Research Type (Original)

Classification and prediction of heart disease using Machine Learning models: A promising approach for medical diagnosis.

Orlando Iparraguirre-Villanueva[1], Cleoge Paulino-Moreno[2*]
[1]Universidad Nacional Tecnológica de Lima Sur, Lima, Perú
[2]Universidad Nacional Ciro Alegría, Huamachuco, Perú
Autor corresponsal: paulinozenaida18@gmail.com

Article info

Keywords:
Machine learning
prediction,
heart diseases
models

Abstract

Background: heart disease is one of the leading causes of death worldwide, claiming 17.9 million lives. They are a major public health problem that affects people regardless of age or gender.

Objective: This work aims to classify and predict heart disease using Machine Learning (ML) models such as Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT) and Logistic Regression (LR).

Methods: We worked with the Cleveland dataset from Kaggle, consisting of 303 patient records and 14 attributes. This research was conducted in different stages, including model understanding, dataset analysis and cleaning, ML model training, and model performance evaluation.

Results: The results showed that the RF and KNN models achieved the highest levels of performance and accuracy with 88.52%, surpassing the other models such as SVM, NB, and LR which obtained 86.89% accuracy, and DT with 78.69%.

Conclusions: In conclusion, the RF and KNN models stand out over the other models for this type of prediction task.

1 Introduction

One of the leading causes of death worldwide is heart disease [1]. The World Health Organization (WHO) reports that approximately 17.9 million people die each year from heart disease [2], and surprisingly, heart disease claims more lives each year than any other factor [3]. The most common heart disease, which affects around 41 million individuals, has experienced a significant increase in different countries [4], and the growing incidence of heart disease in the population is alarming [5]. Heart disease continues to represent a public health challenge on a global scale [6], with 428.7 deaths per 100,000 inhabitants in Haiti, followed by Guyana with 427.7, Suriname with 290, the Dominican Republic with 255.7 and Honduras with 363.3, among others [7]. In Spain, heart diseases are among the five main reasons for death [8]. In Ecuador, ischemic heart disease is the principal cause of death in the population, with approximately 8,779 deaths [9], while in Costa Rica, cardiovascular diseases are among the main reasons for death [10].

Over the next decade, an estimated 23.6 million people will lose their lives due to cardiovascular disease. Early identification of cardiovascular disease is critical in reducing the death rate and burden of disease, as well as early detection of risk factors that allow early warning [11]. It is possible to prevent most heart diseases by addressing associated risk behaviors such as smoking, an unhealthy and high-fat diet, being overweight, lack of physical exercise, and alcohol abuse [12].

Currently, artificial intelligence (AI) can be used in different fields of medicine, including clinical diagnosis [13]. Computational algorithms have a superior ability to detect diseases through medical images with greater accuracy than the human eye [14]. AI-enabled systems have played a significant role in the healthcare field in different countries where the use of AI expanded to explore solutions that could improve the delivery of medical services [15]. Machine Learning (ML) emerged as a decision support tool, where algorithms are used to create models capable of learning from data and recognizing patterns [16]. The ability of AI and ML algorithms to examine complex datasets benefits clinicians by predicting diseases at early stages and also contributes to advanced patient care and improving health outcomes. Accurate classification of cardiac disease can support the physician in making appropriate decisions for patients [17]. Likewise, accurate prediction of heart disease is of vital importance to provide effective treatment to patients before they experience a medical emergency.

Heart diseases represent a significant challenge to public health worldwide. Nowadays, many people suffer from heart disease. Due to this, throughout the last decades, experts have employed various approaches to predict heart diseases. For example, [18] proposed to creation of a detection system for classifier diseases such as SVM, NB, and KNN. The results evidenced that NB is the algorithm with the highest efficiency, with an accuracy of 96.9%. Similarly, [19] developed a novel method for classifying heart disease. They used a method based on data, fuzzy clustering, and modifiers. They concluded that the presented model achieved an accuracy of 80.46% which was lower than the 82.67% obtained by SVM. In the same way, in the paper [20] aimed at predicting heart diseases by using ML and EMR data features, they used NB, LR, RF, and neural networks classifiers along with cross-sectional features (CS) and longitudinal features (LT). RF achieved the highest score of 0.902 in the combination of CS and LT features. Also [21] proposes a detection system for coronary artery disease using RF and XGBoost classifiers. By combining RF and the TPOT classifier, the highest accuracy was achieved with a percentage of 97.52%. Likewise, [22] carried out a comparison between different classifiers to identify heart disease. He evaluated the KNN, DT, and RF algorithms. As a result, it was found that RF obtained an accuracy of 100%, which indicates that it is the most effective classifier. Finally, [23] proposed an automated approach using DL for heart sound signal classification. The RF-MFO-XGB ensemble model was used to perform the classification. The results showed that a classification accuracy of 89.08% was achieved.

This study will classify and predict heart disease using ML models such as RF, KNN, SVM, NB, DT, and LR. This paper is structured into several sections that address each aspect of the research. In section 2, an analysis of previous works related to the topic will be carried out. Section 3, on the other hand, will be devoted to explaining and developing the methodology used in the study. In section 4, an analysis of the results obtained will be made. In section 5, discussions on the results of other previous works will be carried out. Finally, in the sixth part of the paper, conclusions will be presented.

2 Method

In this part of the research, the theoretical foundations of the RF, DT, SVM, LR, NB, and KNN algorithms, as well as the procedure used to classify and predict heart disease, are presented.

2.1 Random Forest

RF has become one of the most successful algorithms in the ML field. It generates a series of decision trees randomly and subsequently combines them[24] using different training datasets and features. These individual models are combined using techniques such as voting or averaging to achieve the result RF has been shown to possess the ability to effectively handle data sets with high dimensionality and multicollinearity [25]. In classification, the RF algorithm stands out above other algorithms showing superior predictive capability [26]. When using the RF model for classification, the Gini index is generally used, and the formula for determining how the nodes of a tree branch are presented in equation (1).

$$Gini = 1 - \sum_{i=1}^{c} (P_i)^2 \qquad (1)$$

Where:

Pi represents the relative class frequency, and C represents the number of classes.

2.2 K-Nearest Neighbors

KNN is a simple but effective classification algorithm [27]. KNN is widely recognized in the ML field due to its effectiveness; it can be employed for both classification and regression on datasets. It is fast and simple to understand, as well as efficient even when working with large datasets [28]. It seeks to identify the K nearest samples for each analyzed sample, to measure the separation between the samples and the evaluated sample, common distance calculation methods are employed, such as Euclidean, Hamming, and Manhattan [29]. In equation (2) the Euclidean distance is presented.

$$\boldsymbol{d(x,y) = \sqrt{\sum_{i=1}^{n} (yi - xi)^2}} \qquad (2)$$

In this equation, "x" and "y" correspond to vectors that symbolize two instances in the feature space, "xi" and "yi" are the components of the vectors "x" and "y" respectively, and "n" represents the number of attributes present in the feature space.

2.3 Support Vector Machine

SVM is an ML that analyzes data and detects samples used in classification tasks [30]. SVM has as its main objective to obtain the separation of classes in the training datasets by generating a surface that maximizes the space between them [31]. It uses a nonlinear transformation to change the original pattern space, of low dimensionality into one of higher dimension, to find the optimal separation hyperplane in the new feature space. The optimal hyperplane is characterized by reaching the maximum margin between the data points, i.e., the largest distance separating the plane from the nearest points in the feature space [32]. Equation (3) shows how to compute the SVM classifier.

$$\left[\frac{1}{n}\sum_{i=1} max\left(0,1-y_i(w^T x_i - b)\right)\right] + \lambda\|w\|^2 \qquad (3)$$

2.4 Naive Bayes Algorithm

The NB classifier employs Bayes' theorem, which is a probabilistic graphical model widely used in real-world scenarios. This model assumes that the attributes of an object are independent or unrelated to each other [33]. Under the assumption that classes are independent, the model looks for an individual relationship between each feature and class attribute[34]. The probability density function is used to represent the classes assigned to the training data. Subsequently, the objects are linked to the class with the highest probability [35]. Equation (4) of NB is presented below. In general, the NB classifier is a powerful tool for analyzing data in various fields, thanks to its ability to model complex relationships and predict results accurately.

$$\boldsymbol{p(c|x)} = \frac{\boldsymbol{p(x|c) * p(c)}}{\boldsymbol{p(x)}} \qquad (4)$$

Specifically, P(c|x) represents the probability that the correct category is c given the characteristics x of the object. In turn, P(x|c) is the probability of observing the features x given that the category is c. P(c) represents the a priori probability of the category c, while P(x) represents the marginal probability of the features x. All these factors must be considered, to classify more accurately.

2.5 Decision Tree Algorithm

Decision trees (DT) are one of the most popular classifiers in use today [36]. The structure of a DT is based on a tree consisting of decision nodes containing labeled questions. The root and internal nodes represent these questions, while the edges lead to leaf nodes that provide solutions associated with each question. At each node, a binary decision is made to separate classes from the full data set [37].

2.6 Logistic Regression

Logistic Regression (LR) is a Deep Learning algorithm that is commonly used for categorization and works with binary variables [38]. It is used to create models that establish a relationship between a binary outcome variable and a set of explanatory variables. LR allows us to estimate the probability that an item belongs to a specific class [39], where the response indicates the success or failure of a particular event. Think of it as a powerful tool that predicts outcomes based on given data [40].

2.7 Dataset

This study uses a Cleveland data set from Kaggle, containing 303 patient entries with a total of 14 characteristics, such as age, sex, type of chest pain, resting blood pressure, cholesterol, fasting blood glucose, electrocardiogram results, maximum achieved heart rate, exercise-induced angina, previous peak, slope, number of major vessels, thalassemia, target, and outcome. In Fig. 1, you can see the diagram describing the sequence of stages in the development of this research.

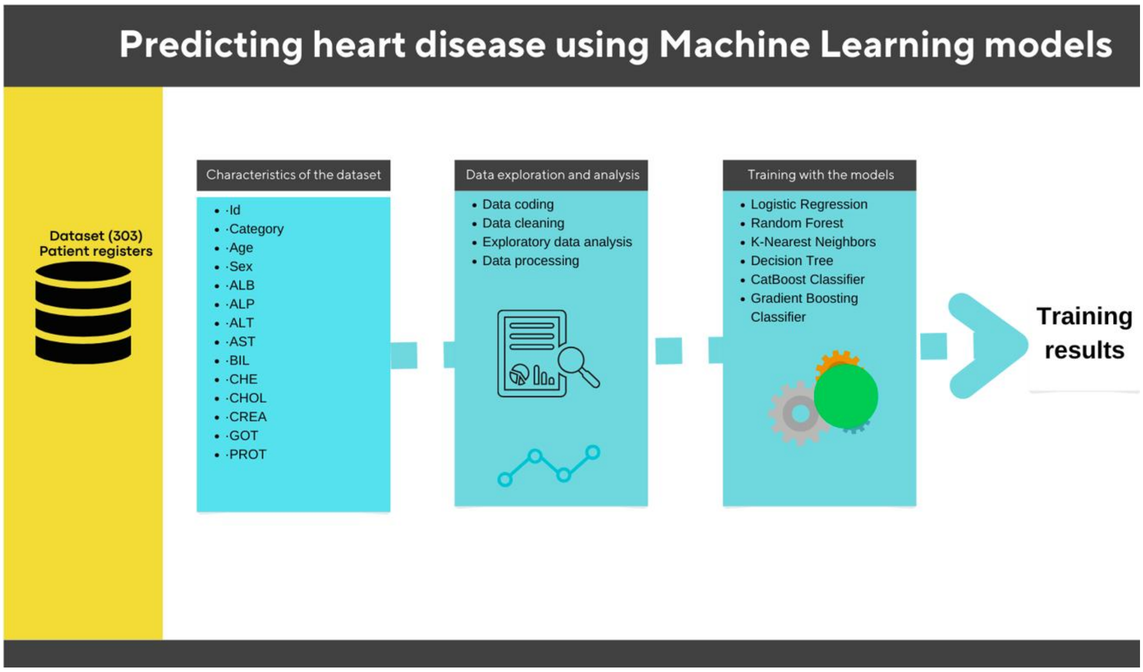

Fig. 1 Development process

2.8 Data processing

The data processing process involves a series of actions to ensure the reliability of this information, such as data cleaning, data transformation, and data analysis. First, we start by importing the required libraries to load the data. In the context of the Python language, data manipulation and analysis require several tools and libraries. In this research, libraries such as Pandas and NumPy were used. Pandas offer flexible and efficient data structures, such as DataFrames, which allow for easy data manipulation and processing, while Numpy is an essential library for numerical analysis in Python. Pandas were used for reading, cleaning, and manipulating data. The NumPy library was used for data analysis, addressing concepts such as mean and media.

In the second step, duplicate, missing, or incorrect information was eliminated. In addition, checks for anomalies and inconsistencies in the dataset were performed to ensure accuracy and reliability. Table I shows the characteristics of the data set.

Table I Type analysis of the dataset

#	Age	Sex	Resting blood pressure	Cholesterol	Fasting blood sugar	Resting electrocardiographic results	The maximum heart rate reached	Exercise-induced angina	Previous peak	…	Chest pain type	Thalassemia	Slope
1	63	1	145	233	1	0	150	0	2.3	…	0	1	1
2	37	1	130	250	0	1	187	0	3.5	…	0	0	1
3	41	0	130	204	0	0	172	0	1.4	…	1	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…
302	57	1	130	131	0	1	115	0	1.2	…	1	3	0
303	57	0	130	236	0	0	174	1	0	…	1	2	0

2.9 Exploratory Data Analysis

This section provides an exploratory analysis of the dataset. Key aspects of the variables and their relationship to heart disease are discussed. The dataset includes information on 303 patients who have suffered myocardial infarctions. The information includes several clinical and demographic variables. The variable "Age" represents the age of the patients and ranges from 29 to 77 years, with a mean of 54 years.

The variable "Sex" represents the sex of the patients, where 1 represents male and 0 female. The variable "Type of chest pain" represents the type of chest pain experienced by the patients, where 1 represents typical angina, 2 represents atypical angina, 3 represents non-anginal pain and 4 represents asymptomatic. The variable "Blood pressure at rest" represents patients' blood pressure at rest, ranging from 94 to 200 mmHg, with a mean of 131 mmHg. The variable "Cholesterol" represents the blood cholesterol level of the patients, ranging from 126 to 564 mg/dl, with a mean of 246 mg/dl. The variable "Fasting blood sugar" represents the blood sugar level of the patients after fasting, where 1 represents more than 120 mg/dl and 0 represents less than or equal to 120 mg/dl. The variable "Electrocardiographic results at rest" represents the results of the electrocardiogram performed on the patients at rest, where 0 represents normal, 1 represents having ST-T wave abnormality and 2 represents showing probable or definite left ventricular hypertrophy. The variable "Maximum heart rate achieved" represents the maximum heart rate achieved by patients during exercise, ranging from 71 to 202 bpm, with a mean of 149 bpm. The variable "Exercise-induced angina" represents whether patients experienced exercise-induced angina, where 1 represents yes and 0 represents no. The variable "Exercise-induced segment depression" ranges from 0 to 6.2, with a mean of 1, and represents the amount of exercise-induced segment depression. Finally, the variable "Number of fluoroscopically stained major vessels" represents the number of fluoroscopically stained major vessels and ranges from 0 to 3.

These aspects are key to analyzing the association with heart disease. A more detailed statistical analysis follows. For example, figure 2(a) shows that people with non-anginal chest pain (cp = 2) have an increased risk of developing heart disease. However, in Figure 2(b), no relationship was found between high fasting blood glucose (fasting blood glucose > 120 mg/dl = 1) and the likelihood of developing heart disease.

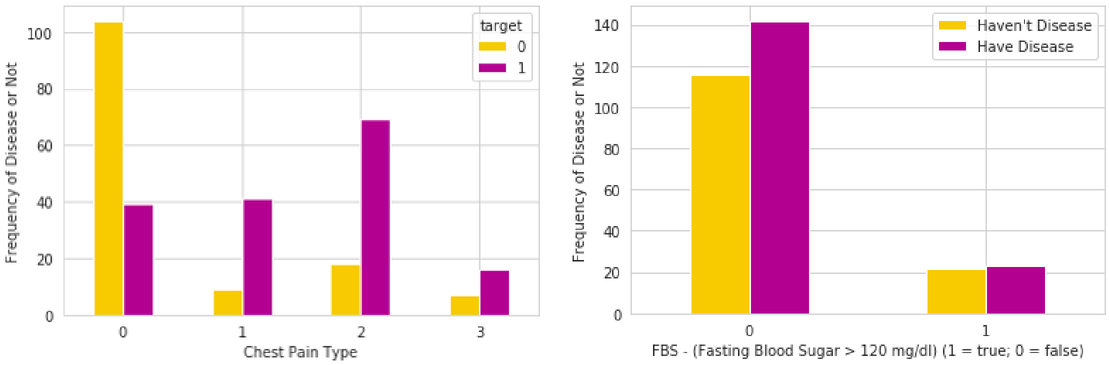

Fig. 2(a)

Fig. 2(b)

Fig. 2 Heart rate: a) according to chest pain; b) according to fasting blood glucose levels

Furthermore, in Figure 3 (a) and Figure 3 (b) belonging to the male gender a flat result (slope = 2) in the stress test slope would increase the possibility of presenting a cardiac condition. A correlation between sex and the slope of the stress test with the possibility of heart disease is evident.

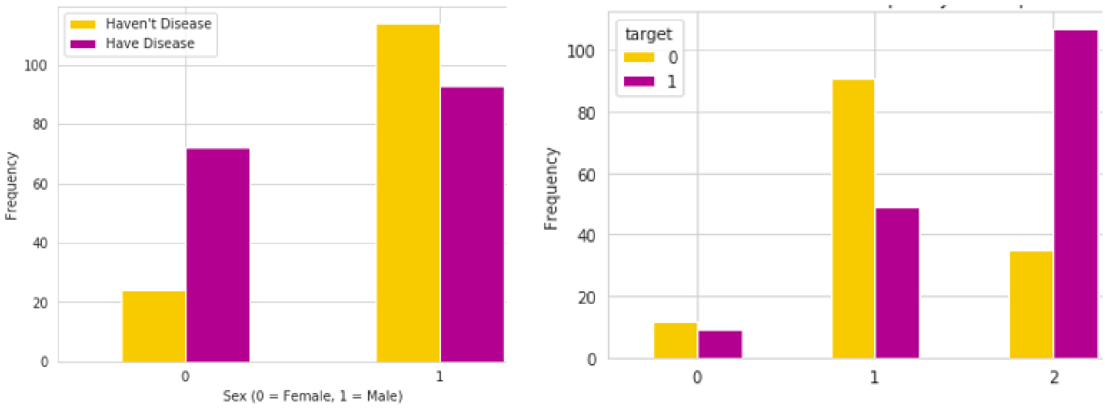

Fig. 3(a)

Fig. 3(b)

Fig. 3 Frequency of heart disease: a) By sex; b) According to slope

2.10 Data Training

The data training stage involves the process of fitting ML models to a data set. In this stage, the models are fitted, aiming to minimize the error between the predictions generated by the model and the data applied in its training. Therefore, data cleaning and data preparation were carried out, including the elimination of missing values and converting categorical variables into dummies. In addition, splits were made in the data, creating separate sets: one for training and another for testing. Eighty percent of the data set was assigned to training and 20% was used to assess performance. Then, ML models (RF, KNN, SVM, NB, DT, and D LR) were selected to compare their performance on the heart disease prediction task. The six models were trained with the training dataset by importing them into the Python sklearn library. Once the models were trained, we evaluated their performance using the test set. Metrics such as accuracy, recall, F1 score, and confusion matrix were calculated to evaluate the models' performance.

3 Results

This section presents the results obtained after applying ML algorithms (RF, KNN, SVM, NB, DT, and LR) to classify and predict heart disease using the Cleveland dataset of the Kaggle platform. This work used six algorithms widely recognized and used in ML. Each algorithm provides a different classification and prediction of heart disease, allowing us to compare their performance and evaluate their effectiveness in diagnosing heart disease. Before the review of the results, a process of data processing and investigation was carried out to ensure their quality and relevance. Exploratory cleaning and analysis techniques were applied to understand the connection between data set characteristics and the existence of heart disease. Throughout this section, the performance of each algorithm in terms of accuracy, throughput, and predictive power is presented. The evaluation metrics used to measure the classification quality are shown. This allows us to determine which algorithm performed best in the heart disease detection task. Analysis of the results is essential to understanding the effectiveness of each algorithm in a specific context. This provides valuable information for future research and clinical applications. Table 2 presents the training results.

Table 2 Evaluation of models

Random Forest				
	Precision (%)	**Recall (%)**	**F1-score (%)**	**Support**
0	88.57	91.17	89.85	27
1	88.46	85.19	86.79	34
accuracy			88.52	61
macro avg	88.51	88.18	88.34	61
weighted avg	88.51	87.84	88.14	61
K-Nearest Neighbour				
	Precision (%)	**Recall (%)**	**F1-score (%)**	**Support**
0	88.57	91.17	89.85	27
1	88.46	85.19	86.79	34
accuracy			88.52	61
macro avg	88.51	88.18	88.34	61
weighted avg	88.51	87.84	88.14	61
Support Vector Machine				
	Precision (%)	**Recall (%)**	**F1-score (%)**	**Support**
0	88.24	88.24	88.24	27
1	88.19	85.19	86.66	34

accuracy			86.89	61
macro avg	88.22	86.72	87.46	61
weighted avg	88.21	86.54	87.36	61
Naive Bayes				
	Precision (%)	**Recall (%)**	**F1-score (%)**	**Support**
0	88.24	88.24	88.24	27
1	88.19	85.19	86.66	34
accuracy			86.89	61
macro avg	88.22	86.72	87.46	61
weighted avg	88.21	86.54	87.36	61
Decision Tree				
	Precision (%)	**Recall (%)**	**F1-score (%)**	**Support**
0	84.37	79.41	81.81	27
1	75.86	81.48	78.57	34
accuracy			78.69	61
macro avg	80.12	80.45	80.28	61
weighted avg	79.63	80.56	80	61
Logistic Regression				
	Precision (%)	**Recall (%)**	**F1-score (%)**	**Support**
0	88.24	88.24	88.24	27
1	88.19	85.19	86.66	34
accuracy			86.89	61
macro avg	88.22	86.72	87.46	61
weighted avg	88.21	86.54	87.36	61

In this study, different levels of accuracy were achieved using different ML models, including RF, KNN, SVM, NB, DT, and LR. The results obtained are as follows: the accuracy of RF and KNN reached 88.52%, DT reached 78.69%, and SVM, NB, and LR reached 86.89%. Based on the information provided in Table II, it can be determined that RF and KNN present the highest average in accuracy, accuracy, and F1-Score. Specifically, the accuracy of RF and KNN models reached 88.51% in precision, 87.84% in accuracy, and 88.14% in F1-Score. Similarly, SVM, NB, and LR obtained equal measures, achieving 88.21% accuracy, 86.54% recall, and 87.36% F1-Score.

Finally, the DT model achieved 79.63% accuracy, 80.56% recall, and 80% F1 score. These results give us a clear picture of the performance of each algorithm on the dataset used in this study. This is in terms of the classification and prediction of heart disease.

4 Discussion

In recent decades, ML has emerged as a promising tool for early and accurate prediction and detection of heart disease. Heart disease is one of the leading causes of death worldwide. Therefore, heart disease prediction becomes a valuable tool to treat patients before complications occur and improve their quality of life. In addition, by identifying those at highest risk, a preventive approach can be applied, which can reduce complications and improve long-term health outcomes. RF, KNN, SVM, NB, DT, and LR models were used in this study. The training results revealed that RF and KNN performed 88.52% each, and obtained the best results in heart disease detection, which does not match the results obtained in the study [18] that evaluated different ML models, and the results indicated that the NB algorithm proved to be the most effective with an accuracy of 96.9%. However, according to the results of [19], it was concluded that SVM was more accurate than other ML models such as KNN, NB, and RF obtaining an accuracy of 82.67%, the result shows a lower accuracy compared to 88.52% obtained by RF and KNN in this study. Coinciding with the work [20], where they used different classifier algorithms, obtaining an accuracy of 90.2%, this result is higher than the 88.52% obtained in this study, it is due to different factors, one of them could be the techniques used. This is related to the results obtained in [21], where they concluded that the combination of RF and the TPOT classifier achieved the highest accuracy with an accuracy of 97.52%. Similarly, this work agrees with [22] where different ML models were compared. These algorithms were KNN, RF, and DT. As a result, they found that RF obtained an accuracy of 100%. This work contributes to the medical community as a tool capable of efficiently predicting cardiac diseases. It is important to point out that the results correlate with previous studies and reinforce the relevance of ML models in the medical field.

The prognosis of heart disease using ML techniques is a valuable tool for medicine; however, the accuracy in the detection of these diseases depends largely on the quality and relevance of the data used during the model training process.

5 Conclusions

After training and comparing the different ML prediction models (RF, KNN, SVM, NB, DT, and LR) for predicting heart disease, the following conclusions were reached. It could be concluded that the KNN and RF models obtained the most outstanding results in terms of accuracy and performance in predicting heart disease, with 88.52% accuracy. This outperforms the other models, such as SVM, NB, and LR, which obtained 86.89% accuracy, and DT, with 78.69%. Therefore, these models stand out as the most accurate predictors of heart disease. This makes them valuable tools for improving the management and treatment of people at risk for these diseases. Their application can be beneficial by providing early and accurate detection of heart disease. This, in turn, enables more timely and effective medical care to be provided to affected patients. In addition, certain attributes have been found to influence an increased likelihood of heart disease. Attributes such as sex, fasting blood glucose, and blood type may play a crucial role as determinants of heart disease.

Although the results obtained in this study are promising, it is pertinent to highlight that the limited number of records in the data set is a significant limitation. This aspect should be addressed in future work to advance the prediction of heart disease.

Finally, ML models can be valuable resources for detecting heart disease. The implementation of ML models such as RF, KNN, SVM, NB, DT, and LR, together with careful variable selection, can greatly increase prediction accuracy. It is expected that soon it will be possible to train these models with large data sets to improve prediction results.

6 References

[1] “Cardiovascular diseases (CVDs).” Accessed: Jul. 13, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] “Cardiovascular diseases.” Accessed: May 28, 2023. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

[3] “Enfermedades cardiovasculares - OPS/OMS | Organización Panamericana de la Salud.” Accessed: May 28, 2023. [Online]. Available: https://www.paho.org/es/temas/enfermedades-cardiovasculares

[4] J. S. Aluru, A. Barsouk, K. Saginala, P. Rawla, and A. Barsouk, “Valvular Heart Disease Epidemiology,” *Med Sci (Basel)*, vol. 10, no. 2, Jun. 2022, doi: 10.3390/MEDSCI10020032.

[5] N. A. M. Zaini and M. K. Awang, “Performance Comparison between Meta-classifier Algorithms for Heart Disease Classification,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, pp. 323–328, 2022, doi: 10.14569/IJACSA.2022.0131039.

[6] "Tamizaje de cardiopatías congénitas," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 6, no. 3, pp. 1548–1558, Jun. 2022, doi: 10.37811/CL_RCM.V6I3.2311.

[7] "La Carga de Enfermedades Cardiovasculares - OPS/OMS | Organización Panamericana de la Salud." Accessed: May 28, 2023. [Online]. Available: https://www.paho.org/es/enlace/carga-enfermedades-cardiovasculares

[8] P. Zueras and E. Rentería, "La esperanza de vida libre de enfermedad no aumenta en España," *Perspectives Demogràfiques*, pp. 1–4, Jan. 2021, doi: 10.46710/CED.PD.ESP.22.

[9] G. Lugmaña, S. Carrera, A. A. Fernández, and D. Andrade, "Registro Estadístico de Defunciones Generales. Elaborado por: Revisado por", Accessed: Jun. 04, 2023. [Online]. Available: www.ecuadorencifras.gob.ec

[10] D. C. Elizondo, "FACTORES DE RIESGO CARDIOVASCULAR," *Revista Ciencia y Salud Integrando Conocimientos*, vol. 4, no. 1, p. undefined-undefined, Jan. 2020, doi: 10.34192/CIENCIAYSALUD.V4I1.108.

[11] L. Veloza, C. Jiménez, D. Quiñones, F. Polanía, L. C. Pachón-Valero, and C. Y. Rodríguez-Triviño, "Variabilidad de la frecuencia cardiaca como factor predictor de las enfermedades cardiovasculares," *Revista Colombiana de Cardiología*, vol. 26, no. 4, pp. 205–210, Jul. 2019, doi: 10.1016/J.RCCAR.2019.01.006.

[12] "Cardiovascular diseases (CVDs)." Accessed: Jun. 04, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[13] P. Délano R., "Inteligencia artificial en otorrinolaringología," *Revista de otorrinolaringología y cirugía de cabeza y cuello*, vol. 79, no. 1, pp. 7–7, Mar. 2019, doi: 10.4067/S0718-48162019000100007.

[14] E. Basáez and J. Mora, "Salud e inteligencia artificial: ¿cómo hemos evolucionado?," *Revista Médica Clínica Las Condes*, vol. 33, no. 6, pp. 556–561, Nov. 2022, doi: 10.1016/J.RMCLC.2022.11.003.

[15] H. da C. Nunes, R. M. C. Guimarães, and L. Dadalto, "Desafíos bioéticos del uso de la inteligencia artificial en los hospitales," *Revista Bioética*, vol. 30, no. 1, pp. 82–93, Mar. 2022, doi: 10.1590/1983-80422022301509ES.

[16] E. José De la Hoz Domínguez *et al.*, "Aprendizaje automático y PYMES: Oportunidades para el mejoramiento del proceso de toma de decisiones," *Investigación e Innovación en Ingenierías*, vol. 8, no. 1, pp. 21–36, Jan. 2020, doi: 10.17081/INVINNO.8.1.3506.

[17] T. N. Nguyen and T. H. Nguyen, "Deep learning framework with ECG feature-based kernels for heart disease classification," *Elektronika ir Elektrotechnika*, vol. 27, no. 1, pp. 48–59, Feb. 2021, doi: 10.5755/J02.EIE.27642.

[18] M. M. Rahma and A. D. Salman, "Heart Disease Classification-Based on the Best Machine Learning Model," *Iraqi Journal of Science*, vol. 63, no. 9, pp. 3966–3976, 2022, doi: 10.24996/IJS.2022.63.9.28.

[19] K. Bahani, M. Moujabbir, and M. Ramdani, "An accurate fuzzy rule-based classification systems for heart disease diagnosis," *Sci Afr*, vol. 14, Nov. 2021, doi: 10.1016/J.SCIAF.2021.E01019/AN_ACCURATE_FUZZY_RULE_BASED_CLASSIFICATION_SYSTEMS_FOR_HEART_DISEASE_DIAGNOSIS.PDF.

[20] Q. Li, A. Campan, A. Ren, and W. E. Eid, "Automating and improving cardiovascular disease prediction using Machine learning and EMR data features from a regional healthcare system," *Int J Med Inform*, vol. 163, p. 104786, Jul. 2022, doi: 10.1016/J.IJMEDINF.2022.104786.

[21] R. Valarmathi and T. Sheela, "Heart disease prediction using hyper parameter optimization (HPO) tuning," *Biomed Signal Process Control*, vol. 70, p. 103033, Sep. 2021, doi: 10.1016/J.BSPC.2021.103033.

[22] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Comput Biol Med*, vol. 136, p. 104672, Sep. 2021, doi: 10.1016/J.COMPBIOMED.2021.104672.

[23] A. Rath, D. Mishra, G. Panda, and M. Pal, "Development and assessment of machine learning based heart disease detection using imbalanced heart sound signal," *Biomed Signal Process Control*, vol. 76, p. 103730, Jul. 2022, doi: 10.1016/J.BSPC.2022.103730.

[24] S. Das, M. S. Imtiaz, N. H. Neom, N. Siddique, and H. Wang, "A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier," *Expert Syst Appl*, vol. 213, p. 118914, Mar. 2023, doi: 10.1016/J.ESWA.2022.118914.

[25] O. Iparraguirre-Villanueva, A. Epifanía-Huerta, C. Torres-Ceclén, J. Ruiz-Alvarado, and M. Cabanillas-Carbonell, "Breast Cancer Prediction using Machine Learning Models," *International Journal of Advanced*

Computer Science and Applications, vol. 14, no. 2, pp. 610–620, Dec. 2023, doi: 10.14569/IJACSA.2023.0140272.

[26] D. Yang *et al.*, “Compressive strength prediction of concrete blended with carbon nanotubes using gene expression programming and random forest: hyper-tuning and optimization,” *Journal of Materials Research and Technology*, vol. 24, pp. 7198–7218, May 2023, doi: 10.1016/J.JMRT.2023.04.250.

[27] Z. XI, Y. LYU, Y. KOU, Z. LI, and Y. LI, “An online ensemble semi-supervised classification framework for air combat target maneuver recognition,” *Chinese Journal of Aeronautics*, vol. 36, no. 6, pp. 340–360, Jun. 2023, doi: 10.1016/J.CJA.2023.04.020.

[28] M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,” *Decision Analytics Journal*, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/J.DAJOUR.2022.100071.

[29] M. M. Rahma and A. D. Salman, “Heart Disease Classification–Based on the Best Machine Learning Model,” *Iraqi Journal of Science*, vol. 63, no. 9, pp. 3966–3976, Sep. 2022, doi: 10.24996/IJS.2022.63.9.28.

[30] D. K. Jana, P. Bhunia, S. Das Adhikary, and A. Mishra, “Analyzing of salient features and classification of wine type based on quality through various neural network and support vector machine classifiers,” *Results in Control and Optimization*, vol. 11, p. 100219, Jun. 2023, doi: 10.1016/J.RICO.2023.100219.

[31] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, and A. Hussain, “A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms,” *Healthcare Analytics*, vol. 3, p. 100185, Nov. 2023, doi: 10.1016/J.HEALTH.2023.100185.

[32] B. Dimitrijevic, R. Asadi, and L. Spasovic, “Application of hybrid support vector Machine models in analysis of work zone crash injury severity,” *Transp Res Interdiscip Perspect*, vol. 19, p. 100801, May 2023, doi: 10.1016/J.TRIP.2023.100801.

[33] J. Luke and Suharjito, “Data Mining of Automatically Promotion Tweet for Products and Services Using Naïve Bayes Algorithm to Increase Twitter Engagement Followers atPT. Bobobobo,” *Procedia Comput Sci*, vol. 59, pp. 254–261, Jan. 2015, doi: 10.1016/J.PROCS.2015.07.550.

[34] M. Shaheen, N. Naheed, and A. Ahsan, “Relevance-diversity algorithm for feature selection and modified Bayes for prediction,” *Alexandria Engineering Journal*, vol. 66, pp. 329–342, Mar. 2023, doi: 10.1016/J.AEJ.2022.11.002.

[35] S. Lahmiri, “A comparative study of statistical machine learning methods for condition monitoring of electric drive trains in supply chains,” *Supply Chain Analytics*, vol. 2, p. 100011, Jun. 2023, doi: 10.1016/J.SCA.2023.100011.

[36] L. Yi and W. Kang, “A New Genetic Programming Algorithm for Building Decision Tree,” *Procedia Eng*, vol. 15, pp. 3658–3662, Jan. 2011, doi: 10.1016/J.PROENG.2011.08.685.

[37] Z. Guo, Y. Shi, F. Huang, X. Fan, and J. Huang, “Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management,” *Geoscience Frontiers*, vol. 12, no. 6, p. 101249, Nov. 2021, doi: 10.1016/J.GSF.2021.101249.

[38] O. Iparraguirre-Villanueva *et al.*, “Comparison of Predictive Machine Learning Models to Predict the Level of Adaptability of Students in Online Education,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 4, pp. 494–503, 2023, doi: 10.14569/IJACSA.2023.0140455.

[39] W. Książek, M. Gandor, and P. Pławiak, “Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma,” *Comput Biol Med*, vol. 134, p. 104431, Jul. 2021, doi: 10.1016/J.COMPBIOMED.2021.104431.

[40] T. M. Jawa, “Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia,” *Alexandria Engineering Journal*, vol. 61, no. 10, pp. 7995–8005, Oct. 2022, doi: 10.1016/J.AEJ.2022.01.047.