Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 201 (2022) 690–695

Procedia Computer Science

www.elsevier.com/locate/procedia

The 3rd International Workshop on Hospital 4.0 (Hospital)
March 22-25, 2022, Porto, Portugal

Predictive Analytics to support diabetic patient detection

Maria João Vaz[a], João Lopes[a]*, Hugo Peixoto[a], Manuel Filipe Santos[a]

[a]*Centro ALGORITMI, University of Minho*

Abstract

The strong growth in the number of diabetics in recent years has become a major health concern. The dependence on sugar consumption has caused a rapid growth in the level of diagnoses and in the number of deaths associated. In this context, the project developed allowed a study on how Diabetes can be detected in a timely manner, through the existence of pre-indicators of the disease, defining factors that may determine its onset. For this study, data are collected from Hospital de Santa Luzia (ULSAM), considering aspects such as patient profile, prescribed drugs and previous diagnoses. The results prove that machine learning models using profile data with medical drugs produced the best results, optimizing the predictive ability of Diabetes.

Keywords: Diabetes; Predictive Analytics; Artificial Intelligence

1. Introduction

Diabetes is considered a chronic non-transmissible disease, resulting from the imbalance between the level of sugar in the blood and the amount of insulin, a hormone produced by the pancreas and responsible for metabolizing glucose [3]. There are three types of Diabetes, including Type 1 Diabetes, where pancreas stops producing insulin, Type 2 Diabetes, usually associated with being overweight and/or obese due to insufficient insulin production relative to the amount of glucose in the blood, and finally Gestational Diabetes, which only occurs during the period of pregnancy [4]. As the clinical history data of patients grows fast, new data mining techniques are increasingly emerging, allowing the possibility of anticipating the detection of the disease. This study will allow us to understand

* Corresponding author. Tel.: +351934373667;
E-mail address: lopesit@outlook.pt

10.1016/j.procs.2022.03.092

how the application of Machine Learning (ML) techniques in this area should be conducted, understanding how the clinical history of patients can help in the early detection of Diabetes. In this way, 403 diabetic and non-diabetic patients are counted, with a total of 2541 and 1894 hospital visits, respectively, related to the diagnosis of a given disease. Information regarding the characterization of the patient was also considered.

2. Background

2.1. Portuguese Reality of Diabetes

Diabetes has progressed over the years. In 2015, this disease was responsible for 4% of deaths in Portugal. Moreover, "it is estimated to affect 13.3% of the population aged 20-79 years" [7]. In 2019, 4.2 million people died with Diabetes and 463 million were diagnosed with Diabetes between 20 and 79 years old. Worldwide, is estimated to increase to 700 million people in the next 15 years [13].

2.2. Related Works

Due to the progress of Diabetes in recent years and the serious complications that result from it, this disease has become a priority in the research and development of intelligent solutions to improve his management.

The study developed by Raj & Sampath (2019) [11] compares the performance of Support Vector Machines (SVM) and Naive Bayes (NB) algorithms in predicting diabetes, considering information from the patient's health status such as age, blood pressure and blood glucose level. The results showed that the SVM predictive model performs quite well with an accuracy of 82% compared to 62.5% for NB method. On the other side, Dutta, Paul, & Ghosh (2018) [6] focused on RF, SVM and LR algorithms in predicting diabetes on women. The data used in the problem modeling were age, glucose level, number of times a woman had been pregnant, insulin values, body mass index, among others. Hence, they demonstrated that the RF model was the one with the best accuracy (around 84%) and that it best suited for predicting diabetes. Finally, a study focused on preprocessing methods such as Principal Component Analysis and Discretization, exploring SVM, DT and NB algorithms using data related blood pressure, body mass index, body temperature and age. The results demonstrated the effectiveness of using feature engineering techniques before predictive models increasing the accuracy of DT model from 75.1 % to 79.01% and the NB model from 75.82% to 79.01%. SVM model accuracy decreased from 76.6% to 75.69% [16].

On this background, it is understood that the use of personal data produces good results with respect to disease prediction. However, we consider that there is a lack of analysis in the use of medical data and associated medical drugs used by the patient. Thus, we intend to combine this information with the patient's basic profile to understand whether this data holds potential for future studies. The next chapters discuss how the project was conducted, the data used, and the results obtained.

3. Background

3.1. DSR and CRISP-DM

Since this work is a research project and, to understand whether it is possible to classify the clinical status of patients, two methodologies were followed: Design Science Research (DSR) as a research methodology, and Cross-Industry Standard Process for Data Mining (CRISP-DM). DSR structures the project in 6 phases: 1. Problem identification and motivation; 2. Definition of solution objectives; 3. Design and development; 4. Demonstration; 5. Evaluation; 6. Communication. These phases provide guidelines for the evaluation and development of research projects [2,8]. CRISP-DM is the process that presents the life cycle of a Data Mining project in a real context. This project includes a set of six phases (1. Business Understanding; 2. Data Understanding; 3. Data Preparation; 4. Modeling; 5. Evaluation; 6. Deployment.) [1]. There are dependencies between them and does not have a rigid structure [17]. Since both methodologies are used concurrently, the relationship between them throughout the project phases is described. These are presented in Table 1.

Table 1 - Crossover of CRISP-DM and DSR

	CRISP-DM Phases					
DSR Phases	1	2	3	4	5	6
1	X					
2	X	X				
3		X	X	X		
4					X	X
5					X	X
6						

3.2. Tools and Algorithms

R programming language was used for the preparation of data and consequent analysis of it. A set of libraries were used to enable the analysis and consequent predictions of the data, highlighting:

- Dplyr's library allows the use of the DataFrame object to provide storage and manipulation of the data organized by columns [14];
- Caret for ML development [9].

4. Case Study

The case description goes through the methodology presented in the CRISP-DM section. As previously mentioned, the main objective is to predict the clinical status of a patient. The provided data has records about diagnostics and drugs, with each record referring to a patient.

4.1. Business Understanding

The first phase focuses on understanding the project's objectives and requirements. The main purpose is predicting the classification of a specific patient on Diabetes, assessing the impact that each variable has on this characterization, predicting the diagnosis of each patient and the clinical support to be provided to patients. Considering that this is a preliminary phase of the study, it will be attempted to understand which lines the project should follow to model future work.

4.2. Data Understanding

Initially, a list of several diabetic and non-diabetic patients was obtained. The diabetic cases were isolated to identify all diagnoses prior to the first diagnosis. Regarding the prescribed medication, the process of obtain more information followed the same logic as the previous one. Thus, for this study, diabetic and non-diabetic patients are counted. It was concluded that the information obtained presents several records for each patient such as the diagnoses identified and the medication prescribed and the respective dates of those events, as well as some personal information such as their age and gender. The analysis allowed us to define a profile of the diabetic patient. Additionally, is possible to understand a set of clinical data available (each clinical condition is reported as an attribute, creating a clinical history for each patient):

- Cardiovascular, respiratory and abdominal diseases, and other relevant criteria such as anxiety states and hospitalizations;
- Medical drugs already prescribed and usually consumed.

4.3. Data Preparation

In this phase, some classic procedures in data processing tasks are performed: normalization of the data type, elimination of nulls or strange characters and possible junctions to cross-reference existing information between patients. A new attribute was created to indicate the diagnosis and medication, according to the history identified in a list which characterizes an order of preference over each case. Therefore, for a given patient, the number of occurrences for that diagnosis and the total number of prescriptions for a given medicine are listed.

4.4. Modeling

Different classification techniques were applied, according to those identified in previous studies, such as: DT, RF, SVM, NB, NN and LR. and Gradient Boosting Machines (GBM). The development of this models proceeds a set of operations used to optimize results: Implementation of Recursive Feature Elimination (RFE), Cross Validation (CV) and Hyperparameter Tuning (HT). The first one is a technique that selects a subset of most relevant features for a dataset and his application had the purpose to understand the importance of some variables and his utility depends on final predictions. CV is a resampling method that offer more reliable results and is very popular because estimate the capacity to make predictions on unused data during the training phase, evaluating his performance [10]. HT with grid search choose a set of optimal parameters for a learning algorithm [5]. Grid search picks out a grid of hyperparameter values and evaluates all of them with the metrics for evaluation [15]. Class balancing is a typical procedure in classification problems but this was not necessary.

The scenarios were chosen to cover all possibilities: Data with patient profile, diagnostics and medical drugs (S1), Data with patient profile and diagnostics (S2), Data with patient profile and medical drugs (S3) and Data with diagnostics and medical drugs (S4). Therefore, 28 models (4 Scenarios × 7 Techniques) were induced to obtain results, presents in next subchapter.

4.5. Evaluation

The developed models are evaluated to understand if they meet the defined organizational objectives, enabling a global analysis of this research with a review of the entire process and determination of future steps. Metrics are used to understand which algorithm has the best results.

Thus, considering all the initial context of the project, knowledge acquired with related work and data available for this study, the following values were defined for the metrics Area under the ROC Curve (AUC), Accuracy (AC), Precision (PR) and Sensitivity (SE):

- AUC and AC $\geqslant$ 75%
- PR and SE $\geqslant$ 70%

5. Results and Discussion

The best result obtained by the algorithm and its respective scenario are shown for the target under study. The table below presents the best results obtained for each scenario.

Table 2 – Evaluation for Diabetes outcome

Model	Metric	S1	S2	**S3**	S4
DT	AUC	0.7397	0.6997	0.5455	0.7231
	AC	0.6918	0.6364	0.5455	0.6364
	PR	0.7000	0.6067	0.5333	0.6364
	SE	0.7091	0.8182	0.7273	0.6364
GBM	AUC	0.5744	0.6457	0.6777	0.4980

	AC	0.5909	0.5545	0.5909	0.5000
	PR	0.6000	0.5444	0.6000	0.4545
	SE	0.5455	0.3636	0.5455	0.4545
NB	AUC	0.6033	0.5455	**0.8347**	0.7479
	AC	0.5909	0.5000	**0.8182**	0.6364
	PR	0.5833	0.5000	**0.8889**	0.6154
	SE	0.6364	0.5454	**0.7273**	0.7273
NN	AUC	0.6694	0.6612	**0.7521**	0.5289
	AC	0.5909	0.5909	**0.7727**	0.5909
	PR	0.5625	0.6250	**0.7500**	0.5833
	SE	0.8182	0.4545	**0.8182**	0.6364
RF	AUC	0.7355	0.6612	0.7190	0.7810
	AC	0.6364	0.5909	0.6364	0.6818
	PR	0.6667	0.6000	0.6667	0.7500
	SE	0.5455	0.5455	0.5455	0.5455
SVM	AUC	0.7190	0.6529	**0.7538**	0.6116
	AC	0.6364	0.6364	**0.7727**	0.6818
	PR	0.6364	0.7143	**0.8000**	0.8333
	SE	0.6364	0.4545	**0.7273**	0.4545
LR	AUC	0.6612	0.6198	**0.7603**	0.5289
	AC	0.5455	0.5000	**0.8182**	0.5909
	PR	0.5333	0.5000	**0.8182**	0.5833
	SE	0.7273	0.3636	**0.8182**	0.6364

For each model, the evaluation values are shown. For instance, AC is the measure of the closeness to a specific value and AUC is the measure across all the possible thresholds. The PR and SE measures allow you to evaluate the relevance of the results obtained. PR is the ratio between the True Positives and all the Positives and SE evaluate the True Positive Rate [12]. It is evident that S3, which uses data about patient profile and associated medical drugs produces the best predictive results. NB, NN, SVM and LR are the most successful models, accomplishing all the evaluation parameters.

6. Conclusions

This research highlights the ability to predict a diabetic patient using ML techniques with clinical data, presenting the entire process of obtaining predictive results.

The results have a distinction of 4 models, according to the metrics used. Given the existing data, the implementation of these techniques and the results obtained exceeded expectations. Furthermore, it is clear how important data is in the process of predicting a diabetic patient, describing the patient's profile and his medication.

In terms of future work, we have defined the limitation of techniques to be used in those that had obtained the best results. At the same time, it is crucial obtain more data to improve the predictive ability of the implemented algorithms.

Acknowledgements

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

[1] Azevedo, Ana & Santos, Manuel. (2008). KDD, semma and CRISP-DM: A parallel overview. 182-185.

[2] Braga, J. Optimization of Surgery Scheduling in Healthcare. Master's Thesis, University of Minho, Guimarães, Portugal, 2021.

[3] Cardiologia, F. -F. Diabetes. Obtido de FPC - Fundação Portuguesa de Cardiologia: http://www.fpcardiologia.pt/saude-docoracao/factores-de-risco/diabetes/

[4] Contreras, I., & Vehi, J. (2018). Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. Journal of Medical Internet Research

[5] Cortez, P. (2014). [BOOK] Modern Optimization with R. In Springer.

[6] D. Dutta, D. Paul and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning," 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2018, pp. 924-928, doi: 10.1109/IEMCON.2018.8614871.

[7] Diabetes facts & figures. Obtido de International Diabetes Federation - Diabetes facts & figures: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

[8] Lopes, J. Adaptive Business Intelligence: Predictive and Optimization Models in Healthcare. Master's Thesis, University of Minho, Guimarães, Portugal, 2020.

[9] Max, Kuhn. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 28. 10.18637/jss.v028.i05.

[10] Misra, Puneet & Singh, Arun. (2020). Improving the Classification Accuracy using Recursive Feature Elimination with Cross-Validation. 11. 659-665.

[11] R. S. Raj, D. S. Sanjay, M. Kusuma and S. Sampath, "Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2019, pp. 41-45, doi: 10.1109/ICATIECE45860.2019.9063792.

[12] Santos, M. F., & Azevedo, C. (2005). DATA MINING - Descoberta de Conhecimento em Bases de Dados. FCA - Editora de Informática

[13] Saúde, S. -S. Acordo de cooperação entre a APDP e o Serviço Nacional de Saúde. Obtido de SNS - Serviço Nacional de Saúde: https://www.sns.gov.pt/noticias/2018/05/14/diabetes/

[14] Singh, Gurpreet & Soman, Biju. (2019). Data Transformation using dplyr package in R. 10.13140/RG.2.2.10397.46565.

[15] Philipp Probst, Anne-Laure Boulesteix, and Bernd Bischl. 2019. Tunability: importance of hyperparameters of machine learning algorithms. J. Mach. Learn. Res. 20, 1 (January 2019), 1934–1965.

[16] V. V. Vijayan and C. Anjali, "Decision support systems for predicting diabetes mellitus — A Review," 2015 Global Conference on Communication Technologies (GCCT), 2015, pp. 98-103, doi: 10.1109/GCCT.2015.7342631.

[17] Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 1–13 April 2000.