International Journal of Data Informatics and Intelligent Computing (IJDIIC)
Vol. 4, No. 4, December 2025, pp. 11~27
DOI: 10.59461/ijdiic.v4i4.229

Machine learning approaches using correlation filters for heart failure diagnosis: a comparative study of supervised techniques

Vitória S. Souza[1], Danielli A. Lima[1]
[1]Federal Institute of Education, Science and Technology of the Triângulo Mineiro (IFTM), Computational Intelligence and Robotics Laboratory (LICRO), Patrocínio Campus, Brazil.

Article Info

Article history:

Received August 28, 2025
Revised October 10, 2025
Accepted October 18, 2025

Keywords:

Classification
Machine Learning
Cardiology
Heart disease
Prediction diagnosis

ABSTRACT

This study addressed how machine learning could be used to detect factors that influenced the probability of survival of patients with heart failure, based on a database with 12 attributes collected from 299 different patients. Along with applying correlation filters, to obtain attributes that may be more important in a certain way for the disease, further assisting in new forms of treatment, and helping to reduce costs for diagnosis. In this study, we evaluated the accuracy of eight data mining algorithms for predicting heart disease using the heart failure dataset. We implement a methodology that includes 100 simulations with 10 correlation filter variations to ensure reliable and robust results. Among the eight classification algorithms, Support Vector Machine and Random Forest provided the best accuracy (84.18%). Considering the averages for all correlation filter variations, the Random Forest algorithm had the highest average (80.07%), and the Probabilistic Neural Network had the worst performance (69.43%). Analysis of other evaluation metrics revealed that our approach using a Multilayer Perceptron with a correlation filter (0.10) was the best alternative with 83.50% accuracy. Therefore, the diagnosis of cardiac insufficiency required only four attributes: creatinine phosphokinase, serum sodium, sex, and hospitalization time. This streamlined approach not only saved time and resources but also enhanced diagnostic efficiency, unlike previous works that use all base attributes for classification. Our findings suggest that data mining techniques can be a useful tool for predicting heart disease, and the proposed method.

Corresponding Author: Danielli A. Lima (e-mail: danielli@iftm.edu.br)

1. INTRODUCTION

The heart is one of the most important organs in the human body, supporting life in humans. Because it is such an important organ, great care is needed, because if you develop a cardiovascular disease that is not well treated, or if there is a setback in the tests or even the person themselves, a fatal event may occur. Cardiovascular diseases can be developed over time or acquired through heredity. The prevention and treatment of these diseases require the use of accurate and reliable techniques for early diagnosis and identification of risk factors. Often, what leads to the development of heart disease is poor nutrition, which is why eating healthy foods is of paramount importance [1]. Examples of some heart diseases are high blood pressure - popularly known as high blood pressure, it is characterized by high levels of blood pressure in the arteries [2]; heart failure-based on a disorder where the heart becomes unable to meet the body's needs, causing restriction of blood flow, congestion of blood in the veins and lungs; acute myocardial infarction - described as myocardial necrosis resulting from abstraction of a coronary artery; among many others heart diseases [3][4].

By applying artificial intelligence (AI) in the field of medicine, together with the analysis of physicians, we can provide significant improvements in the discovery, treatment, and analysis of various diseases. By analyzing a database (DB) collected from patients with heart failure and applying machine learning (ML) algorithms to predict the survival of patients with cardiovascular diseases, we can identify

which factors predominate in the diagnosis, through data science (DC), supported by experts. In this context, data mining (DM) has been widely used to identify patterns and build models for predicting heart disease.

Among data mining techniques, classification is one of the most used to predict the occurrence of heart disease [5][6]. Several classification algorithms can be applied in this context, such as Random Forest, Support Vector Machine, and Naive Bayes, among others. However, choosing the best algorithm for this task can be challenging, due to the complexity of the data and the nature of the problem. Specifically, in this work, we apply eight classification algorithms, namely Decision Tree (DT), Support Vector Machine (SVM), Probabilistic Neural Networks (PNN), Multilayer Perceptron (MLP), Naive Bayes Learner (NBL), Gradient Boosted Learner (GBL), K-Nearest Neighbors (KNN), and Random Forest Learner (RFL) [7][8]. These algorithms are considered different supervised machine learning (SML) techniques, with a focus on classifying heart disease in cases of survival or death. This study hypothesizes that the use of different data mining algorithms can lead to significant variations in the accuracy of predicting heart disease. In addition, it is expected that other metrics, such as Sensitivity, f-measure, and Cohen's kappa, can provide additional information on the performance of algorithms and help in selecting the best model for this task. We believe in the importance of early and accurate prediction of heart disease to provide patients with adequate treatment and improve their quality of life. In addition, choosing the best classification algorithm can help reduce costs with unnecessary tests and treatments, in addition to contributing to the advancement of data mining in health.

The objectives of this research are: (1) to compare the accuracies, using a confusion matrix, of eight data classification algorithms for predicting heart disease based on a public dataset; (2) to evaluate the algorithms' other performance metrics and compare them with accuracy, to understand which technique obtained the best result; and (3) implement a work methodology that allows simulating the variation of the correlation filter and evaluating the influence of this variation on the performance metrics. Finally, (4) the results are discussed and analyzed from the parameters collected by the classification algorithms. Being an existing heart disease, analyzed by [3], our study aims to contribute to the discovery and analysis of the worsening of the disease itself. Based on the methodology used, we intend to identify the possibility of early identification of worsening or not of heart failure. The main difference of this work was the application of correlation filters [9] to reduce the number of parameters used in the classification, thus improving the medical diagnosis. Our article addresses the problem of choosing and evaluating the best machine learning algorithms to predict the clinical outcome of patients with heart failure, seeking to provide physicians and specialists with a method for early diagnosis.

2. LITERATURE REVIEW

This section will present the main definitions of heart diseases and how they can be developed. We will also discuss, in more detail, the attributes referring to the database we use for research. Additionally, we will discuss related works that addressed the use of the same database related to heart failure.

2.1. Definitions of heart disease

Risk behavior for heart disease involves a series of habits and practices that can negatively affect cardiovascular health. Some of the main risk factors include smoking, excessive alcohol consumption, lack of regular physical activity, unhealthy diet, and chronic stress. Cigarette smoking is particularly harmful, as the chemicals in cigarettes can damage artery walls and increase blood pressure, while excessive alcohol consumption can increase the risk of hypertension and heart failure [10]. In addition, lack of regular physical activity can lead to weight gain and high blood pressure, and an unhealthy diet high in fats and sugars can contribute to the development of heart disease [11]. Proper management of these risk behaviors is essential to preventing heart disease and maintaining good cardiovascular health.

Some blood tests can be used to detect serious heart disease. One of the most common is the troponin test, which measures the amount of troponin in the blood. Troponin is a protein released into the blood when heart cells die or become damaged, which can indicate a recent heart attack [12]. Another common test is the B-type natriuretic peptide (BNP), which measures the amount of BNP in the blood [13]. BNP is produced by the heart in response to stress and can indicate heart failure. In addition, the lipid test or lipid profile can be used to measure cholesterol and triglyceride levels in the blood, which can contribute to the development of heart disease [14]. Uric acid can also be measured to identify the risk of heart disease, as high levels are associated with a higher risk of hypertension and coronary artery disease. However, it is important to remember that these blood tests cannot diagnose heart disease alone and must be combined with other tests for a more complete assessment of cardiovascular health.

2.2. Data Science in Cardiovascular Diagnosis

Data science has the potential to help diagnose heart disease in a variety of ways. One is through the analysis of large data sets, such as imaging tests, blood test results, and information about the patient's medical history. Machine learning and artificial intelligence algorithms can be applied to this data to identify patterns and correlations that can help diagnose heart disease. For example, analysis of imaging studies can help identify the presence of plaque in the coronary arteries, while analyzing data from blood tests can help identify biomarkers associated with heart disease.

Additionally, health monitoring data, such as that generated by wearable devices, can be used to identify early signs of heart disease, such as an irregular heartbeat. Data science can also be used to develop risk prediction models, which assess a patient's individual risk of developing heart disease based on various risk factors such as age, family history, smoking, and cholesterol levels. These models can help clinicians identify patients who may be at risk and take preventive measures before the disease develops. In summary, data science offers great potential to improve the diagnosis and treatment of heart disease, improving the accuracy and effectiveness of cardiovascular healthcare.

A study [15] was carried out on the population of Pakistan with heart failure, estimating the survival and mortality rates. We used 200 bootstrap replications, the mean linear predictor slope, the ROC curve, the Cox model, and a nomogram for graphical visualization of survival. According to them, age, ejection fraction, sodium, anemia, blood pressure, and creatine were very significant data for the analysis. With the ROC curve, it was possible to detect that in a longer follow-up time, 81% of the death events were detected, while in a short time, it could only recognize 77%. In addition, the authors highlighted the importance of risk stratification based on clinical and laboratory factors to identify patients with heart failure at higher risk of mortality. This study was relevant because it highlights the importance of survival analysis as a tool to identify risk factors associated with mortality in patients with heart failure. In addition, the study results provide important information to improve risk stratification and management of patients with heart failure in clinical settings. In [3], work already addresses only two clinical parameters for the survival approach of patients with heart failure, using the same basis as [15], which are serum creatinine and ejection fraction, on which the construction of the models was based on machine learning. Ten different methods from different areas of machine learning were applied to predict survival. In the paper [3], the authors described in their study [16] also reached a result that ejection fraction and serum creatinine were the most relevant data for predicting heart failure. Among the five classifiers considered by the author, the decision tree provided the best results, resulting in 80% accuracy.

In the work of [17] authors used to predict patient survival, this study uses new categorization models: Decision Tree (DT), Adaptive Boost Classifier (AdaBoost), Logistic Regression (LR), Random Gradient Classifier (SGD), Random Forest (RF), gradient augmentation classifier (GBM) and an additive tree (ETC), a Naive Bayes Gaussian classifier (GNB) and a support vector machine (SVM). The class asymmetry problem was easily achieved using the synthetic minority noise oxidation (SMOTE) technique. Additionally, the machine learning model is trained on the key features chosen by RF. The results are compared with the results obtained by the machine learning algorithm using the full feature set. Experimental results show that ETC outperforms other models and reaches an accuracy value of 0.92 with SMOTE in predicting survival in cardiac patients. However, for this accuracy, all attributes were used, unlike our work, where we applied filters to reduce the parameter characteristics space.

The paper of [18] presents an IoT-enabled framework secured by Public Key Infrastructure (PKI) titled Cardiac Diagnostics and Demographic Identification (CDF-DI) Resource Systems with meaningful models that recognize various heart disease features related to HF. To achieve this goal, we used statistical and motor fixation techniques to psychoanalyze secondary cardiac data attachment. Elevated levels of Serum Creatinine (SC) and Serum Sodium (SS) can produce kidney problems and are commonly found in patients with HF. The most salient algorithm was the Random Forest (RF), suggesting five key features to stipulate long-suffering survival status with 96% clarity: follow-up months, CS, ejection fraction (EF), creatine phosphokinase (CPK), and platelets. Furthermore, RF selected the fifth prominent features (smoking habits, CPK, platelets, month following, and SC) in category recognition with a clarity of 94%. In addition, the fifth vital characteristics, regarding CPK, SC, subsequent month, platelets, and EF, were significant predictors for the long-suffering age selvage with a clarity of 96%. The Kaplan Meier graph revealed that the elite period mortality occurred in the very advanced age. The recommended resources have possible effects on clinical practice and would support the existing medical sector to recognize the likely survival status of cardiac long-sufferers. The medicine must be stored mainly in the following months: SC, EF, CPK, and platelet score for the survival of the long-suffering in the situation.

The study by [19] makes use of Convolutional Neural Network (CNN) models based on deep learning to evaluate numerical data in the medical field, especially in cases of heart failure. The numerical data is converted into grid images after normalization, and five different CNN models are trained and tested.

The ResNet18 model achieves an accuracy of 95.13%. The results highlight the feasibility of the proposed method, showing that it is possible to classify numerical data in varied fields, such as medicine. They are especially important for identifying patients with heart failure with a high probability of survival and those at risk of worsening the condition. On the other hand, in the work of [20], the random forest learner stood out with an accuracy of 87.21% when using a filter that considered attributes with an area under the curve greater than 0.4, considering values of area under the curve. Additionally, the fuzzy rules learner demonstrated its effectiveness by achieving an accuracy of 84.45% with a filter limit of 0.6, focusing on ejection fraction, serum sodium, time attributes, and class for death events.

In Table 1, we present the main results achieved by the authors to predict the survival of patients with heart failure. Highlighting each method used by them and the main results obtained. From machine learning analysis to the use of frameworks with Convolutional Neural Network models. Providing a quick analysis and understanding of the main results with several analyses for survival in the development of heart disease.

Table 1. Main results of the authors

Author(s)	Main Result
[15]	They studied the population of Pakistan with the disease, highlighting the importance of risk stratification based on clinical factors to identify patients athigh risk of mortality.
[3]	They studied the survival of patients with heart failure using machine learning, focusing on serum creatinine and ejection fraction. With the decision tree, theyachieved an accuracy of 80%.
[17]	With several machine learning models, they achieved superior performance with the ETC model, with an accuracy of 92.62% using the SMOTE technique.
[18]	Using an IoT framework for diagnosing the disease, they highlighted the decision tree algorithm, identifying five characteristics to predict survival with 96%accuracy.
[19]	Used Convolutional Neural Network models along with deep learning to evaluate heart failure data, with a clarity of 95.13% with ResNet18.
[20]	Random Forest achieved 87.21% accuracy using a 0.4 AUC filter. Fuzzy rules achieved 84.45% accuracy, focusing on ejection fraction, sodium, time, anddeath class (AUC filter 0.6).

3. METHOD

In this section, we will present the methodology we used to develop this article and the conclusions that will be presented in the results of this research. The workflows used will be presented, both to exemplify the general research methodology and the methods used in the data mining platform that was used. Furthermore, the algorithms used are presented, and their definitions are assigned. Finally, the correlation filter is presented, and the evaluation metrics are defined so that the experimental results can be presented. With this, we intend to predict a possible worsening of heart failure using machine learning and public data.

First, regarding the methodological approach, the research is considered quantitative, as the attributes are statistically evaluated through data collection for heart diseases that are available online and publicly in the UCI Machine Learning Repository "Heart Failure Clinical Records", available at https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records. Additionally, the nature of the research is considered applied, since data classification algorithms are applied in a structured and specific database. The procedures adopted here were through experiments, in which x = 100 simulations are performed to achieve the best precision and accuracy for the data set studied. The research is explanatory, as we aim to connect the attributes identified through the correlation filter to understand the causes and effects of survival or death of people with heart disease.

3.1. Definition of the database used

There are many cardiovascular diseases, which are diseases that affect the heart and can lead to death in some cases if not treated well. Acquired over time due to poor diet, lack of exercise, or stress. Among the various cardiovascular diseases, we have Acute Myocardial Infarction, characterized by a clot that blocks blood flow to the heart, among others, such as hypertension and rheumatic heart disease [21]. We will use a database with 299 patients, 105 women, and 194 men, aged 40 and not exceeding 95 years old, and having 13 attributes. As shown in Table 2, all variables in the database are detailed, with explanations and

measurements. The data collection took place in 2015, between April and December. As [3] mentions in his study, both had left ventricular systolic dysfunction and previous heart failure.

Each attribute may or may not be relevant for predicting the survival or death of patients, among the 13 attributes one of them is the class DEATH_EVENT which is the classification class, where it will be classified if the patient survived defined as (0), using the string replacer we define it as (YES) and the dead patients defined as (1), and when applying the string replacer we replace with (NO) in relation to the disease.

Table 2. Explanation of the feature and measurement range

Resources	Explanation	Measurement	Range	Median	Mean
Age	Patient's age	Years	[40, 95]	60	64.43
Anemia	Decreased red blood cells or hemoglobin	Boolean	[0, 1]	0	0.43
High pressure	If the patient has hypertension	Boolean	[0, 1]	0	0.43
Creatine phosphokinase	CPK enzyme level in the blood	mcg/L	[23, 7861]	250	1438.29
Diabetes	If the patient has diabetes	Boolean	[0, 1]	0	0.43
Ejection fraction	Blood ejection from the heart	Percentage	[14, 80]	38	40.86
Sex	Woman or man	Binary	[0, 1]	1	0.57
Platelets	Blood platelets	kiloplatelets/ml	[25.01, 850.00]	262000	309585.7
Serum creatinine	Blood creatinine level	mg/dL	[0.50, 9.40]	1.1	2.27
Serum sodium	Blood sodium level	mEq/L	[114, 148]	137	135
Smoker	If the patient smokes	Boolean	[0, 1]	0	0.43
Time	Follow-up period	Days	[4,285]	113	138.43
Death event (Class)	Patient's mortality during follow-up	Boolean	[0, 1]	0	0.32

The other attributes are sex - binary attribute set to (0) - female and (1) - male; anemia, which is characterized by a decrease in red blood cells or hemoglobin [22], the doctor took into account that patients are considered anemic if the hematocrit levels in the blood are less than 36%.

The creatine phosphokinase is an indicator of CPK in the blood. When damage to muscle tissue occurs, CPK flows into the blood. In addition, a high level of CPK in the blood can constitute a heart disease, in the range of (23, 7861) Mcg/L. The serum creatine - based on the level of creatine in the blood, doctors approach serum creatine in the blood to analyze the kidney function; ejection fraction (EF) - defined as the amount of blood that the left ventricle pumps with each contraction, range (14, 80) (%) [23]; age - patient's age; diabetes - if the patient has diabetes, caused by insufficient production of insulin, a hormone that regulates blood glucose boolean value (0, 1) [24].

The high blood pressure characterizes whether or not the patient has hypertension; platelets represent platelets in the blood, also known as thrombocytes, are structures present in the blood that, contrary to what many people think, are not complete cells in range (25.01, 850.00) kilo platelets/mL [25]. The smoking value - represents if the patient smokes, a boolean value (0, 1). Time represents the period of follow-up of the research, ranging from 4,285 days, and serum sodium represents the mineral that serves for the good functioning of the muscles and nerves, and the serum sodium test indicates if the patient has normal levels of sodium in the blood, in a range of 114- 148. As one of the branches of machine learning, supervised learning is based on a model that can henceforth learn from predefined results, making use of the informed data that are well labeled. Thus, being able to train the algorithm to perform some specific tasks, eight different techniques will be applied in this work to predict better results.

3.2. Processing and evaluation of the data used

As shown in Figure 1, data were initially collected from the UCI Machine Learning platform Repository and placed in Excel .xlsx format, and we used the KNIME Analytics Platform to read data from 299 patients. Then, in the data preprocessing phase, for some attributes, the data in string format was transformed into numbers, and numerical data was transformed into strings, which were then performed by a string replacer to make the replacements correctly. Then, in the data preprocessing phase, for some attributes,

string data was transformed into numbers, and numeric data was transformed into strings. With the "number to string" and "string to number" nodes, we can perform the conversion through numerical and text interpretation. They are useful for us to use in some situations where we need to deal with different types of data when analyzing our information, which was then performed by a string substitute to make the substitutions correctly. Next, we performed a filtering process for rows that had inconsistent values for the age attribute; these rows were, in turn, excluded. We then normalize each of the data. The normalization technique allows us to place the values of variables on a single scale, between 0 and 1. This technique is useful so that the values of different variables do not take over the modeling process. With the knime normalization node, we can perform MIN-MAX normalization. It's Formula 1. Where X is the normal value, Xnorm will be our normalized value, Xmin will be the minimum value of the attribute, and Xmax will be the largest value of the attribute. Normalizing the values to within the range of 0 and 1, making the data proportional.

$$Xnorm = \frac{\text{X}-\text{Xmin}}{\text{Xmax}-\text{Xmin}} \quad (1)$$

Denormalization was then performed to do some of the data visualizations and statistics. We leave a node for attributes filtering (if necessary) and apply the linear correlation, which calculates for each pair of selected attributes a correlation coefficient, i.e., a measure of the correlation of the two variables. The correlation filter, in turn, determines which attributes are redundant (i.e., correlated) and filters them out. The output table will contain the reduced set of attributes. After performing all these steps, we apply the X-Partitioner, which represents a cross-validation loop. That is, we reapplied the learning algorithm that was applied with a value of (X = 100) iterations. At the end of the loop, there must be an X-Aggregator to collect the results from each iteration. All nodes in between these two nodes are executed as many times as iterations should be performed. Thus, we chose data mining algorithms for learning and prediction. In this sense, eight algorithms were trained they are DTL, GBL, KNN, MLP, NBL, PNN, RFL, and SVM [26]. Finally, the scorer is calculated, where a confusion matrix is calculated with the number of matches in each cell.

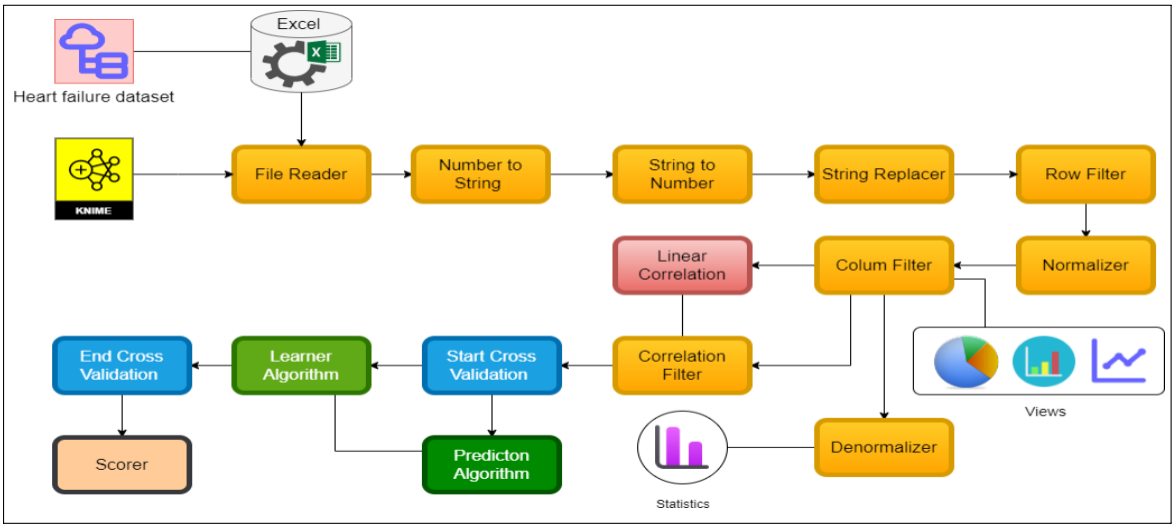

Figure 1. Proposal flowchart for heart disease data mining

With the confusion matrix, we were able to evaluate the performance of the machine learning model. Performing the Comparison of our model with the real results of our data has four parts: (i) False Positive (FP) is when the result is negative but classified as positive, (ii) False Negative (FN) when the result is positive but it classifies as negative, (iii) True Positive (TP) when it is true, and (iv) True Negative (TN) when it is negative. By counting all these terms and obtaining the confusion matrix, it is possible to calculate classification accuracy evaluation metrics.

This is a table where the performance of the classification models is presented. Providing us with some important metrics, such as the accuracy and hit rate that the model obtained, so that we can evaluate its effectiveness. True positive occurs when our model can correctly distinguish an instance as positive when it is true, that is, it "correctly classifies the data, as well as the classification assigned to it in the base", when it hits the classification that is labeled in the data.

3.3. Algorithmic precision assessment

The use of machine learning algorithms has been shown to be effective in assisting in the diagnosis of heart disease. Among the available algorithms, K-nearest neighbor, Decision Tree, Probabilistic Neural Network, Multilayer Perceptron, Random Forest, Naive Bayes, Gradient Boosted, and Support Vector Machine stand out. However, it is important to assess the accuracy of each of these algorithms to ensure their reliability in diagnosing heart disease. In this sense, a methodology is proposed that involves carrying out 100 simulations with variations in the correlation filter to evaluate the precision of each of the eight mentioned algorithms. Next, we describe the methodology to be followed:

1. Perform data preprocessing, including data cleaning and normalization.
2. Separate the database into training (70%) and testing (30%) using the cross-validation technique with X = 100.
3. Apply the Pearson correlation filter to select the most relevant variables in each simulation, varying the correlation threshold in increments.
4. Apply the following machine learning algorithms in each simulation:
   (a) K-Nearest Neighbors (KNN)
   (b) Decision Tree Learner (DTL)
   (c) Probabilistic Neural Network (PNN)
   (d) Multilayer Perceptron (MLP)
   (e) Random forest Learner (RFL)
   (f) Naive Bayes Learner (NBL)
   (g) Gradient Boosted Learner (GBL)
   (h) Support Vector Machine (SVM)
5. Evaluate the accuracy of the models using the accuracy and error rate.
6. Calculate the mean (x) and standard deviation (Std. Dev. s) of the evaluation metrics across all simulations for each algorithm.
7. Compare the results of the different algorithms in terms of performance, taking into account the variance of the correlation filter (CF = 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.50, 0.75, 0.85, 1.00).
8. Evaluate the accuracy of the models using the following metrics:
   (a) sensitivity and specificity
   (b) Cohen's Kappa
   (c) recall and precision
   (d) f-measure
9. Select the algorithm with the best performance in terms of accuracy, considering the variation of the correlation filter (|CF| = 10).

3.4. Correlation Filter

Characterized as a technique widely applied in analyzing the combination of variables in a data set, the correlation filter uses a correlation node to demarcate redundant (correlated) attributes and perform filtering. The output table will contain a reduced set of attributes. This filtering step works as follows: In each attribute of the correlation model, the count of correlated attributes is established based on a threshold value for the correlation coefficient (described in the dialog box). As a result, the attribute that has the most correlated features becomes the one chosen to "survive," and the other correlated attributes are filtered out. This procedure is repeated until no more attributes can be identified. The problem of finding a minimum set of attributes that satisfies the constraints is difficult to solve analytically. However, this applied method is commonly known to be a good approximation. A set of filters with values 1.00, 0.85, 0.75, 0.50, 0.30, 0.25, 0.20, 0.15, 0.10, 0.05 was applied to the eight classification algorithms. When we use filters, they are responsible for defining which database attributes will be most relevant to carry out the classification, that is, which parameters were considered most relevant for the survival analysis. It is possible to observe that some filters used all attributes, and some others performed greater filtering in the specified amount, with a maximum value of nine biomarkers excluded.

3.5. Classification Algorithms

Data mining has proven to be an essential tool in the early detection of heart disease. In this context, the choice of machine learning algorithms is critical to the accuracy of the results. The eight algorithms used in this study are commonly applied in cardiovascular disease prediction and stand out for their ability to extract relevant information from the data and build accurate predictive models. The use of these algorithms together allows a more comprehensive analysis of the data and the selection of the one that best suits the

objectives of the study. Thus, the careful choice of algorithms to be used can be crucial for the prevention and treatment of heart diseases.

SVM: The Support Vector Machine is characterized by training a support vector machine on the initial data, which can draw a hyperplane for each class [27]. One of the main advantages of SVM is its ability to handle high-dimensional data and overlapping classes. Furthermore, the SVM can use kernel functions to transform the data into a more complex feature space and find a more accurate separating hyperplane. However, the proper choice of kernel and SVM parameters is essential for good performance. Furthermore, SVM can be computationally intensive for very large datasets and is not very effective in dealing with noisy or unbalanced data.

KNN: The objective of this algorithm is to classify a new sample based on its proximity to the training data. K-Nearest Neighbors sorts the data based on the training examples that are close in the feature space; all numerical columns, and Euclidean (or Manhattan) distance are used in the implementation [28]. The algorithm works as follows: first, a value is defined for K, which is the number of closest samples that will be used to determine the class of the new sample. Then, to classify a new sample, the algorithm calculates its distance from all training samples. The K closest samples are selected, and the most frequent class among these samples is assigned to the new sample. In this work, the value of K = 9, that is, we use nine nearest neighbors to do the classification.

DTL: In the Decision Tree, several decision points will be created. These points are the “nodes” of the tree, and within each node, the outcome of the decision will be to follow one path or another. The existing path is “branch”. The nodes are responsible for indicating meetings of one or another branch of the flow sequence [8]. The objective is to find the best question to ask at each node to reduce the impurity of the samples in each subset as much as possible. The tree can be pruned, that is, pruned, to avoid overfitting and improve generalization. The decision tree is an intuitive and easy-to-interpret algorithm, which makes it a valuable tool for decision-making in different application areas.

PNN: The Probabilistic Neural Network is a probabilistic neural network, consisting of four layers: the input layer, the pattern layer, the sum layer, and the output layer. In the input layer, samples are encoded into binary arrays. In the patterns layer, a frequency table is constructed for each class of samples. In the summing layer, the frequencies are summed for each class, and the class with the highest sum is chosen as the target class. In the output layer, the posterior probability for each class is calculated and can be used to assess the confidence in the classification. Dealing only with linearly separable problems, its use is suitable for simple classification problems [29].

MLP: The multilayer perceptron is very similar to the perceptron; the difference is that the MLP has several layers of neurons, which may contain hidden layers, interconnected with the desired weights. The input layer receives the input data, and each neuron performs a weighted operation on that data. The hidden layer performs non-linear transformations on the input data, while the output layer generates the final predictions. The connections between the layers are formed by synaptic weights that are adjusted during the neural network training process. Finally, MLP's learning is given through the error backpropagation algorithm.

RFL: The Random Forest creates several decision trees at random. It works by creating multiple decision trees and combining their predictions to generate a more accurate final prediction. Each decision tree is built using a random sample of the original dataset and a random selection of features for each split in the tree. This helps to avoid overfitting and improves model generalizability. When a new data point is entered into the model, it is passed through each decision tree, and the class with the most votes is selected as the final prediction. This means they are used to choose the final result, like a vote [20]. Random Forest is widely used for classification and regression and is particularly useful on large and complex datasets, where it can handle a large number of features and complex interactions between them.

NBL: The Naive Bayes is based on the Bayes Theorem and is responsible for making a probabilistic classification of observations, typifying them in preestablished classes. NBL uses the conditional probability of each feature given each class to make predictions. Naive Bayes assumes that all features are independent of each other, which is an oversimplification that is not always true in reality. However, this assumption simplifies the calculation of conditional probabilities and makes the algorithm computationally efficient. When a new data point is entered into the model, the algorithm calculates the probability of belonging to each class and selects the class with the highest probability as the final prediction.

GBL: The Gradient Boosting learner creates a link of weak models, in which each one has the objective of minimizing the error of the previous models, through the loss function [30]. It works in sequential steps, adding weak models that focus on residual errors left by the previous model. GBL starts with a weak model, such as a decision tree, and adjusts the data weights to highlight difficult cases. It then adds a new model that focuses on residual errors from the previous model, and so on, until a strong model is built. GBL is especially useful for dealing with uneven or rare data, where minority classes can be lost or misclassified.

3.6. Evaluation Metrics

In this section, we will present the results obtained by the model through the classification algorithms. Thus, it is fundamental to understand the measures to verify if the algorithm obtained a good result in question.

Accuracy is measured using four parameters: (i) False Positive (FP) is when the result is negative but classifies as positive, (ii) False Negative (FN) when the result is positive but classifies as negative, (iii) True Positive (TP) when it is true, that is, number of dead, and (iv) True Negative (TN) when it is negative, how many did not die. By counting all these terms and obtaining the confusion matrix, it is possible to calculate accuracy evaluation metrics (A) for the classification, according to equation (2), resulting in a value between 0 and 1.

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

The error (E) is the case where the algorithm could not get it right. In this sense, the error is calculated considering the difference between the total and the accuracy (A) value and is given by equation (3).

$$E = 1.0 - A \tag{3}$$

Cohen's Kappa: Created by Jacob Cohen in 1960, Cohen's Kappa (K) comprises the precision statistic, which is a quantitative measure of reliability that measures the agreement between two evaluators, each of whom performs the classification of different items in different categories, respectively excluded. Its equation is given by (4), where P(0) is defined as the relative acceptance rate, and P(e) is the hypothetical acceptance rate when there is complete agreement between the sets of data, k = 1.0.

$$K = \frac{\text{P(0)}-\text{P(e)}}{1-\text{P(e)}} \tag{4}$$

Sensitivity and Specificity: Both also belong to the precision statistic. Sensitivity can be understood as one of the metrics that aims to evaluate the usefulness of the method in detecting results with vigor, classified as positive. Otherwise, Specificity, unlike Sensitivity, is based on the perception of carrying out the evaluation of methods to identify negative results [31].

Precision and Recall: Precision is nothing more than a metric that has the ability to actually evaluate the number of true positive results (TP) about the sum of all positive predictions. Recall can be used in cases where false negatives are considered to be more harmful than false positive results.

F1-Score: Aiming at using only one metric that unites Sensitivity with precision in order to arrive at a singular number that can actually determine the fitness of our model. When we have a high value, we can conclude that our result (accuracy) was in fact relevant, when its true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values did not have major changes.

3.7. Process steps

In this section, the flowchart developed on the KNIME Analytics Platform will be presented, in which all the data classification algorithms that were used in this work are represented. The KNIME Analytics Platform is based on software that allows us to analyze data and create workflows visually, without employing explicit programming. KNIME allows us to organize all our processes into a visual workflow, connecting graphical building blocks so that we can manipulate the data in the way we prefer — from editing and processing to applying algorithms for decision-making. Using block programming, it is possible to understand data in a more accessible way. KNIME nodes are fundamental in the workflow for analyzing data and building models, since each one performs a specific task, such as reading, processing, and training. They are the building blocks that enable structured data analysis.

The flowchart is based on the learning algorithm presented in the previous section, as well as on the methodology shown in Figure 1. The eight selected data mining algorithms were implemented in the KNIME Analytics Platform. This platform is an open-source data analysis solution that allows the integration of several modules and tools for data processing, analysis, and visualization. KNIME provides a vast library of nodes, including preprocessing, data mining, machine learning, and visualization nodes, which were used in the proposed methodology to preprocess, evaluate, and select the most efficient algorithms for predicting heart disease. The KNIME Analytics Platform facilitated a detailed analysis of the data, offering an open-source, easy-to-use, and user-friendly environment for executing the proposed tasks. A workflow was created for classifying the database, as shown in Figure 2.

To improve the results of the data analysis, filters were applied to all algorithms. These filters were used to determine which base attributes would be redundant for evaluation — that is, which were highly correlated — so that the relevant attributes could be prioritized. As shown in Table 4, ten correlation filters were applied, highlighting in blue the best accuracy results for each algorithm with their respective filters, and in red the worst results.

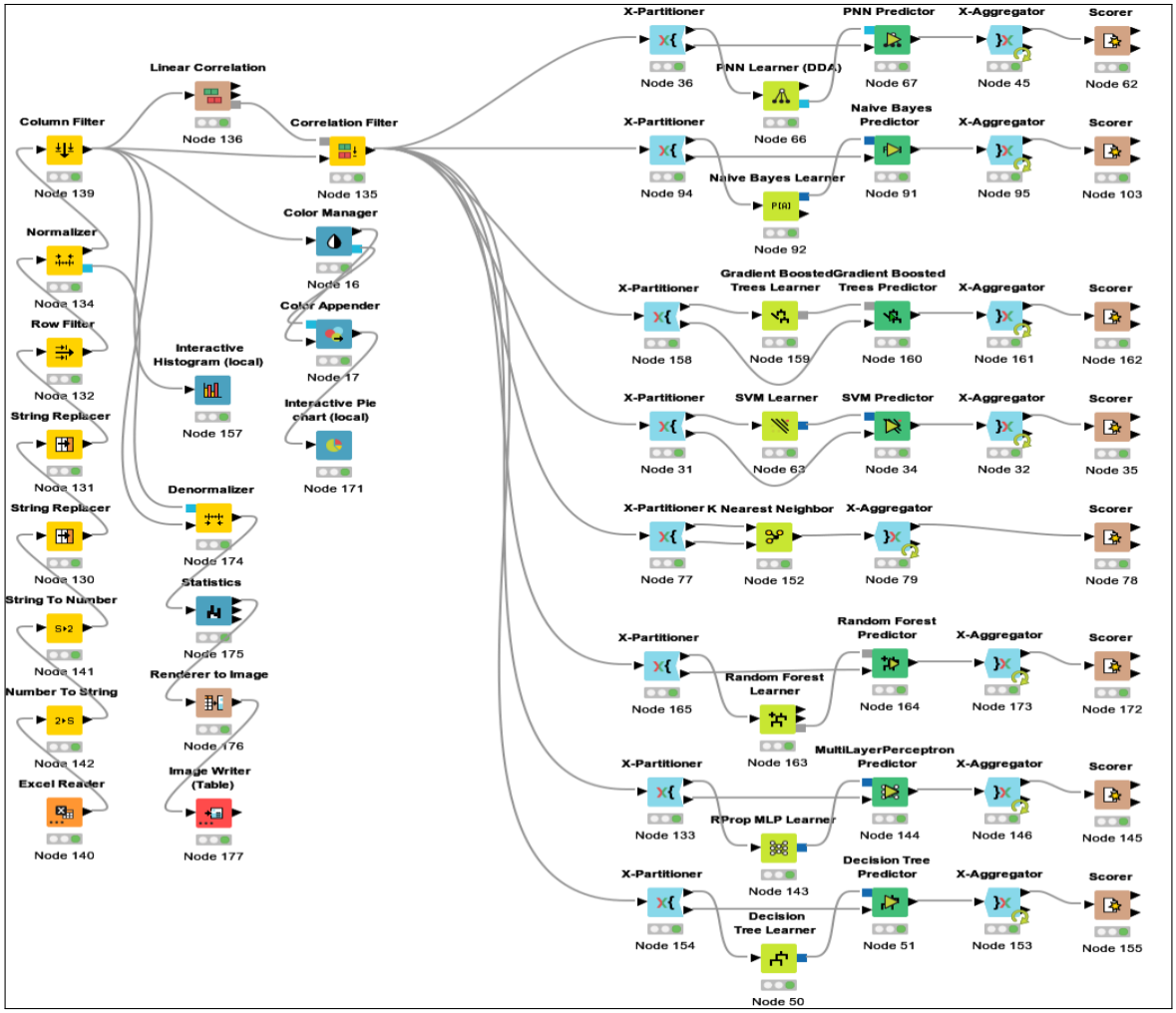

Figure 2. Workflow proposed in the KNIME analytics platform

4. RESULTS AND DISCUSSION

In this section, we will present a visualization of the data from the database under study. In this sense, we will present different types of graphs to understand the base that is being studied. Furthermore, based on the selected algorithms, we will present the classification results for each of the algorithms, considering different correlation filters.

4.1. Patterns and insights of the data used

First, after carrying out the necessary transformations and preprocessing the data, we performed some statistical analysis, as shown in Figure 3. Thus, we obtained a total of 202 patients from the YES class who survived (68.01%) and 95 patients from the NO class who did not survive (31.99%), as shown in the pie chart of Figure 3, left side.

In Figure 3, on the right side, we observe patients who survived by age (YES) who have a median age (60 years) lower than patients who did not survive (NO) with a median age of 65 years. Among those who survived (YES), the lowest and highest values for age are, respectively, 40 and 85 years, with an outlier of a surviving patient of 90 years. For those who did not survive (NO), the lowest and highest values are respectively 42 and 95 years old. However, overall (considering YES and NO), the ages of patients vary between 40 and 95 years old, with a mean age for patients of 63.43 years old. For then, each of the 11 remaining attributes (ex, First, after carrying out the necessary transformations and preprocessing the data,

we performed some statistical analysis, as shown in Figure 4, along with the breakdown of the statistical values demonstrated in Table 3.

Thus, we obtained a total of 202 patients from the YES class who survived (68.01%) and 95 patients from the NO class who did not survive (31.99%), as shown in the pie chart of Figure 3(a). In Figure 3(b), we observe that patients who survived by age (YES) have a median age of 60 years, lower than that of patients who did not survive (NO), with a median age of 65 years. Among those who survived (YES), the lowest and highest values for age are, respectively, 40 and 85 years, with an outlier of a surviving patient of 90 years. For those who did not survive (NO), the lowest and highest values are respectively 42 and 95 years old. However, overall (considering YES and NO), the ages of patients vary between 40 and 95 years old, with a mean age for patients of 63.43 years old.

4.2. Analysis of classification algorithms

In the accuracy comparison subsection, the algorithms were evaluated in terms of their accuracy. On the other hand, in the subsection comparing other performance indicators, such as Sensitivity, specificity, Cohen's kappa, and f-measure, it was observed that each algorithm performed differently in each of these metrics.

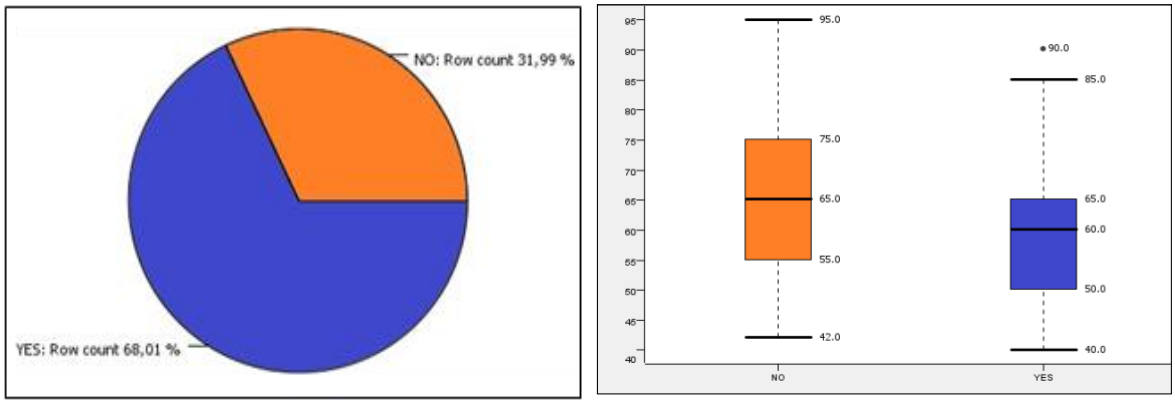

Figura 3. Statistics for the database of 299 patients for class and age

4.2.1. Comparison of model accuracies

Analyzing Table 5, we came to the conclusion that SVM and RFL were the best classifiers, with 84.18% correlation, 0.30, and NBL and PNN were the worst among the eight, with 80.14% when applying the 0.10 filter for the best performance (in blue). Considering the global means for each algorithm, the results obtained in this research showed that the Random Forest algorithm in mean had the best accuracy in predicting heart disease, with 80.07% accuracy, followed by the SVM algorithm with 79.39% mean of accuracy and GBL and MLP which obtained 78.79% and 78.25%, respectively, mean of accuracy considering all filters simulations. But in this case, the MLP algorithm obtained 83.50% accuracy. Among the best performing ranking algorithms, MLP obtained a lower standard deviation (0.051%) compared to the other classification algorithms that obtained good results in the classification, and the correlation filter is 0.10, which requires only four parameters to classify. On the other hand, the Probabilistic Neural Network algorithm had the worst performance, with 69.43% accuracy. Despite PNN having a Standard Deviation (Std. Dev.) of only 0.03945, it is still not enough to say that it is a good algorithm for this problem, since it presented very low averages in all executions.

These results indicate that machine learning algorithms can be useful in detecting heart disease and that Random Forest and Support Vector Machine are promising algorithms for this task. Accuracy isn't the sole metric for evaluating these algorithms. Sensitivity and specificity are also important, depending on the clinical context, as discussed in the next subsection.

4.2.2. Analyzing other evaluation metrics

When we use appropriate evaluation metrics, we can conveniently evaluate whether the performance of a machine learning model in different environments was good or not. Furthermore, we can make sure that these models can be advantageous for a specific proposal. We brought in the Table 5 the respective filters that provided the best results for the tested algorithms, highlighting in red for accuracy we have the worst result 0.801 (NBL) and in blue the best 0.842 (SVM, RFL), for Cohen's kappa the best is the green 0.624 (GBL) and the worst the orange 0.497 (NBL), for the other recall, precision, Sensitivity, specificity, and f-

measure we take into account the YES, NO classes where for the YES class we define the best as being blue and the worst red, and for class NO in green we have the best and in orange the worst for all metrics. Occasionally, the SVM algorithm did not show the best Cohen's Kappa shown in Table 5. This metric measures the agreement between two raters and sorts different items into various categories, in that order, excluding, presenting a value of 0.62% with the 0.30 filter, not much below the best filter found for Cohen's kappa.

That means the GBL algorithm had the best performance for Cohen's Kappa 0.624% (considering the correlation filter equals 0.30). It can be observed that the algorithm NBL obtained a lower performance among all those that had a greater accuracy, collected from Table 4 for the construction of Table 5, with a result of 0.801 for accuracy and 0.497 for the Cohen's kappa coefficient. Considering all metrics, we can see that the SVM had a good performance for the 0.30 correlation filter, considering a good f-measure (combination of precision and recall) equals 88.8% and an excellent accuracy equals 84.2%.

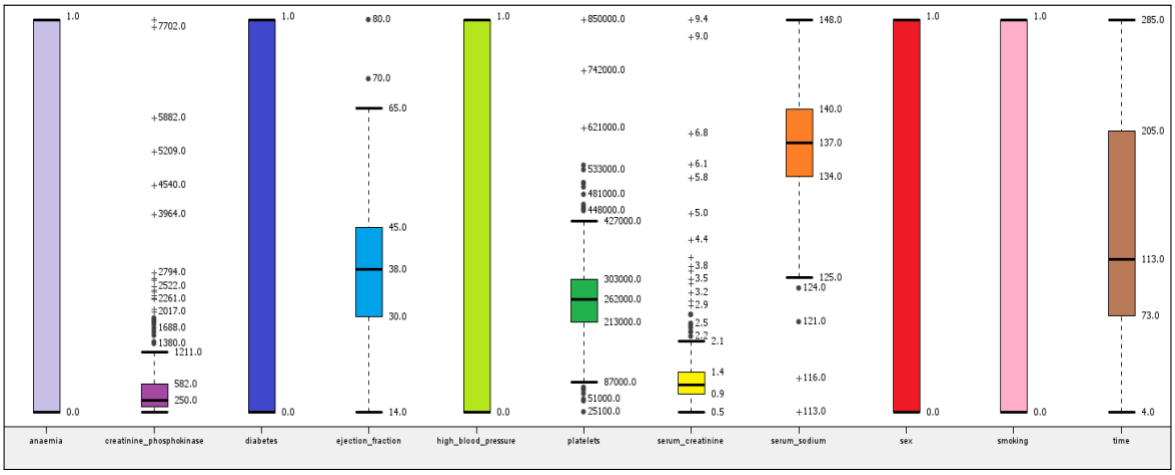

Figure 4. Boxplots for the "Heart Failure" database of 299 patients, considering 11 attributes

Table 3. Statistics for each of the attributes in the database

Statistics	Age	Anaemia	Creatinine phosphokinase	Diabetes	Ejection fraction	High blood pressure	Platelets	Serum creatinine	Serum sodium	Sex	Smoking	Time
Minim	40	0	23	0	14	0	25100	0.5	113	0	0	4
Smallest	40	0	23	0	14	0	87000	0.5	125	0	0	4
Lower Quartile	51	0	118	0	30	0	213000	0.9	134	0	0	73
Median	60	0	250	0	38	0	262000	1.1	137	1	0	113
Upper Quartile	70	1	582	1	45	1	303000	1.4	140	1	1	205
Largest	95	1	1211	1	65	1	427000	2.1	148	1	1	285
Maxim	95	1	7861	1	80	1	850000	9.4	148	1	1	285
Mean	64.43	0.43	1438.29	0.43	40.86	0.43	309585.7	2.27	135	0.57	0.43	138.43

4.3. Discussion and Limitations

This study aims to develop a classification model for diagnosing heart disease based on 12 parameters: age, anemia, creatine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, smoking, sex, and time. The model is based on data mining techniques and aims to identify the most important factors for classifying patients with heart disease. Accuracy and other evaluation metrics were used to evaluate model performance and determine that SVM was the classification

algorithm that performed best. In this sense, we can observe that smoking alone was not relevant for the diagnosis of the disease, considering the correlation filter equals 0.30.

However, considering an algorithm that had a favorable performance for all scenarios, and used a much smaller number of parameters, it was the GBL algorithm, with a correlation filter equal to 0.30, which had an accuracy of 83.8%, with Cohen's kappa of 0.624 (best performer). In addition, GBL also presented an excellent f-measure level of 88.2% for class YES, and for class NO, it is 74.2% (the best performance for class NO). Considering all filter simulations, GBL achieved an overage mean of 78.79%, and MLP an overage mean of 78.25%; this last one is similar to the value of the best-analyzed algorithm, but it used much fewer parameters, that is, only four parameters: creatinine phosphokinase, serum sodium, sex, and time (when correlation filter 0.10).

It was found that the most relevant parameters for the diagnosis of heart disease, considering the database with 299 patients, were the class DEATH_EVENT and the four parameters listed below: (i) creatinine phosphokinase, (ii) serum sodium, (iii) sex, and (iv) time. These were identified through data mining algorithms applied to a dataset of patients with heart disease. Data analysis showed that these parameters have a strong correlation with the diagnosis of heart disease, which can help doctors diagnose and treat the disease more accurately. With this discovery, it is hoped that new diagnostic and treatment techniques can be developed to improve the quality of life of patients with heart disease.

Some possible limitations of this work include the use of a specific and limited database, which may limit the generalization of results to other populations or contexts; the choice of eight data classification algorithms, which may not include other options relevant to the problem; the limitation of the chosen evaluation metrics, which may not reflect all relevant aspects of the performance of the algorithms; and the possible limitations of the methodology used, which may not be the most suitable for evaluating data classification algorithms in other contexts or for other research questions.

Table 4. Accuracies for each of the classification algorithms

Filters	DT	SVM	KNN	MLP	RFL	GBL	NBL	PNN
1.00	81.15%	83.84%	70.71%	80.47%	83.84%	83.17%	76.43%	68.69%
0.85	81.15%	83.84%	70.71%	80.47%	83.84%	83.17%	76.43%	68.69%
0.75	81.15%	83.84%	70.71%	80.47%	83.84%	83.17%	76.43%	68.69%
0.50	81.15%	83.84%	70.71%	80.47%	83.84%	83.17%	76.43%	68.69%
0.30	80.81%	84.18%	72.39%	82.16%	84.18%	83.84%	76.43%	69.02%
0.25	80.81%	84.18%	72.39%	82.16%	84.18%	83.84%	76.43%	69.02%
0.20	63.97%	71.72%	65.99%	70.71%	75.42%	71.38%	70.03%	68.35%
0.15	67.34%	68.01%	67.68%	71.04%	71.04%	67.34%	71.72%	68.01%
0.10	72.73%	82.49%	82.16%	83.50%	83.17%	83.50%	80.14%	80.14%
0.05	62.29%	68.01%	69.02%	71.04%	67.34%	65.32%	72.05%	64.98%
Mean	75.25%	79.39%	71.25%	78.25%	80.07%	78.79%	75.25%	69.43%
StdDev	0.07922	0.07089	0.04321	0.05148	0.06370	0.07581	0.03021	0.03945

However, despite these limitations, this study provides a comparative evaluation of 8 classification algorithms and provides important insights into which metrics should be considered in evaluating the algorithm's performance against different types of errors. The discovery that only four attributes are required for diagnosis could be useful for future studies and could be a promising area for optimizing the diagnosis of heart disease.

Our article presented superior results compared to previous works in the literature, achieving high accuracy (83.50%) by using only four parameters, while other works used all attributes from the database without any filtering. For instance, achieved an accuracy close to 83.30%, and the authors [32] achieved 85.14%, both using all attributes from the database. In [17] work, the authors achieved an interesting result of approximately 92%; however, for this result, nine characteristics (exams) were necessary. The algorithm with the best performance was ETC (Extra Trees Classifier) with SMOTE (Synthetic Minority Oversampling Technique), which showed the highest result in all evaluation measures and achieved 0.9262 accuracy. In this study, we achieved an accuracy of 83.50% using an MLP classifier with only four attributes: creatinine phosphokinase, serum sodium, sex, and time. Conversely, in the paper by [20], the authors employed various classification algorithms, with the random forest learner standing out with an 87.21% accuracy using the RFL

classifier. They used 10 attributes for their results: Anemia, creatine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum sodium, sex, smoking, and time. This approach required more computational processing time and additional laboratory, physical, and financial resources for the exams.

Table 5. Best filters and their metrics

Learner	Filter	Accuracy	Cohen's Kappa	Class	TP	FP	TN	FN	Recall	Precision	Sensitivity	Specificity	F-measure
DT	0.50	0.811	0.571	NO	69	30	172	26	0.726	0.697	0.726	0.851	0.711
				YES	172	26	69	30	0.851	0.869	0.851	0.726	0.86
GBL	0.30	0.838	0.624	NO	69	22	180	26	0.726	0.758	0.726	0.891	0.742
				YES	180	26	69	22	0.891	0.874	0.891	0.726	0.882
KNN	0.10	0.822	0.554	NO	54	12	190	41	0.568	0.818	0.568	0.941	0.671
				YES	190	41	54	12	0.941	0.823	0.941	0.568	0.878
MLP	0.10	0.835	0.592	NO	58	12	190	37	0.611	0.829	0.611	0.941	0.703
				YES	190	37	58	12	0.941	0.837	0.941	0.611	0.886
NBL	0.10	0.801	0.497	NO	49	13	189	46	0.516	0.79	0.516	0.936	0.624
				YES	189	46	49	13	0.936	0.804	0.936	0.516	0.865
PNN	0.10	0.801	0.524	NO	58	22	180	37	0.611	0.725	0.611	0.891	0.663
				YES	180	37	58	22	0.891	0.829	0.891	0.611	0.859
RFL	0.30	0.842	0.623	NO	65	17	185	30	0.684	0.793	0.684	0.916	0.734
				YES	185	30	65	17	0.916	0.86	0.916	0.684	0.887
SVM	0.30	0.842	0.62	NO	64	16	186	31	0.674	0.8	0.674	0.921	0.731
				YES	186	31	64	16	0.921	0.857	0.921	0.674	0.888

Regarding the literature review, [15] studied the survival and mortality rates of heart failure patients in Pakistan, highlighting the importance of risk stratification based on clinical and laboratory factors. In [3], the authors focused on two clinical factors, serum creatinine and ejection fraction, for predicting patient survival. Our objective is not limited to identifying only two attributes from the dataset; instead, we aim to compile a comprehensive list of the most significant attributes present in the data. In [16], the relevance of ejection fraction and serum creatinine was emphasized, using a decision tree with 80% accuracy. In this case, they used only one algorithm, and in this paper, we use eight algorithms to compare, achieving 84%. In the work of [17], the authors employed various classification models, achieving an accuracy of approximately 92% with the ETC algorithm. Otherwise, in our approach, combining the correlation with filters offers a promising way to reduce attributes and achieve efficient classification results. This approach is much more interesting for computational processing because the fewer attributes are used for classification, the faster the processing is in computational terms. Additionally, for the public health system, which is one of the focuses of our work, this reduction of attributes becomes significant for the economy of exams and, consequently, of public money, guaranteeing greater speed in the clinical diagnosis of the patient.

Thus, our article is advantageous both from a clinical and computational perspective, reducing processing time and resources (number of attributes) required for diagnosis. This means that, compared to other techniques, our approach (MLP + correlation filter) is able to perform the diagnosis with similar efficiency while requiring fewer computational resources. This simpler, more efficient approach can be particularly valuable in clinical settings, where speed and accuracy are critical to decision-making. Furthermore, our algorithm presents a great advantage in terms of cost, both for the patient and for the public health system, since it only uses a small set of parameters (4 parameters), which can result in a significant reduction in costs with exams and complementary procedures. This makes our technique even more accessible and can facilitate the early diagnosis of heart disease, increasing the chances of effective treatment and improving patients' quality of life.

Our main focus with these results is to apply them to software for mobile devices, where various parties would benefit, both professionals in the field (doctors and nurses) and ordinary people, who do not have easy access to basic treatment methods. This application will facilitate both diagnosis and how treatment can be carried out. Through the information provided in the application, such as basic test results, the data will be securely transferred to the software to continue the analysis, and with the result in hand, provide insights to healthcare professionals so that they can obtain a better direction in the application of treatment. Ordinary users will be able to predict whether they are predisposed to a worsening of the disease, recommending that they always seek medical treatment. Providing a certain ease in diagnosing and treating the specific condition with a certain cost reduction, focusing on essential attributes. The data used will be clinical and laboratory data, which will then be transferred to the application through a form. Aimed at both common patients and doctors specializing in the field. The software will process the data based on the applied algorithm and will also give the result of whether or not the patient will be able to survive or have the disease worsen. The application processing will occur locally in the software itself, and the data will be stored in the cloud on a Backend as a Service (BaaS) platform for future studies.

5. CONCLUSION

Thus, we know that heart disease has been causing a significant number of deaths worldwide. Many diseases affect the heart, among them are high blood pressure, heart failure, and acute myocardial infarction, which can develop through poor health care, such as poor diet or scarcity of physical activities, or can even develop over time. It is important to take care of the heart because it is responsible for supplying blood and oxygen to our entire body and cells, and it is the main organ that keeps us alive. It is important to prevent heart disease because if one develops it and does not receive proper treatment and care, the heart may not be able to withstand it, and the person may die due to improper care or the fact that the person's body itself is not able to fight that disease. Through data science, together with experts in the field, we can predict the chances that patients have of surviving heart disease.

Therefore, after performing all the analysis of the ML algorithms and having applied the correlation filters, we can conclude based on the tables shown above that the best algorithms that stood out in the study addressing the possibility of survival of patients after an illness were the SVM and RFL algorithms, both of which had the same result, respectively, with 84.18% of accuracy. Therefore, both algorithms performed better with the 0.30 filter, where 12 attributes were considered most relevant to the filter, excluding only one attribute: smoking. On the other hand, PNN was the algorithm with the worst performance (69.43%).

The RFL, despite having obtained the second-best accuracy, got the best performance on average, considering all correlation filters (80.07%) and a good Cohen's Kappa result. GBL proved to be a good alternative, with a performance close to RFL and a shorter processing time. The MLP algorithm demonstrated high performance, achieving 83.5% accuracy, along with high precision, recall, Sensitivity, and specificity values, resulting in 88.6% f-measure. Notably, these results were achieved using a (0.10) correlation filter, implying that only four attributes of the process algorithm (creatinine phosphokinase, serum sodium, sex, and time) were used in the classification. Moreover, considering all correlation filter variations, the average accuracy of MLP was 78.25%, indicating that the algorithm performed well in classification even with a limited number of attributes. The discovery that only four attributes are needed for the diagnosis of heart disease could be useful for future studies, in terms of reducing the number of variables to be collected and analyzed. This can save time and resources and improve diagnostic efficiency. Furthermore, the choice of the best algorithm may depend on the size and quality of the database, as well as the specific characteristics of the studied population. Future work related to this research may include the use of other data preprocessing techniques to improve the quality of information used by mining algorithms and the application of unsupervised learning techniques to identify possible patterns or clusters in the data.

In future works, we also plan to develop an application to show users if they tend to develop some heart disease and the possible care that they should have, without discarding the follow-up of a specialist doctor. Through the analysis of the algorithms of this research, we will apply one of our choices, judged with a good performance, to analyze the data informed by the user. There will be a total of 4 patient attribute pieces of information.

Besides that, we plan to use this patient's data so that we can assemble a newly updated database, with the patient's consent, so that we can find out what else is relevant to the possibility of survival from any current heart disease. Finally, it would also be interesting to investigate the possibility of using other supervised algorithms for the diagnosis of heart diseases, as well as to evaluate the effectiveness of these techniques in Comparison with the traditional supervised learning techniques presented in this study. Therefore, it is important to carry out additional studies to evaluate and compare the performance of these algorithms in different clinical settings and different populations.

DATA AVAILABILITY STATEMENT

The data presented in this study are available on request from the corresponding author.

ACKNOWLEDGEMENTS

The author DAL thanks FAPEMIG and CNPq for financial support. The author VSS thanks FAPEMIG and IFTM for the scientific initiation scholarship.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest in this work.

REFERENCES

[1] H. Kim, L. E. Caulfield, V. Garcia-Larsen, L. M. Steffen, J. Coresh, and C. M. Rebholz, "Plant-Based Diets Are Associated With a Lower Risk of Incident Cardiovascular Disease, Cardiovascular Disease Mortality, and All-Cause Mortality in a General Population of Middle-Aged Adults," *J Am Heart Assoc*, vol. 8, no. 16, Aug. 2019, doi: 10.1161/JAHA.119.012865.

[2] S. Brouwers, I. Sudano, Y. Kokubo, and E. M. Sulaica, "Arterial hypertension," *Lancet*, vol. 398, no. 10296, pp. 249–261, Jul. 2021, doi: 10.1016/S0140-6736(21)00221-X.

[3] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med Inform Decis Mak*, vol. 20, no. 1, p. 16, Dec. 2020, doi: 10.1186/s12911-020-1023-5.

[4] L. Donisi *et al.*, "Bidimensional and Tridimensional Poincaré Maps in Cardiology: A Multiclass Machine Learning Study," *Electronics*, vol. 11, no. 3, p. 448, Feb. 2022, doi: 10.3390/electronics11030448.

[5] D. Chicco, M. J. Warrens, and G. Jurman, "The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment," *IEEE Access*, vol. 9, pp. 78368–78381, 2021, doi: 10.1109/ACCESS.2021.3084050.

[6] O. O. Oladimeji and O. Oladimeji, "Predicting Survival of Heart Failure Patients Using Classification Algorithms," *JITCE (Journal Inf Technol Comput Eng*, vol. 4, no. 02, pp. 90–94, Sep. 2020, doi: 10.25077/jitce.4.02.90-94.2020.

[7] Y. Xue and Y. Zhao, "Structure and weights search for classification with feature selection based on brain storm optimization algorithm," *Appl Intell*, vol. 52, no. 5, pp. 5857–5866, Mar. 2022, doi: 10.1007/s10489-021-02676-w.

[8] D. A. Lima, M. E. A. Ferreira, and A. F. F. Silva, "Machine Learning and Data Visualization to Evaluate a Robotics and Programming Project Targeted for Women," *J Intell Robot Syst*, vol. 103, no. 1, p. 4, Sep. 2021, doi:10.1007/s10846-021-01443-w.

[9] R. S. Dornelas and D. A. Lima, "Correlation Filters in Machine Learning Algorithms to Select Demographic and Individual Features for Autism Spectrum Disorder Diagnosis," *J Data Sci Intell Syst*, vol. 1, no. 2, pp. 105–127, Jun. 2023, doi: 10.47852/bonviewJDSIS32021027.

[10] P. C. Dinas, Y. Koutedakis, and A. D. Flouris, "Effects of active and passive tobacco cigarette smoking on heart rate variability," *Int J Cardiol*, vol. 163, no. 2, pp. 109–115, Feb. 2013, doi: 10.1016/j.ijcard.2011.10.140.

[11] D. Di Raimondo, G. Rizzo, G. Musiari, A. Tuttolomondo, and A. Pinto, "Role of Regular Physical Activity in Neuroprotection against Acute Ischemia," *Int J Mol Sci*, vol. 21, no. 23, p. 9086, Nov. 2020, doi: 10.3390/ijms21239086.

[12] X. Jia *et al.*, "High-Sensitivity Troponin I and Incident Coronary Events, Stroke, Heart Failure Hospitalization, and Mortality in the ARIC Study," *Circulation*, vol. 139, no. 23, pp. 2642–2653, Jun. 2019, doi: 10.1161/CIRCULATIONAHA.118.038772.

[13] T. Nishikimi and Y. Nakagawa, "Potential pitfalls when interpreting plasma BNP levels in heart failure practice," *J Cardiol*, vol. 78, no. 4, pp. 269–274, Oct. 2021, doi: 10.1016/j.jjcc.2021.05.003.

[14] G. Isola, A. Polizzi, S. Santonocito, A. Alibrandi, and S. Ferlito, "Expression of Salivary and Serum Malondialdehyde and Lipid Profile of Patients with Periodontitis and Coronary Heart Disease," *Int J Mol Sci*, vol. 20, no. 23, p. 6061, Dec. 2019, doi: 10.3390/ijms20236061.

[15] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," *PLoS One*, vol. 12, no. 7, p. e0181001, Jul. 2017, doi: 10.1371/journal.pone.0181001.

[16] M. Al Mehedi Hasan, J. Shin, U. Das, and A. Yakin Srizon, "Identifying Prognostic Features for Predicting Heart Failure by Using Machine Learning Algorithm," in *2021 11th International Conference on Biomedical Engineering and Technology*, New York, NY, USA: ACM, Mar. 2021, pp. 40–46. doi: 10.1145/3460238.3460245.

[17] A. Ishaq *et al.*, "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[18] D. Kumar *et al.*, "Cardiac Diagnostic Feature and Demographic Identification (CDF-DI): An IoT Enabled Healthcare Framework Using Machine Learning," *Sensors*, vol. 21, no. 19, p. 6584, Oct. 2021, doi: 10.3390/s21196584.

[19] M. F. Aslan, K. Sabanci, and A. Durdu, "A CNN-based novel solution for determining the survival status of heart failure patients with clinical record data: numeric to image," *Biomed Signal Process Control*, vol. 68, p. 102716, Jul. 2021, doi: 10.1016/j.bspc.2021.102716.

[20] V. S. Souza and D. A. Lima, "Identifying Risk Factors for Heart Failure: A Case Study Employing Data Mining

Algorithms," *J Data Sci Intell Syst*, vol. 2, no. 3, pp. 161–173, Sep. 2023, doi: 10.47852/bonviewJDSIS32021386.

[21] G. A. Roth *et al.*, "Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019," *J Am Coll Cardiol*, vol. 76, no. 25, pp. 2982–3021, Dec. 2020, doi: 10.1016/j.jacc.2020.11.010.

[22] G. C. De Santis, "Anemia," *Med (Ribeirao Preto Online)*, vol. 52, no. 3, pp. 239–251, Nov. 2019, doi: 10.11606/issn.2176-7262.v52i3p239-251.

[23] H. Iwano and W. C. Little, "Heart failure: What does ejection fraction have to do with it?," *J Cardiol*, vol. 62, no. 1, pp. 1–3, Jul. 2013, doi: 10.1016/j.jjcc.2013.02.017.

[24] A. D. Deshpande, M. Harris-Hayes, and M. Schootman, "Epidemiology of Diabetes and Diabetes-Related Complications," *Phys Ther*, vol. 88, no. 11, pp. 1254–1264, Nov. 2008, doi: 10.2522/ptj.20080020.

[25] M. Scherlinger, C. Richez, G. C. Tsokos, E. Boilard, and P. Blanco, "The role of platelets in immune-mediated inflammatory diseases," *Nat Rev Immunol*, vol. 23, no. 8, pp. 495–510, Aug. 2023, doi: 10.1038/s41577-023-00834-4.

[26] H. Lu, S. Uddin, F. Hajati, M. A. Moni, and M. Khushi, "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus," *Appl Intell*, vol. 52, no. 3, pp. 2411–2422, Feb. 2022, doi: 10.1007/s10489-021-02533-w.

[27] A. M. Santos *et al.*, "Semivariogram and Semimadogram functions as descriptors for AMD diagnosis on SD-OCT topographic maps using Support Vector Machine," *Biomed Eng Online*, vol. 17, no. 1, p. 160, Dec. 2018, doi: 10.1186/s12938-018-0592-3.

[28] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification," 2003, pp. 986–996. doi: 10.1007/978-3-540-39964-3_62.

[29] W. Huang, Y. Cui, H. Li, and X. Wu, "Effective Probabilistic Neural Networks Model for Model-Based Reinforcement Learning USV," *IEEE Trans Autom Sci Eng*, vol. 22, pp. 11625–11641, 2025, doi: 10.1109/TASE.2025.3539317.

[30] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif Intell Rev*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.

[31] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Med Res Methodol*, vol. 19, no. 1, p. 64, Dec. 2019, doi: 10.1186/s12874-019-0681-4.

[32] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators," *Appl Sci*, vol. 11, no. 18, p. 8352, Sep. 2021, doi: 10.3390/app11188352.

BIOGRAPHIES OF AUTHORS

Vitória S. Souza holds a postgraduate degree in Strategic Business Management (2025) and a bachelor's degree in systems Analysis and Development (2023) from the Federal Institute of Triângulo Mineiro (IFTM), Patrocínio Campus, Brazil. She worked as a scholarship holder at the Research Support Foundation of the State of Minas Gerais (FAPEMIG), conducting scientific initiation activities in the areas of artificial intelligence and data mining. She collaborates as a researcher at the Laboratory of Computational Intelligence, Robotics, and Optimization (LICRO). Her research interests include artificial intelligence, data science, machine learning, data mining and analysis, as well as the development of decision support systems. She has experience with data mining tools, particularly the KNIME Analytics Platform, and proficiency in the Python programming language. She is interested in scientific collaboration and academic production in the aforementioned areas. She can be contacted at email: vitoriasteffane5@gmail.com.

Danielli A. Lima Danielli Araújo Lima holds a PhD in Computer Science (UFU, 2016), with postdoctoral research in Software Engineering and Neuroscience (USP, 2022). She is a full-time professor at IFTM Campus Patrocínio, collaborating in the Professional Master's in Technological Education (IFTM) and serving as a permanent professor in Public Administration (UFTM). She also earned degrees in Technological Education (IFTM, 2021), Digital Technologies for Education (UFC, 2021), Systems Analysis and Development (IFTM, 2018), and Civil Construction Design (IFPB, 2021). Leader of LabITec and LICRo research groups, she reviews top international journals. Her interests include artificial intelligence, robotics, neuroscience, cellular automata, and complex systems. She can be contacted at email: danielli@iftm.edu.br