Improving the Effectiveness of Heart Disease Diagnosis with Machine Learning

Catarina Oliveira[1[0000−0002−3325−6016]], Regina Sousa[1[0000−0002−2988−196X]], Hugo Peixoto[1[0000−0003−3957−2121]], and José Machado[1[0000−0003−4121−6169]]

Centro Algoritmi, University of Minho, Braga, 4710, Portugal
a88327@alunos.uminho.pt, regina.sousa@algoritmi.uminho.pt, {hpeixoto,jmac}@di.uminho.pt

Abstract. Despite technological and clinical improvements, heart disease remains one of the leading causes of death worldwide. A significant shift in the paradigm would be for medical teams to be able to accurately identify, at an early stage, whether a patient is at risk of developing or having heart disease, using data from their health records paired with Data Mining tools. As a result, the goal of this research is to determine whether a patient has a cardiac condition by using Data Mining methods and patient information to aid in the construction of a Clinical Decision Support System. With this purpose, we use the CRISP-DM technique to try to forecast the occurrence of cardiac disorders. The greatest results were obtained utilizing the Random Forest technique and the Percentage Split sampling method with a 66 percent training rate. Other approaches, such as Naïve Bayes, J48, and Sequential Minimal Optimization, also produced excellent results.

Keywords: Heart Disease · Classification · Data Mining · Machine Learning · Decision Support Systems

1 Introduction

A "heart disease" is a catch-all term for a wide range of conditions that affect the structure and function of the heart. It is important to remember that all heart diseases are Cardiovascular Diseases, but not all CVDs are heart diseases. Coronary heart disease is the most common type of heart disease, killing 360.900 people in 2019. Heart disease is the leading cause of death in the United States for men, women, and people of most racial and ethnic groups[1, 2].

Every year, approximately 659.000 people in the United States die from heart disease, accounting for one out of every four deaths. In the United States, someone dies from cardiovascular disease every 36 seconds, and someone has a heart attack every 40 seconds. In terms of costs, heart disease cost the United States approximately $363 billion per year between 2016 and 2017 [2].

Focusing now on CVDs, which involve not only the heart but also the blood vessels, these were responsible for an estimated 17.9 million deaths in 2019, accounting for 32% of all global deaths, and remain the leading cause of death

globally. Eighty-five percent of these deaths were caused by heart attacks and strokes (cerebrovascular diseases). Tobacco use, unhealthy diet and obesity, physical inactivity, harmful use of alcohol, diabetes, high blood pressure, and others are all risk factors for CVDs that should be considered when performing patients' exams[3, 4].

Because heart diseases claim so many lives each year and cost so much money to countries, it is critical to keep track of people's health in order to make an accurate diagnosis or choose the best treatment available.

This is where Machine Learning (ML) and Data Mining (DM), two features that have revolutionized the Decision Support Systems paradigm in Healthcare, come into play. There are Knowledge-Based Clinical Decision Support Systems (CDSS), which are typically divided into three components: the knowledge base, the inference or reasoning engine, and Non-knowledge-Based CDSS, which use ML to allow the computer to learn from previous experiences or recognize patterns in clinical data [5].

The primary goal of this paper is to determine the presence or absence of a heart disease in patients using data collected from their clinical records and any hidden knowledge they may have. To be successful, the current work required some prior research on this theme as well as existing work on it, as well as familiarity with the CRISP-DM methodology.

2 State of the Art

Since heart diseases have such a big impact in today's society, many are the studies around this theme and around DM techniques allied to Clinical Decision Support Systems. Therefore, in this section, a few studies will be mentioned in order to give the reader a better understanding of what already as been done and studied and the background which inspired this paper.

Pattekari et al. developed a prototype Heart Disease Prediction System using Naïve Bayesian Classification technique and defended that this was the most effective model to predict patients with heart disease. The data source was linked to questionnaires that contemplated many attributes that will be taken into consideration in this paper, such as age, sex, blood pressure, blood sugar, and others. In fact, these medical profiles could predict the likelihood of patients getting a heart disease because they enabled significant relationships between medical factors related to heart disease to be established [6].

Esfahani et al. used a new DM technique for cardiovascular disease detection which consisted in a fusion strategy of the three best classifiers in terms of the result achieved on the F-Measure value. Therefore, Neural Network, Rough Set and Naïve Bayes were combined by a weighted majority vote and achieved an F-Measure of 86.8%, a better result than when comparing with the F-Measure values of each classifier independently (Neural Network alone achieved an F-Measure of 86.1%, Rough Set achieved 85.7% and Naïve Bayes with 84.6%) [7].

Abdullah and Rajalaxmi developed a DM model using Random Forest Classifier in order to improve not only the prediction accuracy, but also, in order to

investigate some events related to (Coronary) Heart Disease. The results showed that this classification was successful in terms of predicting the events and the risk factors related to it and even had better results when compared to Decision Trees, used in other similar studies [13].

Almustafa performed a comparative analysis of different classifiers for the classification of a heart disease dataset for positive and negative diagnosed participants and the results ended up being very promising in terms of accuracy for the K-NN (K=1), Decision Tree J48 and JRip classifiers when compared to others, mentioned earlier, such as Naïve Bayes and SVM [9].

Martins et al. also mentioned in their study that not all metrics had the same importance and that realizing if a patient was correctly diagnosed with CVD (precision) and the amount of diseased patients who were correctly predicted (sensitivity) were more relevant than knowing the amount, of all the patients, who were correctly labeled (accuracy) and the amount of healthy people who were predicted as being healthy (specificity). A threshold was also defined as the combination of the four metrics mentioned, in order to filter the most suitable models [10].

3 Data Mining Approach

The main aim of this study was to develop a solution that would be able to predict the presence of a heart disease in patients through knowledge hidden in their medical records. Indeed this is extremely important due to the problematic in question and because of the impact it has in people's lives and in Healthcare systems globally. In order to conduct this study, WEKA software was used.

In order to achieve such results, the starting point of this work was the CRISP-DM methodology. This methodology counts with a flexible sequence of six phases, such as Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. All this phases allowed the construction of a DM model to be later used and to deal with the real world problems, in this case, related to the prediction of heart diseases. An overview representation of the steps in this work is presented in Figure 1.

3.1 Heart Disease Dataset

The dataset used to develop this work was the result of the combination of 5 different datasets over the common features already available independently [11]. The total number of observations was 1190, however, since there were 272 duplicated observations, the final dataset only counted with 918 observations. Table 1 presents a brief description of the dataset's attributes and the Table 2 counts with an analysis of the same attributes. There were no missing values. Adding to this, the distribution of the class if of 44,7% with no presence of heart desiease, i.e. normal individuals, and 55,3% with presence of heart disease, i.e. not normal. This indicates that the dataset is well balanced.

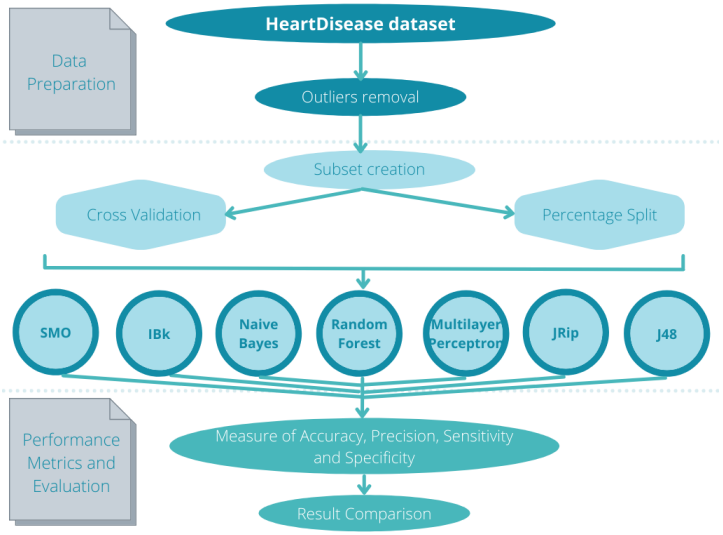

Fig. 1. Overview of the steps in this study.

Table 1. Attribute Description of Heart Disease Dataset

No.	Attribute	Type	Description
1	Age	Numeric	Age of patient [years]
2	Sex	Nominal	Sex of patient [M: male; F: female]
3	ChestPainType	Nominal	[TA: Typical Angina; ATA: Atypical Angina; NAP: Non-Anginal Pain; ASY: Asymptomatic]
4	RestingBP	Numeric	Resting blood pressure [mm Hg]
5	Cholesterol	Numeric	Serum cholesterol [mm/dl]
6	FastingBS	Nominal	Fasting blood sugar [1: if FastingBS > 120 mg/dl; 0: otherwise]
7	RestingECG	Nominal	Results [Normal; ST: ST-T wave abnormality, LHV: probable/definite left ventricular hypertrophy]
8	MaxHR	Numeric	Maximum heart rate achieved [Numeric value between 60 and 202]
9	ExerciseAngina	Nominal	Exercise-induced angina [Y: Yes; N: No]
10	Oldpeak	Numeric	ST [Numeric value measured in depression]
11	ST_Slope	Nominal	The slope of the peak exercise ST segment [Up: upsloping; Flat; Down: downsloping]
12	HeartDisease	Nominal	Output class [1: heart disease; 0: normal]

Table 2. Attribute Analysis of Heart Disease Dataset

No.	Unique	Distinct	Max/Most	Min/Least	Average	Deviation	Distribution
1	3	50	77	28	53.511	9.433	-
2	0	2	-	-	-	-	M(79%); F(21%).
3	0	4	ASY	TA	-	-	ASY(54%); NAP(22.1%); ATA(18.8%); TA(5%).
4	14	67	200	0	132.397	18.514	-
5	66	222	603	0	198.8	109.384	-
6	0	2	-	-	-	-	0(76.7%); 1(23.3%).
7	0	3	Normal	ST	-	-	Normal(60.1%); LVH (20.5%); ST (19.4%).
8	19	119	202	60	136.809	25.460	-
9	0	2	-	-	-	-	N(59.6%); P(40.4%).
10	15	53	-2.6	6.2	0.887	1.067	-
11	0	3	Flat	Down	-	-	Flat(50.1%); Up(43%); Down(6.9%).
12	0	2	-	-	-	-	1(55.3%); 0(44.7%).

3.2 Data Preparation

In this step, there was a need to prepare and clean the data by eliminating duplicated data, removing outliers, dealing with missing values and other in-

consistencies. As mentioned previously in the Data Understanding stage, this dataset had no missing values and the duplicated data had already been removed. Adding to that, no inconsistencies were found, therefore, the main focus was detecting outliers using WEKA's *InterquartileRange* filter and then eliminating these instances.

3.3 Modeling

With the data already prepared, it was possible, in this stage, to define the Data Mining Model (DMM). DMM can be described through a few aspects such as the type of approach (A), the set of scenarios considered (S), the chosen DM techniques (DMT), the sampling methods used (SM), the data approaches followed (DA) and finally the target variable (T). The number of generated simulations can be calculated using Equation 1 [12].

$$DMM_n = A_f \times S_i \times DMT_y \times SM_c \times DA_b \times T_t \tag{1}$$

For this work it was defined that:

- A={Classification}
- T={HeartDisease}
- S={S1,S2}
- DMT={Naïve Bayes (NB), Sequential Minimal Optimization(SMO), RandomForest (RF), JRip, J48, IBk, MultilayerPerceptron(MP)}
- SM={Cross-validation 10 Folds, Percentage Split 66%}
- DA={Without Oversampling and Undersampling}

Where:

- S1={all attributes}
- S2={Age, Sex, ChestPainType, Cholesterol, MaxHR, ExerciseAngina, Oldpeak, St_Slope, HeartDisease}

Therefore, and having Equation 1 in mind, 28 simulations were generated (1 [A] x 1 [T] x 2 [S] x 7 [DMT] x 2 [SM] x 1 [DA]).

In order to generate S2, WEKA's supervised filter *AttributeSelection* (CfsSubsetEval) was used. This filter is responsible for selecting only the most relevant attributes and, therefore, reducing the number of attributes that have to be analyzed.

The DMT chosen were NB, SMO, RF, JRip, J48, IBk and MP. This way it would be possible to evaluate which of the DTMs mentioned in the different papers previously worked best for this situation.

The SM used were cross validation with 10 folds and percentage split with 66%. Percentage Split is helpful for getting a fast impression of a model's performance. According to the literature, a common split value for train and test sets is 66 percent to 34 percent. All other configurations were used as WEKA's default.

In terms of DA and since the class was balanced, there was no need to follow approaches such as *Oversampling* or *Undersampling*.

3.4 Evaluation

Performance metrics play a very important role at this stage, since they are responsible for the validation of the result's reliability obtained with the different algorithms. The performance metrics considered in this study were:

- *Accuracy:* Correctly true positive (TP) classified instances. [10] In a more practical way it is translated to the amount of patients who were correctly labeled out of the total patients in study. This value can be obtained through Equation 2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- *Precision:* It measures the classifier's exactness. [10] It is the amount of patients who really had heart disease out of all the labeled as having it. Precision can be obtained with the help of Equation 3

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- *Sensitivity:* It measures the classifier's completeness. [10] It is the amount of patients who were correctly predicted as having heart disease out of all the patients who had heart disease. This value can be calculated by using Equation 4.

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

- *Specificity:* Correctly true negative (TN) classified instances. The amount of healthy patients who were predicted as so, out of all healthy patients. This value can be obtained by using Equation 5.

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

Similar to what was mentioned in the related works section of this piece, even though all these four performance metrics have a great impact and importance for the result's credibility, for this study and because of its Healthcare related theme, precision and sensitivity, are the key most relevant ones. Indeed, it is much more crucial to detect correctly a patient that has a heart disease when compared to a healthy patient who was wrongly labeled as a carrier of a heart diease, simply because the wrong diagnosis in one might be fatal and in the other, at least most of the time, isn't.

The best values for accuracy, sensitivity, specificity and precision , according to each technique used, are presented in terms of percentage in Table 3.

Despite these individual best values, it was of interest to find the best overall results and, therefore, a threshold was defined. The best results to be considered would be those that had all the performance metrics above the average, that contemplated the 28 simulations, of each performance metric. In a more practical

way, the only situations to achieve the title of overall best results would be those that had an accuracy > 85,7%, a sensitivity > 87,7%, a specificity > 83,1% and lastly a precision > 85,6%. However, because JRip and IBk did not achieve very good results when compared to others, the average for each metric decreased and 11 out of the 28 situations had its metrics above the conditions previously mentioned.

In order to reduce this number and do a better filtering, out of these 11 best situations, an average was calculated for the 4 performance metrics of each situation. The top 5 best results of these averages determined the top 5 overall best results which are represented in Table 4.

Since it was mentioned before that the key most relevant performance metrics were precision and sensitivity, another variation of this filtering was done. While calculating the average, the weight of these parameters was duplicated, in other words, they were considered twice, in order to represent its' importance. Despite this new filtering, the top 5 best results did not change, only the positions in this ranking did, with No.2 switching positions with No. 3 and No.4 switching with No.5, as it is possible to observe when comparing the results between Table 4 and Table 5.

Table 3. Best values for each performance metric according to each technique

DM Technique	Scenario	Sampling Method	Accuracy	Sensitivity	Specificity	Precision
Naïve Bayes	S2	Percentage Split	88,1	-	87,3	-
	S1		-	90,4	-	88,7
Random Forest	S1	Percentage Split	90,4	93,8	85,8	90,4
	S2		-	-		-
SMO	S1	Percentage Split	89,1	90,4	87,3	89,1
JRip	S2	Percentage Split	85,5	91,0	-	85,6
		Cross Validation	-	-	80,4	-
J48	S1	Percentage Split	90,0	93,2	85,8	90,1
IBk	S1	Percentage Split	85,5	85,9	85,1	85,6
MultilayerPerceptron	S2	Percentage Split	87,5	89,8	84,3	87,4

Table 4. Top 5 overall and above threshold results unconsidering the importance of each metric

No.	DM Technique	Scenario	Sampling Method	Metrics' Average	Accuracy	Sensitivity	Specificity	Precision
1	Random Forest	S1	Percentage Split	0,913	0,904	0,938	0,858	0,904
2	J48	S1	Percentage Split	0,902	0,887	0,904	0,866	0,887
3	Naïve Baye	S1	Percentage Split	0,897	0,900	0,932	0,858	0,901
4	SMO	S1	Percentage Split	0,892	0.881	0.887	0.873	0,881
5	Naïve Baye	S2	Percentage Split	0,890	0,891	0,904	0,873	0,891

Table 5. Top 5 overall and above threshold results considering the importance of each metric

No.	DM Technique	Scenario	Sampling Method	Metrics' Average
1	Random Forest	S1	Percentage Split	0,911
2	Naïve Bayes	S1	Percentage Split	0,897
3	J48	S1	Percentage Split	0,896
4	Naïve Bayes	S2	Percentage Split	0,893
5	SMO	S1	Percentage Split	0,890

4 Discussion and Contributions

In this study, we evaluated the possibility of determining the presence of cardiac disease by employing data mining. For this purpose, the CRISP-DM methodology was followed, and WEKA software was used. Two scenarios were evaluated, one where all attributes were taken into account (S1) and the other where attribute selection was performed using WEKA software (S2).

The CfsSubsetEval evaluator assessed feature selection in S2 provided nine attributes. The method calculates each attribute's correlation with the degree of redundancy between the attributes, choosing the ones with the best correlation. Based on the results obtained, scenario (S1), which includes all attributes, gave the best results, unlike scenario (S2), where RestingBP, FastingBS, and RestingECG were excluded. Despite their low correlation, the excluded attributes provided necessary information that would have improved the prediction accuracy of the algorithm. Based on this result, it would be recommended to repeat the task using other methods of attribute selection by WEKA, and compare their respective impacts on the accuracy prediction.

The data set was evaluated by 10-fold cross-validation and standard split-percentage with 66% training and 34% for testing. Split-percentage showed the best results compared to cross-validation. By Cross-validation, the dataset was divided into ten equally sized segments. Then ten iterations took the place of training, followed by testing, ensuring that a different segment of data was used for testing in each iteration. Not all models were able to be evaluated by cross-validation, and it is recommended in future work to evaluate the rest of the methods with it. Using cross-validation, we could better understand and evaluate the prediction accuracy and variance of the dataset in the models.

The evaluation results obtained by split percentage were positive, especially in the S1 scenario. Random-Forest was the technique that showed better results than the rest of the presented methods. It achieved the best overall results with an accuracy of 90,4%, a sensitivity of 93,8%, a specificity of 85,6%, and a precision of 90,4%. NB, J48, and SMO algorithms also achieved excellent results. The advantage of these techniques is that they are not as slow in training as the random forest technique. In the case of larger datasets could be more beneficial [15]. The two techniques that consistently and independently from the situation showed the worst results when compared to others were JRip and IBk. Indeed, JRip implements propositional rules; these are rules that follow an IF - Then

structure, and the lower results of the metrics might be an outcome of a not-so-strong rule [16]. On the other hand, the poor results related to IBk might be a consequence of either a lack of representation or even a not meaningful and efficient distance measure.

5 Conclusions and Future Work

To increase the effectiveness of a heart disease diagnosis and, consequently, to improve the patient's quality life and also to reduce costs in the Healthcare systems through the implementation of a Clinical Decision Support System, different data mining techniques, sampling methods and scenarios were tested.

In this study, RF technique, through a 66% Percentage Split and using S1 achieved the best results and, therefore, should be used in the future. However, NB, J48 and SMO also achieved great results and must be on the table for future works and implementations. In terms of sampling methods, Percentage Split and Cross Validation were studied, and the best results were mainly and curiously associated with Percentage Split.

Future work can be done with a dataset closer to what would be found in reality, in other words, with a larger dataset and also with more instances, in order to assess the previous conclusions in terms of the best techniques and sampling methods. Assessing other ways to differentiate the weights of importance of the 4 performance metrics considered in this study can also be a field for study.The same goes for other performance metrics that weren't studied in this paper.

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. Know the Differences: Cardiovascular Disease, Heart Disease, Coronary Heart Disease, `https://www.nhlbi.nih.gov/sites/default/files/media/docs/Fact_Sheet_Know_Diff_Design.508_pdf.pdf`. Last accessed 27 Dec 2021.
2. Heart Disease Facts, *Centers for Disease Control and Prevention, https://www.cdc.gov/heartdisease/facts.htm. Last accessed 26 Apr 2022.*
3. *Cardiovascular diseases (CVDs).* World Health Organization, `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`.Last accessed 27 Dec 2021.
4. Ahmad T, Munir A, Bhatti SH , Aftab M, Raza MA: Survival analysis of heart failure patients: A case study. *PLoS ONE 12(7): e0181001* **2017**.
5. Berner ES, *Overview of Clinical Decision Support Systems*;2nd edn; Springer; **2007**; vol. 233; 4–8.
6. Pattekari SA, Parveen A, Prediction system for heart disease using Naïve Bayes. *International Journal of Advanced Computer and Mathematical Sciences*; **2012**, Citeseer *3*, 290–294.

7. Esfahani HA, Ghazanfari M, Cardiovascular disease detection using a new ensemble classifier. In *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*; **2007**; IEEE; 1011–1014.https://doi.org/10.1109/KBEI.2017.8324946
8. Abdullah SA, Rajalaxmi RR, A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier. *IJCA Proceedings on International Conference in Recent trends in Computational Methods, Communication and Controls (ICON3C 2012)*; **2012**, *3*, 22–25.
9. Almustafa KM, Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics 21*; **2020**, *278*, `https://doi.org/10.1186/s12859-020-03626-y`
10. Martins B, Ferreira D, Neto C, Abelha A, Machado J, Data mining for cardiovascular disease prediction. *Journal of Medical Systems* **2021**, Springer, *45*.
11. fedesoriano, Heart Failure Prediction Dataset. *kaggle* **2021**, `https://www.kaggle.com/fedesoriano/heart-failure-prediction?select=heart.csv`. Last accessed 20 Dec 2021.
12. Fonseca F, Peixoto H, Miranda F, Machado J, Abelha A, Step towards prediction of perineal tear. *Procedia computer science*; **2017**, Elsevier *113*, 565–570.
13. Abdullah SA, Rajalaxmi RR, A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier. *IJCA Proceedings on International Conference in Recent trends in Computational Methods, Communication and Controls (ICON3C 2012)*; **2012**, *3*, 22–25.
14. Nogueira M, Ferreira D, Neto C, Abelha A, Machado J, Data Mining for the Prediction of Fetal Malformation Through Cardiotocography Data tear. *Information Technology and Systems - ICITS 2021*; **2021**, Springer *1331*, 60–69.
15. Peixoto H, Francisco A, Duarte A, Esteves M, Oliveira S, Lopes V, Abelha A, Machado J, Predicting postoperative complications for gastric cancer patients using data mining.**2019**, Springer-Verlag.
16. Melo I, Medeiros N, Silva I, Lira L, Moraes R, Evaluation of the performance of the JRIP algorithm in the classification of heart disease diagnosis. **2019**, IV National Congress of Research and Teaching in Sciences.
17. G Prashant, Cross Validation in Machine Learning, *Towards Data Science*. **2017**, `https://medium.com/towards-data-science/cross-validation-in-machine-learning-72924a69872f`. Last accessed 18 Mar 2022.