard (HP) Model 5890A equipped with a split/splitless capillary inlet system, a DB-17 fused-silica capillary column (30m x 0.25mm I.D., 0.25 μ m film thickness) and a flame ionization detector, using N₂ as carrier gas (20 Psi, 1.0 ml/min) and a 1: 50 split ratio. The oven temperature was held at 80°C for 2 min, then programmed to 150°C at 30°C/min and to 280°C at 3°C/min. Injector and detector temperatures were set at 265°C and 305°C, respectively. The quantification of the amino acids was calculated from GC analytical patterns considering the total area under the peaks as 100%. Data are shown as mean $\pm$ SD and expressed as mg/L. Quantified amino acids: Aspartic acid (Asp); Glutamic acid (Glu); Serine (Ser); Glycine (Gly); Threonine (Thr); Alanine (Ala); Tyrosine (Tyr); Phenylalanine (Phe); Isoleucine (Iso); Leucine (Leu); Lysine (Lys); Hydroxyproline (Hpro); Proline (Pro); and Cysteine (Cys).

4.4.5 Statistical Analysis and Chemometrics

All processing and data analyses were performed by using the R language through the package *specmine* [8], available in CRAN.

4.4.5.1 Metabolomics Data Processing

The original spectra dataset comprehends a set of 30 wine samples with an interval of peaks from 0.0002 – 12.7966 ppm, having an average of 52346 peaks per sample. All data processing was performed by using the R language through the package *specmine* [8], available in CRAN. The Peak alignment grouped proximal peaks together according to their median position across all samples, using a moving window of 0.03ppm. $1.7E^4$ was the baseline selection limit for the detection of minimum positive values. Typical resonance regions of the deuterium oxide (D₂O)(4.80ppm) and internal standard (TSP)(0.00ppm) signals were removed from the dataset. Peak groups with less than 20% of values were excluded. Taking into account the distinct orders of magnitude of the variables, a generalized logarithmic transformation of the data was performed followed by data standardization. The processed dataset of the 30 CS wine samples was extracted into a file (comma separated values format .csv), where the first column indicates peak positions (ppm) and the second one represents the peak intensities containing a total of 114 peaks per sample, with an interval of peaks comprehending 0.85ppm to 7.37ppm.

The dataset of the amino acids concentration by CG was pre-processed following the same protocol. Missing value imputation was done filling with a constant value of 5e-3. To clean and transform the data, auxiliary functions were used, such as normalization to the mean followed by scaling transformation. The treated datasets were extracted into files (comma separated values format .csv). In the first dataset, the columns indicates amino acids and the rows the amino acids concentration (mg/L) per sample, containing

a total of 14 amino acids concentrations per 18 wine samples (6 groups of samples from different harvest years).

4.4.5.2 Metabolite annotation from ^1H -NMR spectra

In accordance with the need for standardized reporting of metadata or information describing large-scale metabolomics (Sumner et al., 2007) [203]), previously to the ^1H -NMR experiments of the wine samples the minimum reporting standards related to the chemical analysis aspects of metabolomics studies were taken into consideration, in order to obtain accurate, reproducible, and comparable ^1H -NMR data sets. The metabolite annotation was performed through a code developed in R language by the research group [165], [205]. The annotation of metabolites was based on an algorithm developed by the host group [165], by comparing the wines samples spectra with the spectral references available in spectral libraries from free databases Human Metabolome DataBase (HMDB), Biological Magnetic Resonance Data Bank (BMRDB), Spectral Database for Organic Compounds (SDBS) and a list of spectral references previously annotated by the host group. Hypergeometric tests coupled with uniqueness scores was the method selected and the acquisition parameters were taken into account: frequency (500MHz), organism (vine), and solvent (D_2O). The final score for each reference is an average between the matching score, the scores given to the frequency and solvent under which the reference was acquired, and the organism score.

4.4.5.3 Statistical and Machine Learning Data Analysis

The metabolomics datasets (^1H -NMR data and metabolites concentration) were analyzed using statistical univariate techniques, in this case ANOVA and the post hoc Tukey test, to detect eventual statistical differences ($p < 0.01$) derived from the effects of wine production year (as a discrete variable) in the peaks intensity. Further, to understand the relationship between the metabolomics data (^1H -NMR and metabolites concentration - dependent variables) and the age of storage and climatic factors data (independent continuous variables) a linear regression (LR) model was used. The LR models used an adjusted R^2 with a defined threshold $>50\%$ (for NMR data) with p -values < 0.01 .

Unsupervised multivariate statistical models were built, starting with principal components analysis (PCA) followed by two clustering methods, hierarchical clustering analysis (HCA) (present in the article) and K-means clustering (present in supplemental materials). PCA can be used to see if there is a clear discrimination between the samples with different years of wine production in the latent space. Then, two clustering methods were also applied. HCA used an Euclidean distance between samples and the complete linkage method. This parameter was chosen because it tends to find compact clusters of approximately equal diameters and does not force clusters together due to single elements being close to each other, like in single linkage clustering. K-means clustering was performed by separating the samples into six groups ($k=6$) as shown in supplemental material. This allows to compare the obtained clusters with the natural ones, e.g. according to their years of production.

Then, descriptive and classification models were built by applying supervised classification methods.

Regarding the machine learning methods (PLS, RF, KNN), the R package Caret was selected, building a wrapper over a number of different machine learning methods. In both cases, an estimation of the error of the classifier was computed using a 10 times 10-fold cross validation scheme, for different configurations of the internal parameters. Afterward, for each of the methods, the variable importance is estimated for each possible feature (resonance) and the top features are ranked.

Qualitative analyses were also performed taking into account the identified resonance peaks coming from statistical and chemometrics analysis, based on ANOVA and machine learning results. The idea was to verify the overlap of the selected peaks with the quantified metabolites (amino acids).

Multivariate statistical analysis and machine learning approaches to classify musts of Portuguese native grapevine varieties using FT-IR

Abstract

The identification of musts of grapevine varieties has high interest and social-economic value, particularly for Portuguese native wines. These wines possess unique organoleptic properties, resulting from specific grape cultivars. Globally, Portuguese wines made with native grapes had a growth consumption and economic value in the last years.

Fourier Transform Infrared spectroscopy (FT-IR) allows fast and robust spectral recording without sample pre-treatment, but requires spectral signal treatment. Due to the complexity of the generated signals, specific spectral and data pre-processing treatments are needed. Often, these procedures conjugate statistical tools and machine learning approaches to maximize the accuracy of wine must variety classification.

This study aims to assess the capability of common high-speed analytical devices present in the wine industry, such as FT-IR, used to control wine must maturation (grape juices before fermentation), to classify Portuguese native musts of grapevine into categories according to their varieties. Another aim is to describe which specific spectral bands allow discrimination of a given variety and how they influence on the model output.

Our data analysis was applied on one red international must of grapevine variety (Syrah) and four native Portuguese ones (white varieties: Arinto and Viosinho, red varieties: Aragonez and Touriga Nacional) used for wine production from two PDO provenances in the region of Lisbon. The analytical approach was divided into several stages: exploratory and statistical analysis, variety classification, feature analysis, and a comparative analysis between statistical and Machine Learning approach results.

Despite the statistical similitude of the FT-IR spectra among different musts of grapevine varieties (observed through mean spectral differences and ANOVA analysis), Partial Least Squares-Discriminant Analysis (PLS-DA), Logistic Regression (LR) and Support Vector Classifiers (SVC) were the supervised models with highest performances in classifying samples ranging from 0.69 to 0.98 of PR-AUC score for

the different varieties. Arinto and Viosinho were the varieties with the best model classification followed by the red variety of Syrah under this metric.

Finally, features influencing the model output were identified using Shapley values. These features were linked to the correspondent functional chemical groups. Unique features, spectral bands correlated only with the target variety. Aragonez presents two unique features with positive impact on the model output, the bands at 1127 cm^{-1} and 1130 cm^{-1} , related to aliphatic compounds. Syrah has one feature related to other three features, with highest impact for this variety whit a positive impact on the model output, the band at 1161 cm^{-1} , associated to strong vibration in acyl groups of esters compounds. Touriga Nacional possesses three unique features, the bands at 999 cm^{-1} , 1015 cm^{-1} with a positive impact and 1123 cm^{-1} with a negative impact on the model output. These features are related to monosubstituted alkenes. The white varieties have also a unique feature, the band at 2924 cm^{-1} , related to carbon dioxide compounds with a O=C=O bonds pattern vibrations, with a positive impact to classify Viosinho, and a negative one to classify Arinto. In this context, the identified spectral bands are especially significant in wine authentication, providing useful information for musts of grapevine varieties classification of international and Portuguese native varieties.

5.1 Introduction

Wine is one of the most internationally traded agricultural product with a continuous increase in market value. After 1990, wine consumption had an exponential growth worldwide due to the globalization effects [206]. That unprecedented boom led to improvements in the quality and diversity of available wines [206]. Winemaking has a high cost and is a complex process based on the combination of grape varieties, *terroir* effects, harvest timing, harvest methods, fermentation, storage conditions, and ageing, as well as many other conditions capable of influencing the wine quality and authenticity [1].

Since 2017, according to data published by INE (Portuguese National Statistics Institute), Portuguese wines had exponential growth in consumption. This rise also continued during the global pandemic challenges of 2020. The exportation growth led to an increase in the average price. This event also occurred on PDO (Protected Designations of Origin) Portuguese wines made from native grapevine varieties.

Portugal has more than 250 native grapevine varieties and has 31 wine regions with PDO, defined as geographical areas where certain types of grapes are allowed, obeying constraints such as regulated winemaking processes with a maximum of vine yields. These constraints aim to ensure the quality and origin of that beverage, according to the classification of the European Union (OIV, CONGRESS 2011).

The Portuguese native grapevine varieties give rise to variety-based wines with unique organoleptic characteristics such as color, aroma, flavor, and alcohol/ acidity balance [207], [208]. It is paramount to ensure the authenticity of the grapevine must of given varieties to avoid not allowed grapevine must blending or other events such as mislabeling. Thus, the capacity to confirm grapevine must variety in a winemaking process has social and economical importance (INE), by managing and preserving the

heritage, the *terroirs* and the PDO regions.

Several methodologies based on isotopic, molecular and spectroscopic techniques, combined with various chemometric methods, have proved to be helpful in winemaking to discriminate distinct wine grape varieties [209], [210], [211], [212], [213].

In addition, the most “easy to use”, robust, economic, and globally applied analytical technique is Fourier Transform Infrared (FT-IR) spectroscopy, coupled with chemometrics tools, by international organizations, companies, and also in academic research [210], [212].

FT-IR spectroscopy is used in wine research and winemaking, as it provides a simple, fast, cost-effective, and non-destructive way to analyze samples without the need for chemicals or preparatory steps [210], [212]. Nonetheless, this analytical technique has also commonly been used in winemaking for grapevine must maturation control, being the most common tool in the wine industry.

However, there is still a reduced number of publications and applications regarding the differentiation of samples based on wine must variety, using FT-IR [210], [212]. There is even less research work directed to the classification of Portuguese wines produced with native grape varieties, which achieved increased market value in the last years, especially those which are produced in PDO regions (INE).

In this study, machine learning algorithms were applied to a FT-IR dataset coupled with chemometric algorithms aiming to discriminate musts of grapevines according to their varieties. For that, the dataset used came from routine analysis used by industry to control the quality of the wine must before the fermentation process. Thus, these methods permit to reuse the data set for authentication and characterization purposes.

This study is the first one to include the red and white Portuguese grape varieties produced in PDO regions of Portugal, besides an international red variety in a metabolomics dataset, combined with statistical, machine learning tools and model interpretability.

The application of data analysis tools on metabolomics wine must data provides a comprehensive discussion capable of bringing new highlights to the field of wine production and quality, helping to characterize the Portuguese native wines produced in Portugal, and consecutively contributes as another tool that can be used in their authentication process.

The samples of grapevine musts investigated in this work are from Portuguese native grapes, Arinto and Viosinho (white grapes), Aragonez and Touriga Nacional (red grapes), as well as Syrah, an international red grape, from two vineyards of the geographical area of Lisbon district (PDO - Óbidos and, PDO - Alenquer). The grapes used in our study are considered high-quality ones. The major differences between the two white must grapes used in our study are the maturation time, and the bleeding process. Viosinho has a reduced maturation time compared to Arinto, as Viosinho is one of the selected mono-varietal grapes considered to the production and preparation of blending wines, and also considered to add structure and flavor to the fortified Port wines, as well as being used to blend with other grape varieties [208]. The red varieties Aragonez and Touriga Nacional present an average maturation time, but Touriga Nacional grapes produce a wine must with a slightly higher sugar content compared to Aragonez ones [214]. The international Syrah grape variety is well adapted and allowed to be cultivated in the two Portuguese PDO

regions considered in this study. It can be used to produce mono-varietal wines and also to blend with other varieties, more commonly with Touriga Nacional [215].

The study approach is divided into the following stages: exploratory and statistical analysis, variety classification feature analysis, and a comparative analysis between statistical and Machine Learning results. The first objective deals with the complex discrimination of varieties based on the visual analysis of the FT-IR spectra and statistical results to identify and understand the chemical discrepancies between the investigated grape varieties. The second goal is to create accurate classification models of wine must variety using five machine learning approaches. These models are combined with evaluation metrics (Precision-Recall Area Under the Curve, Receiver Operating Area Under the Curve, Mathew Correlation Coefficient, and other metrics), feature selection methods, and preprocessing data treatments. The final goal of the work is to provide access to feature importance and enable structural interpretation of model classification. FT-IR coupled with chemometrics data analysis tools may provide an accurate and reliable method to discriminate grape must varieties based on some identified spectroscopic bands, as shown by the results.

5.2 Results

This study contemplates the analysis of FT-IR metabolomics data set of 96 Portuguese wine must samples from five wine grape varieties produced in Lisbon region, during two consecutive harvests. Two datasets were organized from the samples based on the wine type of must (red and white).

The data were preprocessed as mentioned in Materials and Methods section. Firstly, a exploratory and statistical analysis was conducted, followed by an untargeted and qualitative metabolomics analysis with ML models to classify each grape must variety, followed by analysis of the features with importance on the model classification using Shapley values to better understand the impact and how the features (FT-IR spectroscopic bands) influence on the model performance.

As mentioned in the Materials and Methods section, each group of wine must samples was produced in two places of the same PDO region in Portugal and kept under similar conditions of light, humidity, and temperature over the storage time till analysis. Metabolomics based-approaches were used in all samples during the control maturation time and the resulting data were combined with univariate and multivariate statistical techniques and machine learning tools to answer to the aims of the study.

5.2.1 Exploratory analysis

Visual observation of the FT-IR spectra of grapevine musts samples, does not allow to discriminate the samples according to their grape must variety (Figure 20A and B). This difficulty is expected due to the type of grapevine musts, the proximity of production region and the biological origin of the grapevine varieties, and winemaking process used on the samples of this study.

Looking at the fingerprint region interval $[926-1450\text{ cm}^{-1}]$, it was possible to observe that some band

regions with different intensity levels. This fingerprint region is known due to the bending and structural vibrations, which are particularly sensitive to large wavenumber shifts, and permits the identification of specific functional groups. As an example, it was observed spectral differences at the absorption bands around 960 cm^{-1} and 990 cm^{-1} . These bands are related to aromatic compounds and present an angular deformation at CH sp^3 outside the aromatic ring. Usually, absorptions bands between $900\text{--}1500\text{ cm}^{-1}$, with CO bonding are related to organic acids such as the phenolic ones. In the spectral window at $1000\text{--}1100\text{ cm}^{-1}$ it was also noted some vibrations with different intensity levels associated with CO group present in glucose and fructose molecules. In the $1050\text{--}1150\text{ cm}^{-1}$ spectral window related to alcohols, some CO stretching bands at 1061 cm^{-1} and 1065 cm^{-1} bands were also found.

Another interval region between $1200\text{--}1450\text{ cm}^{-1}$ showed intense bands differences related to ring vibrations and bending CH sp^2 also reported by Murru et al., 2019[212]. Finally, spectroscopic discrepancies were also found at $1800\text{--}1880\text{ cm}^{-1}$, and $2900\text{--}3000\text{ cm}^{-1}$ regions assigned, respectively, to alkenes with stretching of double bond CO sp^2 and alkanes with stretching of CH group at $2950\text{--}2960\text{ cm}^{-1}$ [216], [217].

Besides visual inspection of the FT-IR spectra, in order to explore and identify metabolomic differences concerning the grapevine musts varieties, a statistical analysis was performed.

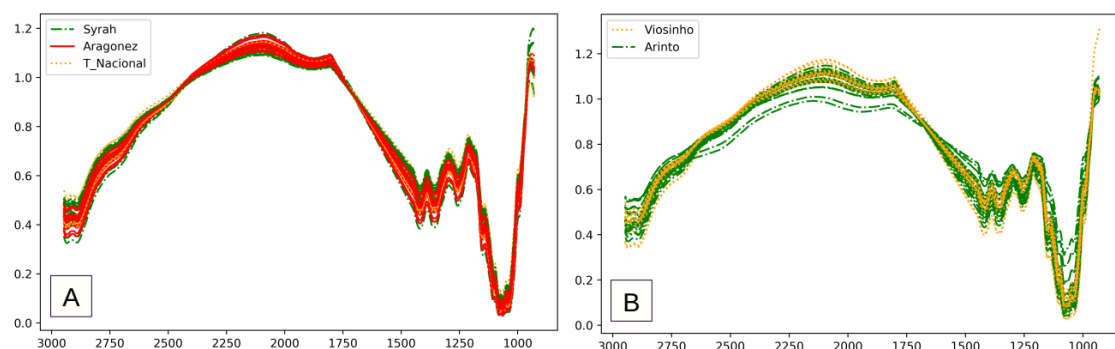


Figure 20: **FT-IR spectra plot for wine must samples grape varieties without water region ($1450\text{--}1800\text{ cm}^{-1}$)**. A: Plot of FT-IR spectra of red musts of grapevine varieties samples. B: Plot of FT-IR spectra of white musts of grapevine varieties samples.

5.2.2 Statistical analysis

To understand if there were statistical differences between the spectral information of each grape variety group an ANOVA followed by the t-Test and Tukey's HSD test were applied to the FT-IR dataset to determine spectral differences between each grape variety pair with the information regarding the features (i.e., spectral bands) which allow performing linear discrimination of each grape variety pair.

This analysis gives an indicative result to estimate the ML algorithms performance to classify each grape must variety. By knowing which variety pair have more features with the ability to discriminate according to the target variety, one will have an idea of which varieties will be easier to achieve a good classification score and the same as the inverse results. With these results one can test and select the

best feature selection, evaluation and scoring methods to the models variety classification. By using ML algorithms combined with feature selection, a sub-data set of features is used to classify each variety, but here the selected features can have non-linear relationships between other features. Thus, the other important reason to perform ANOVA analysis is to check if the identified features with linear capability to discriminate varieties are the same as the calculated important features used on model variety classification.

In Table 13 the ANOVA results with t-test shows the spectral bands (cm^{-1}) identified with the lowest corrected p values (< 0.01) for each pair variety as well as the functional chemical groups related to the identified spectroscopic bands.

Table 13: **ANOVA results. Wavenumber (cm^{-1}) with the lowest corrected p, values (< 0.01).**

The first column shows the t-test, which consists of the pairs of grapevine musts varieties that were significantly different in terms of means for each feature, the second one the respective spectroscopic bands, the third column includes the chemical groups related to the wavenumbers identified.

Tukey test (pairs of grapevine musts varieties)	Wavenumber (cm^{-1})	Chemical Groups
Aragonez vs Touriga Nacional	2203, 2207, 2211	Spectral window 2010-2342 cm^{-1}: Wavenumbers related to sp C-X triple bonds and stretching of carbon-carbon. (3 bands)
	1802, 1806	Spectral window 1802-1806 cm^{-1}: related to aromatics compounds with single CH bend. Wavenumbers at 1770-1800 cm^{-1} and 1735-1810 cm^{-1} are assigned to conjugated acids and small rings of esters, respectively. (2 bands)
	988, 992, 995, 999, 1003, 1007, 1011	Spectral window 988-1011 cm^{-1}: Wavenumbers at 600-1000 cm^{-1} and 1000-1350 cm^{-1} are related to alkenes sp ² bend and amines with N-C bend, respectively. Wavenumbers at 1000-1260 cm^{-1} are related to alcohols with C-OH bend. And the wavenumbers 988 cm^{-1} , 992 cm^{-1} , 995 cm^{-1} and 999 cm^{-1} are related to aromatics compounds with single C-H bend. (7 bands)
Aragonez vs Syrah	1123*, 1127*, 1165, 1169, 1173, 1177, 1181, 1184, 1188, 1192, 1196, 1200, 1204, 1208, 1211	Spectral window 1123-1200 cm^{-1}: Wavenumbers at 1000-1350 cm^{-1} are related to amines with N-C bend. Wavenumbers at 1000-1260 cm^{-1} are assigned to alcohols with C-OH bend. Wavenumbers at 1040-1250 cm^{-1} are related to ethers. Wavenumbers 1123 cm^{-1} and 1127 cm^{-1} are related to aliphatic compounds. (15 bands, 2 unique features)
Syrah vs Touriga Nacional	992, 995, 999, 1003, 1007, 1011, 1015*	Spectral window 992-1015 cm^{-1}: Wavenumbers at 600-1000 cm^{-1} are related to alkenes sp ² bend, aromatics compounds with single C-H bend. (7 bands, 1 unique feature)
Arinto vs Viosinho	995, 999, 1003	Spectral window 992-1003 cm^{-1}: Wavenumbers at 600-1000 cm^{-1} are related to alkenes sp ² bend, assigned to aromatics compounds with single C-H bend. (3 bands)

The results based on the spectral absorbance profiles reveal that Aragonez vs Syrah was the pair of grape must variety with higher discrepant FT-IR bands i.e., 15 signals. Interestingly, all of these bands were detected in the region at 1123-1200 cm^{-1} , which is related to amines with N-C bend, alcohols with C-OH bend, and ethers. The spectral bands at 1123 cm^{-1} and 1127 cm^{-1} are related to aliphatic compounds and these features will also be selected as important ones in the red varieties classification models. The variety pair Aragonez vs Touriga Nacional was the second one with more features, i.e., 12 FT-IR bands with linear capability to discriminate the pair variety. The majority of these bands was found to be at 988-1011 cm^{-1} region that has been related to the alkenes sp² bend and associated to aromatics compounds with single C-H bend. The remain features were detected at 2010-2342 cm^{-1} spectral window, which is related to stretching bonds of carbon, and also at 1802-1806 cm^{-1} , where the bands have been assigned to aromatics compounds and to conjugated acids. FT-IR bands of the spectral windows at 1735-1810 cm^{-1} are related to sp² C-X double bonds related to ester, ketones and anhydrides compounds.

The variety pair Syrah vs Touriga Nacional showed the less number of discriminant features, only 7

spectral bands at 992-1015 cm^{-1} region. These features have been related to alkenes sp^2 bend and to aromatics compounds with single C-H bend.

Considering the variety pair of white grapevine musts varieties, Arinto vs Viosinho, the ANOVA results revealed only 3 features, with spectroscopic bands at 995-1003 cm^{-1} , which have been associated to alkenes sp^2 bend and to aromatics compounds with single C-H bend.

The interesting results from ANOVA were the identification of some resonances with capability to distinguish the variety pairs Touriga Nacional vs Aragonez (12 bands), Touriga Nacional vs Syrah (7 bands), and Arinto vs Viosinho (3 bands).

These results allow us to perceive that some FT-IR spectral windows can linearly discriminate grapevine musts varieties. They also reveal that the majority of the FT-IR spectral signals have a common expression in all varieties, as previously noted by the visual inspection of the FT-IR spectroscopic profiles.

To overcome these constraint, machine learning models were constructed with the main goal to classify the grapevine musts varieties samples according to the variety.

5.2.3 Supervised Data Analysis: Musts of grapevine variety classification

To overcome the challenge to positively predict wine grape must varieties, machine learning models were constructed with the main goal to classify the musts of grapevine samples according to the variety.

Firstly, unsupervised models were also executed (PCA, t-SNE, HCA), but they did not show a good cluster separation regarding the grapevine musts varieties investigated. The overall results showed overlap between all varieties, as expected based on the previous statistical analyses and, for this reason, these results are not presented and not discussed in this chapter. The results concerning the non-supervised approach are shown in the supplemental material (Figures S.37, 38, 39, 40, 41, 42) for the interested reader.

5.2.4 Variety classification

Based on the previous results of the univariate and the multivariate analyses, it was observed that by using statistical tools there were only a few number of features which allow to linear discriminate grapevine musts varieties. Besides, looking to the results from the unsupervised cluster analyses performed there were a high overlap between the grapevine musts samples from all varieties investigated.

To overcome the difficulty of FT-IR complex matrix data, five supervised classification models, GBC, LR, PLS-DA, RF, and SVC were calibrated with hyper-parameters optimization and evaluated, leading to 15 and 5 models for red and white grape varieties, respectively. The performance of all produced models is presented in the supplementary material (in Supplementary Tables) for the interested reader.

As previously mentioned, the main objective of this exploratory study is to build classification models with the highest performance and accuracy to classify each wine must variety, and then use interpretability tools to understand how the features with higher importance explain those models.

However, according to the literature, to achieve the most accurate classification and to analyze the importance of features without misleading interpretation, several treatments of data should be careful conducted:

- Water region of the FT-IR spectra of grapevine musts samples should be eliminated to avoid noises which influence the models performance. In our data, the water region was eliminated according to the reference of the WineScan manufacturer.
- The treatment of FT-IR spectral signals of grapevine musts samples need to be included to avoid noises and other misclassified points [210], [211], [213], [218]–[220]. In our study, after testing several methods combinations, SNV plus standardization achieved the best results on model performance.
- To improve the model classification performance, feature selection methods should be used to reduce the problem dimensionality and to facilitate and improve the computation of feature importance [221]. In our study, two filter feature selection methods were used. Two methods allowed to detected and treat constant, quasi-constant, duplicated and correlated features. Only correlated features were detected and treated. The selected method was used to find groups of correlated features and to select from each group the feature with highest variance. As result, the feature selection method returned a sub-dataset of features that are uncorrelated with each other based on the utilized threshold. As mentioned above, the similarity of samples among different grape varieties generated a high number of correlated features. However, in our study if all correlated features were eliminated, only a reduce number of variables would be available and useful information could be lost. Thus, to overcome this constraint, several thresholds of variance were tested among with the default 0.80 (0.70, 0.75, 0.90, and 0.95). The use of Pearson correlation coefficient with the threshold of 0.90 was the method that achieved the best results on the models performance. This method reduce the 433 features to 20 features for the red varieties and 433 to 17 features for the white varieties. As consequence, the computation of features importance and the interpretability of the model output magnitude was also improved in all variety models, by eliminating features with a correlation equal or higher than 90%. The Tables 25 and 26 present in the supplemental material, shows each group of correlated features and the selected feature with the highest variance for red and white grape varieties.
- To evaluate and compared the models performance, there are several available approaches and different scoring metrics. In authentication problems as our study, it is necessary to adequate the validation strategy to evaluate the model performance to ensure that the model is able to generalize well for every possible target. Binary classification model was the select approach to predict the target variety and due to the reduced number of samples, repeated nested k-fold cross-validation was the selected method to infer the model performance (details are shown in the Materials and Methods section). To compare each model with one another, we selected different scoring metrics.

Since the main aim of the study is to classify a target must variety from any other source. This mean to classify the entire possible positive class target and to take attention to sensitivities.

Precision-Recall curve (PR curve) and Receiver Operating Characteristic curve (ROC curve) are well suitable approaches to evaluate binary classification outcomes. However, for class imbalance, as the Aragonez variety, PR curve is more sensitive than ROC one, by changing the curve shape more drastically. Other metrics are suitable for our problem as Matthews correlation coefficient and Balanced Accuracy. These metrics also provide accurate evaluations for imbalance classes in binary classification models. However, these metrics require a defined decision threshold.

To evaluate and compare our models results, several metrics (as shown in the supplementary material (in Supplementary Tables) were tested, but focus was putted on the PR-AUC, ROC curve and also at the MCC, with a threshold calculated based on the PR threshold mean. Because this is an exploratory study with contributions to wine authentication, the score metric on the PR threshold mean was computed instead of the default threshold, due to the capability of PR curve metric to better test all possible positive class target with attention to the sensitivity of the model. Nevertheless, the selected threshold can be changed according to the sensitivity and cost established or required for the classification model, according to the defined wine authentication regulations.

Based on those assumptions, the models from the cross-validation with higher score metric of PR-AUC were chosen to classify each given wine must variety. Table 14 presents the results of classification models with cross-validation for all varieties with the highest performance scores.

The results using FT-IR spectral data set as input features show that PLS-DA, LR, and SVM were the models with the highest performance to classify the varieties of our studied grapevine musts samples.

For Syrah variety the best model was PLS-DA, reaching 0.97 of PR-AUC. The red Portuguese native varieties Aragonez and Touriga Nacional achieved reduced performances on their classification, as the models with the higher performances were different from the Syrah. For Touriga Nacional, the best model was LR reaching 0.88 of PR-AUC, and for Aragonez the best model was SVC, reaching 0.69 of PR-AUC.

The white varieties also achieved higher classification performance with PLS-DA model, reaching 0.98 and 0.94 of PR-UC for Arinto and Viosinho varieties, respectively.

Table 14: **Performance scores of the best classification models for the grapevine musts varieties studied.** All models had pre-process methods(SNV plus Scaling) with the application of Feature Selection (Drop of constant features with 90% of correlation). The MCC evaluation metric followed a threshold based on the calculated PR Threshold mean. Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC:Matthews Correlation Coefficient; BA: Balanced Accuracy; PR Threshold: Precision-Recall mean Threshold; SD: Standard Deviation.

Target Variety	ML Model	PR-AUC $\pm$ SD	ROC-AUC $\pm$ SD	MCC $\pm$ SD	BA $\pm$ SD	PR Threshold $\pm$ SD
Arinto	PLS-DA	0.98 $\pm$ 0.008	0.97 $\pm$ 0.013	0.92 $\pm$ 0.04	0.96 $\pm$ 0.02	0.72 $\pm$ 0.06
Viosinho	PLS-DA	0.94 $\pm$ 0.03	0.97 $\pm$ 0.015	0.92 $\pm$ 0.04	0.96 $\pm$ 0.02	0.86 $\pm$ 0.03
Aragonez	SVC	0.69 $\pm$ 0.07	0.87 $\pm$ 0.04	0.49 $\pm$ 0.11	0.74 $\pm$ 0.05	0.55 $\pm$ 0.06
Syrah	PLS-DA	0.97 $\pm$ 0.01	0.95 $\pm$ 0.02	0.88 $\pm$ 0.04	0.94 $\pm$ 0.02	0.54 $\pm$ 0.04
Touriga Nacional	LR	0.88 $\pm$ 0.05	0.95 $\pm$ 0.01	0.75 $\pm$ 0.07	0.88 $\pm$ 0.04	0.57 $\pm$ 0.05

In our analyses the classification models were built for both white wine must varieties together due to the different number of samples for each wine must variety. Thus, since the class sizes were not perfectly balanced, some differences in the models performance was expected when considering the target grapevine must variety. Here it can be observed that despite the differences of the classes, metrics evaluation as BA and MCC considered the same weight to both classes, but PR-AUC adjust that imbalance size with an adjusted class classification threshold, leading to a different performance according to the selected variety target. ROC-AUC, as expected, did not adjust the model evaluation. These results reinforce the sensibility of PR-AUC for binary classification models with non-balanced classes, as in our study.

Despite the complexity of the data and confounding influences, the results reveal that PLS-DA, LR and SVC were the models with better performance to classify the grapevine musts varieties, with good PR-AUC scores reaching 69-98%. It was also observed that the use of feature selection in our analytical approach contribute to build more accurate models (by using a sub-data set of features with less correlation between themselves than the original data set). This approach combined with the proper model evaluation and the use of criterious score metrics lead to a selection of models with more precise performances and contribute to more accurate model's interpretation.

5.2.5 Features analysis

ML models are generally considered 'black boxes', but the ability to easily interpret the results of a variety classification model would improve our understanding of the factors underlying the differences between each wine variety, rather than merely classify then. This information can be used in wine industry to improve variety authentication process and to avoid events of fraud. Studies using FT-IR data set and specifically of grapevine musts samples seem to be scarce addressing to the issue of variety classification models interpretation, and also with a reduced number of works to classify grapevine musts samples according to the variety [210], [209]. Our work followed this direction, in order to build different ML models using FT-IR spectral profiles from wine must samples as input data to find out which bands (i.e., features) are the most predictive to classify each variety and how they are related to the model performance.

Besides, by determining which features contributed most to a certain wine variety classification would help to uncover potential biomarker bands associated to functional chemical groups, i.e., chemical signatures.

To assess the feature importance, the SHapley Additive exPlanations (SHAP) value estimation method was adopted from the field of game theory. This method can be applied for classification problems, as in our study, by computing the average marginal contribution of a feature value across all possible coalitions [57], [56], [58].

By computing the Shapley values of the best variety classification models, it was analyzed the selected features with the higher impact on the model performance their contributions to each variety classification. An important note to be considered is that to compute Shapley values, KernelShap need to use a reduce number of features due to the time required to work, and also correlated features should be previous eliminated to avoid to weighting the features wrongly, leading to an incorrect interpretation of the model performance. As mentioned above in the variety classification section and bellow in the Materials and Methods section, correlated features were identified and treated.

The computed mean Shap values for the models with the higher performance of all varieties resulted in 20 features for red grapevine musts varieties and 17 ones for white grapevine musts varieties.

The identified features with importance for the red and white classification models are present in all models; however, their impact and order of importance are different for each variety model. In Figures 21, 22, 23, 24 the Shapley values are shown for the models with the higher performance for each red and white varieties.

5.2.6 Features analysis of the best models performance for each Red Variety

For the red varieties classification, Shapley values calculated an average range of impact from 0-0.30. The results for the mean Shap values and their average impact on the model output are shown to each feature for every red varieties models classification at Figures 21, 22.

Syrah

Considering the best classification model of the Syrah variety, PLS-DA, the calculated Shapley values reached an average impact of 0.30 on the model output. The feature with highest influence on the model classification was the band at 1161cm^{-1} , reaching an average impact around 0.30 on the model output. That spectral band is related to strong vibration in acyl groups (C-O bond) of esters and it was computed by Shapley values of the best model for Aragonéz variety. That feature also presented a higher impact to Aragonéz variety with SVC model, with an average impact on the model output of 0.25. In the Figures 21 A and C, it can be observed that the feature at 1161 cm^{-1} classify the variety Syrah with high values, and a positive impact for the PLS-DA model. Similarly, that band also classify the Aragonéz variety with the high shap values with a negative impact on the SVC model. This means that different values of this important feature are the way to classify two red different varieties.

The feature with the second highest influence on the model classification of Syrah variety was the spectral band at 999 cm^{-1} , reaching an average impact around 0.20 on the model output. That band has been assigned to the presence of monosubstituted alkene. This feature was also computed by Shapley values of the best model of Touriga Nacional variety, and showed the higher impact by LR model, with an average impact on the model output around 0.25. This feature also influenced both models with different impacts. Besides, looking at the calculated Shaply values at the Figure 21 A and B, it can be observed that the signal at 999 cm^{-1} classify the Syrah variety with high values, showing a negative impact on the PLS-DA model. Similarly, that feature classify the variety Touriga Nacional with high values, with a positive impact for the model of LR. This means that the computed Shapley values explains how the feature at 999 cm^{-1} has different impacts to classify two red varieties. Another FT-IR spectral band that showed the same effect on both Syrah and Touriga Nacional classification models was found at 1015 cm^{-1} . This feature is associated to sp^2 C-H bends patterns for alkenes and can be related to C-O bend of alkoxy group from esters or alcohols compounds. To classify Syrah variety the high values of the feature produced a negative impact on the PLS-DA model output, as its high values showed positive impact on the model LR output to classify Touriga Nacional.

The followed important features to classify Syrah variety can be observed also in the other varieties classification models, some of them have the same importance order, as 1146 cm^{-1} with Aragonez and 2897 cm^{-1} with Touriga Nacional. Other features with different orders and impacts can be observed in Figure 21 A, B, C and Figure 22 A, B, C.

Touriga Nacional

Considering the classification of Touriga Nacional variety, the model with higher performance was the LR and the calculated Shapley values reached an average impact around 0.25 on the model output. The feature with highest influence on the model classification was the signal at 999 cm^{-1} , as previously noted in the Syrah variety classification, reaching a positive and average impact on model around 0.25 as observed in Figure 21 B, Figure 22 B.

The spectral band at 1015 cm^{-1} achieved the second position, being a feature not correlated to another one in the FT-IR data set, with a positive impact on the Touriga Nacional LR model of 0.14. This feature is associated to the sp^2 C-H bends patterns for alkenes, and can be related to C-O bend of alkoxy group from esters, or alcohols compounds. This feature is also present in other models varieties. However, it showed a higher impact on the Touriga Nacional classification compared to the other red wine must variety models. Looking at the other red wine must varieties, for Aragonez and Syrah varieties this feature had less weight in the classification, and in both models, their shap high values has a negative impact on the outputs. Taking into account the FT-IR data set of our study, this feature may be studied in future works as a potential biomarker to classify the grapevine musts samples of Touriga Nacional variety.

Another and not correlated feature in the FT-IR data set was the spectral signal at 1123 cm^{-1} , with the 8th position in importance in the LR model output. This feature also achieved the highest position in the classification models of all red grape varieties. This FT-IR band has been associated to the sp^2 C-H bends patterns for alkenes and can be related to C-O bend of alkoxy group from esters or alcohols compounds.

This unique feature showed the higher weight when compared to the other models varieties, but a negative impact to classify Touriga Nacional. Looking at the other red varieties, for Aragonez and Syrah ones this feature had less weigh, and in both cases their shap high values have a positive impact on the models outputs. It is worth mentioning that the spectral band at 1123 cm^{-1} can be further study to understand its potential as a biomarker for the classification of grapevine musts samples of the Touriga Nacional variety.

The features utilized to classify Touriga Nacional are also present in the other wine must varieties, with distinct impacts, and order, as it can be observed in Figure 21 A, B, C and Figure 22 A, B, C.

Aragonez

Considering the classification of Aragonez variety, the model with higher performance was the SVC and the calculated Shapley values reached an average impact around 0.25 on the model output. The feature with highest influence on the model classification was the signal at 1161 cm^{-1} , as previously found in the Syrah classification, reaching an average impact on 0.25, a value lower in comparison to that noted in the Syrah classification. As mentioned above, for the Aragonez variety the high values of that feature had a positive impact on the classification of its wine must samples.

As previously seen in the models, other important features to classify Aragonez variety can also be found in the other grapevine musts variety classification models, with the same/or not order and impact, as shown in Figure 21 C, and Figure 22 C.

As noted for Touriga Nacional variety, there were two unique features in the highest position for all red varieties classification. In the rank of the 2nd position, the spectral band at 1127 cm^{-1} . Was found and has been assigned to the sp² C-H bends patterns for alkenes and can be related to the C-O bend of alkoxy group from esters or alcohols compounds. This feature (1127 cm^{-1}) showed a positive and higher impact (i.e., 0.17) in Touriga Nacional wine must variety than in the remaining models varieties. Looking at the other red wine must varieties, for Touriga Nacional and Syrah this feature showed shap values with negative impact on the models output and with lesser weigh in both models. In the 5th position it was found the FT-IT signal at 1130 cm^{-1} , associated to the sp² C-H bends patterns for alkenes and also related to C-O bend of alkoxy group from esters or alcohols compounds. This unique feature had shap values with the higher weight and positive impact when compare to the other models varieties.

Looking at the other red wine must varieties, for Touriga Nacional and Syrah this feature have showed less weigh values, and in both cases their shap high values had a negative impact on the models outputs.

As mentioned for the Touriga Nacional, taking into account the FT-IR data of our study, i.e., the spectral bands at 1123 cm^{-1} and 1127 cm^{-1} , might be investigated in future works as potential biomarkers to classify grapevine musts samples of Aragonez variety.

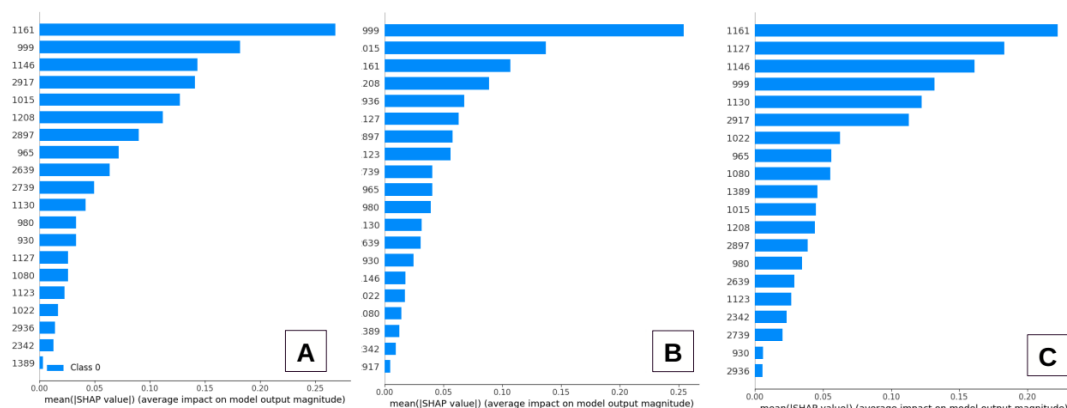


Figure 21: **Top 20 most important features ranked by mean absolute SHAP values.** The features are ordered according to their importance of the best model classification for each red wine must variety. The Mean Shap values and their average impact on model output are shown for each feature. A: SHAP values and the average impact of feature's on model of PLS-DA to classify Syrah. B: SHAP values and the average impact of feature's on model of LR to classify Touriga Nacional. C: SHAP values and the average impact of feature's on model of SVC to classify Aragonéz.

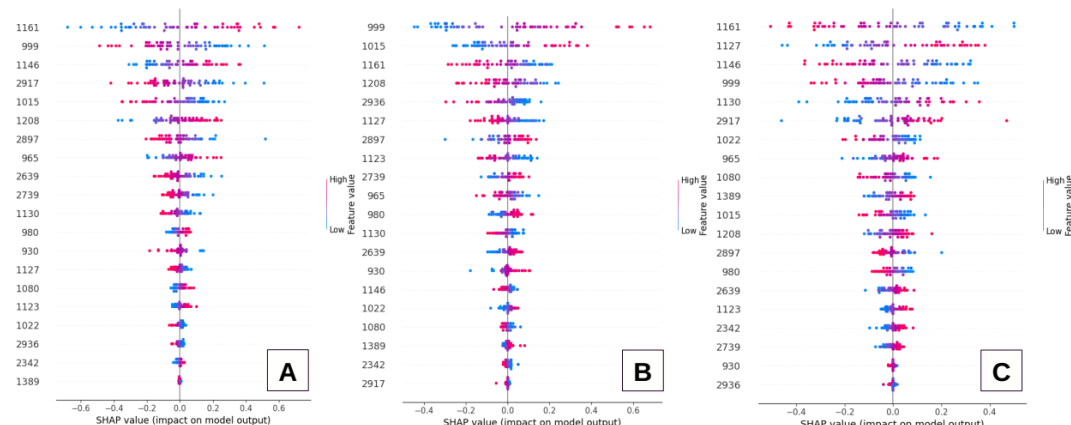


Figure 22: **The 20 most important features are shown and ordered according to their importance to the best model classification for each red wine must variety.** The Shap values and their impact on model output are shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of PLS-DA to classify Syrah. B: SHAP values and impact of feature's on model of LR to classify Touriga Nacional. C: SHAP values and the impact of feature's on model of SVC to classify Aragonéz.

5.2.6.1 Features analysis of the best models performance to each White variety

For the white wine must varieties classification, Shapley values calculated 17 important features with impact on the models output, with an average range of 0-0.30 (30% of impact). The results for the mean Shap values and their average impact on model output are shown for each feature of Arinto and Viosinho varieties with higher PR-AUC classification models in Figures 23 and 24.

Arinto vs Viosinho Considering the classification of these varieties, the model with higher performance was the PLS-DA and the calculated Shapley values reached an average impact around 1.75 on the model output.

The feature with highest influence on the model classification was the signal at 988 cm^{-1} (PLS-DA), associated to the presence of monosubstituted alkenes. It achieved a 1.75 of the impact model (PLS-DA, SVC) and considering the target varieties in study, the impact of this feature was different for each grape genotype. Thus, to classify the Arinto variety, that feature showed the higher shap value and positive impact on the model output and the inverse to classify the Viosinho variety.

The spectral band at 1177 cm^{-1} was the 2nd position in importance, with a impact on the classification model around 1.40. Such a signal has been associated to strong vibration of acyl groups (C-O bond) of esters or anhydrides compounds.

The remaining features identified presented lesser impact on the model output and considering the target variety, the impact on the model was the inverse of the other variety classification models.

To classify Arinto and Viosinho, a unique feature has been detected, i.e., 2924 cm^{-1} , at the 15th position and, therefore, with a reduce impact on the model output. This spectral band is related to carbon dioxide compounds with a O=C=O bonds pattern vibrations and showed a reduced impact in the classification model. For the classification of Arinto, the feature with a high shap value had a negative impact on the model output, as the opposite was found to classify the Viosinho variety. This feature was the only one correlated to the target variety and to infer on its potential to be used as biomarker, other white varieties should be added to the data set and the same analysis performed again.

An important consideration refers to the PLS algorithm that was constructed for continuous regression settings and does not possess any explicit restrictions, forcing the output to be 1 or 0. Thus, the output can possess any value as herein shown in the computed shapley values, with a range of 0-1.75. However, if the model is well calibrated, the result will be close to one of the binary outputs. In our case, some of the samples of the same variety presented features with different values, what can influence the computation of the shapley values.

To sum up, the 20 features with highest Shapley values from the best red varieties classification models revealed four unique features, the bands at 1015 cm^{-1} , 1123 cm^{-1} , 1127 cm^{-1} , and 1130 cm^{-1} . Interestingly, for the white varieties Arinto and Viosinho, only the band at 2924 cm^{-1} , from the 17 features with importance for the model varieties classification, was identified. It was also noted that there are four important features with impact in the best models classification of red and white wine must varieties, e.g., the spectral bands at 930 cm^{-1} , 980 cm^{-1} , 1080 cm^{-1} , and 2639 cm^{-1} .

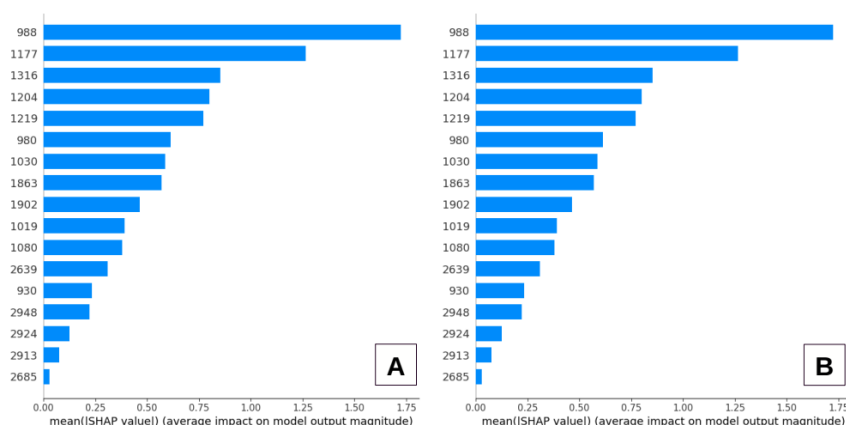


Figure 23: **Top 17 most important features ranked by mean absolute SHAP values.** The features are ordered according to their importance of the best model classification for each white wine must variety. The Mean Shap values and their average impact on model output are shown for each feature. A. SHAP values and the average impact of feature's on model of PLS-DA to classify Arinto. B. SHAP values and the average impact of feature's on model of PLS-DA to classify Viosinho.

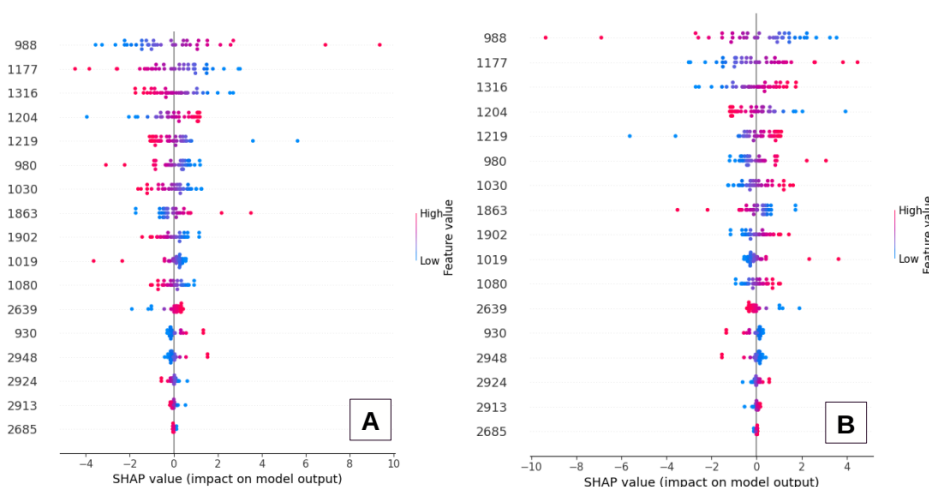


Figure 24: **The 17 most important features are shown and ordered according to their importance to the best model classification for each white wine must variety.** The Shap values and their impact on model output are shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A. SHAP values and impact of feature's on model of PLS-DA to classify Arinto. B. SHAP values and impact of feature's on model of PLS-DA to classify Viosinho.

5.2.7 Comparative analysis between Statistical, Machine Learning approach and Shapley values results

Table 15 contains the features computed by shapley values with highest impact on each wine grape must variety, and some of them are unique features, as previous mentioned, these features are spectral bands

correlated only with the target variety. It is worth mentioning that some of the features with importance were previously identified by ANOVA. These results allow to understand that statistical tools are useful to discriminate wine grape must varieties, but the use of ML algorithms (with adequate evaluation methods, features treatments, and score metrics), combined with the Shapley values computation, gives more information to discriminate and classify grapevine musts varieties.

Table 15: **Comparative analysis between Statistical, Machine Learning approach and Shapley values results.** Note: Arg: Aragonez; TN: Touriga Nacional; SY: Syrah; LR: Logistic Regression; SVC: Support Vector Classification; PLS-DA: Partial Least Squares-Discriminant Analysis; (-): Negative impact on the model output; (+): Positive impact on the model output.

Feature	ANOVA Group Varieties	ML Variety Classification and Shapley values	Correlated Features
999	TN vs SY	TN variety: LR model [0.25 impact, (+) impact, 1st position (other varieties have reduced impact, lower ranking order, and (-) impact)]	3 correlated features
1015 - unique feature	and TN vs Arg TN vs SY	TN variety: LR model [0.14 impact, (+) impact, 2nd position (other varieties have reduced impact, lower ranking order, and (-) impact)]	0 correlated features
1123 - unique feature	Arg vs SY	TN variety: LR model [0.07 impact, (-) impact, 8th position (other varieties have reduced impact, lower ranking order, and (+) impact)]	0 correlated features
1127 - unique feature	Arg vs SY	Arg variety: SVC model [SVC model: 0.17 impact, (+) impact, 2nd position (other varieties have reduced impact, lower ranking order, and (-) impact)]	0 correlated features
1130 - unique feature	—	Arg variety: SVC model [0.12 impact, (+) impact, 5th position (other varieties have reduced impact, lower ranking order, and (-) impact)]	0 correlated features
1161	—	SY variety: PLS-DA model [0.30 impact, (+) impact, 1st features position (other varieties have reduced impact, lower ranking order, and (-) impact)]	3 correlated features
2924 - unique feature	—	Arinto variety: PLS-DA model [reduce impact, (-) impact, 15th position (Viosinho variety has the same impact order but a (+) impact)]	0 correlated features

In this context, after performing a comparative analysis based on the ANOVA, ML and Shapley values, it can be observed that all varieties, except Syrah, possess unique features with distinct impacts and are related to their chemical compounds.

- Aragonez presents two unique features with positive impact on the model output, the bands at 1127 cm^{-1} and 1130 cm^{-1} , related to alphatic compounds.
- Syrah does not present any unique feature. However, it has one feature correlated to other 3 ones, with the highest impact ranking for this variety classification and with a positive impact on the model output, the FT-IR band at 1161 cm^{-1} . This signal has been associated to strong vibration of acyl groups (C-O bond) in esters compounds.
- Touriga Nacional possesses two unique features, 1015 cm^{-1} , with a positive impact on the model output, and the signal at 1123 cm^{-1} with a negative impact on the model output. These features

are related to monosubstituted alkenes with C-O bend of alkoxy groups from esters or alcohols compounds.

- Viosinho and Arinto have also a unique feature to distinguish them, the signal at 2924 cm^{-1} , assigned to the O=C=O bonds pattern vibrations.

5.3 Discussion

The ANOVA analysis identified Syrah and Aragonez, as the variety pair with more features with the capability to discriminate samples. However, despite the statistical similarities of FT-IR spectra of the varieties investigated, ANOVA analysis identified some features able to discriminate all the red grape variety pairs, and three features able to distinguish the white wines varieties.

As expected, the similarity of the FT-IR spectra of the samples used in our study does not allow the discrimination of the grapevine musts by varieties using multivariate statistical analyses. Three unsupervised models, i.e., PCA, HCA, and t-SNE were tested, but did not achieve a good separation of the data related to the varieties. As mentioned above, the results are presented in the supplemental material.

Five supervised machine learning models were tested for the variety classification task using PR-AUC as evaluation metric. These different models have been evaluated to understand and then select the model with the best behavior or the most adequate one to classify each variety.

In the end of the analyses, the best approaches to classify each variety were the supervised model combined with feature selection and with adequate metric scores. With the tested pipeline, the main goal to classify each grapevine musts variety was achieved, and by using the Shapley values to help interpret the model outputs and feature impacts. There are no other works with model interpretation in the field of wine variety classification. Despite of being an exploratory analysis, the results found are interesting and can be extended to more wine varieties and a greater number of samples. In further works, our pipeline can be reproduce on other metabolomics data set from different techniques to compare the models performance and features with importance.

Other previous studies (mentioned bellow) have developed models for grape variety classification based on must, which achieved a performance with accuracy above 80%. However, these did not take attention to the class size and in some works the signal spectra treatment was not mentioned. Although distinct results can be found in the literature, they are not directly comparable to ours as the carried spectral technique, grape varieties, samples type, and the classification pipeline are different. For example Arana et al., 2005 classified two white grape varieties (Viura and Chardonnay) using NIR data, achieving 97.2% accuracy using PLS regression model. In our work, we also applied binary classification to the white wine grape varieties as Arana and collaborators, but we also include three more red grapevine musts varieties. Another difference in our work is that all grape varieties were produced in PDO regions of Portugal.

Cozzolino et al. 2012, have used NIR and MIR-ATR spectroscopies to perform a two-case classification of wine variety for Chardonnay and Riesling. They attain 86% accuracy using Linear Discriminant Analysis

to classify the grape varieties. As in our work, one of the grape variety was produced in two places, which lead to a drop in the model performance. More recently, Murru et al., 2019 performed a multiclass variety classification using Artificial Neural Networks with AUC measure, having a success rate of 83-95%, using also grape varieties data, based on FTIR-ATR.

The study closest to our work is a multi-classification model of white grape must varieties based on FT-IR (Roussel et al., 2003) where authors used pre-processing techniques based on Genetic Algorithms combined with a PLS-DA, a multivariate classification achieving 90.4% accuracy with unbalanced classes.

In this context, with our data, after performing several models combined with different treatment and selection methods, Syrah, Arinto and Viosinho were the varieties that have the highest discrimination. This fact may be explained due to the biological origin of the varieties and the size of the classes. The computation of Shapley values allowed to understand the impact of the important features identified, but also in a comparative analysis it also allow identifying unique features. All varieties, except for Syrah, possess unique features with distinct impacts on the classification models and are related to their chemical composition.

The overall synopsis of the process and the main results are shown in Figure 25.

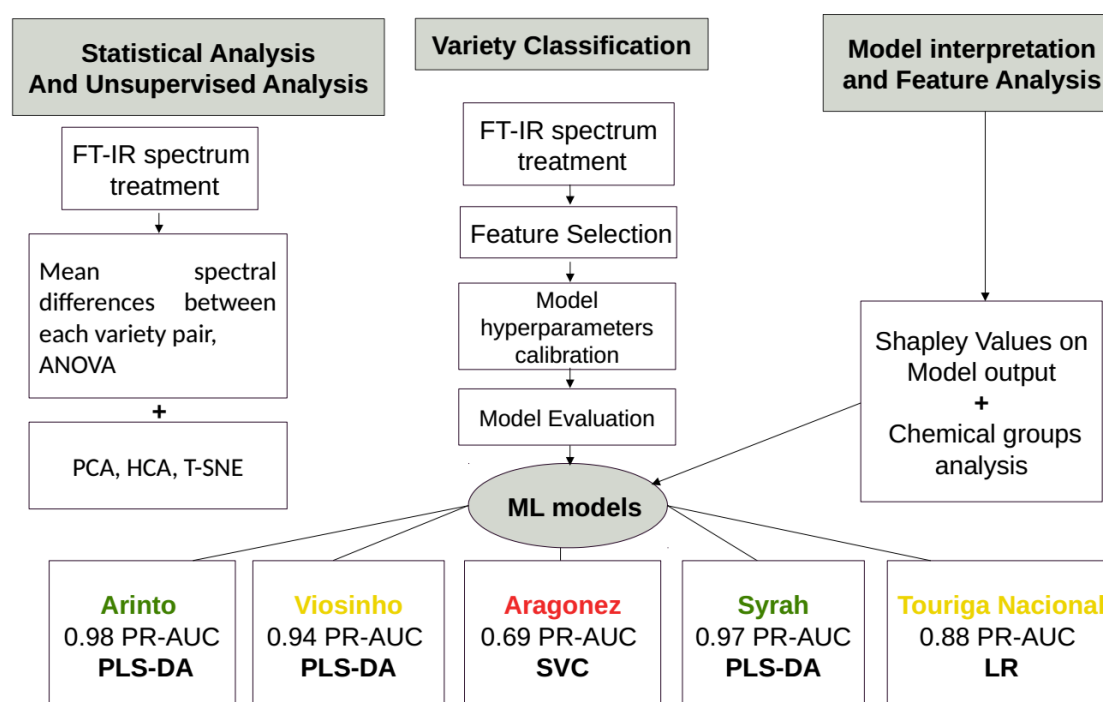


Figure 25: **Synopsis of the results with higher performance and the data analysis on FT-IR spectrum.** The squares present the classification model and the metric score of best model performance to each wine must variety.

5.4 Conclusions

Our study addressed two main objectives: the variety classification and the feature analysis with classification models interpretation, by reusing the available data avoiding other costs to explore Portuguese grapevine musts varieties.

Despite some statistical similarities detected in the FT-IR spectra of the grape varieties, the PLS-DA, LR and SVM models resulted to be good to classify the grapevine musts varieties investigated. The calculation of Shapley values allowed identifying the most influential features and their effect on the models output. We also relate these identified features, i.e., the FT-IR bands, with their correspondent functional chemical groups.

These outcomes highlight the importance and contribution of machine learning models, evaluation metrics, and optimization approaches on chemometrics to food analysis and authentication. To infer the influence of maturation, geographic origin, and other factors, further studies should be carried out with more data related to these issues, because our results regarding the grapevine musts variety classification could be affected/influenced by some other factors not herein considered.

The identified spectral regions in this work raise further questions, namely if other experimental techniques would detect the same chemical groups or correlated compounds. Follow-up studies should focus on verifying if the same FT-IR bands or spectral regions are identified as relevant features in the wine musts varieties classification based on different techniques to validate our case study.

To enrich the classification models of our pipeline more data from other musts of grapevine varieties should be added. Still, the proposed pipeline can serve as a basis for future works in the authentication of grapevine musts varieties, while reducing costs and time by reusing FT-IR data from routine analysis on grapevine musts varieties control.

5.5 Materials and Methods

5.5.1 Samples

Ninety six samples of grapevine musts, resulting from the crushing of berries during the control of maturation, from five varieties (16 Arinto, 22 Viosinho, 26 Syrah, 18 Touriga Nacional, and 14 Aragonez), were kindly supplied by the Enology Laboratory of INIAV-Polo de Inovação de Dois Portos, to be used for grapevine musts variety classification using FT-IR spectroscopy data and machine learning models. The samples were produced in two consecutive harvests of 2017 and 2018, in two Portuguese PDO regions around Lisbon district, (Alenquer: 39° 3' 59.50"N, 9° 6' 53.74"W, and Óbidos: 39° 11' 45.69"N, 9° 9' 54.74"W and, 39° 21' 7.57"N, 9° 1' 6.26"W). The control of maturation process, conducted by INIAV, used the same conditions over the grapevine musts according to the wine type (red or white), following reference methods.

5.5.2 Methods

5.5.2.1 FT-IR spectral measurement

There was no sample preparation to perform the FT-IR analysis. Before each experiment, the samples were kept at room temperature. A WineScan FT 120 instrument (Foss Electric, Denmark) was used to obtain the FT-IR spectra. The instrument was equipped with a model 5027 autosampler (64 tray, 40 ml-cups). A sample volume of 7 ml (standard setting) was pumped through the cuvette (optical path length 37 mm), which is located at the heater unit of the instrument. The temperature of the samples was set to 40°C. Analysis time took 30 s/sample. Cleaning was automatically programmed to occur every 5 min. The instrument was zeroed before any set of analyses with the zeroing solution (S-6060, Foss Electric). The instrument was standardized before the initial calibration with FT-IR equalizer solution (537811, Foss Electric) and repeated at least once a month.

Samples were scanned from 926 to 5012 cm^{-1} at 4 cm^{-1} intervals. The number of scans generated per sample, the selection of wavenumbers, and the processing of spectra have been fixed by the manufacturer and are not accessible to change by the user. For every wine must sample, two replicate spectra were acquired. A detailed discussion regarding the analytical procedure to carry out the FT-IR measurements can be found in the work reported by Patz and co-workers [220], [213].

5.5.3 Statistical Analysis and Chemometrics

All processing and data analyses were performed by using the Python language, and the libraries pandas, mumpy, ResearchPy, Scipy, statsmodels, and scikit-learn.

5.5.3.1 Spectral Data Processing and Data pre-processing

Two datasets were organized from the samples based on the wine type of must (red and white). All samples that had replicates were transformed by computing the respective spectra mean values. The water region [1451-1799 cm^{-1}] was removed from all spectra due to noise as advised by the Winescan manufacturer. No missing values and no outliers were detected. Prior to data analyses, it was necessary to used appropriate pre-processing data and spectral data signal treatments to minimize physical effects as presented bellow [33]. SNV plus Standardization was the selected treatment utilized to perform the data analyses. So, multiplicative interferences in baseline shift and curvilinearity were first corrected by SNV [218]. This was followed by z-Score Standardization by removing the mean band values and dividing by the corresponding band standard deviation for each band individually.

5.5.3.2 Data analysis

Exploratory and statistical analysis

A preliminary analysis included statistical univariate techniques and unsupervised multivariate statistical models to understand the samples relationships among the different varieties according to the wine type (red or white wine type).

To understand if there are statistical differences between different combinations of wine must varieties a parametric test was conducted. ANOVA and t-test, between each varieties pair were carried to identify which specific spectroscopic bands (cm^{-1}) could be useful to discriminate between the investigated classes. The paired differences with p-value < 0.01 were considered to had statistically significant resonances differences.

The unsupervised multivariate statistical models applied to the FT-IR dataset in this study were Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). PCA is a widely used tool for interpreting and visualizing the main features of FT-IR data for classifying samples. In this study, PCA was used to investigate if there was a clear discrimination between the samples with different grapevine musts varieties in the latent space. HCA used an Euclidean distance between samples and the complete linkage method. This alternative was chosen because it tends to find compact clusters of approximately equal diameters and does not force clusters together due to single elements being close to each other, like in single linkage clustering. In its turn, t-SNE is a method which applies a non-linear dimensionality reduction technique to keep the similar data points close in a lower-dimensional space. This method uses a student t-distribution to compute the similarity between two points in lower-dimensional space.

Variety classification and features analysis

To classify each variety a one-vs-all approach was followed, employing binary classification to separate a target class from any other set of grapevine musts. The main objective was not to identify the grapevine musts present in the data set, but to classify a target variety from any other source. Since this is an exploratory work with potential to add more varieties in further studies, the binary classification is the best approach to apply into this scenario.

The following models were used: Random Forest Classifier (RF), Gradient Boosting Classifier (GBC), Logistic Regression (LR), Partial Least Squares-Discriminant Analysis (PLS-DA), and C-Support Vector Classification (SVC). Several performance metrics were tested and compared (Area Under the Precision-Recall Curve (PR-AUC), Area Under the Receiver Operating Characteristic Curve (ROC-AUC), Matthews Correlation Coefficient (MCC), Balanced Accuracy (BA), Precision, Recall, Accuracy), being the PR-AUC the selected metric to evaluate the models performance.

Due to the reduced number of samples, repeated nested k-fold cross-validation was utilized to infer the model performance. The outer loop was executed with $k = 10$, while the inner cross-validation loop was executed with $k = 5$. The inner cross-validation loop serves to fine-tune the model hyper-parameters, while the outer loop to estimate the model performance. It is important to be aware that all the models

are trained on the inner loop step.

The whole process was repeated 10 times by reshuffling the data set. The mean and standard deviation of performance of the best models were computed (concerning the outer loop). Table 16 shows the hyper-parameters tested for each model.

Table 16: **Hyper-parameters used by the distinct models.**

Model	Hyper-parameters
Random Forest Classifier	N of Estimators: 10, 50, 100; Max N of Features: auto, 2, 5; Max. of depth: none, 2, 5, 10; Class weighth: balanced.
SVC C-Support Vector Classification – Support Vector Machine Classification	C: 0.1, 1, 10, 100; Gamma: 1, 0.1, 0.01, 0.001, 0.0001; Gamma: scale, auto; Kernel: rbf, linear; Class weighth: balanced
Gradient Boosting Classifier	N of Estimators: 10, 50, 100; Learning rate: 0.1, 0.01, 0.001, 0.0001; Max N of Features: auto, 2, 5; Max. of depth: none, 2,5,10; Class weighth: balanced
Logistic Regression	C: 0.00001, 0.001, 0.1, 1, 10, 100; Class weighth: balanced
Partial least squares-discriminant analysis	N of Components: 1-11.

For each training set and test set, the following preprocessing was carried on: each model was calibrated using the treated raw spectra with the SVN method, using the aforementioned preprocessing pipeline. The calibration was performed with explicit feature selection. The explicit feature selection method consisted in dropping correlated features above a given threshold (in this work 90%). First, the correlation matrix of the spectral resonances was computed. Next, the correlation matrix was further processed by forming clusters of bands with correlation values above the threshold. For each cluster, the band with the highest variance is kept and the remaining ones were dropped.

Feature analysis

To assess the feature importance, the best models from each class were chosen based on the superior performance in the aforementioned cross-validation scheme for each target variety. Afterwards, each band influence on the target classes was inferred using the Shapley values. These values are the average marginal contribution of a feature value across all possible coalitions and can be applied for classification problems, as in our study [56] [58]. For binary classification, SHAP values for classes 0 and 1 are symmetrical. The selected feature contributes to a certain amount towards class 1, at the same time reduces the probability of being class 0 by the same amount.

Machine learning approaches to classify honeys based on flora and geographic origin from Santa Catarina, Brazil using ^1H -NMR

Abstract

The identification of the flora and geographic origin of honey has high interest and economic value, particularly for the honey with Geographical Indication (GI), which possess unique organoleptic properties, resulting from unique fauna and flora available in native forests and protected areas. Globally, honeys produced in Brazil had an exponential growth in production and consumption since 2020.

Nuclear Magnetic Resonance (NMR) allows fast and robust spectral recording without sample pre-treatment, but requires spectral signal processing. Due to the complexity of the generated signals, specific data pre-processing methods need to be applied, which are typically followed by a conjugation of statistical tools and machine learning approaches to maximize the accuracy of classification models of honey as to their flora and geographic origin.

This exploratory study aims to assess the capability of high-speed analytical devices present in the authentication and quality-control honey industry, such as ^1H -NMR, used to classify Santa Catarina (SC) kinds of honey, according to their region of production (from the eleven agroecological regions) and flora group origin. Another aim is to describe which specific features allow for the discrimination of a given region and flora group and how they influence the model output.

Our data analysis contemplated 56 honey samples from the eleven agroecological regions of Santa Catarina state, southern Brazil. These samples have been produced in geographic regions where seven flora types are predominant for bee foraging, as follows: “Uva do Japão” (*Hovenia dulcis*), “Silvestre”, Eucalyptus (*Eucalyptus spp*), “Uva mix”, and Other (representing several multifloras).

The devised approach encompasses the following stages: (i) exploratory and statistical analysis of the ^1H -NMR dataset; (ii) origin classification (region and flora) and feature analysis, and (iii) comparative cross-analysis between the results of geographic and flora classifications.

Despite the statistical similitude of the ^1H -NMR data from different types of honeys, machine learning

models reached good discrimination performances.

Logistic Regression (LR), Support Vector Classifiers (SVC), Random Forest (RF), and Partial Squares-Discriminant Analysis (PLS-DA) were the supervised models with the highest performance ranging from 0.60 to 0.91 for the Area Under the Curve of Receiver Operating Characteristic (ROC-AUC) for the different geographic regions origin. PLS-DA and SVC were the supervised models with the highest performances ranging from 0.56 to 0.72 of ROC-AUC for the different flora origins.

Finally, features (i.e., ^1H -NMR resonances) influencing the model output were identified using Shapley values. These features were linked to the correspondent molecules found in the honey investigated, and a comparative analysis based on the feature importance results for region of production and flora origin was conducted. It was observed that there are shared important features for the sample classification as to their geographic regions and flora groups.

These results thus highlight some common features (resonances) with influence either on flora and the correspondent producer region of honey at the same time, which are especially significant in honey authentication, providing useful information for honeys flora and geographic classification of honeys produced in Santa Catarina, Brazil.

6.1 Introduction

Honey is a natural food produced by bees from the nectar of flowers (floral honey), but it can also be obtained from living plant exsudates or from excretions of plantsucking insects producing honeydew, such as the “Bracatinga” one, a typical honeydew produced in southern Brazil [222], [223].

Honey is one of the most internationally consumed agricultural food, with a continuous increase in market value, and since 2009, its production and consumption have increased, generating strong interest in the honey trade globally [224], [225].

According to the statistical report by M. Shahbandeh, the average annual production volume of natural honey in the world reached a peaked in 2017, with 1.88 million metric tons and has reduced since then with the registered production of around 1.77 million metric tons in 2020. Despite the challenges of the Covid-19 pandemic in 2020, the interest and the consumption of natural honey has grown and its market value has also increased, reaching around eight billion dollars in that year [226].

Certified honey is a food product resulting from biological flora and fauna areas without any chemical treatments. Brazil has the largest extension of natural forest in the world, with rich native flora, good resources for the development of the bees, and tropical weather in most of its areas, without pesticides and/or other chemical residues [227].

The high biodiversity of flora and fauna of several protected geographical areas offers the conditions to produce honey with high quality and in quantity. With the regulation of standard requirements established by CODEX Alimentaris and the Brazilian government, the Brazilian honeys increased their commercialization around the world and in UE markets. The certifications issued to that food allow controlling its quality,

safety, and authenticity. Nowadays, Brazil is considered one of the biggest producers of honey in the world, with a potential to increase its production [226].

Santa Catarina (SC - southern Brazil) is not among the largest Brazilian states in geographic area, but it has been ranked as an high honey producer in the country, overcoming the median national production of over 7.5 million tons against the national average of 6.5 million of tonn [228]. Besides, despite the honey production numbers of that state, the most important is the high quality of the honey produced, which has been internationally awarded many times.

SC state has, over its eleven agroecological regions, a high flora and *terroir* producing more than 100 types of honey with distinct colors, aromas, flavors, and consistency. Figure 1 presents the agroecological regions of Santa Catarina and the produced honey types, which have been awarded in several International Federation of Beekeepers' Associations Conferences (APIMONDIA), achieving gold medals at the years of 1979, 2007, 2013, 2015, 2017, and more recently in 2019 during the 46th Apimondia Congress, in Canada.

From the vast honey types, there are four produced in larger quantities in SC:

- “Silvestre” honey is the most produced in Santa Catarina. This honey is multifloral, resulting from a wide variety of flora, and varies according to the region and the predominant flowering at the time of production. The darker color indicates a higher concentration of mineral salts, while lighter honeys are usually milder.
- “Eucalyptus” (*Eucalyptus spp*) honey is produced from March to May, mainly in the south of the state where the population of this plant species is higher. It presents an amber color and a stronger flavor.
- “Uva-japão” (*Hovenia dulcis*) honey is produced mainly in the west region of the state in November and December, during the flowering season of this tree. It shows a light color and an extremely mild flavor, and hardly crystallizes when stored at room temperature.
- “Bracatinga” (*Mimosa scabrella*), honeydew: This honeydew honey is produced by exudates mixed with enzymes produced by mealybugs (*Tachardiella sp.*), small sucking insects that attack the trunk of the Bracatinga tree, releasing exudates that are foraged by *Apis mellifera* bees. Its production occurs only in even years, i.e., every two years due to the plant-sucking insect's cycle, between March to May. The “Bracatinga” is a native tree of the high and colder regions in southern Brazil, as in the North Plateau of Santa Catarina. “Bracatinga” honeydew has Geographical Indication (GI) registration and it has been produced in restricted geographical regions in SC, with specific climate and soil characteristics where the donor plant *M. scabrella* is found.
- “Canudo-de Pito” (*Senna bicapsularis*), honey has unique features and was firstly awarded in APIMONDIA congress in 1979. It is a monofloral honey with reduced time of production due to the short flowering cycle of the nectar donor plant, i.e., from December 10th to January 10th. Besides, in the SC state, the populations of “Candudo-de-Pito” trees have a restricted occurrence, mostly in

regions with altitudes over 1000 m, as observed in the agroecological region of Serrano Plateau of São Joaquim city (28° 17' 38"S, 49° 55' 54"W, 1354 m altitude). Due to its high quality and geographic location limitation that make this honey unique, more recently a series of efforts has been made pursuing the GI registration of that honey in Brazil.

Honey production is a complex process resulting from multiple interactions of flora, fauna, and environmental factors. The fraud in honey represents the third most prevalent in the food industry. Nowadays, consumers and producers are more conscious regarding sustainable production systems and the food product's authenticity. So, it is imperative to ensure the honey authenticity for consumers and beekeepers and also to value their products. Therefore, as a requirement, it is crucial to follow standards of production with information regarding flora, geographical, and climatic conditions included, as well as other extraction procedures and storage conditions.

There are several appropriate analytical techniques and methodologies to test quality, detect adulteration events, and also to certificate flora and geographic origins of honey [125], [229], [227], [225]. In addition, the most advanced, comprehensive, and globally used analytical technique is nuclear magnetic resonance (NMR) spectroscopy, coupled with chemometrics tools, by international organizations, companies, and in academic research [125], [229], [225]. NMR provides a simple, fast, and non-destructive way to analyze honey samples without the need for chemicals or preparatory steps [125], [229], [227], [225].

Nevertheless, there is a reduced number of publications and applications to honey classification based on the flora and geographic origin [130], [227], [223], [230], [231]. Even less research work is directed to the classification of Brazilian honey, one of the most important producers of that food worldwide [230].

In this study, machine learning algorithms were applied to a ¹H-NMR dataset aiming at to discriminate honeys according to their flora and geographic origins. The application of data analysis tools on honey metabolomics studies provides instruments allowing comprehensive discussions capable of bringing new highlights to the field of honey characterization, quality, and authentication.

The samples were produced by family bees of *Apis mellifera* spread over the eleven agroecological regions of Santa Catarina state (southern Brazil), which include the most produced honey types Eucalyptus, "Silvestre", "Uva do Japão" and multi-floral. All samples came from controlled and registered apiaries with geographic locations. Figure 26 shows a map of the agroecological regions of Santa Catarina and their most common honey produced.

The approach followed in this work is divided into stages: exploratory and statistical analysis of the ¹H-NMR dataset, variety classification, feature analysis, and a comparative analysis between the flora and geographic classification models built

The first objective deals with the discrimination of regions and group plants based on the visual and statistical differences of the data, to understand the major differences between the different varieties; the second one aimed to achieve the most accurate and efficient classification models for each production honey region and flora origin using five different machine learning approaches (SVM, PLS-DA, RF, GBC,

and Logistic Regression (LR)), combined with the evaluation metrics Area Under the Curve of Receiver Operating Characteristic (ROC-AUC), Mathews Correlation Coefficient (MCC), feature selection methods and preprocessing data treatments. One final aim is to provide access to feature importance and enable structural interpretation of model-based classification. $^1\text{H-NMR}$ coupled with chemometrics data analysis tools may provide an accurate and reliable method to discriminate honeys flora and geographic origins based on some chemical signatures, i.e., resonances, as shown by our results.

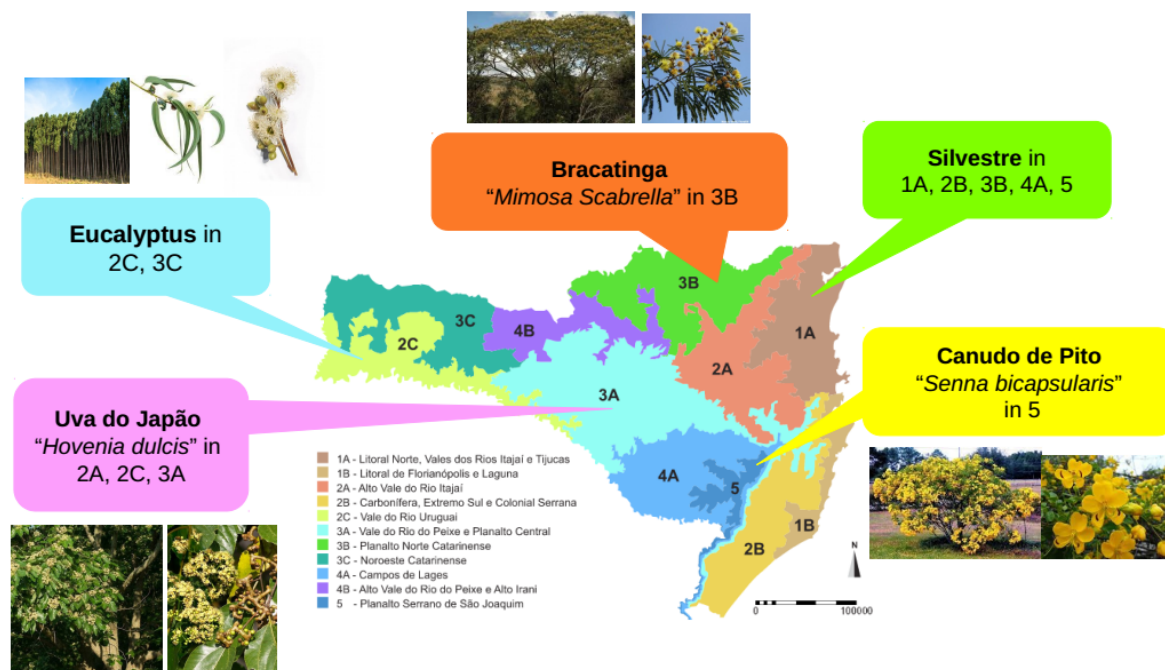


Figure 26: **The eleven agroecological regions of Santa Catarina Estate and the most common flora honeys.** Adapted from <https://ciram.epagri.sc.gov.br/apicultura/monitor-colmeias.jsp>.

6.2 Results

This study contemplates the analysis of a metabolomics data set of 56 Brazilian honey samples from the eleven agroecological zones of the Santa Catarina state (SC), harvested in the 2018-2019 production season. The metabolomics dataset was organized based on the geographic production region and the flora origin of the samples.

The samples were produced in the 11 agroecological regions (1A, 1B, 2A, 2C, 3A, 3B, 3C, 4A, 4B, and 5) of SC state (Figure 26) and the information due to the flora group was included. In the dataset, samples were organized as to their flora information in main groups, i.e., flora composition (Eucalyptus, "Silvestre", "Uva do Japão", Uva Mix, and Other) to include the different flora origins available in all samples. Despite having honey samples of "Bracatinga" and "Canudo de Pito" their reduced number of samples did not allow to categorize them as separate flora groups, being these samples added to the category "Other".

The geographic and flora information of the honey samples is presented in Table 17 and Table 18.

Every honey sample came from a controlled production system and every apiary is registered with its geographic location and according to its flora available to the bees in the surroundings of the hive. Santa Catarina estate has a large variety of native flora which are present in a large number of regions. Knowing the existing flora in the area of each apiary allows identifying the majority floral group that can be used by the bees to produce honey. However, due to the large diversity of flora, the presence of other minority nectar and pollen plant donors cannot be excluded from other flora origins. Based on this information, our samples are not considered pure mono-floral honey.

As explained in the introduction, the main goal of the study is to classify the honey samples according to each of the eleven individual regions due to the IG importance and the economic value associated with honey products. Nonetheless, it is known that there is a vast flora diversity within the geographic regions, which led to an analysis focusing on the botanic origin of the honeys samples. We also compared both analyses to verify the existing connections between the geographic and flora origins of the honeys from the SC.

The data were preprocessed as mentioned in detail in the methods section. Then, an exploratory and statistical analysis was conducted, followed by a non-targeted metabolomics analysis with ML models to classify each region and flora group, preceded by an analysis of the features (i.e., $^1\text{H-NMR}$ resonances) with importance on the model's classification using Shapley values, to understand the impact and how such features influence on the model performance. In the end, we compared the Shapley values of geographic and flora models classification to observe the existing link between both influences on our honey samples produced in SC.

The exploratory and statistical analyses on geographic and flora classifications are presented together to simplify their comprehension. The supervised analyses are presented separately due to a large amount of information.

Table 17: **Geographical information of the honey samples.**

	1A	1B	2A	2B	2C	3A	3B	3C	4A	4B	5	Total(N)
Samples (N)	4	3	7	3	6	7	6	4	5	4	7	56

Table 18: **Flora information of the honey samples.** Uva mix samples are multiflora honeys with presence of "Uva do Japão" flora. Other honey samples are multiflora honeys with different types of flora as cinnamon, fruit trees, native trees, "Canudo de Pito" tree, "Bracatinga" tree, and other plants from each region.

Flora	Samples (N)	From regions
Uva do Japão	11	2A, 2C, 3A, 3B, 3C, and 4A
Eucalyptus	8	1A, 1B, 2A, 2C, and 3C
Silvestre	7	1A, 2B, 3A, 3B, 4A, and 5
Uva Mix	4	2A, 2B, and 3A

Other	26	1A, 1B, 2A, 2B, 3A, 3B, 4A, 4B, and 5
Total(N)	56	

6.2.1 Exploratory analysis

After preprocessing the ^1H -NMR spectra data, visual analysis of the spectra was performed and certain visual analysis of the spectra was performed and certain discrepancies were noted spread over the aliphatic (0.00 to 3.00 ppm), anomeric (3.00 to 5.50 ppm), and aromatic (5.50 to 8.00 ppm) regions. Still, in this way, was not possible to clearly discriminate the samples according to their geographic region or flora groups as shown in Figure 27 (A and B represent the honey samples by regions and flora origins, respectively).

In the region of the sugar resonances (Figure 27 (A and B)), there are dominant signals very similar between all honey samples, namely the resonances related to glucose (around 3.22ppm - βC2H and 5.23ppm - βC1H), fructose (around 4.11ppm - C3H), and sucrose (around 4.22 ppm and 5.42ppm - d, C1H). As expected, this was the region with high-intensity peaks since those carbohydrates are the major constituents of honeys.

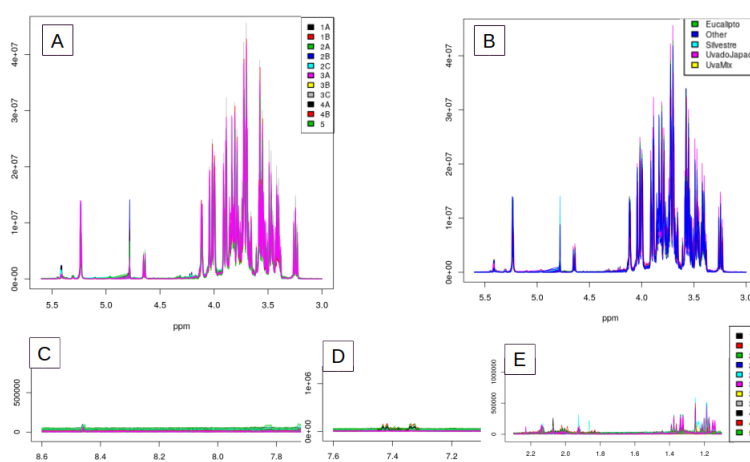


Figure 27: ^1H -NMR spectra spectra plot for honey samples based on region production(A) and flora origin (B) at interval of 3.0ppm-5.5ppm. C: Zoom plot of ^1H -NMR spectra of all honey samples at interval of 7.7ppm-8.6ppm. D: Zoom plot of ^1H -NMR spectra of all honey samples at interval of 7.6ppm-7.1ppm. E: Zoom plot of ^1H -NMR spectra of all honey samples at interval of 1.1ppm-2.3ppm.

The resonances with less intensity were related to the other compounds groups detected both in the aromatic and aliphatic regions, such as phenylalanine (7.3 ppm and 7.4 ppm) in Figure 27(D), formic acid, (7.8 ppm and 8.4 ppm) in Figure 27(C), ethanol (1.2 ppm), acetic acid (2.0 ppm), and proline (1.90 ppm and 2.10 ppm) in Figure 27(E), the main amino acid in honey. To note that this is not a systematic annotation analysis, rather a comparative one based on the available information of current literature [232], [130], [229], [83].

Taking into account the ^1H -NMR honey profiles were very similar among all samples, further statistical analyses were performed to identify and extract relevant information from the metabolomics dataset.

6.2.2 Statistical analysis

Two ANOVA analyses (followed by post-hoc Tukey tests) were performed to discriminate the honey samples from SC by Region and by Flora group.

The results based on the univariate statistical analyses reveal that there is only one feature with the capability to linearly discriminate regions, i.e. the resonance at 5.41 ppm. Such a ^1H -NMR signal has been attributed to sucrose and allowed to discriminate 2A region samples from region 3A, and 2A region samples from region 5 samples. Considering the ANOVA results of flora groups, no features with the capability to linearly discriminate the samples according to their flora origin were detected.

Interestingly, the aforementioned results indicate a reduced number of resonances with the capability of sample discrimination considering the geographic or the flora origin of honey, reinforcing the chemical similarities of that food matrix between all samples, despite they have been originated from sites with quite distinct flora and climate and soil conditions.

6.2.3 Multivariate data analysis

To extract and identify more relevant patterns, three unsupervised multivariate statistical methods were applied to the complex metabolomics matrix of honey samples.

Firstly, principal component analysis (PCA) was performed, followed by the clustering method hierarchical clustering analysis (HCA), and the t-distributed Stochastic Neighbor Embedding (t-SNE) method, in order to detect relevant features to classify samples from the metabolomics dataset, and also pinpointing relevant information that may help to characterize the studied variables (geographic regions and flora). The chosen parameters are presented in the Methods section.

The results from the unsupervised models did not show a good sample separation regarding the investigated variables. The overall results showed sample overlaps between all production regions and all flora. These results were expected based on the previous visual analysis of the ^1H -NMR spectra and univariate statistical analyses. For this reason, these results are not further discussed. Still, the results concerning the non-supervised approaches are all shown in the supplementary material (Figures S.43, 44, 45, 46, 46, 48) for the interested reader.

6.2.4 Geographic Region and Flora classification

Univariate analysis showed a reduced number of features capable of discriminating honey samples and the multivariate analyses revealed a high overlap between the honey samples. These results were observed both for the region and flora group discrimination.

To overcome the difficulty of classifying honey samples from such a complex dataset matrix, five supervised classification models, GBC, LR, PLS-DA, RF, and SVC were calibrated with hyper-parameters optimization and evaluated, leading to 55 models for region of production and 25 models for floral groups. The performance of all produced models is presented in the supplementary material (Table 28 and Table 29) for the interested reader.

As mentioned in the introduction section, the main objective of this exploratory study is to build classification models with the highest performance and accuracy to classify each sample regarding its production region and flora origin, and then use interpretability tools to understand how the features with higher importance explain those models.

However, according to the literature, to achieve the most accurate classification and also to analyze the important features without misleading interpretation, several treatments should be carefully conducted:

- Solvents region of the $^1\text{H-NMR}$ spectra signal of honey samples should be eliminated to avoid noises that influence the models' performance, in some cases can increase the models' classification with misleading results. In our data, the D_2O region was eliminated according to the literature [130].
- The treatment of the spectra signal of honeys needs to be included to avoid noises and other misclassified points [130]. In our study, after testing several method combinations, baseline correction, missing values treatment, followed by standardization achieved the best results on model performance.
- To improve the model classification performance, feature selection methods should be used to reduce the problem of dimensionality and to facilitate and improve the computation of feature importance [221]. In our study, two filter feature selection methods were used. Two methods allowed the detection and treatment of constant, quasi-constant, duplicated, and correlated features. Correlated features were detected and treated by eliminating features with a correlation equal to or higher than 90%. The selected method was used to find groups of correlated features and to select from each group the feature with the highest variance. As result, the feature selection method returns a sub-dataset of features that are uncorrelated with each other based on the utilized threshold. The parameters of these methods are shown in the methods section. The feature selection allows reducing the 36 features (resonances, ppm) to 29 features. Table 19, shows each group of correlated features and the selected feature with the highest variance for the target variable.
- An adequate validation strategy to evaluate the model performance is necessary to ensure that the model can generalize well for every possible target. In our study, we selected the binary classification model approach to predict the target variable and, due to the reduced number of samples, with leave-one-out cross-validation to infer the model performance (details are shown in the methods section) in the training set and a test set to validate the generalization capabilities with unforeseen data. To compare each model with another, different scoring metrics were used. The main aim of the study is to correctly classify target honey by Region and Flora origin from any other source.

Table 19: **Feature with highest variance and the correspondent correlated features group.**

First column presents the feature with the higher variance correlated to a set of peaks given a predefined threshold. These features are the selected to the new sub-dataset used at all models classification. And these features were used by Shapley method to interpret their impact on each model classification. Second column are the correspondent correlated features to each feature. Note: * are features with high variance and with lower threshold correlation (0.90) with other features.

Feature with highs variance (Peak (ppm))	Correlated features group (Peak (ppm))
1.34*	1.34
2.07*	2.07
3.23*	3.23
3.26*	3.26
3.39*	3.39
3.41	3.41, 3.44
3.47*	3.47
3.50*	3.50
3.53*	3.53
3.56*	3.56
3.58*	3.58
3.69	3.66, 3.69
3.72	3.81, 4.02, 4.12
3.75*	3.75
3.78*	3.78
3.84*	3.84
3.90*	3.90
3.99*	3.99
4.04*	4.04
4.17*	4.17
4.21*	4.21
4.31*	4.31
4.64	4.65, 4.64
4.96	5.38, 4.96
5.12*	5.12
5.24*	5.24
5.31*	5.31
5.41*	5.41

In our case, the Area Under the Curve of Receiver Operating Characteristic (ROC-AUC) is suitable to evaluate binary classification outcomes for imbalanced classes as in our study. This metric is less sensitive than Area Under the Curve of Precision and Recall (PR-AUC). Other metrics suitable for our problem are Matthews correlation coefficient and Balanced Accuracy. These metrics also provide accurate evaluations for imbalanced classes in binary classification models, however, they require a defined decision threshold. To evaluate our models' results we tested several metrics (as shown in the supplemental material (Table 28 and Table 29)).

However, because this is an exploratory study with contributions to honeys classification, we focus the scoring metric on the ROC-AUC and also at the MCC, with a threshold calculated based on the ROC threshold mean due to the main importance given to the global importance of the classifications. Nevertheless, the selected threshold can be changed according to the sensitivity and cost established or required for the

classification model, in accordance with the defined honey production regulations.

Based on those assumptions, the models with higher score metric of ROC-AUC in the test set were chosen to classify each given Region and Flora group.

6.2.4.1 Geographic Regions classification

Table 20 presents the results of classification models for all regions origins with the highest performance scores.

Of the eleven agroecological regions of SC state, the Region 1B “Litoral de Florianópolis e Laguna” was the honey production zone with the best model classification performance with the LR reaching 0.91 of ROC-AUC.

In Region 3C “Noroeste Catarinense” the highest performance score was achieved with the LR model 0.84 score of ROC-AUC, and Region 5 “Planalto Serrano de São Joaquim” with 0.69 score of ROC-AUC with the SVC model. These two regions were the ones with higher classification model scores followed by the other five regions performances.

Region 4A “Campos de Lages” achieved the higher score of classification with a performance of 0.73 with SVC model. Region 3B “Planalto Norte Catarinense” with the RF model achieved 0.70 of ROC-AUC performance classification. Region 4B “Alto Vale do Rio do Peixe e Alto Irani” with PLS-DA achieved a performance classification score of 0.69 ROC-AUC. Region 1A “Litoral Norte, Vale dos Rios Itajai and Tijucas” reaching 0.65 score of ROC-AUC with LR model classification. And Region 2C “Vale do Rio Uruguai” with the PLS-DA model achieved a classification performance 0.60 of ROC-AUC.

Table 20: **Performance scores for best models classification for all Regions.** All models had pre-process methods with the application of Feature Selection (Drop of constant features with 90% of correlation). The MCC evaluation metric followed a threshold based on the calculated ROC Threshold mean. Rows with background color gray are the regions with higher classification models. Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; ROC Threshold: Receiver Operating Characteristic.

Target Region	ML Model	ROC-AUC	PR-AUC	MCC	BA	ROC Threshold
1A	LR	0.65	0.12	0.15	0.52	0.08
1B	LR	0.91	0.23	0.30	0.71	0.10
2A	RF	0.48	0.12	-0.07	0.45	0.09
2B	GBC	0.43	0.06	-0.09	0.43	0.00
2C	PLS-DA	0.60	0.26	0.30	0.61	0.33
3A	PLS-DA	0.47	0.10	-0.02	0.49	0.15
3B	RF	0.70	0.17	0.18	0.65	0.06
3C	LR	0.84	0.47	0.20	0.67	0.03
4A	SVC	0.73	0.13	0.16	0.64	0.10
4B	PLS-DA	0.69	0.12	0.09	0.58	0.11
5	SVC	0.82	0.23	0.41	0.78	0.12

6.2.4.2 Flora groups classification

Although it is known that SC state has several native plants in the majority of its regions, spread in the Atlantic Rainforest and Araucaria biomes, for instance, creating a unique type of flora, there are some flora species in each region that allow producing honey with some peculiar traits. Here, instead of classifying each honey sample by the geographic region of production, we looked at the type of biological origin. Our data set resulted in honey samples with a vast combination of flora (different types of trees, bushes, and small flowers), but there are some samples where the major flora origin is known and registered by the producers. Because our samples came from controlled honey production system where the geographic localization and the major pollen donor plants are known for each apiary, we reduced the flora information to five groups of flora. Each group is designated with the main composition of existing flora origins. Table 21 presents the results of classification models with validation for all flora groups with the highest performance scores.

The “Uva do Japão” was the flora group with the best model classification performance, with the PLS-DA reaching 0.72 of ROC-AUC. There were other flora group with similar performance score for the models’ classification the group Other (monofloral honey samples) with the SVC model with achieving a performance of classification of 0.67 of ROC-AUC. “Silvestre” as native flora group with the SVC model reached the classification performance of 0.56 ROC-AUC.

These three flora groups were the ones with higher classification model scores followed by the other two flora groups with reduced performances under 0.50 of ROC-AUC (in the Test set), not achieving a positive classification for Uva Mix and Eucalyptus flora groups.

Table 21: **Performance scores for best models classification for all Flora honeys.** All models had pre-process methods with the application of Feature Selection (Drop of constant features with 90% of correlation). The MCC evaluation metric followed a threshold based on the calculated ROC Threshold mean. Rows with background color gray are the regions with higher classification models. Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; ROC Threshold: Receiver Operating Characteristic.

Target Flora	ML Model	ROC-AUC	PR-AUC	MCC	BA	ROC Threshold
Uva do Japão	PLS-DA	0.72	0.41	0.22	0.64	0.21
Eucalyptus	RF	0.44	0.09	-0.05	0.46	0.04
Silvestre	SVC	0.56	0.14	0.09	0.57	0.12
Uva Mix	SVC	0.47	0.07	0.04	0.54	0.07
Other	SVC	0.67	0.70	0.37	0.67	0.57

In our analyses there are some considerations to note that can influence the classification of the models’ performance:

- Seasonality, altitude, and climatic changes are some of the influences on food product production like honey.

- Pollution, fertilizers, and human activities like fires also configure critical pressures on honey bees production [233], [234], [235], [236].
- The vast diversity of native plants, animals, and or insects interactions are other influences factors to difficult the honey classification. An example of a beneficial relationship is the “Bracatinga” honeydew. This honeydew is the result of plant-sucking insects which infest the tree species named “Bracatinga” (*Mimosa scabrella*). This honey has an IG classification with production in Santa Catarina state since plant specie is more prevalent in highlands in SC.

To better understand which of these influence factors have more impact on the model’s Santa Catarina region classification and how these influences can modulate the region classification further studies should be carried out.

However, despite the complexity of the data and confounding influences, the results reveal that LR, SVC, RF, and PLS-DA, were the models with better performance to classify each Geographic Regions classification, with ROC-AUC scores reaching 60-91%. PLS-DA, and SVC were the models with better performance to classify each flora group of honey, with good ROC-AUC scores reaching 56-72%.

6.2.5 Features analysis of models classification

ML models are generally considered ‘black boxes, but the ability to easily interpret the results of a variety of classification models would improve our understanding of the factors underlying the differences between each production honey region, rather than merely classify them. This information can be used in the geographic honey authentication process to avoid events of fraud. Using the $^1\text{H-NMR}$ dataset of the honey samples produced in southern Brazil, it is worth mentioning there are no studies addressed to the classification models interpretation, despite a large number of works to classify honeys according to the geographic and biological origin [125],[229], [237], [83]. Our work followed this direction, to find out which resonances are the most predictive to classify each production region, and how these features are related to the model performance. Thus, by determining which important features (resonances, ppm) contributed most to a certain region would help to uncover potential biomarkers resonances or to identify groups of chemical compounds to improve honey classification.

To assess the feature importance from the $^1\text{H-NMR}$ dataset, we used the Shapley Additive exPlanations (SHAP) algorithm, a value estimation method from the field of game theory. This method can be applied for classification problems, as in our study, by computing the average marginal contribution of a feature value across all possible coalitions. [57], [56], [58].

By computing the Shapley values of the best classification models for production region, it was analyzed the selected features with the higher impact on the model performance and how they contribute to each variety classification. An important note to consider is that to compute Shapley values, KernelShap needs to use a reduced number of features due to the time required to work, and also correlated features

should be previously eliminated to avoid weighting the features wrongly, leading to an incorrect interpretation of the model performance. As mentioned above in geographic region classification topic and below in the methods section, correlated features were identified and treated.

6.2.5.1 Features analysis of the best models performance for Geographic Regions classification

The computed mean Shap values resulted in twenty features an average range of impact from 0-0.04. The results for the mean Shap values and their average impact on model output are shown in Figures 28, 29, 30, 31, 32, 33. The first seven computed mean Shap values with high importance and their impact on the models are shown in Table 22, for the calculated regions. By looking at these results the impact and order of features are different in each region model. 1A and 4A shared the feature 3.9ppm both with the first order of importance. However, in 1A their impact is higher and is positive compared to region 4A where the shap value for the feature is reduced and with a negative impact on the model. Regions 1B and 5 also shared a common feature with the highest importance in the model performance, in both models' region classifications feature 5.41ppm has a positive impact but the impact is higher in 1B than in the model to classify region 5. The remaining region classification models have different features at the top of important features with different impacts on their classification models.

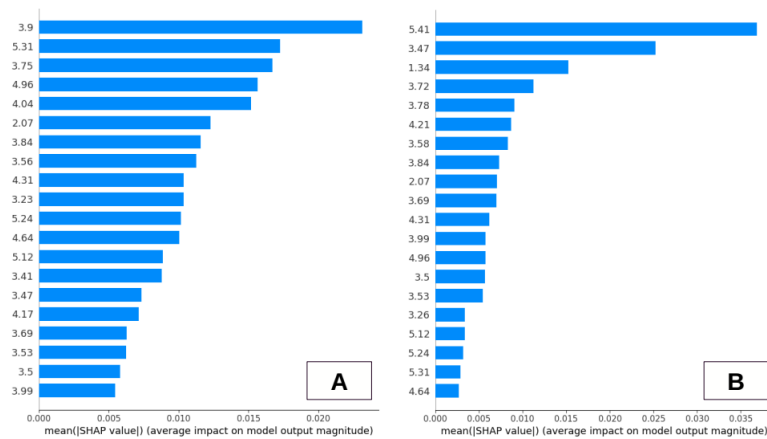


Figure 28: **Top 20 most important features ranked by mean absolute SHAP values. The features are ordered according to their importance of the best model classification for each production region. The Mean Shap values and their average impact on model output is shown for each feature.** A: SHAP values and the average impact of feature's on model of LR to classify 1A. B: SHAP values and the average impact of feature's on model of LR to classify 1B.

Table 22: **Best seven Mean absolute SHAP values and their impact on model output of the best models classification for Geographic Regions.** Note: (-): Negative impact on the model output; (+): Positive impact on the model output.

	LR	LR	PLS-DA	RF	LR	SVC	PLS-DA	SVC
Order	1A	1B	2C	3B	3C	4A	4B	5
1st	3.9	5.41	3.69	5.31	3.56	3.9	3.99	5.41
mean value	0.025	0.040	0.030	0.025	0.035	0.0005	0.05	0.009
impact	(+)	(+)	(-)	(+)	(+)	(-)	(+)	(+)
2nd	5.31	3.47	3.47	5.41	3.99	3.84	5.41	3.5
mean value	0.018	0.027	0.022	0.020	0.027	0.00045	0.035	0.0089
impact	(-)	(+)	(-)	(+)	(-)	(-)	(-)	(+)
3rd	3.75	1.34	4.17	3.75	1.34	5.41	2.07	3.56
mean value	0.017	0.018	0.022	0.018	0.023	0.00044	0.030	0.0069
impact	(-)	(-)	(çç)	(+)	(-)	(+)	(+)	(-)
4th	4.96	3.72	2.07	3.9	3.58	3.47	3.56	3.84
mean value	0.0165	0.0125	0.020	0.0125	0.022	0.00043	0.025	0.0069
impact	(-)	(-)	(+)	(+)	(-)	(+)	(-)	(+)
5th	4.04	3.78	3.56	3.84	5.41	4.96	3.9	3.26
mean value	0.016	0.0120	0.020	0.012	0.022	0.00025	0.025	0.0068
impact	(-)	(-)	(+)	(-)	(-)	(-)	(-)	(+)
6th	2.07	4.21	3.9	4.17	4.31	2.07	1.34	3.9
mean value	0.0125	0.011	0.017	0.011	0.021	0.00022	0.024	0.0067
impact	(-)	(-)	(-)	(-)	(+)	(-)	(-)	(-)
7th	3.84	3.58	3.72	4.31	3.9	3.23	3.23	3.58
mean value	0.0120	0.010	0.016	0.010	0.021	0.00021	0.020	0.005
impact	(+)	(+)	(-)	(+)	(+)	(-)	(+)	(+)

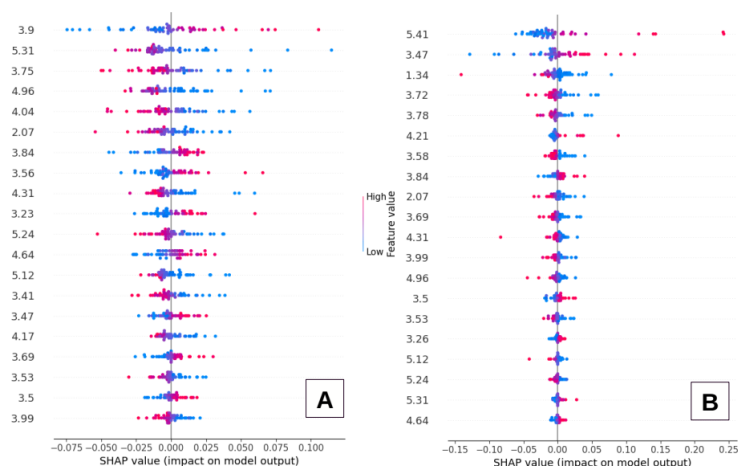


Figure 29: **The 20 most important features are shown and ordered according to their importance to the best model classification for each production region.** The Shap values and their impact on model output is shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of LR to classify 1A. B: SHAP values and impact of feature's on model of LR to classify 1B.

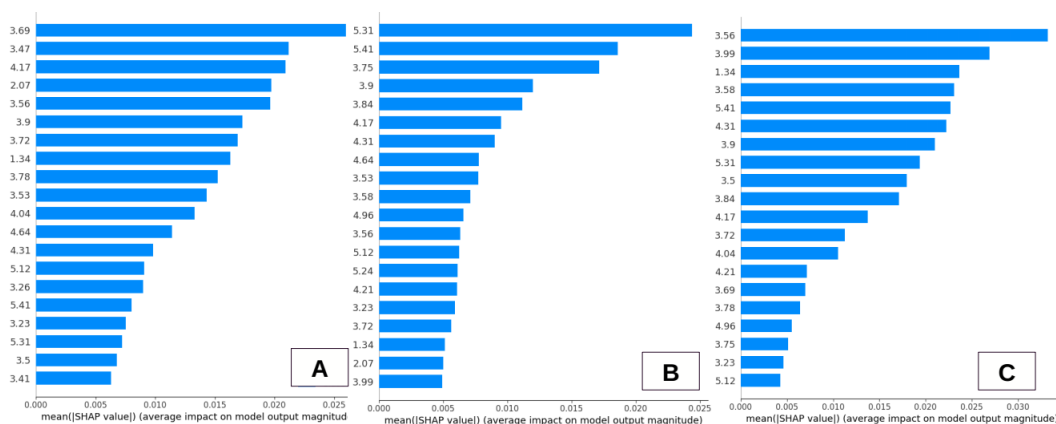


Figure 30: **Top 20 most important features ranked by mean absolute SHAP values.** The features are ordered according to their importance of the best model classification for each each production region. The Mean Shap values and their average impact on model output is shown for each feature. A: SHAP values and the average impact of feature's on model of PLS-DA to classify 2C. B: SHAP values and the average impact of feature's on model of RF to classify 3B. C: SHAP values and the average impact of feature's on model of LR to classify 3C.

6.2.5.2 Features analysis of the best models performance for Flora groups classification

Regarding flora groups with highest ROC-AUC classification models, Shapley values calculated an average range of impact from 0-0.025.

The results for the mean Shap values and their average impact on model output are shown for each feature of Flora groups with higher ROC-AUC classification models in Figures 34, 35. The computed mean Shap values resulted in twenty features. The impact and order of importance are different in each flora

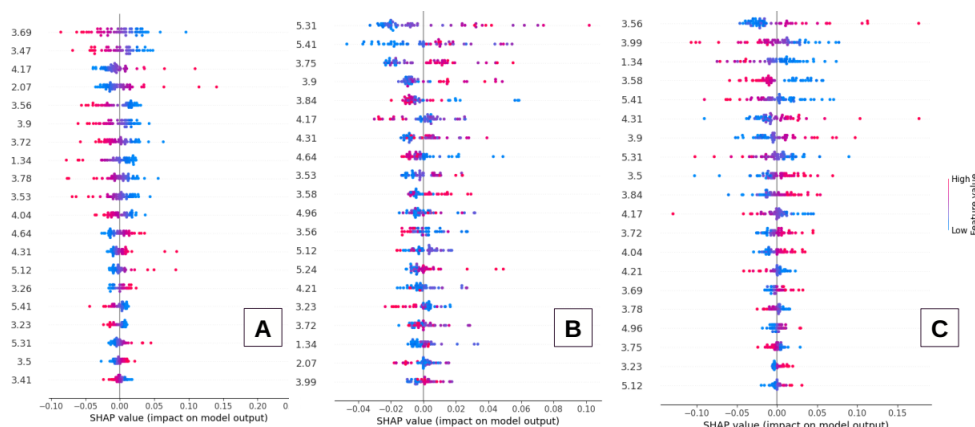


Figure 31: **The 20 most important features are shown and ordered according to their importance to the best model classification for each production region.** The Shap values and their impact on model output is shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of PLS-DA to classify 2C. B: SHAP values and impact of feature's on model of RF to classify 3B. C: SHAP values and the impact of feature's on model of LR to classify 3C.

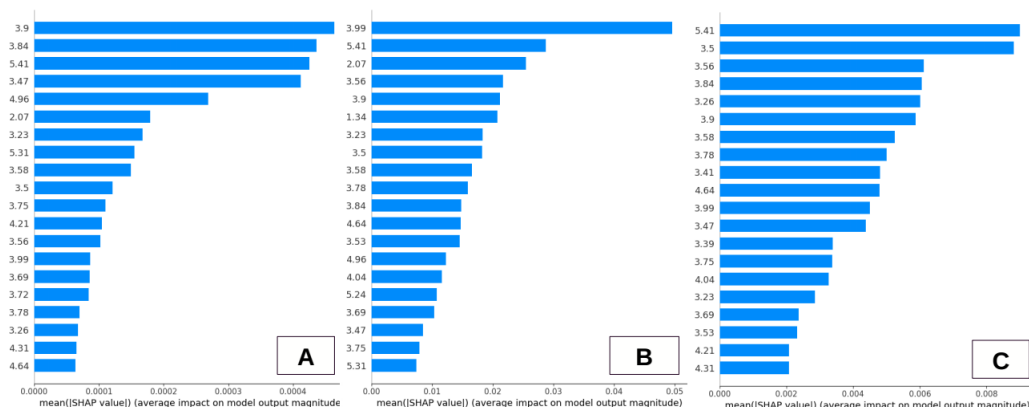


Figure 32: **Top 20 most important features ranked by mean absolute SHAP values.** The features are ordered according to their importance of the best model classification for each each production region. The Mean Shap values and their average impact on model output is shown for each feature. A: SHAP values and the average impact of feature's on model of SVC to classify 4A. B: SHAP values and the average impact of feature's on model of PLS-DA to classify 4B. C: SHAP values and the average impact of feature's on model of SVC to classify 5.

model.

The first seven computed mean Shap values with high importance and their impact on the models are shown in Table 23, for the calculated Flora groups.

No common feature with importance at the first seven positions was observed in Flora group classification models.

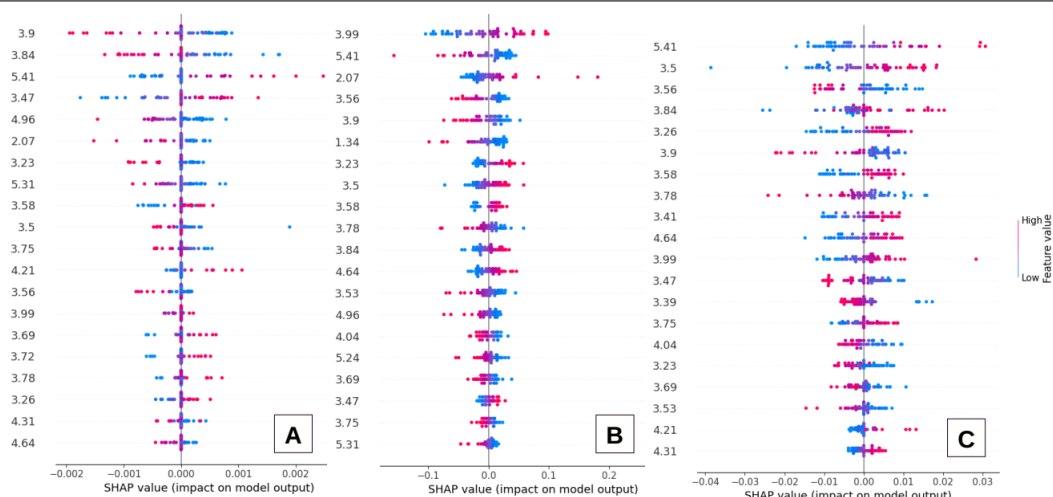


Figure 33: **The 20 most important features are shown and ordered according to their importance to the best model classification for each production region.** The Shap values and their impact on model output is shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of SVC to classify 4A. B: SHAP values and impact of feature's on model of PLS-DA to classify 4B. C: SHAP values and the impact of feature's on model of SVC to classify 5.

Table 23: **Best seven Mean absolute SHAP values and their impact on model output of the best models classification for Flora groups.** Note: (-): Negative impact on the model output; (+): Positive impact on the model output.

	PLS-DA	SVC	SVC
Order	UvadoJapão	Other	Silvestre
1st	2.07	3.41	4.04
mean value	0.030	0.016	0.013
impact	(+)	(+)	(+)
2nd	3.84	5.41	3.23
mean value	0.029	0.015	0.012
impact	(+)	(+)	(+)
3rt	4.31	3.84	3.26
mean value	0.027	0.0145	0.010
impact	(+)	(+)	(-)
4th	5.41	3.5	4.96
mean value	0.022	0.0145	0.009
impact	(-)	(+)	(+)
5th	5.12	3.99	5.31
mean value	0.021	0.0140	0.008
impact	(+)	(-)	(+)
6th	3.69	3.26	3.56
mean value	0.021	0.012	0.0075
impact	(-)	(+)	(-)
7th	3.75	3.9	3.58
mean value	0.020	0.012	0.0075
impact	(+)	(-)	(+)

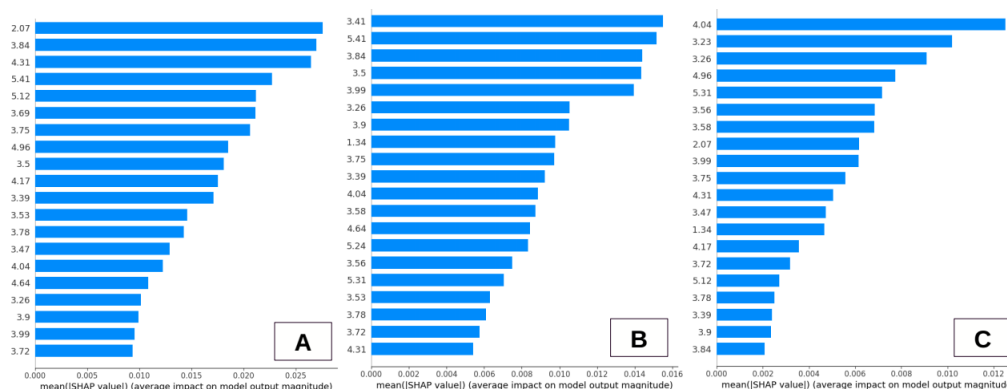


Figure 34: **Top 20 most important features ranked by mean absolute SHAP values. The features are ordered according to their importance of the best model classification for flora origin.** The Mean Shap values and their average impact on model output is shown for each feature. A: SHAP values and the average impact of feature's on model of PLS-DA to classify "Uva do Japão". B: SHAP values and the average impact of feature's on model of SVC to classify Other. C: SHAP values and the average impact of feature's on model of SVC to classify "Silvestre".

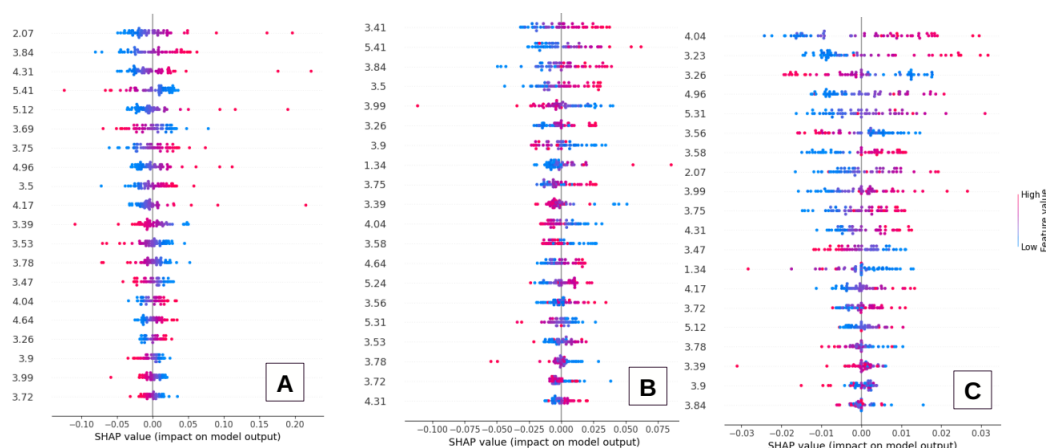


Figure 35: **The 20 most important features are shown and ordered according to their importance to the best model classification for flora origin.** The Shap values and their impact on model output is shown for each feature. Each point represents a sample in the sub-dataset, and the color of the points represents the value of a particular feature for that sample. A: SHAP values and impact of feature's on model of PLS-DA to classify "Uva do Japão". B: SHAP values and impact of feature's on model of SVC to classify Other. C: SHAP values and the impact of feature's on model of RF to classify "Silvestre".

6.2.6 Comparative analysis between the results of Geographic and flora classifications

Some of the features with importance for sample classification were previously identified by ANOVA as able to discriminate two geographic regions, e.g. 3B vs 5, being the resonance at 5.41 ppm the feature in

this case. This resonance has been assigned to the disaccharide sucrose and has a highest shap values impact in 1B, 3B, 4B and 5 regions models classification output.

This results allow to understand that statistical tools has some capacity to discriminate honey samples, but the use of ML algorithms (with adequate evaluation methods, features treatments, and score metrics), combined with the Shapley values computation, give more information to discriminate and classify honey samples.

In this context, after performing the computation of Shapley values for each model classification, we conducted a comparative analysis based on the feature importance results. Interestingly, it was observed that some features with importance for the sample classification in some regions were also considered features with importance in the flora group classification.

- "Uva do Japão" and regions 2C, 3C and 4A. The flora group of "Uva do Japão" presented one feature with importance, which is also present in the region 2C, the resonance at 3.69ppm (can be associated to ethanol (CH_2)). This feature, is only present in the first positions of shap values for the Flora group of "Uva do Japão". The resonance at 4.31ppm is another feature with importance present in region 3C and is associated with acid latic (CH).

Three other features with importance for "Uva do Japão" flora are present in the shap values model of region 4A, the resonance of 3.84ppm, the resonance of 5.41ppm related to sucrose, and the resonance of 4.96ppm.

- The combination of features with importance present in floral and regions also happens in "Silvestre" with regions 1A and 4A.

The resonances of 5.31ppm (associated with α -glucose), 4.96ppm, 4.04ppm, and 2.07ppm (related to proline) are present in the shap values with high importance in Flora "Silvestre" and region 1A models classification.

The resonances of 2.07ppm and 4.96ppm are present not only in region 1A but in also in region 4A.

All of these observations are from features in the top 7 of the features with the highest importance for Region and Flora model classifications.

Considering that "Uva do Japão", and "Silvestre" are plant species and arrangements of these present in several regions, these results of features with importance in the flora group present in more than one region were expected.

The overall synopsis of the process and the main results are shown in Figure 36.

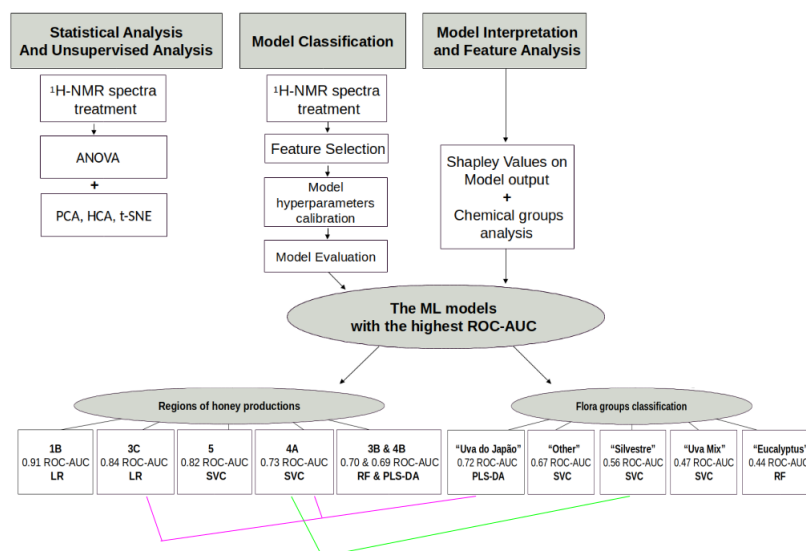


Figure 36: **Synopsis of the results whit higher performance and the data analysis on $^1\text{H-NMR}$ spectra for the regions honeys productions and flora.** The squares present the classification model and the metric score of the best model performance to flora origins. And the colored lines represent the links between flora and production regions.

6.3 Discussion

Reduced significant statistical differences were found in our honey samples based on regions and/or flora composition. Only two features (resonances, ppm) can discriminate samples according to the ANOVA analysis. However, despite the statistical similarities of the honey $^1\text{H-NMR}$ spectra, the supervised machine learning models were able to classify honey samples with good performance scores for each region and each flora group. The computed Shapley values also allowed a good interpretation by understanding the impact of the features with importance on the models' output as the comparative analysis highlighted possible features relation regarding flora and region origins.

In the literature, previous studies have developed models for flora and geographic origin classification which achieved good performances, with honey samples from different producing regions worldwide [148], [126], [83].

There are some studies focused on the authentication and traceability of honey using metabolomics data from NMR-based approaches [148], [126], [125], [229], [238], [239], [83], [240], [237], [241], [242]. However, the majority of them did not use all spectra focusing on specifying honey chemical compounds. A NMR approach covers crucial aspects of honey quality and authenticity on more than 800 commercial samples of monofloral and multifloral honeys from more than 60 different flora origins and more than 35 different countries. This work presents a large and systematic NMR profiling study of honey samples to establish a general methodology for the targeted and non-targeted analysis of honey authenticity [229].

A recent approach by Noviyanto et al., 2020 [243] also computed different supervised models to

classify different types of honey. The authors achieved promising results as k-Nearest Neighbour (kNN) with 84.9% and SVC with 91.7% using Balanced Accuracy as a metric to evaluate the models performance. Considering the honeys produced in Brazil, there are some reports with a focus on flora, and geographic origins but also with research on targeted analyses, as the example of Bracatinga honeydew honey [130], [223], [230], [231] with a increasing interest due to the IG classification of the “Bracatinga” honeydew.

One of the first works using Brazilian honeys is the study presented in 2012 by Boffo and others [130], with a focus on flora classification and with components annotation by the use of $^1\text{H-NMR}$ approach. The authors have used PCA and HCA methods and supervised models, the KNN, Soft independent modeling of class analogy (SIMCA), and PLS-DA for flora classification. The results showed good classification only using PLS-DA, with correctly classified 72.2% of the honey samples. A work more related to our study, with a focus on flora and geographic classification of honey Zhang and their collaborators in 2020, also used a non-target approach to metabolomics honey data using $^1\text{H-NMR}$ spectroscopy [83]. The authors studied eight different honey produced in different regions in China and achieved good classification considering the flora classification of the samples, in a easier way to classify them than geographic origin, due to the associated influence factors. For that, it was computed PCA analysis and Orthogonal Partial Least Squares Discrimination Analysis (OPLS-DA) in a multi-classification problem.

In this context, to the best of our knowledge no other works have been published recently in literature regarding t model interpretation in the field of honey classification, so that this is the first work to include geographic and flora origins for honeys produced in Santa Catarina state, which is the biggest honey production in Brazil. Our work also included several types of models to achieve the best classification model performances. By doing this approach, we have been able to selected the best model classification, that differed regarding the type of honey sample for each flora and geographic origin.

Despite being preliminary, the results herein shown are interesting, mainly regarding the classification of regions 1A, 2C, 3C and 4A, which are related to the “Uva do Japão” and “Silvestre” honeys. These two honeys had already received international recognition increasing their market value.

Finally, our work contributes to characterizing and understanding the authentication process for honeys produced in Brazil.

Based on our pipeline, further studies can be carried out on other geographic and flora origins with a greater number of samples. In further works, our pipeline can be reproduced on other metabolomics datasets from different techniques to compare the models’ performance and highlight features importance.

6.4 Conclusions

Our exploratory study addressed two main objectives: the geographic region and flora classification of Brazilian honey and the feature analysis with classification models interpretation.

Despite some statistical similarities of the honey $^1\text{H-NMR}$ spectra, the LR, SVC and PLS-DA models resulted to be suitable models to classify the samples from SC state based on their region of production

and their flora group origin. The calculation of Shapley values allowed us to identify the most influential features and their influence on the model output. We also related the link between identified features of geographic and flora regions with the correspondent compounds groups.

These outcomes highlight the importance and contribution of machine learning models, evaluation metrics, and optimization approaches on chemometrics to food authentication, particularly in the case of honeys produced in Brazil. To infer the influence of geographic influences such as altitude, rainfall, soil, and other influence factors further studies should be carried out, expanding the comprehension of the effect of environmental factors on the quality of that bee product. Follow-up studies should also focus on verifying if there the same relations hold between other flora origins with different geographic production regions to validate our case study.

To enrich the classification models of our pipeline more data from other honeys should be added. Still, the proposed pipeline can serve as a basis for future works in the authentication of honeys, aiming to contribute to other authentication processes of interest.

6.5 Materials and Methods

6.5.1 Samples

Metabolomics data from fifty six from samples of honey samples were kindly provide by the Federal University of Santa Catarina (UFSC), Brazil, to be used for flora and geographic classification using machine learning models. The information regarding the main flora composition of samples and the region of honey production is show in Table 17 and Table 18. The honey samples were produced in the years of 2018 to 2019, during the seasons of spring, summer and fall, in all eleven agroecological regions of Santa Catarina estate in Brazil. All samples came from registered apiaries with geolocation, information regarding the type of flora available at the close area and the name of producers. The honey production follows the regulated and standard procedures without the use of pesticides, antibiotics and other chemical residues on fauna and flora. All samples were firstly submitted to a quality control analyses as describe in “Regulamento Técnico da Instrução Normativa nº 11/2000”. The metabolomics data of the samples kindly shared to performed our data analysis, came from a collaboration between Federal University of Santa Catarina (UFSC), Estate University of Santa Catarina (UESC), Federation: “Federação das Associações de Apicultores e Meliponicultores de Santa Catarina”(FAASC) and the company: “Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina (EPAGRI). These collaborations are involved in works to perform quality control tests according to the legislation of honeys from Santa Catarina.

6.5.2 Methods

6.5.2.1 1D Nuclear Magnetic Resonance spectroscopy (^1H -NMR)

A 50 mg honey sample was added 700 μL D_2O , mixed (1 min, vortex) and filtered on a cellulose membrane under reduced pressure. Before ^1H -NMR analysis, 3-trimethylsilyl propionic-2, 2, 3, 3- d_4 acid sodium salt (TSP, 98 atoms% D, 0.024g%) was added to samples as internal reference standard. The ^1H -NMR analysis of the samples was performed at Brazilian Biosciences National Laboratory (Brazilian Center for Research in Energy and Materials, Campinas, SP), in a NMR spectrometer operating at a Larmor frequency of 499.726 MHz for ^1H . The ^1H -NMR spectra were recorded on a Varian Inova spectrometer, using the VNMRJ software. All experiments were performed at 300 K and non-spinning. For the high-resolution ^1H -NMR spectra 32K data points (time domain) were collected, with an acquisition time 4.8 s, delay time 1.0 s, 4 dummy scans, and total number of scans 32. Chemical shifts for protons were reported in parts per million (ppm) downfield from TSP at $\delta(^1\text{H})$ 0.00 ppm and referenced to that internal reference standard.

6.5.3 Statistical Analysis and Chemometrics

Pre-processing spectral signal was performed by R language through the package *specmine* [8], available in CRAN.

The remaining processing and data analyses were performed by using the Python language, and the libraries pandas, mumpy, ResearchPy, Scipy, Statsmodels, and Scikit-learn.

6.5.3.1 Spectral Data Processing and Data pre-processing

The Peak alignment grouped proximal peaks together according to their median position across all samples, using a moving window of 0.03ppm. 2E4 was the baseline selection limit for the detection of minimum positive values. Typical resonance of the deuterium oxide (D_2O)(4.80ppm) signal was removed from the dataset. No missing values and no outliers were detected. The processed dataset of the 56 honey samples was extracted into a file (comma separated values format .csv), where the first column indicates peak positions (ppm) and the second one represents the peak intensities containing a total of 36 peaks per sample, with an interval of peaks comprehending 1.34ppm to 5.41ppm. To note the the raw spectra without any treatment comprehends an interval of resonances between 0ppm to 13.78ppm. After spectral data processing, it was followed by z-score standardization, a selected treatment utilized to perform the data analyses.

Two datasets were organized from the samples based on region of production and flora origin. To note that both datasets had the same samples but are organized according to class classification (production region or flora origin). All samples that had replicates were transformed by computing the respective spectra mean values.

6.5.3.2 Data analysis

Exploratory and statistical analysis A preliminary analysis included statistical univariate techniques and unsupervised multivariate statistical models to understand the samples relationships among the different classes according to the region of production and according to the flora origin. To understand if there are statistical differences between different combinations of must varieties an ANOVA with Tukey post hoc-test, between the each region pair and each flora pair, was carried to identify which specific resonances (ppm) could be utilized to discriminate between these classes. The unsupervised multivariate statistical models used in this study were Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA) and T-Distributed Stochastic Neighbor Embedding (t-SNE), alternative approaches to interpret ^1H -NMR data. PCA is widely used tool for interpreting and visualizing the main features, as for classifying samples and in this study was used to see if there is a clear discrimination between the samples with different classes in the latent space. HCA used an Euclidean distance between samples and the complete linkage method. This alternative was chosen because it tends to find compact clusters of approximately equal diameters and does not force clusters together due to single elements being close to each other, like in single linkage clustering. t-SNE is a method which applies a non-linear dimensionality reduction technique to keep the similar data points close in a lower-dimensional space. This method uses a student t-distribution to compute the similarity between two points in lower-dimensional space.

Region and Flora classification and features analysis To classify each class a one-vs-all approach was followed, employing binary classification to separate a target class from any other set of honeys samples. The main objective is to classify a target class from any other source. Since this is an exploratory work with potential to add more classes (not regions but floras) in further studies, the binary classification is the best approach to apply into this scenario.

The following models were used: Random Forest Classifier (RF), Gradient Boosting Classifier (GBC), Logistic Regression (LR), C-Support Vector Classification (SVC) and Partial Least Squares-Discriminant Analysis (PLS-DA). Several performance metrics were tested and compared, being the Receiver Operator Characteristic Under the Curve (ROC-AUC) the selected metric to evaluate the models performance. Due to the reduced number of samples, leave-one-out cross-validation was utilized to train and validate the model. Each model used hyperparameters is described in Table 24.

Table 24: **Hyper-parameters used by the distinct models.**

Model	Hyper-parameters
Random Forest Classifier	N of Estimators: 100; Max N of Features: sqrt; Max. of depth: None; Class weighth: balanced.
SVC C-Support Vector Classification – Support Vector Machine Classification	kernel: rbf; gamma: scale; C: 1; Class weighth: balanced
Gradient Boosting Classifier	N of Estimators: 100; Learning rate: 0.1; Max N of Features: None; Max. of depth: 3; Class weighth: balanced
Logistic Regression	C: 1.0;
Partial Least Squares-Discriminant Analysis	N of Components: 2

For each training, and validation set, the following preprocessing was carried: each model was calibrated using the treated raw spectra with the peak alignment, baseline selection and plus standardization using the aforementioned preprocessing pipeline. The calibration was performed with explicit feature selection. The explicit Feature Selection method consisted in dropping correlated features above a given threshold (in this work 90%). First, the correlation matrix of the spectral resonances was computed. Next, the correlation matrix was further processed by forming clusters of resonances with correlation values above the threshold. For each cluster, the resonance with the highest variance is kept and remaining ones were dropped. The calibrated models were evaluated in the test set, and the spectra treatment was similar (using the mean and standard deviation values computed in the train dataset).

Feature analysis To assess the feature importances, the best models from each class were chosen based on the best obtained performance in the aforementioned cross-validation schema for each target class (Region and Flora). Afterwards, each resonance influence on the target classes was inferred using the Shapley values. These values are the average marginal contribution of a feature value across all possible coalitions and can be applied for classification problems, as in our study [56], [58]. For binary classification, SHAP values for classes 0 and 1 are symmetrical. The selected feature contributes to a certain amount towards class 1, at the same time reduces the probability of being class 0 by the same amount.

Conclusions and Future Work

The overall aim of this thesis was the development of pipeline approaches of statistical and machine learning models using metabolomics datasets to solve/improve different issues of food authentication problems.

A variety of models were built and evaluated for three different authentication problems of two natural food products with a high-value market and PDO importance, mainly, wine, wine musts, and honey samples.

Although the models developed in this work are still far from being used in food authentication processes, the work described here will be useful to guide the development of future computational methods for the design of novel and more accurate classification approaches.

The developed work analyzed wine, wine musts and honey samples obtained from the Federal University of Santa Catarina and wine musts samples obtained from the Instituto Nacional de Investigação Agrária e Veterinária, Dois Polos, Portugal. Although the limited sample size of the metabolomics datasets posed some challenges, the results shed light on the potential of these tools to be use in the authentication of food products. Additionally, this study represents the first investigation focusing on honey samples from Santa Catarina Estate and the first to employ wine musts samples from Portuguese grape varieties grown in PDO regions of Portugal. Further studies are needed to expand on these preliminary findings.

In the following sections, the main contributions of this thesis will be summarized, and suggestions for further improvement will be proposed.

7.1 Main contributions

In Chapters 2 and 3, a review of the literature, regarding the state-of-the-art on metabolomics-based approaches for authentication of natural food products using statistical tools and machine learning algorithms, was performed.

Multiple concepts across the metabolomics field were described and analyzed, including the main statistical and machine learning approaches used in food authentication processes. It became evident

that further development of novel model-based methods was needed to understand and overcome several hurdles still encountered in the reviewed studies.

One of the main contributions of this work was the analysis of different methods tested on the classification performance of ML-based approaches to metabolomics data. A variety of preprocessing methods and modeling strategies were tested, in an attempt to understand which approaches produce the best results for each model classification. These contributions are present in the studies in chapters 4, 5, and 6.

Chapter 4, performed the annotation of the metabolites present in wine samples with different ages of bottle storage, from different harvest years, and produced under similar climatic conditions. To classify each group of wine samples of the same age, different models were tested to select the one with higher performance with results between 80-92% of accuracy. To predict the influence of age on bottle storage and the climatic influences Linear Regression model was utilized. The used pipeline allows us to identify the resonances influenced by each of the climatic factors and age of bottle storage. This work is one of the few with a focus on the age of bottle storage using a dataset with a large time interval (years) and with the climatic influences over the production years. Our work proves the efficacy to combine metabolomics data combined with ML to classify problems related to harvest years and to predict the age of bottle and climatic influences in wine samples.

In Chapter 5, ML was used to classify different Portuguese native wine musts varieties.

During the development of a solution for this classification problem, it was necessary to deal with different wine must varieties samples resulting from the grape maturation process, from different but close production places. The implementation of feature selection methods combined with different models was assessed. This approach also included the test of different pre-processing treatments and different metrics evaluation, to understand which models were more suitable for the classification problem of the study. This pipeline is a possible approach to improve each variety model classification performance. This work also explored the interpretability of the model, being the first study in the authentication of wine musts samples, to use this method. With the use of the interpretation model, it was clear that based on our data that there are some resonance bands related to functional chemical groups which are associated with each variety classification.

In Chapter 6, a similar analysis to chapter 5 was performed for the problem of botanical and geographic classification of honey produced in eleven agroecological regions from Santa Catarina, Brazil. Here the major difficulty was the number of regions with different sample sizes per region. Binary classification models were tested to deal with a large number of regions with different sample sizes. The implementation of feature selection methods combined with different models was well-suited to this type of classification task. And the use of different metrics evaluation, to understand which models were more adequate to the problem of the study was assessed. Here, it was selected the ROC-AUC metric, instead of MCC or PR-AUC, because it allows to understand the generalization of the model to each scenario. This pipeline was a good approach to improve each region's model classification performance as to each botanical honey origin. As in the previous exploratory study, model interpretability, was explored. Based on the

available metabolomics data, it was possible to understand which resonances are associated with each region and botanical origin, and which resonances are related between regions and flora. These results can be used in future works to explore the authentication of honey based on both classifications and in examples of origin certifications. Until the moment of this dissertation submission, this is the first study of honey classification based on region and biological origin from Santa Catarina Brazil.

The work of the three food products (wine, wine must, and honey) confirmed that ML algorithms are well-suited to the classification of authentication problems.

The most important issues to pay attention are the quality of the data and what pre-treatments are more adequate for each metabolomics data. The use of feature selection methods, tested more than one model and evaluation metrics to get the most adequate model, able to interpret each authentication problem.

Despite the ability of ML models to learn relevant features directly from raw data, the use of prior biological knowledge to select features is still useful to reduce the size of high-dimensional omics datasets, particularly the presence of solvents, water, and other elements present in IR and NMR datasets.

The studies of food authentication have a reduced size of datasets and used similar and a reduced number of models to perform predictions or classifications. In this work, despite the reduced sample size for each study, several models were used. Distinct models may learn different aspects of the problem, even in cases with a reduced sample size. When tested more and different models combined with the proper validation model method, it may increase the performance of the final classification model.

In this thesis, the issue of model interpretability was also addressed. Using the feature importance of the Shapley values method, we were able to demonstrate that the decisions made by our ML models are driven by relevant resonances related to functional chemical groups.

The work undertaken in Chapter 4 contributed to extending and validating the Specmine software package, mainly giving support to metabolites annotation and metabolomics spectral data visualization developed by our research group.

With these, reliable and reproducible analysis pipelines on a scientific level, the main objectives for this work were accomplished, although there are still many improvements that could be done in future works.

7.2 Future work

The insights gained from this research can be used as a starting point for the development of better classification models for food authentication in the future. In this section, we suggest ways to extend and improve upon the work described in this thesis.

A major limitation of our work was the reduced size of all datasets, with reduce number of samples not allowing to explore more issues of authentication process and limitation the validation of the models.

With this being said, the developed data analyses with the data available conditions allow us to perform reliable and reproducible analysis pipelines, on a scientific level.

There are still, however, many aspects that could be improved in the web data analysis despite the size of dataset, including the improvement of already existing models implementation.

The future work could encompass:

- Combining multiple metabolomics datasets to have access to more training data, for example by using existing integrative databases;
- Combining multiple models could be a good strategy to improve performance. Distinct models may learn different aspects of the problem, and thus combining them may increase the generalization capacity of the final model;
- Preprocessing methods and evaluation metrics improve models performance in Chapters 4, 5 and 6. Thus, we recommend the use of more elaborate preprocessing and evaluation metrics and in the future food authentication studies.
- Prior knowledge for feature selection proved to be beneficial in Chapter 5 and 6. In the future, other ways of leveraging this knowledge should be explored, including the use of alternative resonances for feature selection and methods that can incorporate chemical knowledge directly within the models.
- Finally, the development of a more general and easy-to-use pipeline to food authentication processes with ML and Model interpretability could be applied to other wine varieties, different honeys and other food products in general. These pipelines could be give important contributions to the field of food authentication and would allow our work to be more widely used by the community.

Bibliography

- [1] D. I. Ellis, H. Muhamadali, S. A. Haughey, C. T. Elliott, and R. Goodacre, "Point-and-shoot: Rapid quantitative detection methods for on-site food fraud analysis—moving out of the laboratory and into the food supply chain", *Analytical Methods*, vol. 7, no. 22, pp. 9401–9414, 2015.
- [2] D. J. Shaw, "World food security", *A History since*, 1945.
- [3] M. Castro-Puyana, R. Pérez-Míguez, L. Montero, and M. Herrero, "Application of mass spectrometry-based metabolomics approaches for food safety, quality and traceability", *TrAC Trends in Analytical Chemistry*, 2017.
- [4] A. Zhang, H. Sun, P. Wang, Y. Han, and X. Wang, "Modern analytical techniques in metabolomics analysis", *Analyst*, vol. 137, no. 2, pp. 293–300, 2012.
- [5] O. Fiehn, "Metabolomics—the link between genotypes and phenotypes", in *Functional Genomics*, Springer, 2002, pp. 155–171.
- [6] P. O. Larsen and M. Von Ins, "The rate of growth in scientific publication and the decline in coverage provided by science citation index", *Scientometrics*, vol. 84, no. 3, pp. 575–603, 2010.
- [7] U. Roessner and J. Bowne, "What is metabolomics all about?", *Biotechniques*, vol. 46, no. 5, p. 363, 2009.
- [8] C. Costa, M. Maraschin, and M. Rocha, "An r package for the integrated analysis of metabolomics and spectral data", *Computer methods and programs in biomedicine*, vol. 129, pp. 117–124, 2016.
- [9] J. M. Cevallos-Cevallos, J. I. Reyes-De-Corcuera, E. Etxeberria, M. D. Danyluk, and G. E. Rodrick, "Metabolomic analysis in food science: A review", *Trends in Food Science & Technology*, vol. 20, no. 11, pp. 557–566, 2009.
- [10] C. A. Georgiou and G. P. Danezis, *Food authentication: Management, analysis and regulation*. John Wiley & Sons, 2017.
- [11] S. G. Villas-Boas, J. Nielsen, J. Smedsgaard, M. A. Hansen, and U. Roessner-Tunali, *Metabolome analysis: An introduction*. John Wiley & Sons, 2007.

- [12] K. O'Shea and B. B. Misra, "Software tools, databases and resources in metabolomics: Updates from 2018 to 2019", *Metabolomics*, vol. 16, no. 3, pp. 1–23, 2020.
- [13] K. Böhme, P. Calo-Mata, J. Barros-Velázquez, and I. Ortea, "Recent applications of omics-based technologies to main topics in food authentication", *TrAC Trends in Analytical Chemistry*, vol. 110, pp. 221–232, 2019.
- [14] R. Ramautar, A. Demirci, and G. J. de Jong, "Capillary electrophoresis in metabolomics", *TrAC Trends in Analytical Chemistry*, vol. 25, no. 5, pp. 455–466, 2006.
- [15] D. S. Wishart, "Metabolomics: Applications to food science and nutrition research", *Trends in Food Science & Technology*, vol. 19, no. 9, pp. 482–493, 2008.
- [16] S.-A. Sansone, T. Fan, R. Goodacre, J. L. Griffin, N. W. Hardy, et al., "The metabolomics standards initiative", *Nature biotechnology*, vol. 25, no. 8, pp. 846–848, 2007.
- [17] S. G. Villas-Bôas, U. Roessner, M. A. Hansen, J. Smedsgaard, and J. Nielsen, "Microbial metabolomics: Rapid sampling techniques to investigate intracellular metabolite dynamics—an overview", *Metabolome Analysis: An Introduction*, pp. 203–214, 2006.
- [18] A. Alonso, S. Marsal, and A. Julià, "Analytical methods in untargeted metabolomics: State of the art in 2015", *Frontiers in bioengineering and biotechnology*, vol. 3, p. 23, 2015.
- [19] J. L. Markley, R. Brüschweiler, A. S. Edison, H. R. Eghbalnia, R. Powers, et al., "The future of nmr-based metabolomics", *Current opinion in biotechnology*, vol. 43, pp. 34–40, 2017.
- [20] A. Owen, "Fundamentals of uv-visible spectroscopy", 1996.
- [21] E. Smith and G. Dent, *Modern raman spectroscopy: A practical approach*. John Wiley & Sons, 2013.
- [22] E. Acuna and C. Rodriguez, *The treatment of missing values and its effect on classifier accuracy*. Springer, 2004, pp. 639–647.
- [23] V. Hodge and J. Austin, "A survey of outlier detection methodologies", *Artificial intelligence review*, vol. 22, no. 2, pp. 85–126, 2004.
- [24] J. L. Schafer, *Analysis of incomplete multivariate data*. CRC press, 1997.
- [25] A. R. Kapil, *Methods of missing value treatment and their effect on the accuracy of classification models*, 2018.
- [26] C. B. Costa, "Development of an integrated computational platform for metabolomics data analysis and knowledge extraction", PhD thesis, 2014.
- [27] G. A. Morris, "Nmr data processing", *Encyclopedia of Spectroscopy and Spectrometry*, pp. 125–133, 2017.
- [28] O. M. Kvalheim, F. Brakstad, and Y. Liang, "Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise", *Analytical Chemistry*, vol. 66, no. 1, pp. 43–51, 1994.

-
- [29] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: Improving the biological information content of metabolomics data", *BMC genomics*, vol. 7, no. 1, pp. 1–15, 2006.
- [30] K. H. Liland, "Multivariate methods in metabolomics—from pre-processing to dimension reduction and statistical analysis", *TrAC Trends in Analytical Chemistry*, vol. 30, no. 6, pp. 827–841, 2011.
- [31] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures.", *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.
- [32] K. Varmuza and P. Filzmoser, *Introduction to multivariate statistical analysis in chemometrics*. CRC press, 2016.
- [33] Å. Rinnan, F. Van Den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra", *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.
- [34] S. Kus, Z. Marczenko, and N. Obarski, "Derivative uv-vis spectrophotometry in analytical chemistry", *Chem. Anal*, vol. 41, no. 6, pp. 889–927, 1996.
- [35] J. Godzien, A. G. de la Fuente, A. Otero, and C. Barbas, *Metabolite annotation and identification*. Elsevier, 2018, vol. 82, pp. 415–445.
- [36] P. Patil, "What is exploratory data analysis", *Toward Data Science*, 2018.
- [37] P. Filzmoser, K. Hron, and C. Reimann, "Univariate statistical analysis of environmental (compositional) data: Problems and possibilities", *Science of the Total Environment*, vol. 407, no. 23, pp. 6100–6108, 2009.
- [38] T. W. MacFarland and J. M. Yates, *Introduction to nonparametric statistics for the biological sciences using r*. Springer, 2016.
- [39] R. B. Darlington and A. F. Hayes, *Regression analysis and linear models: Concepts, applications, and implementation*. Guilford Publications, 2016.
- [40] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.", *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [41] M. Rocha, P. Cortez, and J. M. Neves, *Análise inteligente de dados: Algoritmos e implementação em java*. 2008.
- [42] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics", *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [43] V. Kumar and S. Minz, "Feature selection", *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.
- [44] R. Kohavi, *Special issue on applications of machine learning and the knowledge discovery process*. Kluwer, 1998.

- [45] D. R. Cox, "The regression analysis of binary sequences", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.
- [46] M. Barker and W. Rayens, "Partial least squares for discrimination", *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 17, no. 3, pp. 166–173, 2003.
- [47] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning.[sl]*, 2012.
- [48] M. Amer and S. Todorovic, "Sum product networks for activity recognition. pattern analysis and machine intelligence", *IEEE Transactions on*, 2015.
- [49] R. J. Tallarida, "Quantitative methods for assessing drug synergism", *Genes & cancer*, vol. 2, no. 11, pp. 1003–1008, 2011.
- [50] A. M. Martinez and A. C. Kak, "Pca versus lda", *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 228–233, 2001.
- [51] V. Cortes, "Cortes c., vapnik v", *Support-vector networks, Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [52] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. Springer, 2009, vol. 2.
- [53] S. B. Kotsiantis, "Decision trees: A recent overview", *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261–283, 2013.
- [54] J. H. Friedman, "Stochastic gradient boosting", *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.
- [55] L. S. Shapley, "A value for n-person games", *Classics in game theory*, vol. 69, 1997.
- [56] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", *Advances in neural information processing systems*, vol. 30, 2017.
- [57] L. S. Shapley, H. Kuhn, and A. Tucker, "Contributions to the theory of games", *Ann Math Stud*, vol. 28, no. 2, pp. 307–17, 1953.
- [58] A. Joseph, "Parametric inference with universal function approximators", *arXiv preprint arXiv:1903.04209*, 2019.
- [59] L. M. Reid, C. P. O'donnell, and G. Downey, "Recent technological advances for the determination of food authenticity", *Trends in Food Science & Technology*, vol. 17, no. 7, pp. 344–353, 2006.
- [60] C. Georgiou and G. Danezis, "Elemental and isotopic mass spectrometry", *Advanced Mass Spectrometry for Food Safety and Quality*, vol. 68, pp. 131–243, 2015.
- [61] P. Emilia et al., "The authenticity and traceability of food – consumers protection forms", *The Annals of The University of Oradea*, p. 658, 2013.
- [62] M áJohn Dennis, "Recent developments in food authentication", *Analyst*, vol. 123, no. 9, 151R–156R, 1998.

- [63] "Reporting food fraud.", 2015.
- [64] D. G. Sotirchos, G. P. Danezis, and C. A. Georgiou, "Introduction, definitions and legislation", *Food Authentication: Management, Analysis and Regulation*, pp. 1–18,
- [65] P. Brereton, *New analytical approaches for verifying the origin of food*. Elsevier, 2013.
- [66] D. T. Braden, *Food fraud and adulterated ingredients*. 2014.
- [67] B. Wilson, *Swindled: The dark history of food fraud, from poisoned candy to counterfeit coffee*. Princeton University Press, 2008.
- [68] P. Shears, "Food fraud—a current issue but an old problem", *British Food Journal*, vol. 112, no. 2, pp. 198–213, 2010.
- [69] G. P. Danezis, A. S. Tsagkaris, F. Camin, V. Brusic, and C. A. Georgiou, "Food authentication: Techniques, trends & emerging approaches", *TrAC Trends in Analytical Chemistry*, vol. 85, pp. 123–132, 2016.
- [70] G. Downey, *Advances in food authenticity testing*. 2016.
- [71] E. Polshin, B. Aernouts, W. Saeys, F. Delvaux, F. R. Delvaux, et al., "Beer quality screening by ft-ir spectrometry: Impact of measurement strategies, data pre-processings and variable selection algorithms", *Journal of food engineering*, vol. 106, no. 3, pp. 188–198, 2011.
- [72] M. Lin, M. Al-Holy, S.-S. Chang, Y. Huang, A. G. Cavinato, et al., "Rapid discrimination of alicyclobacillus strains in apple juice by fourier transform infrared spectroscopy", *International journal of food microbiology*, vol. 105, no. 3, pp. 369–376, 2005.
- [73] P. Santos, E. Pereira-Filho, and L. Rodriguez-Saona, "Rapid detection and quantification of milk adulteration using infrared microspectroscopy and chemometrics analysis", *Food chemistry*, vol. 138, no. 1, pp. 19–24, 2013.
- [74] S. Kumar, P. Panchariya, B. Prasad, and A. Sharma, "Discrimination of indian tea varieties using uv-vis-nir spectrophotometer and pattern recognition techniques", *International Journal of Computer Science and Communication Engineering*, vol. 2, no. 2, pp. 15–19, 2013.
- [75] U. T.C. P. Souto, M. J. C. Pontes, E. C. Silva, R. K. H. Galvão, M. C. U. Araújo, et al., "Uv–vis spectrometric classification of coffees by spa–lda", *Food Chemistry*, vol. 119, no. 1, pp. 368–371, 2010.
- [76] O Barbosa-García, G Ramos-Ortiz, J. Maldonado, J. Pichardo-Molina, M. Meneses-Nava, et al., "Uv–vis absorption spectroscopy and multivariate analysis as a method to discriminate tequila", *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 66, no. 1, pp. 129–134, 2007.
- [77] S. Masoum, C. Malabat, M. Jalali-Heravi, C. Guillou, S. Rezzi, et al., "Application of support vector machines to 1h nmr data of fish oils: Methodology for the confirmation of wild and farmed salmon and their origins", *Analytical and bioanalytical chemistry*, vol. 387, no. 4, pp. 1499–1510, 2007.

- [78] G. Papotti, D. Bertelli, M. Plessi, and M. C. Rossi, "Use of hr-nmr to classify propolis obtained using different harvesting methods", *International journal of food science & technology*, vol. 45, no. 8, pp. 1610–1618, 2010.
- [79] D. R. Rudell, J. P. Mattheis, and E. A. Curry, "Prestorage ultraviolet- white light irradiation alters apple peel metabolome", *Journal of agricultural and food chemistry*, vol. 56, no. 3, pp. 1138–1147, 2008.
- [80] G. McDougall, I. Martinussen, and D. Stewart, "Towards fruitful metabolomics: High throughput analyses of polyphenol composition in berries using direct infusion mass spectrometry", *Journal of Chromatography B*, vol. 871, no. 2, pp. 362–369, 2008.
- [81] W. Pongsuwan, T. Bamba, K. Harada, T. Yonetani, A. Kobayashi, *et al.*, "High-throughput technique for comprehensive analysis of japanese green tea quality assessment using ultra-performance liquid chromatography with time-of-flight mass spectrometry (uplc/tof ms)", *Journal of agricultural and food chemistry*, vol. 56, no. 22, pp. 10 705–10 708, 2008.
- [82] H.-S. Son, G.-S. Hwang, H.-J. Ahn, W.-M. Park, C.-H. Lee, *et al.*, "Characterization of wines from grape varieties through multivariate statistical analysis of 1 h nmr spectroscopic data", *Food Research International*, vol. 42, no. 10, pp. 1483–1491, 2009.
- [83] J. Zhang, H. Chen, C. Fan, S. Gao, Z. Zhang, and L. Bo, "Classification of the botanical and geographical origins of chinese honey based on 1h nmr profile with chemometrics", *Food Research International*, vol. 137, p. 109 714, 2020.
- [84] A. J. Siddiqui, S. G. Musharraf, M. I. Choudhary, *et al.*, "Application of analytical methods in authentication and adulteration of honey", *Food chemistry*, vol. 217, pp. 687–698, 2017.
- [85] M. Amargianitaki and A. Spyros, "Nmr-based metabolomics in wine quality control and authentication", *Chemical and Biological Technologies in Agriculture*, vol. 4, no. 1, pp. 1–12, 2017.
- [86] O. HLGMUTH, "Wine-omics", *Nature*, vol. 455, p. 2, 2008.
- [87] M. Anastasiadi, A. Zira, P. Magiatis, S. A. Haroutounian, A. L. Skaltsounis, *et al.*, "1h nmr-based metabonomics for the classification of greek wines according to variety, region, and vintage. comparison with hplc data", *Journal of Agricultural and Food Chemistry*, vol. 57, no. 23, pp. 11 067–11 074, 2009.
- [88] K. Ali, F. Maltese, A. M. Fortes, M. S. Pais, Y. H. Choi, *et al.*, "Monitoring biochemical changes during grape berry development in portuguese cultivars by nmr spectroscopy", *Food Chemistry*, vol. 124, no. 4, pp. 1760–1769, 2011.
- [89] A. Springer, J. Riedl, S. Esslinger, T. Roth, M. Glomb, *et al.*, "Validated modeling for german white wine varietal authentication based on headspace solid-phase microextraction online coupled with gas chromatography mass spectrometry fingerprinting", *Journal of agricultural and food chemistry*, vol. 62, no. 28, pp. 6844–6851, 2014.

- [90] M. Caruso, F. Galgano, M. A. Castiglione Morelli, L. Viggiani, L. Lencioni, *et al.*, “Chemical profile of white wines produced from ‘greco bianco’ grape variety in different Italian areas by nuclear magnetic resonance (nmr) and conventional physicochemical analyses”, *Journal of agricultural and food chemistry*, vol. 60, no. 1, pp. 7–15, 2011.
- [91] M. Martin-Pastor, E. Guitian, and R. Riguera, “Joint nmr and solid-phase microextraction–gas chromatography chemometric approach for very complex mixtures: Grape and zone identification in wines”, *Analytical chemistry*, vol. 88, no. 12, pp. 6239–6246, 2016.
- [92] P. L. Pisano, M. F. Silva, and A. C. Olivieri, “Anthocyanins as markers for the classification of Argentinean wines according to botanical and geographical origin. chemometric modeling of liquid chromatography–mass spectrometry data”, *Food chemistry*, vol. 175, pp. 174–180, 2015.
- [93] R. S. Jackson, *Wine science: Principles and applications*. Academic press, 2008.
- [94] A. Cuadros-Inostroza, P. Giavalisco, J. Hummel, A. Eckardt, L. Willmitzer, *et al.*, “Discrimination of wine attributes by metabolome analysis”, *Analytical Chemistry*, vol. 82, no. 9, pp. 3573–3580, 2010.
- [95] R. Consonni, L. R. Cagliani, V. Guantieri, and B. Simonato, “Identification of metabolic content of selected amarone wine”, *Food chemistry*, vol. 129, no. 2, pp. 693–699, 2011.
- [96] G. Imparato, E. D. Paolo, A. Braca, and R. Lamanna, “Nuclear magnetic resonance profiling of wine blends”, *Journal of agricultural and food chemistry*, vol. 59, no. 9, pp. 4429–4434, 2011.
- [97] E. Ferrari, G. Foca, M. Vignali, L. Tassi, and A. Ulrici, “Adulteration of the anthocyanin content of red wines: Perspectives for authentication by fourier transform-near infrared and 1h nmr spectroscopies”, *Analytica Chimica Acta*, vol. 701, no. 2, pp. 139–151, 2011.
- [98] D. Cozzolino, W. Cynkar, N. Shah, and P. Smith, “Can spectroscopy geographically classify sauvignon blanc wines from Australia and New Zealand?”, *Food Chemistry*, vol. 126, no. 2, pp. 673–678, 2011.
- [99] R. Riovanto, W. U. Cynkar, P. Berzaghi, and D. Cozzolino, “Discrimination between Shiraz wines from different Australian regions: The role of spectroscopy and chemometrics”, *Journal of agricultural and food chemistry*, vol. 59, no. 18, pp. 10 356–10 360, 2011.
- [100] P. Mazzei, N. Francesca, G. Moschetti, and A. Piccolo, “Nmr spectroscopy evaluation of direct relationship between soils and molecular composition of red wines from Aglianico grapes”, *Analytica Chimica Acta*, vol. 673, no. 2, pp. 167–172, 2010.
- [101] L. Liu, D. Cozzolino, W. Cynkar, R. Damberg, L. Janik, *et al.*, “Preliminary study on the application of visible–near infrared spectroscopy and chemometrics to classify Riesling wines from different countries”, *Food Chemistry*, vol. 106, no. 2, pp. 781–786, 2008.

- [102] K. Ali, F. Maltese, R. Toepfer, Y. H. Choi, and R. Verpoorte, "Metabolic characterization of palatinate german white wines according to sensory attributes, varieties, and vintages using nmr spectroscopy and multivariate data analyses", *Journal of Biomolecular NMR*, vol. 49, no. 3-4, pp. 255–266, 2011.
- [103] E. Ioannou-Papayianni, R. I. Kokkinofa, and C. R. Theocharis, "Authenticity of cypriot sweet wine commandaria using ft-ir and chemometrics", *Journal of food science*, vol. 76, no. 3, 2011.
- [104] Y.-l. Zhang, J.-b. Chen, Y. Lei, Q. Zhou, S.-q. Sun, *et al.*, "Discrimination of different red wine by fourier-transform infrared and two-dimensional infrared correlation spectroscopy", *Journal of Molecular Structure*, vol. 974, no. 1, pp. 144–150, 2010.
- [105] H. Smyth, D. Cozzolino, W. Cynkar, R. Damberg, M. Sefton, *et al.*, "Near infrared spectroscopy as a rapid tool to measure volatile aroma compounds in riesling wine: Possibilities and limits", *Analytical and bioanalytical chemistry*, vol. 390, no. 7, pp. 1911–1916, 2008.
- [106] M. Martelo-Vidal, F. Domínguez-Agís, and M. Vázquez, "Ultraviolet/visible/near-infrared spectral analysis and chemometric tools for the discrimination of wines between subzones inside a controlled designation of origin: A case study of rías baixas", *Australian Journal of Grape and Wine Research*, vol. 19, no. 1, pp. 62–67, 2013.
- [107] S. M. Azcarate, M. Á. Cantarelli, R. G. Pellerano, E. J. Marchevsky, and J. M. Camiña, "Classification of argentinean sauvignon blanc wines by uv spectroscopy and chemometric methods", *Journal of food science*, vol. 78, no. 3, 2013.
- [108] M. Koda, K. Furihata, F. Wei, T. Miyakawa, and M. Tanokura, "Nmr-based metabolic profiling of rice wines by f 2-selective total correlation spectra", *Journal of agricultural and food chemistry*, vol. 60, no. 19, pp. 4818–4825, 2012.
- [109] G. Papotti, D. Bertelli, R. Graziosi, M. Silvestri, L. Bertacchini, *et al.*, "Application of one-and two-dimensional nmr spectroscopy for the characterization of protected designation of origin lambrusco wines of modena", *Journal of agricultural and food chemistry*, vol. 61, no. 8, pp. 1741–1746, 2012.
- [110] F. Shen, D. Yang, Y. Ying, B. Li, Y. Zheng, *et al.*, "Discrimination between shaoxing wines and other chinese rice wines by near-infrared spectroscopy and chemometrics", *Food and bioprocess technology*, vol. 5, no. 2, pp. 786–795, 2012.
- [111] E. Lopez-Rituerto, F. Savorani, A. Avenoza, J. H. Busto, J. M. Peregrina, *et al.*, "Investigations of la rioja terroir for wine production using 1h nmr metabolomics", *Journal of agricultural and food chemistry*, vol. 60, no. 13, pp. 3452–3461, 2012.

- [112] R. Godelmann, F. Fang, E. Humpfer, B. Schutz, M. Bansbach, *et al.*, "Targeted and nontargeted wine analysis by 1h nmr spectroscopy combined with multivariate statistical analysis. differentiation of important parameters: Grape variety, geographical origin, year of vintage", *Journal of agricultural and food chemistry*, vol. 61, no. 23, pp. 5610–5619, 2013.
- [113] E. Villagra, L. S. Santos, B. G. Vaz, M. N. Eberlin, and V. F. Laurie, "Varietal discrimination of chilean wines by direct injection mass spectrometry analysis combined with multivariate statistics", *Food chemistry*, vol. 131, no. 2, pp. 692–697, 2012.
- [114] L. Vaclavik, O. Lacina, J. Hajslova, and J. Zweigenbaum, "The use of high performance liquid chromatography–quadrupole time-of-flight mass spectrometry coupled to advanced data mining and chemometric tools for discrimination and classification of red wines according to their variety", *Analytica chimica acta*, vol. 685, no. 1, pp. 45–51, 2011.
- [115] J. Zhang, L. Li, N. Gao, D. Wang, Q. Gao, *et al.*, "Feature extraction and selection from volatile compounds for analytical classification of chinese red wines from different varieties", *Analytica chimica acta*, vol. 662, no. 2, pp. 137–142, 2010.
- [116] L. Louw, K. Roux, A. Tredoux, O. Tomic, T. Naes, *et al.*, "Characterization of selected south african young cultivar wines using ftmir spectroscopy, gas chromatography, and multivariate data analysis", *Journal of agricultural and food chemistry*, vol. 57, no. 7, pp. 2623–2632, 2009.
- [117] C. J. Bevin, R. G. Damberg, A. J. Fergusson, and D. Cozzolino, "Varietal discrimination of australian wines by means of mid-infrared spectroscopy and multivariate analysis", *Analytica Chimica Acta*, vol. 621, no. 1, pp. 19–23, 2008.
- [118] A. C. Pereira, M. S. Reis, P. M. Saraiva, and J. C. Marques, "Madeira wine ageing prediction based on different analytical techniques: Uv–vis, gc-ms, hplc-dad", *Chemometrics and Intelligent Laboratory Systems*, vol. 105, no. 1, pp. 43–55, 2011.
- [119] H.-S. Son, G.-S. Hwang, K. M. Kim, E.-Y. Kim, F. van den Berg, *et al.*, "1h nmr-based metabolomic approach for understanding the fermentation behaviors of wine yeast strains", *Analytical Chemistry*, vol. 81, no. 3, pp. 1137–1145, 2008.
- [120] A. L. Fudge, K. L. Wilkinson, R. Ristic, and D. Cozzolino, "Synchronous two-dimensional mir correlation spectroscopy (2d-cos) as a novel method for screening smoke tainted wine", *Food chemistry*, vol. 139, no. 1, pp. 115–119, 2013.
- [121] M. Urbano, M. D. L. De Castro, P. M. Pérez, J. García-Olmo, and M. A. Gomez-Nieto, "Ultraviolet–visible spectroscopy and pattern recognition methods for differentiation and classification of wines", *Food Chemistry*, vol. 97, no. 1, pp. 166–175, 2006.
- [122] S. Esslinger, J. Riedl, and C. Fauhl-Hassek, "Potential and limitations of non-targeted fingerprinting for authentication of food in official control", *Food research international*, vol. 60, pp. 189–204, 2014.

- [123] M. D. Mandal and S. Mandal, "Honey: Its medicinal property and antibacterial activity", *Asian Pacific Journal of Tropical Biomedicine*, vol. 1, no. 2, pp. 154–160, 2011.
- [124] S. Soares, J. S. Amaral, M. B. P. Oliveira, and I. Mafra, "A comprehensive review on the main honey authentication issues: Production and origin", *Comprehensive Reviews in Food Science and Food Safety*, vol. 16, no. 5, pp. 1072–1100, 2017.
- [125] M. Ohmenhaeuser, Y. B. Monakhova, T. Kuballa, and D. W. Lachenmeier, "Qualitative and quantitative control of honeys using nmr spectroscopy and chemometrics", *ISRN Analytical Chemistry*, vol. 2013, 2013.
- [126] E. Schievano, M. Stocchero, E. Morelato, C. Facchin, and S. Mammi, "An nmr-based metabolomic approach to identify the botanical origin of honey", *Metabolomics*, vol. 8, no. 4, pp. 679–690, 2012.
- [127] M. Kortessniemi *et al.*, "Nmr metabolomics of foods—investigating the influence of origin on sea buckthorn berries, brassica oilseeds and honey", 2016.
- [128] S. Gok, M. Severcan, E. Goormaghtigh, I. Kandemir, and F. Severcan, "Differentiation of anatolian honey samples from different botanical origins by atr-ftir spectroscopy using multivariate analysis", *Food chemistry*, vol. 170, pp. 234–240, 2015.
- [129] T. Woodcock, G. Downey, J. D. Kelly, and C. O'Donnell, "Geographical classification of honey samples by near-infrared spectroscopy: A feasibility study", *Journal of agricultural and food chemistry*, vol. 55, no. 22, pp. 9128–9134, 2007.
- [130] E. F. Boffo, L. A. Tavares, A. C. Tobias, M. M. Ferreira, and A. G. Ferreira, "Identification of components of brazilian honey by 1 h nmr and classification of its botanical origin by chemometric methods", *LWT-Food Science and Technology*, vol. 49, no. 1, pp. 55–63, 2012.
- [131] A. Cavazza, C. Corradini, M. Musci, and P. Salvadeo, "High-performance liquid chromatographic phenolic compound fingerprint for authenticity assessment of honey", *Journal of the Science of Food and Agriculture*, vol. 93, no. 5, pp. 1169–1175, 2013.
- [132] L. Chen, J. Wang, Z. Ye, J. Zhao, X. Xue, *et al.*, "Classification of chinese honeys according to their floral origin by near infrared spectroscopy", *Food chemistry*, vol. 135, no. 2, pp. 338–342, 2012.
- [133] E. Schievano, E. Peggion, and S. Mammi, "1h nuclear magnetic resonance spectra of chloroform extracts of honey for chemometric determination of its botanical origin", *Journal of agricultural and food chemistry*, vol. 58, no. 1, pp. 57–65, 2009.
- [134] E. Etzold and B. Lichtenberg-Kraag, "Determination of the botanical origin of honey by fourier-transformed infrared spectroscopy: An approach for routine analysis", *European Food Research and Technology*, vol. 227, no. 2, pp. 579–586, 2008.
- [135] M. Lolli, D. Bertelli, M. Plessi, A. G. Sabatini, and C. Restani, "Classification of italian honeys by 2d hr-nmr", *Journal of agricultural and food chemistry*, vol. 56, no. 4, pp. 1298–1304, 2008.

- [136] X. Y. Liang, X. Y. Li, and W. J. Wu, "Classification of floral origins of honey by nir and chemometrics", in *Advanced Materials Research*, Trans Tech Publ, vol. 605, 2013, pp. 905–909.
- [137] A. Agila and S. Barringer, "Application of selected ion flow tube mass spectrometry coupled with chemometrics to study the effect of location and botanical origin on volatile profile of unifloral american honeys", *Journal of food science*, vol. 77, no. 10, 2012.
- [138] L. Castro-Vázquez, M. Díaz-Maroto, M. González-Viñas, and M. Pérez-Coello, "Differentiation of monofloral citrus, rosemary, eucalyptus, lavender, thyme and heather honeys based on volatile composition and sensory descriptive analysis", *Food Chemistry*, vol. 112, no. 4, pp. 1022–1030, 2009.
- [139] G. Beretta, E. Caneva, L. Regazzoni, N. G. Bakhtyari, and R. M. Facino, "A solid-phase extraction procedure coupled to 1 h nmr, with chemometric analysis, to seek reliable markers of the botanical origin of honey", *Analytica chimica acta*, vol. 620, no. 1, pp. 176–182, 2008.
- [140] J. A. Donarski, S. A. Jones, M. Harrison, M. Driffield, and A. J. Charlton, "Identification of botanical biomarkers found in corsican honey", *Food chemistry*, vol. 118, no. 4, pp. 987–994, 2010.
- [141] K. D. Labsvards, V. Rudovica, R. Kluga, J. Rusko, L. Busa, M. Bertins, I. Eglite, J. Naumenko, M. Salajeva, and A. Viksna, "Determination of floral origin markers of latvian honey by using irms, uhplc-hrms, and 1h-nmr", *Foods*, vol. 11, no. 1, p. 42, 2022.
- [142] B. Lozano-Torres, M. C. Martínez-Bisbal, J. Soto, M. J. Borrás, R. Martínez-Mañez, and I. Escriche, "Monofloral honey authentication by voltammetric electronic tongue: A comparison with 1h nmr spectroscopy", *Food Chemistry*, vol. 383, p. 132 460, 2022.
- [143] K. A. Aliferis, P. A. Tarantilis, P. C. Harizanis, and E. Alissandrakis, "Botanical discrimination and classification of honey samples applying gas chromatography/mass spectrometry fingerprinting of headspace volatile compounds", *Food Chemistry*, vol. 121, no. 3, pp. 856–862, 2010.
- [144] J. A. F. Pierna, O. Abbas, P. Dardenne, and V. Baeten, "Discrimination of corsican honey by ft-raman spectroscopy and chemometrics/discrimination du miel de corse par spectroscopie ft-raman et chimiométrie", *Biotechnologie, Agronomie, Société et Environnement*, vol. 15, no. 1, p. 75, 2011.
- [145] P. Oliveri, V. Di Egidio, T. Woodcock, and G. Downey, "Application of class-modelling techniques to near infrared data for food authentication purposes", *Food Chemistry*, vol. 125, no. 4, pp. 1450–1456, 2011.
- [146] S. Hennessy, G. Downey, and C. P. O'Donnell, "Attempted confirmation of the provenance of corsican pdo honey using ft-ir spectroscopy and multivariate data analysis", *Journal of agricultural and food chemistry*, vol. 58, no. 17, pp. 9401–9406, 2010.

- [147] T. Woodcock, G. Downey, and C. P. O'Donnell, "Near infrared spectral fingerprinting for confirmation of claimed pdo provenance of honey", *Food Chemistry*, vol. 114, no. 2, pp. 742–746, 2009.
- [148] R. Consonni and L. R. Cagliani, "Geographical characterization of polyfloral and acacia honeys by nuclear magnetic resonance and chemometrics", *Journal of Agricultural and Food Chemistry*, vol. 56, no. 16, pp. 6873–6880, 2008.
- [149] J. A. Donarski, S. A. Jones, and A. J. Charlton, "Application of cryoprobe 1h nuclear magnetic resonance spectroscopy and multivariate analysis for the verification of corsican honey", *Journal of agricultural and food chemistry*, vol. 56, no. 14, pp. 5451–5456, 2008.
- [150] S. Hennessy, G. Downey, and C. O'Donnell, "Multivariate analysis of attenuated total reflection–fourier transform infrared spectroscopic data to confirm the origin of honeys", *Applied spectroscopy*, vol. 62, no. 10, pp. 1115–1123, 2008.
- [151] I Stanimirova, B Üstün, T Cajka, K Riddelova, J Hajslova, *et al.*, "Tracing the geographical origin of honeys based on volatile compounds profiles assessment using pattern recognition techniques", *Food Chemistry*, vol. 118, no. 1, pp. 171–176, 2010.
- [152] S. Li, Y. Shan, X. Zhu, X. Zhang, and G. Ling, "Detection of honey adulteration by high fructose corn syrup and maltose syrup using raman spectroscopy", *Journal of food composition and analysis*, vol. 28, no. 1, pp. 69–74, 2012.
- [153] D. Bertelli, M. Lolli, G. Papotti, L. Bortolotti, G. Serra, *et al.*, "Detection of honey adulteration by sugar syrups using one-dimensional and two-dimensional high-resolution nuclear magnetic resonance", *Journal of agricultural and food chemistry*, vol. 58, no. 15, pp. 8495–8501, 2010.
- [154] X. Zhu, S. Li, Y. Shan, Z. Zhang, G. Li, *et al.*, "Detection of adulterants such as sweeteners materials in honey using near-infrared spectroscopy and chemometrics", *Journal of Food Engineering*, vol. 101, no. 1, pp. 92–97, 2010.
- [155] T. Gallardo-Velázquez, G. Osorio-Revilla, M. Zuñiga-de Loa, and Y. Rivera-Espinoza, "Application of ftir-hatr spectroscopy and multivariate analysis to the quantification of adulterants in mexican honeys", *Food Research International*, vol. 42, no. 3, pp. 313–318, 2009.
- [156] N. Subari, J. Mohamad Saleh, A. Y. Md Shakaff, and A. Zakaria, "A hybrid sensing approach for pure and adulterated honey classification", *Sensors*, vol. 12, no. 10, pp. 14 022–14 040, 2012.
- [157] L. D. Falcão, G. de Revel, M. C. Perello, A. Moutsiou, M. C. Zanús, *et al.*, "A survey of seasonal temperatures and vineyard altitude influences on 2-methoxy-3-isobutylpyrazine, c13-norisoprenoids, and the sensory profile of brazilian cabernet sauvignon wines", *Journal of Agricultural and Food Chemistry*, vol. 55, no. 9, pp. 3605–3612, 2007.
- [158] C Roullier-Gall, V David, D Hemmler, P Schmitt-Kopplin, and H Alexandre, "Exploring yeast interactions through metabolic profiling", *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

- [159] A. Gilinsky, S. K. Newton, and R. F. Vega, "Sustainability in the global wine industry: Concepts and cases", *Agriculture and agricultural science procedia*, vol. 8, no. 1, pp. 37–49, 2016.
- [160] D. A. Magdas, B. I. Cozar, I. Feher, F. Guyon, A. Dehelean, *et al.*, "Testing the limits of ft-raman spectroscopy for wine authentication: Cultivar, geographical origin, vintage and terroir effect influence", *Scientific Reports*, vol. 9, no. 1, pp. 1–8, 2019.
- [161] T. Karbowiak, K. Crouvisier-Urien, A. Lagorce, J. Ballester, A. Geoffroy, *et al.*, "Wine aging: A bottleneck story", *npj Science of Food*, vol. 3, no. 1, pp. 1–7, 2019.
- [162] C. Roullier-Gall, S. S. Heinzmann, J.-P. Garcia, P. Schmitt-Kopplin, and R. D. Gougeon, "Chemical messages from an ancient buried bottle: Metabolomics for wine archeochemistry", *npj Science of Food*, vol. 1, no. 1, pp. 1–7, 2017.
- [163] M. Gil, C. Reynes, G. Cazals, C. Enjalbal, R. Sabatier, *et al.*, "Discrimination of rosé wines using shotgun metabolomics with a genetic algorithm and ms ion intensity ratios", *Scientific reports*, vol. 10, no. 1, pp. 1–7, 2020.
- [164] T. W.-M. Fan and A. N. Lane, "Structure-based profiling of metabolites and isotopomers by nmr", *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 2, no. 52, pp. 69–117, 2008.
- [165] M. M.M. R. Sara Cardoso Débora Cabral, "Nmrfinder: A novel method for 1h-nmr metabolite identification), e 'manuscript submitted for publication", *Manuscript submitted in Metabolomics*, 2020.
- [166] J. Zhu, B. Hu, J. Lu, and S. Xu, "Analysis of metabolites in cabernet sauvignon and shiraz dry red wines from shanxi by 1h nmr spectroscopy combined with pattern recognition analysis", *Open Chemistry*, vol. 16, no. 1, pp. 446–452, 2018.
- [167] A. Cuadros-Inostroza, S. Ruíz-Lara, E. González, A. Eckardt, L. Willmitzer, *et al.*, "Gc–ms metabolic profiling of cabernet sauvignon and merlot cultivars during grapevine berry development and network analysis reveals a stage-and cultivar-dependent connectivity of primary metabolites", *Metabolomics*, vol. 12, no. 2, p. 39, 2016.
- [168] B. Zoecklein, K. C. Fugelsang, B. H. Gump, and F. S. Nury, *Wine analysis and production*. Springer Science & Business Media, 2013.
- [169] E. Soufleros, E. Bouloumpasi, C. Tsarchopoulos, and C. Biliaderis, "Primary amino acid profiles of greek white wines and their use in classification according to variety, origin and vintage", *Food Chemistry*, vol. 80, no. 2, pp. 261–273, 2003.
- [170] M Ibern-Gómez, C Andrés-Lacueva, R. Lamuela-Raventós, and A. Waterhouse, "Rapid hplc analysis of phenolic compounds in red wines", *American Journal of Enology and Viticulture*, vol. 53, no. 3, pp. 218–221, 2002.

- [171] R. Sladkovsky, P. Solich, and M. Urbánek, "High-performance liquid chromatography determination of phenolic components in wine using off-line isotachophoretic pretreatment", *Journal of Chromatography A*, vol. 1040, no. 2, pp. 179–184, 2004.
- [172] M. S. P. Nikfardjam, L. Márk, P. Avar, M. Figler, and R. Ohmacht, "Polyphenols, anthocyanins, and trans-resveratrol in red wines from the hungarian villány region", *Food Chemistry*, vol. 98, no. 3, pp. 453–462, 2006.
- [173] L. Minuti and R. Pellegrino, "Determination of phenolic compounds in wines by novel matrix solid-phase dispersion extraction and gas chromatography/mass spectrometry", *Journal of Chromatography A*, vol. 1185, no. 1, pp. 23–30, 2008.
- [174] V. M. Burin, L. L. F. Costa, J. P. Rosier, and M. T. Bordignon-Luiz, "Cabernet sauvignon wines from two different clones, characterization and evolution during bottle ageing", *LWT-Food Science and Technology*, vol. 44, no. 9, pp. 1931–1938, 2011.
- [175] V. M. Burin, S. G. Arcari, Bordignon-Luiz, M. T, and L. L. F. Costa, "Determination of some phenolic compounds in red wine by rp-hplc: Method development and validation", *Journal of chromatographic science*, vol. 49, no. 8, pp. 647–651, 2011.
- [176] F. d. S. Dias, J. M. David, and J. P. David, "Determination of phenolic acids and quercetin in brazilian red wines from vale do são francisco region using liquid-liquid ultrasound-assisted extraction and hplc-dad-ms", *Journal of the Brazilian Chemical Society*, vol. 27, no. 6, pp. 1055–1059, 2016.
- [177] M. García-Falcón, C Pérez-Lamela, E Martínez-Carballo, and J Simal-Gándara, "Determination of phenolic compounds in wines: Influence of bottle storage of young red wines on their evolution", *Food Chemistry*, vol. 105, no. 1, pp. 248–259, 2007.
- [178] C. V. da Silva Padilha, G. A. Miskinis, M. E.A. O. de Souza, G. E. Pereira, D. de Oliveira, *et al.*, "Rapid determination of flavonoids and phenolic acids in grape juices and wines by rp-hplc/dad: Method validation and characterization of commercial products of the new brazilian varieties of grape", *Food chemistry*, vol. 228, pp. 106–115, 2017.
- [179] M. A. Krstonošić, J. C. Hogervorst, M. Mikulić, and L. Gojković-Bukarica, "Development of hplc method for determination of phenolic compounds on a core shell column by direct injection of wine samples", *Acta Chromatographica*, vol. 32, no. 2, pp. 134–138, 2020.
- [180] G. Antalick, K. S, J. W. Blackman, C. Meeks, A. Deloire, *et al.*, "Influence of grape composition on red wine ester profile: Comparison between cabernet sauvignon and shiraz cultivars from australian warm climate", *Journal of agricultural and food chemistry*, vol. 63, no. 18, pp. 4664–4672, 2015.

-
- [181] X.-K. Zhang, Y.-B. Lan, B.-Q. Zhu, X.-F. Xiang, C.-Q. Duan, *et al.*, “Changes in monosaccharides, organic acids and amino acids during cabernet sauvignon wine ageing based on a simultaneous analysis using gas chromatography–mass spectrometry”, *Journal of the Science of Food and Agriculture*, vol. 98, no. 1, pp. 104–112, 2018.
- [182] K. A. Veselkov, J. S. McKenzie, and J. K. Nicholson, “Multivariate data analysis methods for nmr-based metabolic phenotyping in pharmaceutical and clinical research”, *eMagRes*, pp. 323–334, 2007.
- [183] D. A. Freedman, *Statistical models: Theory and practice*. cambridge university press, 2009.
- [184] P. Arapitsas, G. Speri, A. Angeli, D. Perenzoni, and F. Mattivi, “The influence of storage on the “chemical age” of red wines”, *Metabolomics*, vol. 10, no. 5, pp. 816–832, 2014.
- [185] H. Hopfer, S. E. Ebeler, and H. Heymann, “The combined effects of storage temperature and packaging type on the sensory and chemical properties of chardonnay”, *Journal of agricultural and food chemistry*, vol. 60, no. 43, pp. 10 743–10 754, 2012.
- [186] R. Ghidossi, C. Poupot, C. Thibon, A. Pons, P. Darriet, *et al.*, “The influence of packaging on wine conservation”, *Food Control*, vol. 23, no. 2, pp. 302–311, 2012.
- [187] A. G. Marco and C. A. Azpilicueta, “Amine concentrations in wine stored in bottles at different temperatures”, *Food chemistry*, vol. 99, no. 4, pp. 680–685, 2006.
- [188] N Loscos, P Hernández-Orte, J Cacho, and V Ferreira, “Evolution of the aroma composition of wines supplemented with grape flavour precursors from different varieties during accelerated wine ageing”, *Food chemistry*, vol. 120, no. 1, pp. 205–216, 2010.
- [189] J. Wirth, C. Morel-Salmi, J. M. Souquet, J. Dieval, O Aagaard, *et al.*, “The impact of oxygen exposure before and after bottling on the polyphenolic composition of red wines”, *Food Chemistry*, vol. 123, no. 1, pp. 107–116, 2010.
- [190] T. G. Cerdán and C. Ancín-Azpilicueta, “Effect of oak barrel type on the volatile composition of wine: Storage time optimization”, *LWT-Food Science and Technology*, vol. 39, no. 3, pp. 199–205, 2006.
- [191] G. Han, M. Ugliano, B. Currie, S. Vidal, J.-B. Diéval, *et al.*, “Influence of closure, phenolic levels and microoxygenation on cabernet sauvignon wine composition after 5 years’ bottle storage”, *Journal of the Science of Food and Agriculture*, vol. 95, no. 1, pp. 36–43, 2015.
- [192] C. Dallas and O. Laureano, “Effects of ph, sulphur dioxide, alcohol content, temperature and storage time on colour composition of a young portuguese red table wine”, *Journal of the Science of Food and Agriculture*, vol. 65, no. 4, pp. 477–485, 1994.
- [193] L. Bisson, *Yeast and biochemistry of ethanol formation*, 1996.

- [194] I. H. Gutiérrez, E. S.-P. Lorenzo, and A. V. Espinosa, "Phenolic composition and magnitude of copigmentation in young and shortly aged red wines made from the cultivars, cabernet sauvignon, cencibel, and syrah", *Food Chemistry*, vol. 92, no. 2, pp. 269–283, 2005.
- [195] S. Lafon-Lafourcade, "Wine and brandy", *Food and feed production with microorganisms*, 1983.
- [196] P. Lehtonen, "Determination of amines and amino acids in wine—a review", *American Journal of Enology and Viticulture*, vol. 47, no. 2, pp. 127–133, 1996.
- [197] S. Rochfort, V. Ezernieks, S. E. Bastian, and M. O. Downey, "Sensory attributes of wine influenced by variety and berry shading discriminated by nmr metabolomics", *Food Chemistry*, vol. 121, no. 4, pp. 1296–1304, 2010.
- [198] G. V. Jones, "Climate and terroir: Impacts of climate variability and change on wine", *Fine wine and terroir-the geoscience perspective*, no. 9, pp. 1–14, 2006.
- [199] G. V. Jones and F. Alves, "Impact of climate change on wine production: A global overview and regional assessment in the douro valley of portugal", *International Journal of Global Warming*, vol. 4, no. 3-4, pp. 383–406, 2012.
- [200] G. E. Pereira, J.-P. Gaudillere, P. Pieri, G. Hilbert, M. Maucourt, *et al.*, "Microclimate influence on mineral and metabolic profiles of grape berries", *Journal of Agricultural and Food Chemistry*, vol. 54, no. 18, pp. 6765–6775, 2006.
- [201] C. Van Leeuwen, P. Friant, X. Chone, O. Tregoat, S. Koundouras, *et al.*, "Influence of climate, soil, and cultivar on terroir", *American Journal of Enology and Viticulture*, vol. 55, no. 3, pp. 207–217, 2004.
- [202] M. A. F. Conceição and J. Tonietto, "Climatic potential for wine grape production in the tropical north region of minas gerais state, brazil", *Revista Brasileira de Fruticultura*, vol. 27, no. 3, pp. 404–407, 2005.
- [203] L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, *et al.*, "Proposed minimum reporting standards for chemical analysis", *Metabolomics*, vol. 3, no. 3, pp. 211–221, 2007.
- [204] K. R. Kim, J. H. Kim, E.-j. Cheong, and C.-m. Jeong, "Gas chromatographic amino acid profiling of wine samples for pattern recognition", *Journal of Chromatography A*, vol. 722, no. 1-2, pp. 303–309, 1996.
- [205] M. Maraschin, A. Somensi-Zeggio, S. K. Oliveira, S. Kuhnen, M. M. Tomazzoli, *et al.*, "Metabolic profiling and classification of propolis samples from southern brazil: An nmr-based platform coupled with machine learning", *Journal of natural products*, vol. 79, no. 1, pp. 13–23, 2016.
- [206] K. Anderson and V. Pinilla, *Wine globalization: A new comparative history*. Cambridge University Press, 2018.
- [207] D. Ferreira, N. Moreira, I. Vasconcelos, J. Silva, and P. Pinho, "Varietal thiols in white wines from portuguese grape varieties", 2011.

-
- [208] A. Teixeira, V. Martins, H. Noronha, J. Eiras-Dias, and H. Gerós, "The first insight into the metabolite profiling of grapes from three vitis vinifera l. cultivars of two controlled appellation (doc) regions", *International Journal of Molecular Sciences*, vol. 15, no. 3, pp. 4237–4254, 2014.
- [209] M. Baleiras-Couto and J. Eiras-Dias, "Detection and identification of grape varieties in must and wine using nuclear and chloroplast microsatellite markers", *Analytica Chimica Acta*, vol. 563, no. 1-2, pp. 283–291, 2006.
- [210] S. Roussel, V. Bellon-Maurel, J.-M. Roger, and P. Grenier, "Authenticating white grape must variety with classification models based on aroma sensors, ft-ir and uv spectrometry", *Journal of Food Engineering*, vol. 60, no. 4, pp. 407–419, 2003.
- [211] D. Cozzolino, W. Cynkar, N. Shah, and P. Smith, "Varietal differentiation of grape juice based on the analysis of near-and mid-infrared spectral data", *Food Analytical Methods*, vol. 5, no. 3, pp. 381–387, 2012.
- [212] C. Murru, C. Chimeno-Trinchet, M. E. Díaz-García, R. Badía-Laiño, and A. Fernández-González, "Artificial neural network and attenuated total reflectance-fourier transform infrared spectroscopy to identify the chemical variables related to ripeness and variety classification of grapes for protected. designation of origin wine production", *Computers and Electronics in Agriculture*, vol. 164, p. 104922, 2019.
- [213] K. Kumar, A. Giehl, R. Schweiggert, and C.-D. Patz, "Multidimensional scaling assisted fourier-transform infrared spectroscopic analysis of fruit wine samples: Introducing a novel analytical approach", *Analytical Methods*, vol. 11, no. 32, pp. 4106–4115, 2019.
- [214] N. F. da Silva, "Herdade do esporão, alentejo-controlo de maturação e vinificação de tintos e rosés", 2020.
- [215] P. A.d. C. Fernandes, "Comportamento agronómico e enológico das castas touriga nacional e syrah em seis regiões portuguesas", PhD thesis, 2009.
- [216] B. Smith, *Infrared spectral interpretation: A systematic approach*. CRC press, 2018.
- [217] Beauchamp, *SpectroscopyIntroduction, howpublished = "*
<https://www.cpp.edu/~psbeauchamp/pdf/IRinfoTables12-31-09.pdf>*", year = , note = "[online; accessed*
 .
- [218] R. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra", *Applied spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.
- [219] O. Anjos, A. J. Santos, L. M. Estevinho, and I. Caldeira, "Ftir-atr spectroscopy applied to quality control of grape-derived spirits", *Food Chemistry*, vol. 205, pp. 28–35, 2016.

- [220] X. F. Cadet, O. Lo-Thong, S. Bureau, R. Dehak, and M. Bessafi, "Use of machine learning and infrared spectra for rheological characterization and application to the apricot", *Scientific Reports*, vol. 9, no. 1, pp. 1–12, 2019.
- [221] S. Galli, <https://trainindata.medium.com/>, [Online; accessed January - 2019], 2019.
- [222] E. Crane, "Honey from honeybees and other insects", *Ethology Ecology & Evolution*, vol. 3, no. sup1, pp. 100–105, 1991.
- [223] M. S. Azevedo, S. K. T. Seraglio, G. Rocha, C. B. Balderas, M. Piovezan, L. V. Gonzaga, D. de Barcellos Falkenberg, R. Fett, M. A. L. de Oliveira, and A. C. O. Costa, "Free amino acid determination by gc-ms combined with a chemometric approach for geographical classification of bracatinga honeydew honey (mimosa scabrella bentham)", *Food Control*, vol. 78, pp. 383–392, 2017.
- [224] L. Pippinato, S. Blanc, T. Mancuso, and F. Brun, "A sustainable niche market: How does honey behave?", *Sustainability*, vol. 12, no. 24, p. 10 678, 2020.
- [225] R. Phipps, "International honey market report 2021", *American Bee Journal*, vol. 141, no. 6, pp. 1–6, 2021.
- [226] M. Shahbandeh, *Global production volume of natural honey 2000-2020, howpublished = "*
<https://www.statista.com/statistics/755215/natural-honey-production-volume-worldwide/>, year
= 2021, note = "[online; accessed january - 2022]
.
- [227] F. Marquele-Oliveira, D. B. Carrão, R. O. de Souza, N. U. Baptista, A. P. Nascimento, E. C. Torres, G. de Padua Moreno, A. F. M. Buszinski, F. G. Miguel, G. L. Cuba, *et al.*, *Fundamentals of brazilian honey analysis*, 2017.
- [228] M. Shahbandeh, *Instituto Brasileiro de Geografia e Estatística (IBGE)*, <https://www.ibge.gov.br/>, [Online; accessed January - 2022], 2021.
- [229] M. Spiteri, E. Jamin, F. Thomas, A. Rebours, M. Lees, K. M. Rogers, and D. N. Rutledge, "Fast and global authenticity screening of honey using 1h-nmr profiling", *Food Chemistry*, vol. 189, pp. 60–66, 2015.
- [230] G. Bergamo, S. K. T. Seraglio, L. V. Gonzaga, R. Fett, and A. C. O. Costa, "Physicochemical characteristics of bracatinga honeydew honey and blossom honey produced in the state of santa catarina: An approach to honey differentiation", *Food research international*, vol. 116, pp. 745–754, 2019.
- [231] B. Silva, L. V. Gonzaga, H. F. Maltez, K. B. Samochvalov, R. Fett, and A. C. O. Costa, "Elemental profiling by icp-ms as a tool for geographical discrimination: The case of bracatinga honeydew honey", *Journal of Food Composition and Analysis*, vol. 96, p. 103 727, 2021.

- [232] G. del Campo, J. Zuriarrain, A. Zuriarrain, and I. Berregi, "Quantitative determination of carboxylic acids, amino acids, carbohydrates, ethanol and hydroxymethylfurfural in honey by ¹h nmr", *Food Chemistry*, vol. 196, pp. 1031–1039, 2016.
- [233] S. A. Drivelos, G. P. Danezis, S. A. Haroutounian, and C. A. Georgiou, "Rare earth elements minimal harvest year variation facilitates robust geographical origin discrimination: The case of pdo "fava santorinis"", *Food Chemistry*, vol. 213, pp. 238–245, 2016.
- [234] T. Song, D. Das, Q. Hu, F. Yang, and J. Zhang, "Alternate wetting and drying irrigation and phosphorus rates affect grain yield and quality and heavy metal accumulation in rice", *Science of the Total Environment*, vol. 752, p. 141 862, 2021.
- [235] S. Kratz, J. Schick, and E. Schnug, "Trace elements in rock phosphates and p containing mineral and organo-mineral fertilizers sold in germany", *Science of the Total Environment*, vol. 542, pp. 1013–1019, 2016.
- [236] G. P. Danezis and C. A. Georgiou, "Elemental metabolomics: Food elemental assessment could reveal geographical origin", *Current Opinion in Food Science*, vol. 44, p. 100 812, 2022.
- [237] E. Schievano, M. Sbrizza, V. Zuccato, L. Piana, and M. Tessari, "Nmr carbohydrate profile in tracing acacia honey authenticity", *Food chemistry*, vol. 309, p. 125 788, 2020.
- [238] X. Zheng, Y. Zhao, H. Wu, J. Dong, and J. Feng, "Origin identification and quantitative analysis of honeys by nuclear magnetic resonance and chemometric techniques", *Food Analytical Methods*, vol. 9, no. 6, pp. 1470–1479, 2016.
- [239] N. Bilandžić, M. Sedak, M. okić, A. G. Bošković, T. Florijančić, I. Bošković, M. Kovačić, Z. Puškadija, and M. Hruškar, "Element content in ten croatian honey types from different geographical regions during three seasons", *Journal of Food Composition and Analysis*, vol. 84, p. 103 305, 2019.
- [240] C. He, Y. Liu, H. Liu, X. Zheng, G. Shen, and J. Feng, "Compositional identification and authentication of chinese honeys by ¹h nmr combined with multivariate analysis", *Food Research International*, vol. 130, p. 108 936, 2020.
- [241] X. Song, S. She, M. Xin, L. Chen, Y. Li, Y. Vander Heyden, K. M. Rogers, and L. Chen, "Detection of adulteration in chinese monofloral honey using ¹h nuclear magnetic resonance and chemometrics", *Journal of Food Composition and Analysis*, vol. 86, p. 103 390, 2020.
- [242] D. Gerginova, S. Simova, M. Popova, M. Stefova, J. P. Stanoeva, and V. Bankova, "Nmr profiling of north macedonian and bulgarian honeys for detection of botanical and geographical origin", *Molecules*, vol. 25, no. 20, p. 4687, 2020.
- [243] A. Noviyanto and W. H. Abdulla, "Honey botanical origin classification using hyperspectral imaging and machine learning", *Journal of Food Engineering*, vol. 265, p. 109 684, 2020.

Supplementary Figures

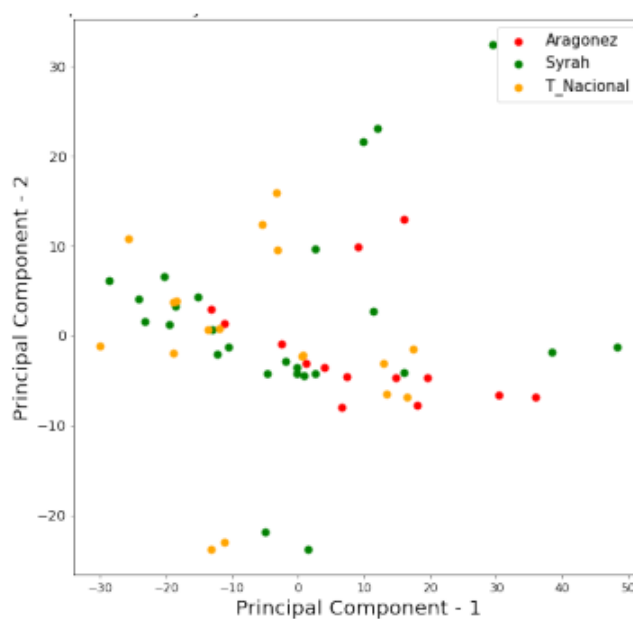
A.1 Supplementary Figures for Chapter 5

Figure 37: **Principal Components Analysis (PCA) 2D scores plot (PC1, PC2) from FT-IR spectra of red wine musts samples of varieties Aragonez, Syrah and Touriga Nacional (Touriga Nacional).** Explain variance of Principal Componets 1 and 2: PC1 = 69.8% PC2 = 21.6%.

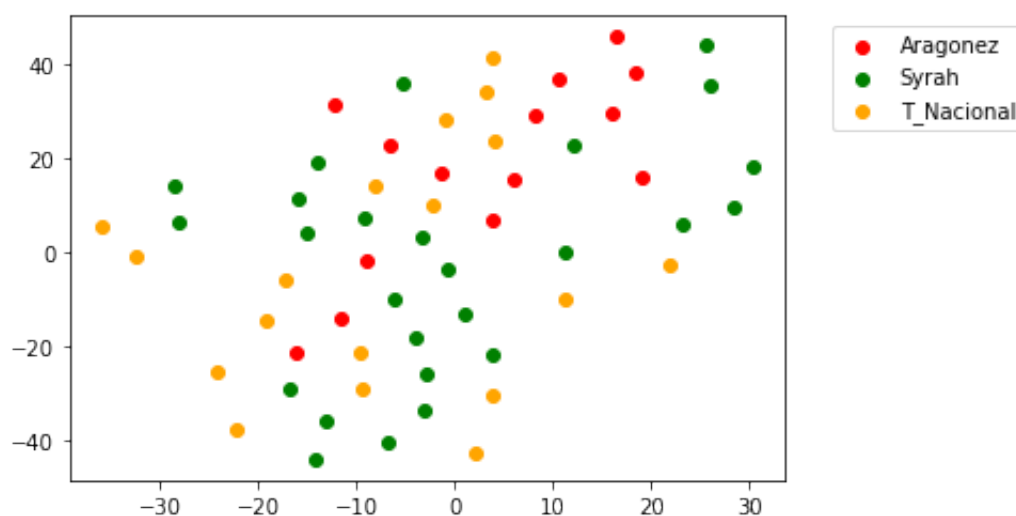


Figure 38: **t-distributed Stochastic Neighbor Embedding (t-SNE)** from FT-IR spectra of red wine musts samples of varieties Aragonez, Syrah and Touriga Nacional (Touriga Nacional)

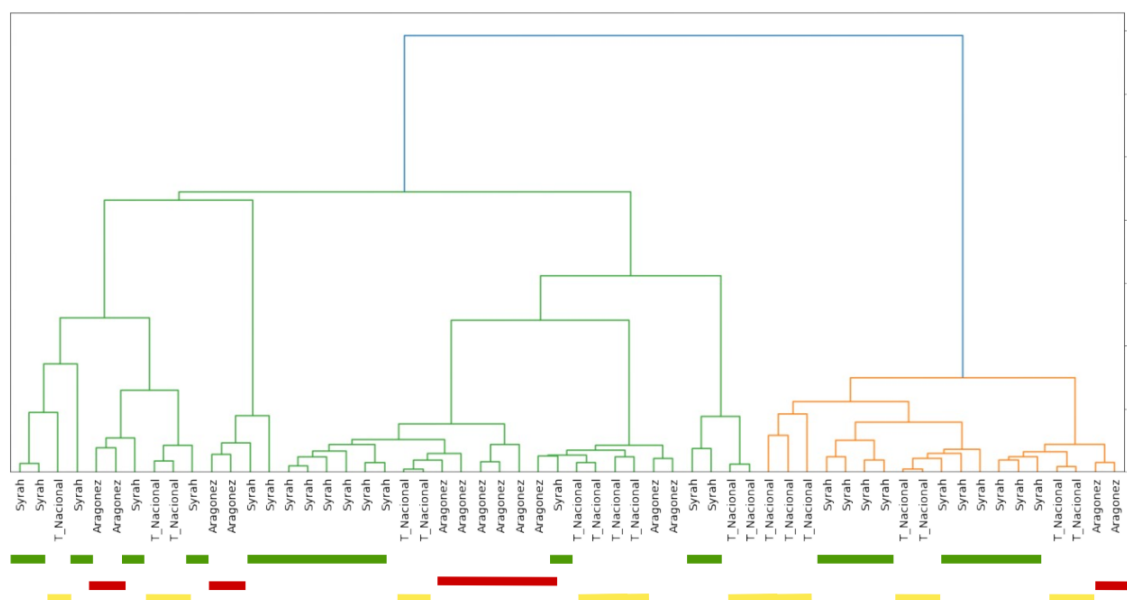


Figure 39: **Dendrogram plot of the hierarchical clustering analysis (HCA)**, with Euclidean distance between samples, from FT-IR spectra of red wine musts samples of varieties Aragonez, Syrah and Touriga Nacional (Touriga Nacional).

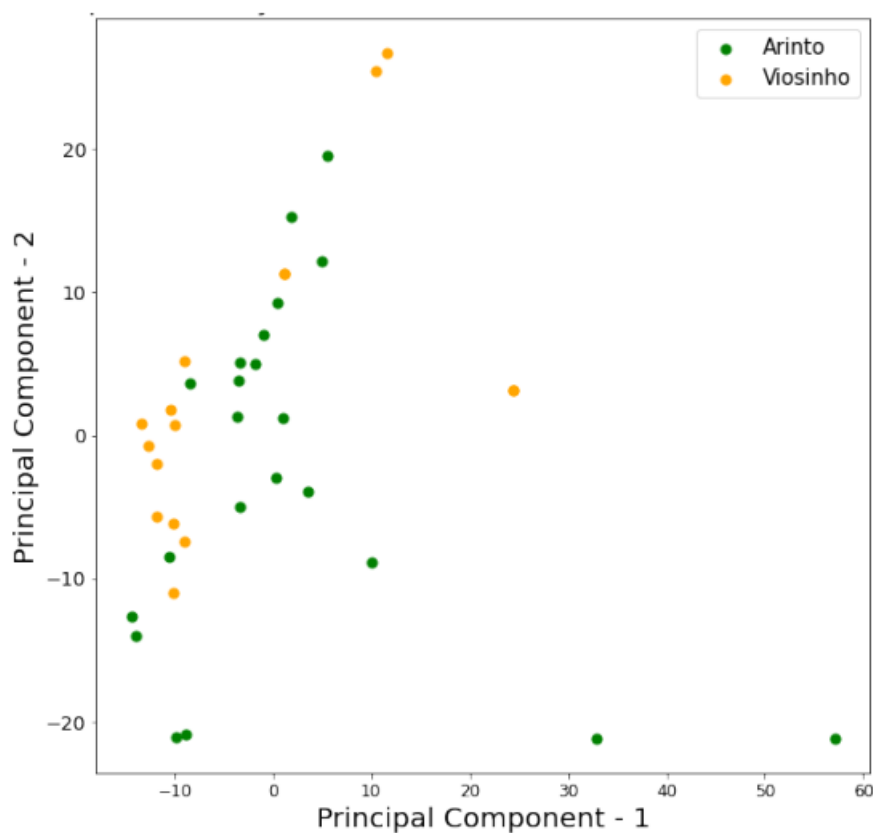


Figure 40: **Principal Components Analysis (PCA) 2D scores plot (PC1, PC2) from FT-IR spectra of white wine musts samples of varieties Arinto and Viosinho.** Explain variance of Principal Components 1 and 2: PC1 = 43.4% PC2 = 30.4%.

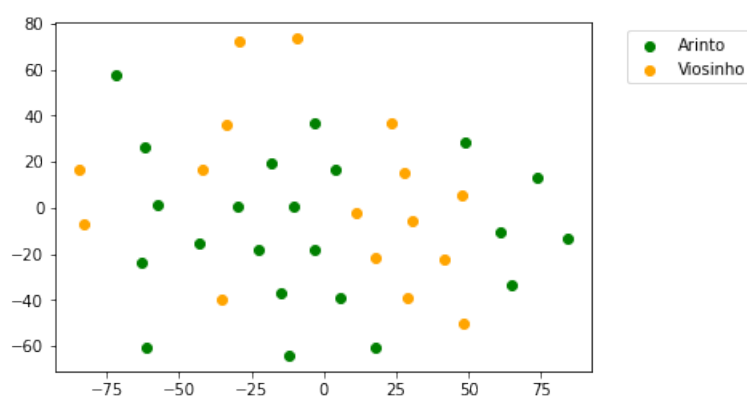


Figure 41: **t-distributed Stochastic Neighbor Embedding (t-SNE) from FT-IR spectra of white wine musts samples of varieties Arinto and Viosinho.**

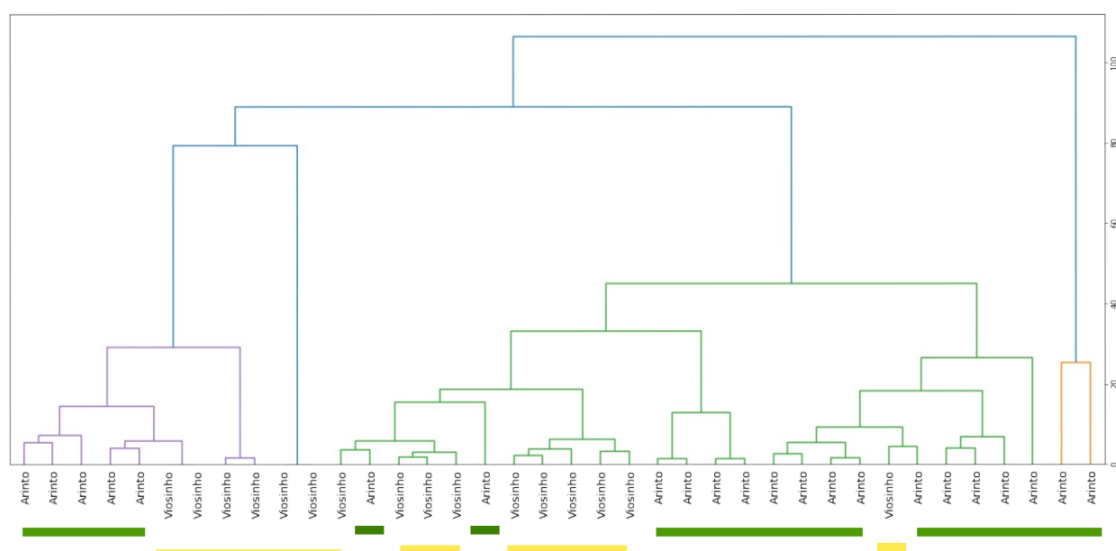


Figure 42: **Dendrogram plot of the hierarchical clustering analysis (HCA), with Euclidean distance between samples, from FT-IR spectra of white wine musts samples of varieties Arinto and Viosinho.**

A.2 Supplementary Figures for Chapter 6

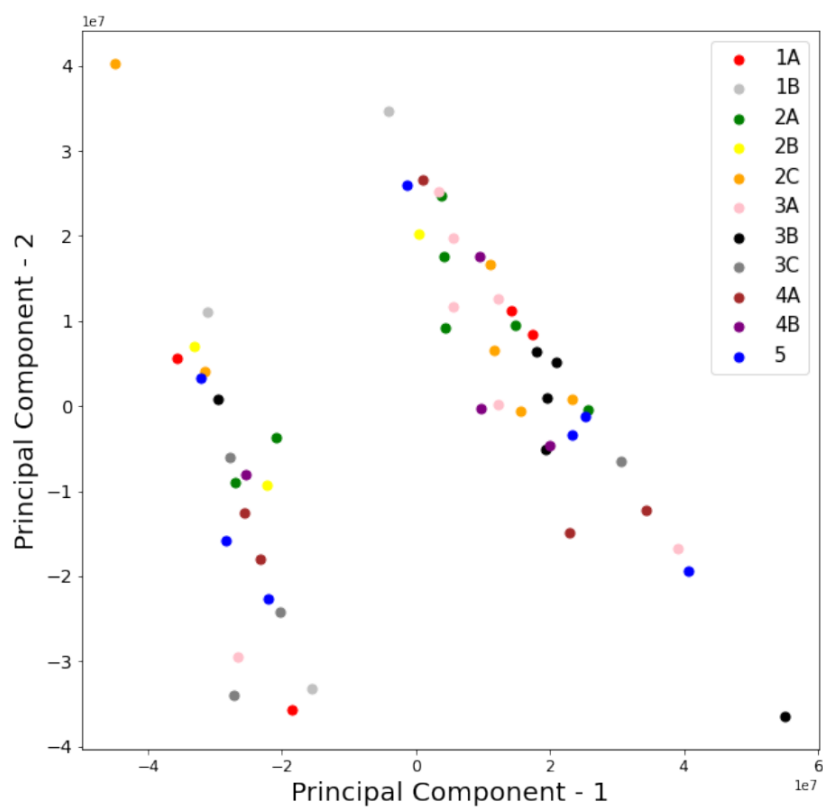


Figure 43: **Principal Components Analysis (PCA) 2D scores plot (PC1, PC2) from ¹H-NMR spectrum of honey samples of all SC Regions.** Explain variance of Principal Components 1 and 2: PC1 = 52%; PC2 =29%;PC3=0.5%.

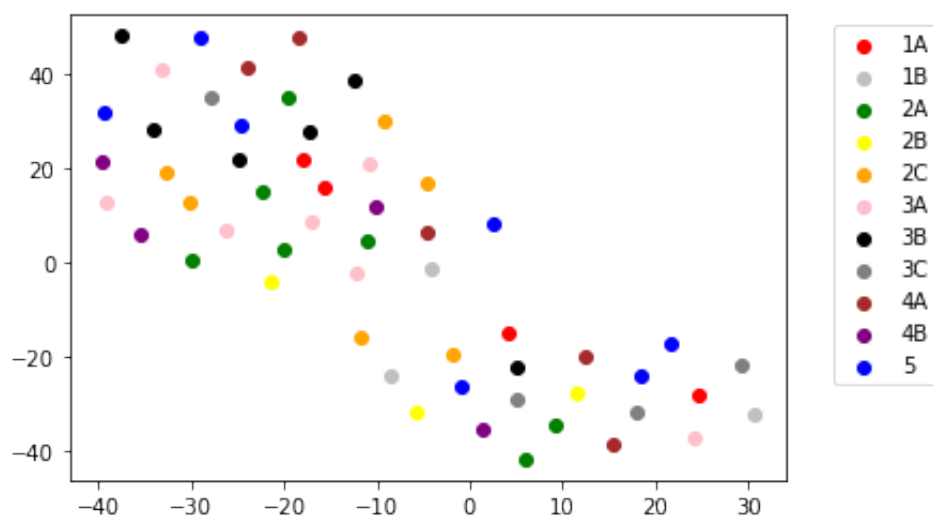


Figure 44: **t-distributed Stochastic Neighbor Embedding (t-SNE)** from ^1H -NMR spectrum of honey samples of all SC Regions.

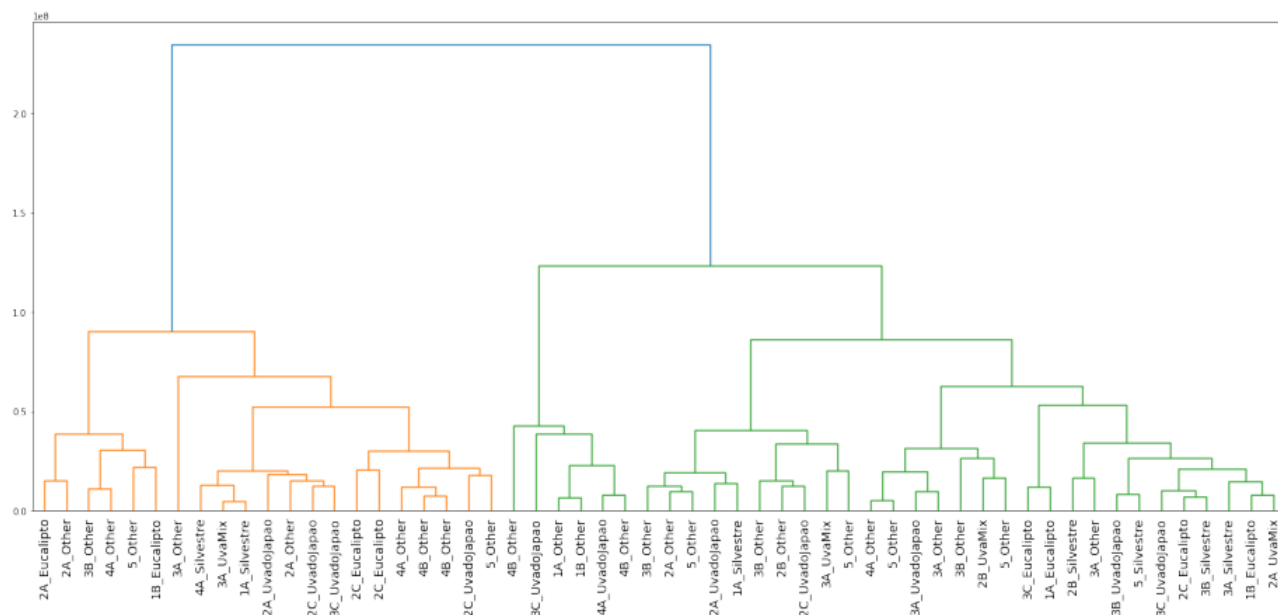


Figure 45: **Dendrogram plot of the hierarchical clustering analysis (HCA)**, with Euclidean distance between samples, from ^1H -NMR spectrum of honey samples of all SC Regions.

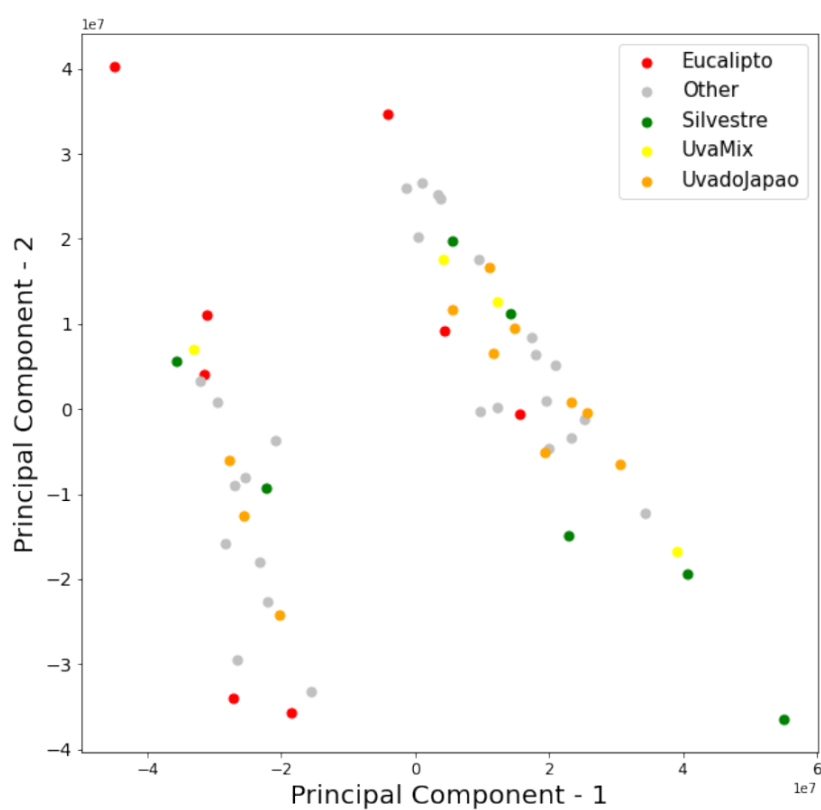


Figure 46: **Principal Components Analysis (PCA) 2D scores plot (PC1, PC2) from ^1H -NMR spectrum of honey samples of Flora origins.** Explain variance of Principal Components 1 and 2: PC1 = 52%; PC2 = 29%; PC3 = 0.5%.

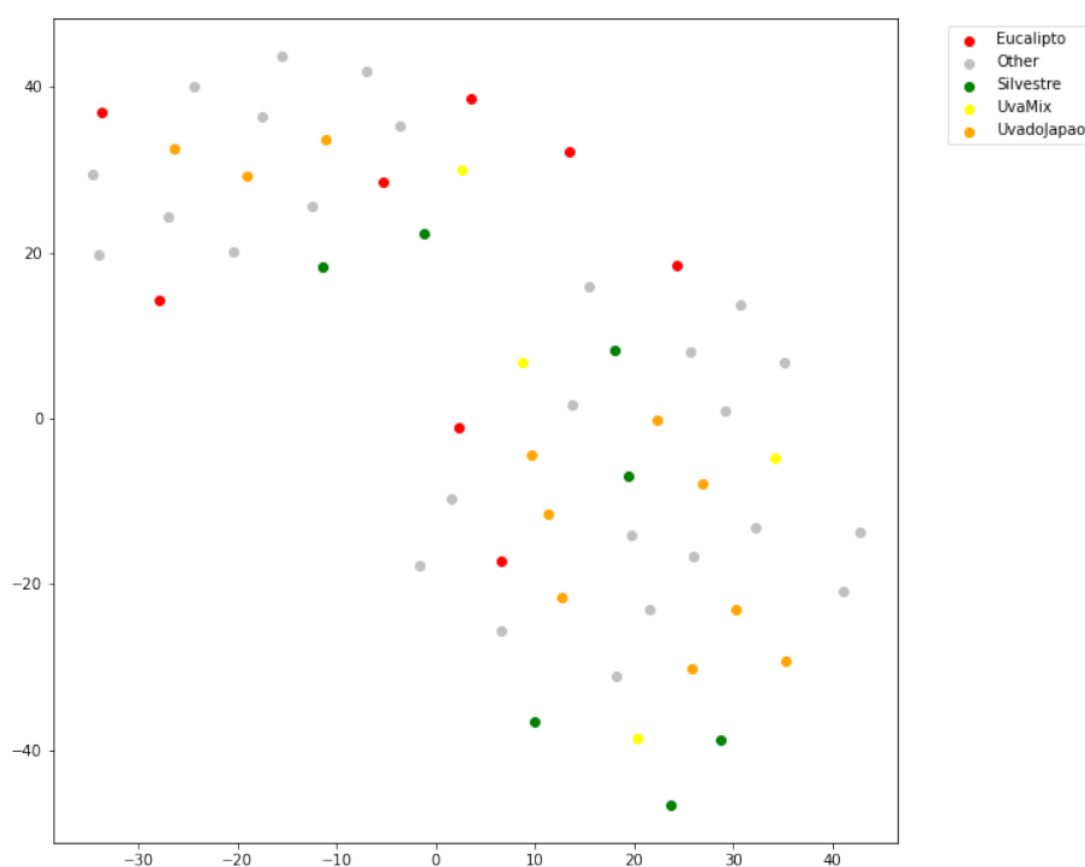


Figure 47: **t-distributed Stochastic Neighbor Embedding (t-SNE)** from ^1H -NMR spectrum of honey samples of Flora origins.

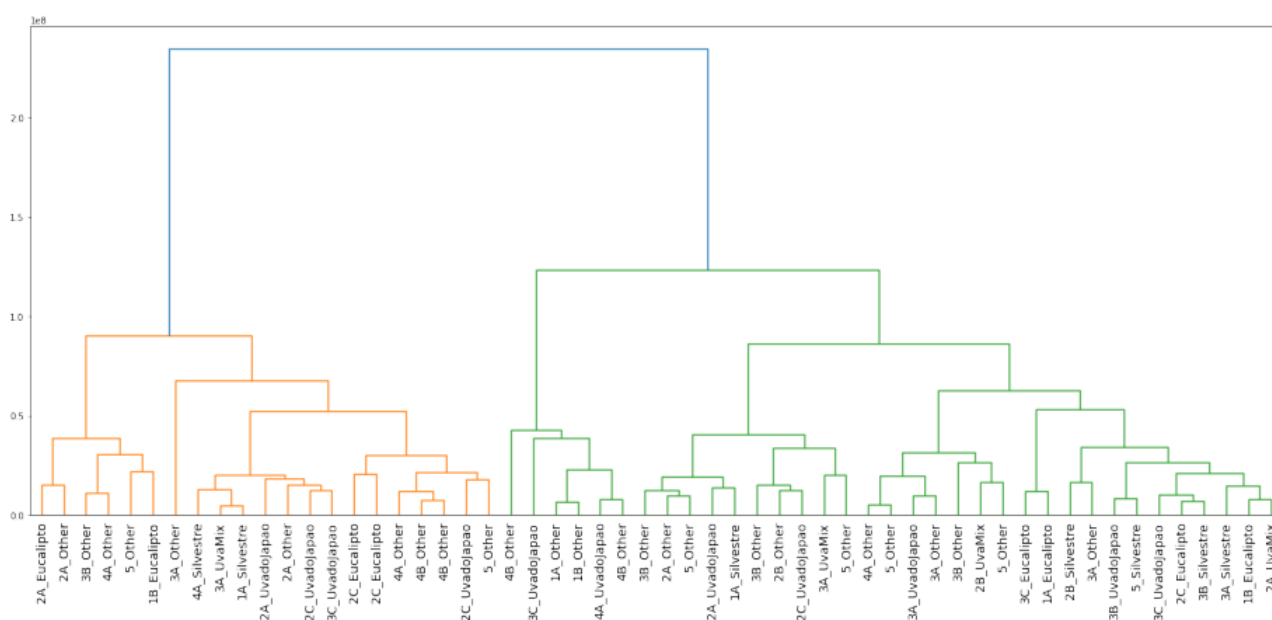


Figure 48: **Dendrogram plot of the hierarchical clustering analysis (HCA), with Euclidean distance between samples, from ¹H-NMR spectrum of honey samples of Flora origins.**

Supplementary Tables

B.1 Supplementary Tables for Chapter 5

Table 25: **Feature with highest variance and the correspondent correlated features group for Red varieties.** First column presents the feature with the higher variance correlated to a set of FT-IR bands given a predefined threshold. This features are the selected to the new sub-dataset used at all red models classification. And these features were used by Shapley method to interpret their impact on each wine grape must variety model classification. Second column are the correspondent correlated features to each feature. Note: 1015*, 1123*, 1127*, 1130* are the features computed by Shapley value with zero correlations to other features, in our work designated as unique features. The wasenumbers at 930**, 980**,1080** and 2639** were features computed by Shapley values for red and white varieties.

Feature with highs variance Wavenumber (cm ⁻¹)	Correlated features group Wavenumber (cm ⁻¹)
930**	2419, 2442, 2461, 2438, 957, 949, 945, 2473, 2458, 2407, 2454, 2465, 953, 941, 937, 2400, 930, 2404, 2427, 2411, 2431, 2415, 2446, 961, 934, 2469,2434, 2450, 2423
965	2388, 2361, 2357, 2944, 2365, 2373, 2384, 2353, 2392, 2369, 2948, 2380, 965, 2396, 972, 968, 976, 2940, 2377
980**	984, 980, 1875, 992, 1435, 1871, 988, 1443, 1863, 1890, 1879, 2350, 1867, 1447, 1883, 1431, 1887, 1439
999	1011, 1007, 1003, 995, 999
1022	1022, 1019
1080**	2168, 2052, 2222, 2245, 2010, 2164, 2149, 1169, 1119, 1316, 2018, 1948, 1200, 2099, 1100, 2045, 2809, 1061, 2909, 1975, 2079, 2187, 2172, 2249, 2191, 2203, 2076, 2238, 1096, 1088, 2130, 1053, 1092, 2812, 1038, 1308, 2180, 1304, 1196, 1111, 1181, 2230, 2218, 2214, 1983, 2095, 1042, 1971, 2241, 1026, 2091, 2068, 2049, 1339, 2160, 2029, 2002, 1964, 1941, 2041, 2022, 1165, 1944, 2014, 1960, 1998, 2110, 1937, 1103, 2087, 1995, 1968, 2025, 2153, 1331, 2106, 1034, 2211, 2037, 2060, 1030, 1933, 1204, 1335, 2056, 1991, 1057, 1323, 1987, 1173, 1046, 2141, 2114, 2195, 1188, 2157, 2226, 1076, 1049, 1929, 2253, 2064, 2006, 1073, 1312, 1327, 2176, 1956, 1080, 2118, 1952, 1069, 1192, 2122, 1184, 1115, 2072, 2207, 2126, 1319, 2137, 1177, 2145, 1300, 2184, 1979, 2199, 2033, 2083, 2103, 2133, 1107, 2234, 1084, 1065
1146	1138, 1146, 1142, 1134
1161	1154, 1157, 1161, 1150

1208	265, 1281, 2315, 2330, 1262, 1215, 2299, 1273, 2272, 1894, 1296, 1821, 1292, 1208, 1806, 2284, 1860, 1250, 1246, 1285, 2288, 1258, 1802, 1898, 1925, 1825, 2265, 2269, 1277, 2257, 2296, 1227, 1809, 1836, 1813, 1848, 1856, 1921, 1906, 1817, 2319, 1844, 1238, 2261, 1914, 2326, 2280, 2323, 1254, 1223, 1917, 1840, 1211, 1219, 1289, 2303, 1829, 1852, 1231, 1235, 2292, 2311, 1833, 1242, 2276, 1902, 1269, 2307, 1910
1389	1416, 1385, 1400, 1354, 1362, 1381, 1408, 1373, 1424, 1370, 1366, 1393, 1389, 1358, 1346, 1377, 1404, 1397, 1343, 1420, 1350, 1412, 1427
2342	2334, 2342, 2338, 2346
2639**	2539, 2508, 2627, 2535, 2596, 2608, 2569, 2523, 2573, 2492, 2554, 2623, 2488, 2635, 2515, 2527, 2519, 2477, 2562, 2496, 2485, 2593, 2616, 2620, 2558, 2500, 2600, 2504, 2542, 2639, 2550, 2631, 2577, 2531, 2566, 2585, 2512, 2604, 2589, 2546, 2612, 2481, 2581
2739	2805, 2894, 2666, 2716, 2755, 2886, 2674, 2735, 2731, 2766, 2789, 2801, 2670, 2720, 2832, 2704, 2770, 2839, 2843, 2647, 2774, 2820, 2878, 2724, 2708, 2782, 2677, 2866, 2859, 2890, 2658, 2855, 2824, 2847, 2739, 2882, 2685, 2863, 2778, 2758, 2870, 2662, 2743, 2701, 2785, 2874, 2643, 2697, 2751, 2654, 2836, 2851, 2797, 2712, 2828, 2762, 2793, 2650, 2681, 2816, 2747, 2689, 2728, 2693
2897	2901, 2897, 2905, 2913
2917	2924, 2917, 2921
2936	2928, 2932, 2936
1015*	1015
1123*	1123
1127*	1127
1130*	1130

Table 26: **Feature with highest variance and the correspondent correlated features group for White varieties.** First column presents the feature with the higher variance correlated to a set of FT-IR bands given a predefined threshold. This features are the selected to the new sub-dataset used at all white models classification. And these features were used by Shapley method to interpret their impact on each wine grape must variety model classification. Second column are the correspondent correlated features to each feature. Note: 2924* is the only feature computed by Shapley value with zero correlations to other features, in our work designated as unique features. The wasenumbers at 930**, 980**, 1080**, 2639** were features computed by Shapley values for red and white varieties.

Feature with highs variance Wavenumber (cm ⁻¹)	Correlated features group Wavenumber (cm ⁻¹)
930*	941, 2388, 2396, 2392, 2369, 2404, 953, 937, 2407, 934, 945, 949, 2373, 2380, 2400, 961, 2377, 2384, 930, 957
980*	2415, 2442, 2419, 2461, 2427, 2446, 968, 2450, 984, 976, 980, 965, 2431, 2458, 2411, 2438, 2434, 2423, 2454, 972
988	1416, 995, 1127, 1408, 1435, 1412, 1424, 1346, 1335, 1339, 1157, 1146, 1169, 1362, 1447, 1350, 1389, 1161, 1150, 1173, 1385, 1007, 1397, 1343, 1130, 1154, 1439, 1370, 1393, 1373, 1404, 1431, 1003, 1165, 1443, 1381, 999, 992, 1011, 1134, 1354, 1142, 1400, 1015, 1366, 1358, 988, 1138, 1377, 1427, 1420
1019	2546, 1285, 2492, 2496, 1289, 2554, 2608, 1254, 1265, 2569, 2612, 1258, 1269, 2481, 1238, 2531, 2581, 2558, 2577, 2523, 2488, 2589, 1273, 2469, 2566, 1119, 1123, 1223, 2485, 2593, 2504, 1277, 2500, 2596, 2519, 1242, 2635, 1262, 2631, 2620, 2616, 2508, 2573, 1227, 2550, 2512, 2585, 1019, 2477, 1022, 2562, 2527, 2473, 2627, 1235, 2515, 2623, 2539, 1281, 1292, 2600, 2542, 2535, 2465, 1231, 2604
1030	2168, 2126, 2253, 2184, 2230, 2238, 2218, 2191, 2176, 2234, 2122, 1115, 2211, 2149, 2199, 2203, 1026, 2187, 2180, 2118, 2145, 2901, 2157, 2133, 2114, 2153, 2141, 2160, 2195, 2222, 2164, 2137, 2130, 2245, 2241, 2226, 1030, 2905, 2249, 2207, 2172, 2214

1080*	1069, 2068, 2079, 1084, 2087, 1046, 2052, 1038, 1100, 2025, 1053, 1096, 1111, 1192, 2037, 2110, 2033, 1073, 2022, 2083, 2091, 2095, 2099, 2056, 2103, 2106, 2072, 2018, 1107, 1034, 1065, 2049, 1061, 1088, 1092, 1103, 2041, 1049, 2029, 1076, 2045, 2060, 1042, 1080, 2076, 1188, 2064, 1057
1177	1979, 1983, 1968, 1975, 1319, 1323, 2010, 1995, 1991, 1331, 1987, 2002, 2014, 1998, 1184, 1960, 2006, 1964, 1181, 1177, 1327, 1971
1204	1304, 1204, 1200, 1196, 1312, 1308
1219	2319, 1809, 1848, 1219, 1829, 2311, 1825, 2288, 2315, 1852, 2296, 2261, 1821, 1813, 2272, 2276, 2257, 2303, 1806, 2323, 2292, 1215, 1840, 1246, 1300, 1250, 1802, 1860, 1817, 1856, 1296, 1211, 2284, 2280, 1833, 1836, 1844, 2265, 2269, 2307, 1208, 2299
1316	1948, 1925, 1956, 1952, 1929, 1937, 1944, 1316, 1921, 1941, 1933
1863	1890, 1894, 1863, 1887, 2326, 2921, 1875, 1883, 1879, 1867, 2917, 2330, 1871, 2334, 1898
1902	2338, 2350, 2342, 1906, 2346, 1910, 1914, 1902, 1917
2948	2361, 2357, 2365, 2353, 2940, 2948, 2944, 2936, 2932, 2928
2639*	2666, 2643, 2662, 2650, 2670, 2677, 2647, 2654, 2639, 2681, 2658, 2674
2685	2793, 2886, 2859, 2770, 2689, 2801, 2728, 2855, 2812, 2724, 2716, 2809, 2751, 2820, 2697, 2774, 2836, 2897, 2816, 2782, 2785, 2755, 2758, 2739, 2863, 2731, 2890, 2701, 2704, 2720, 2874, 2747, 2778, 2882, 2824, 2847, 2805, 2693, 2832, 2766, 2735, 2797, 2828, 2712, 2843, 2894, 2708, 2762, 2870, 2878, 2866, 2851, 2839, 2743, 2685, 2789
2913	2913, 2909
2924*	2924

Table 27: **Supplemental Table: Performance scores for all models classification for all wine musts varieties.** All models had pre-process methods (SNV plus Scaling) with the application of Feature Selection (Drop of constant features with 90% of correlation). Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; PR Threshold: Precision-Recall Threshold; SD: Standard Deviation.

Target Variety	ML Model	PR-AUC±SD	ROC-AUC±SD	MCC±SD	BA±SD	PR Threshold±SD
Arinto	PLS-DA	0.98±0.008	0.97±0.013	0.92±0.04	0.96±0.02	0.72±0.06
Arinto	SVC	0.90±0.04	0.87±0.05	0.70±0.11	0.85±0.06	0.67±0.04
Arinto	RF	0.89±0.04	0.86±0.03	0.56±0.07	0.78±0.04	0.56±0.04
Arinto	GBC	0.89±0.04	0.86±0.04	0.61±0.08	0.81±0.04	0.56±0.04
Arinto	LR	0.77±0.04	0.70±0.05	0.37±0.09	0.78±0.04	0.60±0.04
Viosinho	PLS-DA	0.94±0.03	0.97±0.015	0.92±0.04	0.96±0.02	0.86±0.03
Viosinho	SVC	0.92±0.04	0.91±0.05	0.81±0.09	0.90±0.05	0.70±0.04
Viosinho	GBC	0.90±0.03	0.89±0.04	0.74±0.08	0.87±0.04	0.56±0.04
Viosinho	RF	0.84±0.04	0.86±0.03	0.56±0.07	0.78±0.04	0.56±0.04
Viosinho	LR	0.70±0.05	0.70±0.05	0.36±0.09	0.68±0.05	0.45±0.03
Aragonez	SVC	0.69±0.07	0.87±0.04	0.49 ±0.11	0.74±0.05	0.55±0.06
Aragonez	GBC	0.67±0.06	0.77±0.04	0.41±0.08	0.70±0.04	0.33±0.03
Aragonez	RF	0.60±0.05	0.76±0.04	0.32±0.09	0.66±0.04	0.50±0.04
Aragonez	LR	0.57±0.06	0.82±0.03	0.38±0.08	0.69±0.04	0.37±0.03
Aragonez	PLS-DA	0.53±0.06	0.80±0.03	0.34±0.08	0.67±0.04	0.61±0.06
Syrah	PLS-DA	0.97±0.01	0.95±0.02	0.88 ±0.04	0.94±0.02	0.54±0.04
Syrah	LR	0.92±0.03	0.93±0.02	0.76±0.07	0.88±0.03	0.48±0.05
Syrah	SVC	0.92±0.04	0.94±0.03	0.81±0.07	0.91±0.04	0.57±0.05
Syrah	GBC	0.78±0.04	0.79±0.03	0.48±0.05	0.74±0.03	0.48±0.03
Syrah	RF	0.75±0.04	0.75±0.04	0.32±0.08	0.66±0.04	0.47±0.02
Touriga Nacional	PLS-DA	0.82±0.05	0.91±0.04	0.66±0.08	0.83±0.04	0.61±0.06
Touriga Nacional	SVC	0.82±0.05	0.90±0.04	0.64±0.08	0.82±0.04	0.53±0.05
Touriga Nacional	GBC	0.68±0.06	0.80±0.04	0.39±0.07	0.70±0.04	0.42±0.03

Touriga Nacional	RF	0.65±0.06	0.76±0.04	0.31±0.09	0.65±0.04	0.48±0.03
-------------------------	----	-----------	-----------	-----------	-----------	-----------

B.2 Supplementary Tables for Chapter 6

Table 28: **Supplemental Table: Performance scores for all models classification for all Regions.** All models had pre-process methods with the application of Feature Selection (Drop of constant features with 90% of correlation). Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; PR Threshold: Precision-Recall Threshold; SD: Standard Deviation.

Target Region	ML Model	ROC-AUC	PR-AUC	MCC	BA	ROC Threshold
1A	LR	0.65	0.12	0.15	0.52	0.08
1A	PLSDA	0.58	0.09	0.01	0.51	0.02
1A	SVC	0.55	0.08	0.09	0.58	0.07
1A	RF	0.44	0.07	-0.09	0.42	0.04
1A	GBC	0	0.04	0	0.5	2
1B	LR	0.91	0.23	0.30	0.71	0.10
1B	LOGREG	0.91	0.23	0.3	0.71	0.1
1B	RF	0.79	0.11	0.16	0.65	0.05
1B	PLSDA	0.73	0.09	0.11	0.61	0.14
1B	GBC	0.49	0.15	-0.04	0.49	0.11
1B	SVC	0.3	0.04	-0.26	0.21	0.04
2A	RF	0.48	0.12	-0.07	0.45	0.09
2A	PLSDA	0.45	0.12	-0.05	0.47	0.17
2A	GBC	0.39	0.23	0.26	0.58	1
2A	LR	0.37	0.1	-0.13	0.41	0.15
2A	SVC	0.18	0.08	-0.29	0.31	0.11
2B	GBC	0.43	0.06	-0.09	0.43	0.00
2B	SVC	0.41	0.05	-0.05	0.45	0.05
2B	RF	0.31	0.04	-0.23	0.24	0.03
2B	PLSDA	0.17	0.03	0	0.5	1.43
2B	LR	0.04	0.03	-0.6	0.04	0
2C	PLS-DA	0.60	0.26	0.30	0.61	0.33
2C	LR	0.52	0.33	-0.05	0.46	0.03
2C	RF	0.52	0.15	-0.06	0.46	0.13
2C	GBC	0.45	0.15	0.16	0.58	0
2C	SVC	0.27	0.07	-0.16	0.37	0.09
3A	PLS-DA	0.47	0.10	-0.02	0.49	0.15
3A	SVC	0.46	0.09	-0.03	0.48	0.08
3A	RF	0.39	0.08	-0.18	0.38	0.16
3A	GBC	0.23	0.09	-0.08	0.47	1
3A	LR	0.2	0.06	-0.34	0.25	0.01
3B	RF	0.70	0.17	0.18	0.65	0.06
3B	LR	0.65	0.14	0.15	0.61	0.09
3B	PLSDA	0.64	0.13	0.11	0.58	0.17
3B	GBC	0.2	0.07	-0.17	0.39	0.01
3B	SVC	0.01	0.06	-0.88	0.02	0.08
3C	LR	0.84	0.47	0.20	0.67	0.03
3C	PLSDA	0.75	0.19	0.22	0.62	0.29
3C	RF	0.37	0.06	-0.16	0.38	0.17
3C	SVC	0.03	0.04	-0.37	0.35	0.04

3C	GBC	0	0.04	0	0.5	2
4A	SVC	0.73	0.13	0.16	0.64	0.10
4A	PLSDA	0.19	0.05	-0.36	0.18	0.04
4A	RF	0.13	0.05	-0.41	0.23	0.01
4A	LR	0.08	0.04	-0.46	0.21	0
4A	GBC	0	0.04	0	0.5	2
4B	PLS-DA	0.69	0.12	0.09	0.58	0.11
4B	LOGREG	0.68	0.4	0.08	0.57	0.02
4B	RF	0.6	0.39	0	0.5	0.3
4B	GBC	0.3	0.08	-0.09	0.46	0
4B	SVC	0.15	0.05	-0.33	0.25	0.06
5	SVC	0.82	0.23	0.41	0.78	0.12
5	RF	0.79	0.2	0.29	0.72	0.08
5	PLSDA	0.65	0.39	0.16	0.58	0.18
5	LR	0.51	0.33	-0.07	0.45	0.01
5	GBC	0.21	0.07	-0.14	0.42	0

Table 29: **Supplemental Table: Performance scores for all models classification for all Flora honeys.** All models had pre-process methods with the application of Feature Selection (Drop of constant features with 90% of correlation). Note: PR-AUC: Area Under the Precision-Recall Curve; ROC-AUC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient; BA: Balanced Accuracy; PR Threshold: Precision-Recall Threshold; SD: Standard Deviation.

Target Flora	ML Model	ROC-AUC	PR-AUC	MCC	BA	ROC Threshold
Uva do Japão	PLS-DA	0.72	0.41	0.22	0.64	0.21
Uva do Japão	RF	0.69	0.3	0.22	0.64	0.22
Uva do Japão	RF	0.68	0.4	0.17	0.61	0.07
Uva do Japão	GBC	0.4	0.16	0	0.5	0
Uva do Japão	SVC	0.33	0.14	-0.14	0.42	0.21
Eucalyptus	RF	0.44	0.09	-0.05	0.46	0.04
Eucalyptus	LR	0.42	0.09	-0.07	0.44	0.06
Eucalyptus	PLSDA	0.38	0.08	-0.13	0.4	0.1
Eucalyptus	SVC	0.25	0.07	-0.21	0.36	0.09
Eucalyptus	GBC	0.22	0.08	-0.12	0.44	0.19
Silvestre	SVC	0.56	0.14	0.09	0.57	0.12
Silvestre	GBC	0.49	0.15	0.12	0.58	0
Silvestre	RF	0.48	0.11	-0.17	0.38	0.09
Silvestre	PLSDA	0.43	0.11	0	0.5	0.11
Silvestre	LR	0.42	0.11	-0.09	0.43	0.04
Uva Mix	SVC	0.47	0.07	0.04	0.54	0.07
Uva Mix	PLSDA	0.21	0.04	0	0.5	1.56
Uva Mix	RF	0.04	0.04	0	0.5	1.29
Uva Mix	LR	0.03	0.04	-0.68	0.04	0.01
Uva Mix	GBC	0	0.04	0	0.5	2
Other	SVC	0.67	0.70	0.37	0.67	0.57
Other	RF	0.36	0.39	-0.02	0.49	0.44
Other	GBC	0.35	0.46	-0.08	0.47	0.08
Other	LR	0.19	0.33	-0.4	0.3	0.43
Other	PLSDA	0.15	0.32	-0.49	0.27	0.55