Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 210 (2022) 218–223

Procedia Computer Science

www.elsevier.com/locate/procedia

The 3rd International Workshop on Healthcare Open Data, Intelligence and Interoperability (HODII)
October 26-28, 2022, Leuven, Belgium

Data Mining Models for Automatic Problem Identification in Intensive Medicine

Inês Quesado[a], Julio Duarte[a]*, Álvaro Silva[b], Maria Manuel[b], César Quintas[b]

[a]*Algoritmi/LASI research center, University of Minho, Portugal*
[b]*Centro Hospitalar Universitário do Porto, Portugal*

Abstract

This paper aims to support medical decision making on predicting the diagnosis of COVID-19. Thus, a set of Data Mining (DM) models was developed using prediction techniques and classification models. These models try to understand whether the vital signs of patients have a correlation with a diagnosis. To achieve the objective of the paper, initially, the data was acquired and collected from several data sources such as bedside monitors and electronic nursing records from the Intensive Care Unit of the Santo António Hospital. Secondly, the data was transformed so that it could be used in DM models. The models were induced using the following algorithms: Decision Trees, Random Forest, Naive Bayes, and Support Vector Machine. The analysis of the sensitivity, specificity, and accuracy were the metrics used to identify the most relevant measures to predict COVID-19 diagnosis. This work demonstrates that the models created had promising results.

Keywords: Data Mining; Intensive Medicine; Intensive Care Unit; Vital Signs; Classification

* Corresponding author.
E-mail address: jduarte@di.uminho.pt

1877-0509
10.1016/j.procs.2022.10.140

1. Introduction

Intensive Medicine is an area of medicine that confuses on the prevention, diagnosis, and treatment of patients with critical health problems. This medicine is fundamentally applied in Intensive Care Units (ICUs) [1]. Every day in ICUs, large amounts of clinical data are produced, stored, and analyzed. Some of them can be read directly and others need to be cross-referenced and analyzed to be a source of information. All this makes it sometimes extremely difficult and complex for health professionals to interpret and understand all available data in real time and decide on the best course of action.

The main purpose of this paper is to support medical decision making on predicting the diagnosis of COVID-19. Considering the objective of the paper, it is important to mention that the data to be used during this study were captured during a pandemic period (COVID-19). Thus, the study was directed in that direction since a significant part of the data were conditioned to a whole covid symptomatology that massively occupied the intensive care units.

This article is divided in five sections. The first section is the Introduction, that describes a general introduction to the problem. The second section is called Background where is presented some theoretical fundamentals. The third section is the Main Focus of the Chapter, where it is described all the work carried out in order to achieve the intended objectives. The fourth section discusses the results of the study. Lastly, in the fifth chapter concerns the conclusion regarding to the study.

2. Background

2.1. Intensive Medicine and Intensive Care Units

Intensive Medicine (IM) is a heterogeneous, complex, and evolving medical specialty. According to the Direção Geral de Saúde (DGS), IM can be defined as an area which specifically addresses the prevention, diagnosis, and treatment of acute and potentially reversible potentially reversible, in patients who present with failure of one or more vital functions, imminent or established [2]. These patients are usually admitted to intensive care units (ICU) so that they can maintain their physiological functions through various life support devices until they are able to do so autonomously. The Intensive Care Unit (ICU) is a unit where specialized care is provided for the treatment of patients with complex health conditions, usually organ failure and, consequently, serious life-threatening conditions [3].

2.2. Data Mining

Data Mining is the process of discovering interesting patterns, model, and other kinds of knowledge in large data sets. Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material [4]. According to Vercellis (2009), DM refers to the overall process that comprises the collection and analysis of data, the development of inductive learning models, and the adoption of practical decisions and consequent actions based on the acquired knowledge [5]. The discovery of patterns in the data may be in the form of business rules, affinities, correlations, or terms of prediction models [6]. There are several DM techniques that can be grouped in different ways, namely: classification and regression algorithms, clustering and association algorithms, visualization, and time series prediction. For the study at hand, the classification techniques were used.

2.1. Predictive Business Analytics

According to Delen, Business Analytics is the art and science of identifying or discovering new knowledge, from large volumes of data with a varied typology, using sophisticated Machine Learning, mathematical and statistical models, to support decision making [7]. One of the categories of analytics is Predictive Analytics, which uses statistical models and predictive methods to show what might happen [8].This analytics uses a variety of statistical techniques framed in Data Mining, predictive models, and Machine Learning algorithms in order to analyze historical and current data, identifying predictive variables and thus predicting future or otherwise unknown events [9].

3. Study Description

3.1. Methods and Tools

This study was developed according to the guidelines provided by the Cross Industry Standard Process for Data Mining (CRISP-DM), a methodology often used in solving data mining related problems. In the development phase, the following tools were used: Spyder, a python programming tool, for analyzing, understanding, and preparing data, and the RapidMiner tool for building predictive models

3.2. Business Understanding

The increasing availability of data generated by the different systems present in Intensive Care Units has brought about a huge challenge, due to the excess of non-relational information among them, bringing increased complexity to medical decision making. The need for increasingly accurate decisions in extremely difficult and important scenarios, such as the situations experienced in an ICU, has increased the search for solutions that intensivists in reading the patient's condition. In this context, Intensive Care Medicine has become an area of focus for Data Mining applications Mining applications because of its environment and the characteristics of the data they can provide.

3.3. Data Understanding

The study began with the data extraction from the ICU of the Centro Hospitalar Universitário do Porto, specifically the reading of vital signs of patients whose hospitalisation was between March 2020 and October 2021, as well as the problems associated with ICU patients in the same period. These data were organised into two datasets: Vital Signs, which contained information about the vital signs of patients coming from the ICU, and SOAP, which contains all the information regarding the diagnosis of the patients. The Vital Signs dataset contained records for 70 patients, while the SOAP dataset contained information for 47 patients. In tables 1 and 2 we can see the description of the Vital Signs and SOAP dataset variables, respectively.

Table 1. Vital Signs Dataset

Dataset	Variable	Description
Vital Signs	NUM_SEQUENTIAL	Sequential number of admission
	TSTAMP	Data collection timestamp (2-minute interval)
	DATEHOUR_ENTRY	Date and time of admission
	MNDRY_BLD_PULS_RATE_ART_ABP	Pulse rate
	MDC_PRESS_BLD_ART_ABP_SYS	Systolic blood pressure
	MDC_PRESS_BLD_ART_ABP_MEAN	Mean arterial pressure
	MDC_PRESS_BLD_ART_ABP_DIA	Diastolic Blood Pressure
	MDC_PRESS_BLD_ART_SYS	Systolic Intra-arterial Pressure
	MDC_PRESS_BLD_ART_MEAN	Mean Intra-arterial Pressure
	MDC_PRESS_BLD_ART_DIA	Diastolic Intra-arterial Pressure
	MDC_TEMP	Body Temperature
	MDC_PULS_OXIM_SAT_O2	Oxygen saturation
	MDC_PULS_OXIM_PULS_RATE	Pulse rate determined from a pulse oximeter
	MDC_ECG_CARD_BEAT_RATE	Heart Rate

Table 2. SOAP Dataset

Dataset	Variable	Description
SOAP	NUM_SEQUENTIAL	Sequential number of admission
	NUM_PROCESS	Patient process number
	DATE_DAY	SOAP record day
	EPISODE	Patient's hospitalization number
	SOR	Subjective and Objective Report
	RPI	Treatment

3.4. Data Preparation

Since the present study is to make predictions using data mining techniques (classification approach) it was necessary to make some transformations in the data collected.
In the Vital Signs dataset new columns were added with the minimum, maximum and average of the fundamental variables for this study. After that, a column referring to the time of each record was also created. From these columns, a pivoting by minimum, average, maximum of variables and time was performed. Having said this, and given the large number of nulls, two cases were defined in order to minimize the impact of the results.

- Case 1: Go through each column and if there are no records before or after the record in question, set the value to -1.
- Case 2: Go through each column and for a given patient, fetch the first record you find immediately after the null record and fill it in.

For both cases, if there are records before or after the record in question, the average of both is taken to fill the null value. Next, a new column was added with the problems coming from the SOAP dataset. Finally, in the new dataset a new column named 'COVID-19' was created, which through binary values distinguishes the COVID-19 problems from the remaining problems by assigning the values 1 and 0, respectively. After the data treatment and with the merge between the datasets to create the final dataset, the data reduced to only 13 patients.

Table 3. Final Dataset

Variable	Description	Max	Avg	Min
NUM_SEQUENTIAL	Sequential number of admission	-	-	-
PROBLEM	Name of the problem	-	-	-
MNDRY_BLD_PULS_RATE_ART_ABP	Pulse rate	452	84	16
MDC_PRESS_BLD_ART_ABP_SYS	Systolic blood pressure	360	123	1
MDC_PRESS_BLD_ART_ABP_MEAN	Mean arterial pressure	360	84	1
MDC_PRESS_BLD_ART_ABP_DIA	Diastolic Blood Pressure	360	62	1
MDC_TEMP	Body Temperature	40	36	15
MDC_PULS_OXIM_SAT_O2	Oxygen saturation	100	95	8
COVID-19 PNEUMONIA	Diagnose (0 - Yes, 1 - No)	-	-	-

3.5. Modeling

This phase focused on getting models to translate business goals through the application of data mining techniques. For the present study, the classification models were built using the RapidMiner Studio 9.10. At this stage it was important to select the Data Mining techniques that will best suit rating approach. The techniques selected to solve this classification problem were: Decision Tree (DT), Random Forest (RF), Naive Bayes (NB) and Support Vector Machine (SVM). The developed models were induced using the sampling method 10 Folds Cross Validation. The

scenarios were created based on two factors: the way the null values were filled in and the target. The target contained data that did not allow the model to be generated. So, to solve this situation: two new versions of the dataset were created. The first version, where the different problems were eliminated and a second version where the different problems were grouped in a new category.

- Scenario 1: Case 1 without a new category;
- Scenario 2: Case 1 with a new category;
- Scenario 3: Caso 2 without a new category;
- Scenario 4: Caso 2 with a new category;

These models, that follow the classification approach, can be represented by:

DMM = {4 Scenarios, 4 techniques, 1 Target, 1 Sampling Method} where:

Target (dependent) = COVID-19 Pneumonia;
Input (independent) = {MNDRY_BLD_PULS_RATE_ART_ABP, MDC_PRESS_BLD_ART_ABP_SYS MDC_PRESS_BLD_ART_ABP_MEAN, MDC_PRESS_BLD_ART_ABP_DIA, MDC_TEMP, MDC_PULS_OXIM_SAT_O2};
Sampling Method = $\{10-folds\ CV1\}$

4. Results

All scenarios were studied in detail to understand which models had the best results. To evaluate the models, it was designed a confusion matrix for each one of the models. The confusion matrix allows to determine the number of.

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives
ACC = Accuracy

Sens = Specificity
Spec = Specificity
Prec = Precision
(TP+TN) / (TP+TN+FP+FN) = ACC. (1)

TP/(TP+FN) = Sens. (2)
TN/(TN+FP) = Espec. (3)
TP/(TP+FP) = Precision (4)

Using the CMX it was possible calculate some measures. To this study it was used: accuracy (ACC), sensitivity (Sens), specificity (Espec), Precision (Prec), Area Under the ROC Curve (AUC). In medicine, it is usual to use sensitivity and specificity analysis for error measures [10].

Table 4. Metrics for Decision Tree and Random Forest

	Decision Tree					Random Forest				
	ACC	AUC	Prec	Sens	Spec	ACC	AUC	Prec	Sens	Spec
Scenario 1	66.74%	0.697	83.46%	45,24%	89.67%	62.60%	0.635	72.55%	44.95%	81.40%
Scenario 2	68.68%	0.704	45.56%	42,68%	92.94%	69.92%	0.651	88.21%	43.63%	94.45%
Scenario 3	66.52%	0.686	85.23%	43.03%	91.60%	64.10%	0.631	77.77%	43.44%	86.15%
Scenario 4	69.98%	0.698	89.66%	42.68%	95.45%	70.70%	0.650	91.26%	43.36%	96.22%

Table 5. Metrics for Naïve bayes and Support Vector Machine

	Naïve Bayes					Support Vector Machine				
	ACC	AUC	Prec	Sens	Spec	ACC	AUC	Prec	Sens	Spec
Scenario 1	56.60%	0.390	55.59%	80.08%	31.56%	66.38%	0.692	77.56%	49.40%	84.52%
Scenario 2	54.61%	0.409	51.95%	79.54%	31.36%	69.46%	0.703	80.89%	47.83%	89.65%
Scenario 3	54.82%	0.624	53.86%	88.11%	19.31%	66.38%	0.692	77.56%	49.40%	84.52%
Scenario 4	52.26%	0.637	50.35%	88.07%	18.85%	69.07%	0.699	81.85%	46.21%	90.40%

In terms of model accuracy, all techniques have very similar results, so other metrics should be considered to evaluate the same. In terms of the Area Under the ROC Curve (AUC) metric, the AUC value varies between 0 and 1.0, and the higher the AUC, the better the model's ability to predict the correct category. Thus, the best results are present in the Decision Tree technique. In the context of clinical decision, for the COVID-19 prediction, and considering that 0 is the COVID-19 and 1 is other diagnosis, the assessment must be in favor of sensibility, as it measures the proportion of positives that are correctly identified. This way, the Naïve Bayes technique with scenario 3 is the one that presents the best results in terms of sensitivity; however, it is the scenario within the same technique that presents the worst specificity. All the other techniques have a higher specificity and, on the other hand, a lower sensitivity, which indicates that with the variables in question, there is no discriminatory capacity in relation to the diagnosis of COVID-19. It can also be concluded that the results do not differ much between the different scenarios, which presupposes that the data transformation had no impact on the performance of the models.

5. Conclusion

This study explored the prediction of COVID-19 diagnosis using only vital signs collected from a set of data sources such as bedside monitors and electronic nursing records from the Intensive Care Unit of the Hospital Santo António. This work contributed with a set of classification models that can help health professionals in the decision making. Based on the above results it can be concluded that the models created are promising for further research. Thus, future work could make the models created more assertive by adding a set of new variables that could allow greater discriminatory ability of the covid-19 diagnostic. As an example, a recently published study used the sound characteristics present in breathing and coughing samples for the detection of Covid-19. The results of that study were very interesting and promising [11].

Acknowledgements

The work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: DSAIPA/DS/0084/2018.

References

[1] T. Guimarães, I.Quesado, I.Tavares, M.Passos, J.Duarte, M. F. Santos, A.Silva, “Knowledge Extraction From ICU Data Using Data Visualization,” 2022, pp. 129–149.

[2] J. A. Paiva, A. Fernandes, C. Granja, F. Esteves, J. Miguel, J. Ribeiro, J. Nóbrega, J.Vaz, P. Countinho , “Rede Nacional de Especialidade Hospitalar e de Referenciação - Medicina Intensiva,” *República Port. - Saúde*, 2017, [Online]. Available: https://www.sns.gov.pt/wp-content/uploads/2017/08/RNEHR-Medicina-Intensiva-Aprovada-10-agosto-2017.pdf.

[3] J. Ramon, Fierens D, Güiza F, Meyfroidt, G, Blockeel, H, Bruynooghe, M, Van Den Berghe, G, “Mining data from intensive care patients,” *Adv. Eng. Informatics*, vol. 21, no. 3, pp. 243–256, 2007, doi: 10.1016/j.aei.2006.12.002.

[4] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*. Elsevier Science, 2022.

[5] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*. Wiley, 2009.

[6] R. Sharda, E. Turban, D. Delen, J. Aronson, and T. P. Liang, *Business Intelligence and Analytics: Systems for Decision Support*. Pearson, 2013.

[7] D. Delen, *Prescriptive Analytics: The Final Frontier for Evidence-Based Management and Optimal Decision Making*. Pearson Education, 2019.

[8] R. Vidgen, S. N. Kirshner, and F. B. Tan, *Business Analytics: A Management Approach*. Palgrave Macmillan Limited, 2019.

[9] D. Bentley, *Business Intelligence and Analytics*. Larsen and Keller Education, 2017.

[10] K. Cios, G. Moore, 2002. Uniqueness of medical data mining. In Artificial Intelligence in Medicine 26, 1–24.

[11] N. A. Given, “Audio feature ranking for sound-based COVID-19 patient detection.”.