Journal of Medical Systems (2021) 45: 6
https://doi.org/10.1007/s10916-020-01682-8

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Data Mining for Cardiovascular Disease Prediction

Bárbara Martins[1] · Diana Ferreira[2] · Cristiana Neto[2] · António Abelha[2] · José Machado[2]

Received: 1 October 2020 / Accepted: 24 November 2020 / Published online: 5 January 2021

Abstract

Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels and are a major cause of disability and premature death worldwide. Individuals at higher risk of developing CVD must be noticed at an early stage to prevent premature deaths. Advances in the field of computational intelligence, together with the vast amount of data produced daily in clinical settings, have made it possible to create recognition systems capable of identifying hidden patterns and useful information. This paper focuses on the application of Data Mining Techniques (DMTs) to clinical data collected during the medical examination in an attempt to predict whether or not an individual has a CVD. To this end, the CRossIndustry Standard Process for Data Mining (CRISP-DM) methodology was followed, in which five classifiers were applied, namely DT, Optimized DT, RI, RF, and DL. The models were mainly developed using the RapidMiner software with the assist of the WEKA tool and were analyzed based on accuracy, precision, sensitivity, and specificity. The results obtained were considered promising on the basis of the research for effective means of diagnosing CVD, with the best model being Optimized DT, which achieved the highest values for all the evaluation metrics, 73.54%, 75.82%, 68.89%, 78.16% and 0.788 for accuracy, precision, sensitivity, specificity, and AUC, respectively.

Keywords Cardiovascular disease · Health information systems · Decision support systems · Data mining · CRISP-DM · Classification

Introduction

Every time a patient attends a hospital, data are collected on demographic data, clinical history, symptoms, diagnosis, medication, treatment, among others. Hence, the urgency to deal with the rapidly growing volumes of digital data [5]. Health Information Systems (HIS) are crucial since they provide powerful information for decision making [8]. It is therefore essential that there is no loss or misinformation at any time affecting decisions taken on the basis of the data provided [3]. Thus, Data Mining comes in handy for discovering patterns in large datasets, involving methods that are at the intersection of machine learning, statistics, and database systems [7].

The aim of this work is to satisfy the urgent need to extract useful knowledge hidden in clinical data, particularly to develop a solution able to predict the presence/absence of CVDs through Data Mining. Consequently, the process time between the patient admission and the diagnosis will be reduced and an immediate and adequate treatment can be given to the patient.

According to the World Health Organization (WHO), CVDs are the number one cause of death worldwide, taking an estimated 17.9 million lives each year [1]. CVDs are a group of disorders of the heart and blood vessels and include coronary heart diseases, cerebrovascular disease, rheumatic disease, and other conditions [2]. Patients at risk of CVD may demonstrate raised blood pressure, glucose, and lipids as well as overweight and obesity. Most CVDs can be

This article belongs to the Topical Collection: *Health Information Systems & Technologies*
Guest Editors: Álvaro Rocha and Joaquim Gonçalves

✉ José Machado
jmac@di.uminho.pt

Bárbara Martins
a81824@alunos.uminho.pt

Diana Ferreira
diana.ferreira@algoritmi.uminho.pt

Cristiana Neto
cristiana.neto@algoritmi.uminho.pt

António Abelha
abelha@di.uminho.pt

[1] University of Minho, Campus of Gualtar, Braga 4710, Portugal

[2] Algoritmi Research Center, University of Minho, Campus of Gualtar, Braga 4710, Portugal

prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet, physical inactivity and harmful use of alcohol [9]. Identifying those at highest risk of CVD and ensuring that they receive appropriate treatment can prevent premature deaths [1].

This disease is a development issue in low-and middle-income countries due to the fact that these countries often do not benefit from integrated primary health care, early detection and treatment programs for people with risk factors. To a better understanding of how this disease in handled in European Union (EU), Fig. 1 schematises the distribution costs of CVDs in EU member countries by percentage of primary care, outpatient care, Accident & Emergency (A&E), inpatient care, medications, production losses due to mortality, production losses due the morbidity and, lastly, informal care [10].

The paper is organized as follows: Section "Methodology" presents the CRISP-DM methodology and its stages, Section "Discussion" presents the discussion of the results and lastly, Section "Conclusions and future work" concludes the study and outlines the future work.

Methodology

The CRISP-DM methodology was the one used in this study because it enables the replication of Data Mining projects and assists in their planning and management, being user-friendly and revealing the maturity of the Data Mining [6].

Business understanding

This stage strives to understand the intentions and requirements of the project from the business perspective and converts this knowledge into a Data Mining problem definition [4].

Given the high number of CVD-related deaths, it has become essential to detect and understand its main risk factors to try to mitigate this worldwide problem. Thus, the objective is to develop a solution able to predict the presence/absence of CVDs in patients relying on medical data collected during examination, reducing the process time between the patient admission and its diagnosis and consequently providing an immediate and adequate treatment to the patient. This is a classification problem, in which softwares as Weka and RapidMiner will be used in the next steps.

Data understanding

The dataset used in this study was retrieved from the Kaggle data repository and is related with the detection of CVD cases, including 70000 registers of patients and 12 attributes considered relevant for identifying the disease [11].

All attributes were collected during the patient's medical examination and grouped into 3 categories:

– *Objective data* - factual information;
– *Examination data* - results from medical exams;
– *Subjective data* - information given by the patient.

Table 1 presents a description of each attribute.

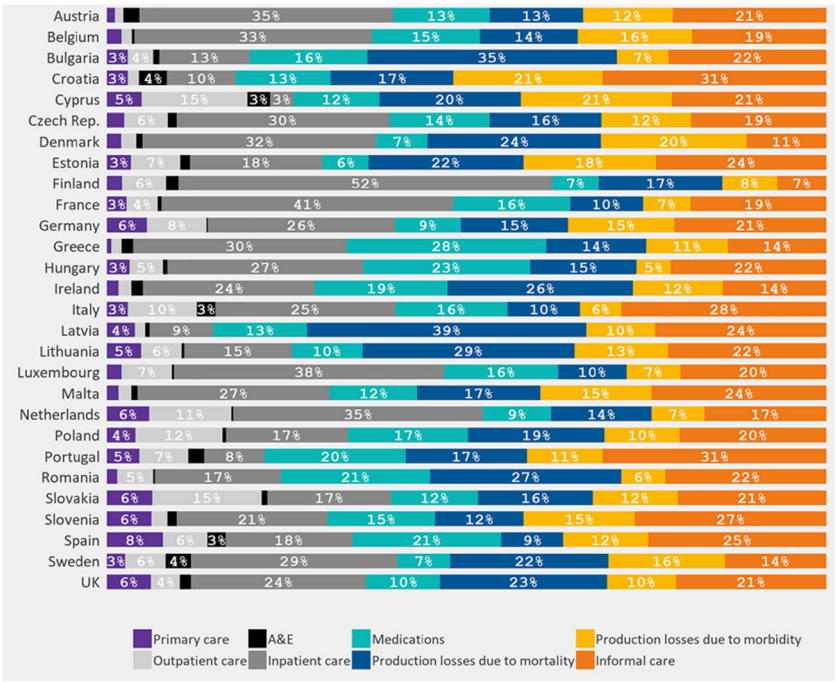

Fig. 1 Distribution of costs of CVD in the EU member countries by category (2015) [10]

Table 1 Characteristics of the attributes

Attribute	Description	Type
id	Patient's unique identifier	Objective
age	Patient's age in days	Objective
height	Patient's height in cm	Objective
weight	Patient's weight in kg	Objective
gender	Patient's gender[1]	Objective
ap_hi	Systolic blood pressure	Examination
ap_lo	Diastolic blood pressure	Examination
cholesterol	Patient's cholesterol[2]	Examination
gluc	Patient's glucose	Examination
smoke	Whether the patient smokes or not[3]	Subjective
alco	Whether the patient consumes alcohol or not[3]	Subjective
active	Whether the patient is physically active or not[3]	Subjective
cardio	Presence of CVD[3]	Target

[1] Values: 1 = male, 2 = female
[2] Values: 1 = normal, 2 = above normal, 3 = well above normal
[3] Values: 0 = false, 1 = true

The target of the study is the *cardio* attribute, which gives information about the existence or absence of CVD. If *cardio* is "0" the patient is healthy, in contrast, if it is "1" the patient has CVD.

It was then necessary to analyze whether there was any class disparity that could influence the way algorithms would learn. Figure 2 shows the balanced data distribution for the target attribute with 50.2% of type "yes" (34820 registers) and 49.8% of type "no" (34605 registers).

Data preparation

At first, data was integrated and the data-cleaning process was applied. In more detail, it was analysed the existence of duplicated data, missing values, outliers, and inconsistencies. It was found that there were no duplicated data, missing values or inconsistencies. The outliers were removed with the filters *InterquartileRange* and *RemoveWithValues* of WEKA, in which the first detects outliers and extreme values based on the interquartile ranges and the other one removes instances according to a certain value.

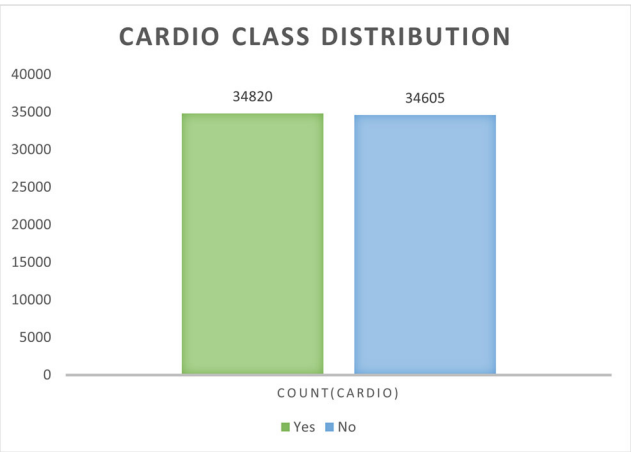

Fig. 2 Class distribution of the target variable *cardio*

Afterwards, it was time to proceed to the data transformation. Since the attribute *age* was measured in days, it was converted into years with the expression "age/365". By default, most attributes were imported as Integer, however attributes *active*, *alco*, *cardio* and *smoke* correspond to the binominal type and *cholesterol* to the polynominal type. Hence, these attributes were transformed accordingly.

Finally, the feature weights were evaluated using different criteria, such as *Weight by Correlation*, which provides a correlation with the label attribute (cardio) in which very high correlation values could be damaging to the results. Since none of the attributes were highly correlated, they were all kept to proceed to the next phase. Other weight measurements, such as *Gini Index*, *Information Gain* and *Information Gain Ratio*, have allowed the definition of certain thresholds for which different scenarios have been created to be used in the Modeling stage. Table 2 presents the weight values of the attributes.

Modeling

This stage consists in selecting the modeling technique, generating the test design, building the model and, lastly, assessing its performance [6].

Primarily, the modeling techniques were selected based on the operator *Compare ROCs* that calculates the

Table 2 Weights' distribution

Feature weight							
Correlation		Gini index		Information gain		Information gain ratio	
age_years	0.239	ap_hi	0.095	ap_hi	0.142	ap_hi	0.160
weight	0.181	ap_lo	0.062	ap_lo	0.094	ap_lo	0.107
cholesterol	0.164	cholesterol	0.024	cholesterol	0.037	age_years	0.067
gluc	0.089	age_years	0.020	age_years	0.028	height	0.061
ap_lo	0.065	weight	0.012	weight	0.017	weight	0.057
ap_hi	0.054	gluc	0.004	gluc	0.006	cholesterol	0.034
active	0.036	active	0.001	active	0.001	gluc	0.010
smoke	0.016	height	0.000	height	0.000	active	0.001
height	0.012	smoke	0.000	smoke	0.000	smoke	0.000
alco	0.007	alco	0.000	alco	0.000	alco	0.000
gender	0.007	gender	0.000	gender	0.000	gender	0.000

Receiver Operating Characteristic (ROC) curves. This curve represents in the *X* axis the True Positive Rate (number of positive instances correctly classified / total of positive instances) and in the *Y* axis the False Positive Rate (number of negative instances incorrectly classified / total of negative instances). Figure 3 shows the output of this operator using Cross Validation. The DMTs that were compared correspond to k Nearest Neighbour(k-NN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Gradient Boosted Tree (GBT), Rule Induction (RI), Deep Learning (DL), Generalized Linear Model (GLM) and, lastly, Logistic Regression (LR). Comparing the ten curves,

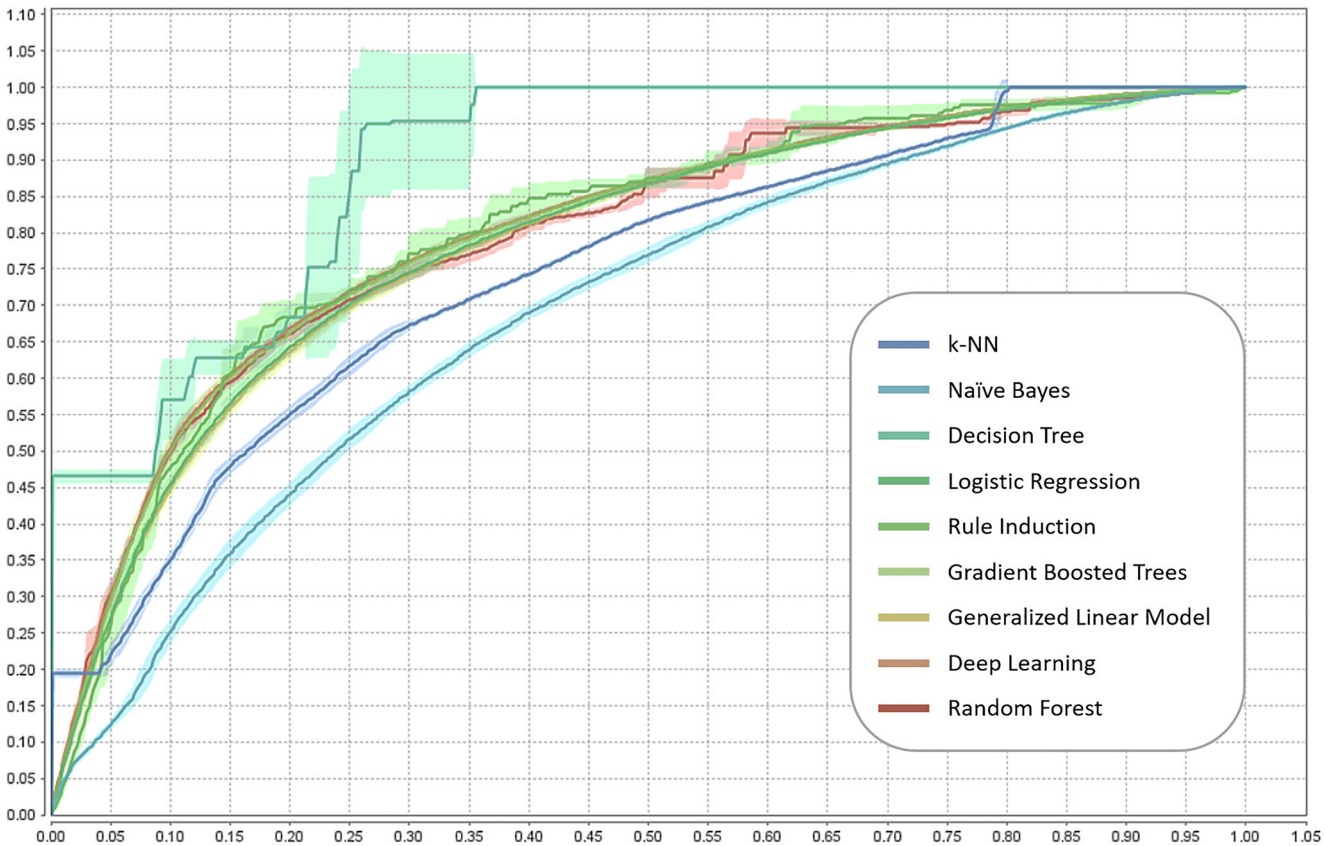

Fig. 3 ROC curves using cross validation

DT had the best performance and, in contrast, k-NN and NB had an aleatory behaviour corresponding to the least effective classifiers. Thus, only DT, RI, RF and DL were used for the next steps. Since DT was the technique with the best results, in order to obtain higher results, some parameters, namely (*criterion*, *maximal_depth*, *confidence* and *minimal_leaf_size*), were optimized using the *Optimized Parameters* operator, which is a nested operator that executes the subprocess for all the combinations of values of the parameters and then delivers the optimal parameter values.

For each DMT, two sampling methods were tested: Cross Validation using 10 folds, where all data is used for testing and for training, and Split Validation with a split ratio of 70%, which only splits the data into one training and one test set. Since the dataset was balanced, there was no need to apply sampling techniques, so the Data Approach (DA) used was the one without sampling techniques. In order to evaluate which attributes were the most relevant to predict the existence of CVD, several scenarios were created:

- S1: All attributes
  {*All attributes*}
- S2: Excluding attributes with weights of 0.000
  {*age_years, height, weight, gender, ap_hi, ap_lo, cholesterol, gluc*}
- S3: Including attributes with weights above 0.010
  {*age_years, weight, ap_hi, ap_lo, cholesterol*}.

As mentioned before, there was only one target variable named *cardio*.

In short, a Data Mining Model (*DMM*) can be described as belonging to an Approach (*A*), being composed by a Scenario (*S*), a Data Mining Technique (*DMT*), a Sampling Method (*SM*), a Data Approach (*DA*) and a Target (*T*).

$$DMM = \{A, S, DMT, SM, DA, T\}$$

Specifically in this work, A = {*Classification*}; S ={*S1, S2, S3*}; *DMT* = {*DT, DT optimized, RI, RF, DL*}; *SM* = {*Cross Validation, Split Validation*}; *DA* = {*Without Undersampling and Oversampling*}; T = {*cardio*}.

Evaluation

In this stage, the performance of each DMM was assessed through some metrics to assure the evaluation of the quality of the models, guaranteeing the reliability of the results. These metrics are numerical measures, obtained from the confusion matrix, that quantify the performance of a given classifier. This matrix presents the number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) [4]. In this project, the following evaluation metrics were used:

- *Accuracy* - correctly TP classified instances (1)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- *Precision* - measure of a classifier's exactness (2)

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- *Specificity* - correctly TN classified instances (3)

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

- *Sensitivity* - measure of a classifier's completeness (4)

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

To understand the metrics' meaning in this context, they can be answered to the following questions:

- *Accuracy*: How many people were correctly labeled out of all the people?
- *Precision*: How many of those labeled has having CVD had actually CVD?
- *Specificity*: Of all the people who are healthy, how many of those were correctly predicted?
- *Sensitivity*: Of all people who have CVD, how many of those were correctly predicted?

In addition, the Area Under the ROC Curve (AUC) has also been included, which represents the degree of separability of the model, ranging from 0 to 1, the higher the AUC, the better the model is at distinguishing between classes.

Tables 3 and 4 present the models that achieved the best results of *accuracy*, *precision*, *sensitivity* and *specificity*.

Overall, it can be noticed that the best values of *accuracy*, *precision*, *sensitivity* and *specificity* correspond to 73.54%, 80.80%, 72.09% and 86.73%, respectively. In order to choose the most suitable model, a threshold was introduced, which combines the four metrics: *accuracy* $\geq$ 71%, *precision* $\geq$ 75%, *sensitivity* $\geq$ 68% and, finally, *specificity* $\geq$ 77%. The threshold values were defined to find a balance between all the metrics according to the results obtained.

Table 5 exhibits the models that achieved the threshold previously defined and Fig. 4 displays the corresponding confusion matrices.

Table 3 Best results obtained by DMT using cross validation

Cross validation								
DMT	Accuracy		Precision		Sensitivity		Specificity	
DT	S3	72.55%	S1; S2	73.55%	S3	70.77%	S1; S2	75.24%
DT optimized	S1	73.14%	S1	75.35%	S3	69.22%	S1	77.63%
RI	S1	72.98%	S3	77.32%	S1	68.87%	S3	81.82%
RF	S1	71.95%	S2	80.11%	S3	61.39%	S2	85.75%
DL	S2	71.46%	S2	67.86%	S2	81.33%	S2	61.66%

Discussion

The analysis of the obtained results allows to understand which of the SMs presented the highest performance scores. Generally, *Split Validation* showed better results than *Cross Validation*. These two SM are similar, being *Split Validation* identical to one iteration of the *Cross Validation*. However, they are usually used with different intents. *Cross validation* is commonly used when it is wanted a more thoroughly tested models, impacting the computation of the process when data is very large - which is the case of this particular study - but being a suitable validation in the case of life-or-death. In contrast, *Split Validation* is adequate for very large data and very complex preprocessing processes, due to the fact that it can accept some uncertainty about the robustness of the model. Therefore, this last SM probably showed better scores since it could have neglected some valuable information when computing information.

After assessing the scenarios applied on the models, it can be noticed some variation among them. Comparing S1 to S2, depending on the DMT, performance values have worsened in almost every situation - with the exception of the DT algorithm, where the performance values remained the same in both SM, and the DL algorithm, where the values improved. Thus, removing attributes such as *smoke*, *alco* and *gender*, even though they had weight values of 0, as the dataset only has 13 attributes, could affect the prediction negatively. Proceeding to the same comparison between S2 and S3, in most cases the results don't improve, except in DT and in some metrics of RF. Finally, comparing S1 with S3 - the two most contrasting scenarios - it is verified a similar behaviour of S1-S2 comparison: DT metric values have improved and in the other algorithms some values have improved and others have worsened. Therefore, making assumptions among DMTs is quite challenging when values vary in such an uncertain way.

Table 5 shows that the DMT that performed best was DT, as expected, given the results presented by DT in Fig. 3. In order to obtain higher results, some parameters were optimized, and once again, the expected was obtained, i.e, the results improved for all the evaluation metrics. Yet, the values did not enhance significantly. Other parameters may have more impact when improving the process, but due to the large number of instances, the time of computing these optimizations would increase exponentially. In addition, S1 was the scenario that stood out among the three, although S2 and S3 also made it to the best results table. Finally, the SM that performed best was the *Split Validation*, which is consistent with the conclusions outlined above. Regarding the evaluation metrics, it was obtained values of approximately 0.7 for Accuracy and Sensitivity and values of approximately 0.8 for Precision, Specificity and AUC. Figure 4 shows that, although satisfactory, the results obtained are not ideal and are not sufficient to be implemented in clinical settings where excellent performance is required given the considerable number of FN and FP.

Table 4 Best results obtained by DMT using split validation

Split validation								
DMT	Accuracy		Precision		Sensitivity		Specificity	
DT	S3	73.02%	S1; S2	76.18%	S3	68.24%	S1; S2	80.40%
DT optimized	S1	73.54%	S1	75.82%	S3	72.09%	S1	78.16%
RI	S1	73.22%	S3	76.67%	S1	70.77%	S3	80.37%
RF	S1	71.91%	S2	80.80%	S1	57.38%	S2	86.73%
DL	S2	71.94%	S3	68.85%	S1	81.85%	S2	64.18%

Table 5 Best models achieving the defined threshold

DMT	S	SM	Accuracy	Precision	Sensitivity	Specificity	AUC
DT	S3	SV	73.02%	75.31%	68.24%	77.77%	0.757
DT opt	S1	CV	73.14%	75.30%	68.61%	77.63%	0.782
DT opt	S1	SV	73.54%	75.82%	68.89%	78.16%	0.788
DT opt	S2	SV	73.51%	75.71%	68.97%	78.01%	0.789

For this particular study, the metrics used to compare the performance of the models don't have the same importance. Realizing if a person was correctly diagnosed as having CVD (precision) and how many of the diseased people were correctly predicted (sensitivity) are more relevant than knowing how many, of all the people, were correctly labeled (accuracy) and how many of the healthy people were predicted as being healthy (specificity). In other words, the most relevant criteria to choose the most suitable models correspond to precision and sensitivity. The reason behind this consideration is the fact that, in this particular case, it is preferable to diagnose a person with CVD when that is not true (FP) than to predict that a person don't have CVD when, in fact, it has (TN). Thus, and due to the critical nature of this problem, FN must be avoided at all costs and models that can best identify TP, and consequently can achieve better sensitivity results, should be prioritized. Hence, a threshold was defined to filter the models that could ensure better results.

In the end, it was possible to conclude that the most suitable model, out of the 30 models created, was DMM = {*Classification, S1, DT optimized, Split Validation, Without Under and Oversampling, cardio*}. This DMM achieved the highest values of Precision (75.82%), Accuracy (73.54%) and Specificity (78.16%) and the second highest values of Sensitivity (68.89%) and AUC (0.788).

Conclusions and future work

In conclusion, this project proved that by using real data from Electronic Health Records (EHR), it is viable to use DMMs to predict the existence or not of CVDs. For some of the constructed DMMs, it was possible to achieve scores over 70%, which is considered as being satisfactory. The best induced method corresponded to the first scenario of attributes, with the DT optimized technique and using the *Split Validation* method. Results could be more indicative of CVDs if more attributes were added, improving the variability and density of information.

In this particular case, the number of false positives must be minimized as much as possible. So, although the results obtained were satisfactory, there is still some future work to be done before implementing a decision-making support system based on one of the models induced during this work in a hospital environment, such as applying more robust algorithms, like neural networks, to boost the performance of the model. This technique presents a considerable manipulation range given the parametrization that can be applied.

In addition, it could be suggested to include datasets with more diverse data and from different hospital facilities of various regions to identify patterns in the data at national level and thus verify the generalization of the model.

n = 20 827	Predicted No	Predicted Yes
Actual No	8124	2322
Actual Yes	3297	7084

a

n = 69 425	Predicted No	Predicted Yes
Actual No	27030	7790
Actual Yes	10861	23744

b

n = 20 827	Predicted No	Predicted Yes
Actual No	8165	2281
Actual Yes	3230	7151

c

n = 20 827	Predicted No	Predicted Yes
Actual No	8149	2297
Actual Yes	3221	7160

d

Fig. 4 Confusion matrices for the best models achieving the defined threshold

Additionally, more experiments could be done on using different parameters and DMTs.

Author Contributions Conceptualization, B.M.; Methodology, B.M., D.F., and C.N.; Software, B.M.; Validation, D.F., and C.N.; Writing: B.M, and D.F.; Project Administration, A.A, and J.M.; Funding Acquisition, A.A, and J.M.

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

Compliance with Ethical Standards

Conflict of interests Bárbara Martins, Diana Ferreira, Cristiana Neto, António Abelha, and José Machado declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Cardiovascular diseases (cvds). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
2. Anderson, K.M., Odell, P.M., Wilson, P.W., and Kannel, W.B., Cardiovascular disease risk profiles. *American Heart Journal* 121(1):293–298, 1991.
3. Brito, C., Esteves, M., Peixoto, H., Abelha, A., and Machado, J., A data mining approach to classify serum creatinine values in patients undergoing continuous ambulatory peritoneal dialysis. *Wirel. Netw*:1–9, 2019.
4. Ferreira, D., Silva, S., Abelha, A., and Machado, J., Recommendation system using autoencoders. *Appl. Sci.* 10(16):5510, 2020.
5. Jothi, N., Husain, W. et al., Data mining in healthcare–a review. *Procedia Computer Science* 72:306–313, 2015.
6. Neto, C., Brito, M., Lopes, V., Peixoto, H., Abelha, A., and Machado, J., Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients. *Entropy* 21(12):1163, 2019.
7. Parva, E., Boostani, R., Ghahramani, Z., and Paydar, S., The necessity of data mining in clinical emergency medicine; a narrative review of the current literatrue. *Bulletin of Emergency & Trauma* 5(2):90, 2017.
8. Sousa, R., Ferreira, D., Abelha, A., and Machado, J., Step towards monitoring intelligent agents in healthcare information systems. In: *World Conference on Information Systems and Technologies*, pp. 510–519: Springer, 2020.
9. Thomas, H., Diamond, J., Vieco, A., Chaudhuri, S., Shinnar, E., Cromer, S., Perel, P., Mensah, G.A., Narula, J., Johnson, C.O. et al., Global atlas of cardiovascular disease. *Global heart* 13(3), 2018.
10. Timmis, A., Townsend, N., Gale, C.P., Torbica, A., Lettino, M., Petersen, S.E., Mossialos, E.A., Maggioni, A.P., Kazakiewicz, D., May, H.T. et al., European society of cardiology: cardiovascular disease statistics 2019. *European Heart Journal* 41(1):12–85, 2020.
11. Ulianova, S., Cardiovascular disease dataset. https://www.kaggle.com/sulianova/cardiovascular-disease-dataset, 2019.