Heart Disease Prediction using Machine Learning

Aadib Malim, Kaif Kalokhe, Nouman Kupe, Aquib Mulla
Student
Department of Artificial inntelligence and Machine Learning
Anjuman I Islam Abdul Razzaq Kalsekar Polytechnique,
Panvel, India – 410206

Kirti Karande
Assistant Professor
Department of Artificial inntelligence and Machine Learning
Anjuman I Islam Abdul Razzaq Kalsekar Polytechnique,
Panvel, India – 410206

Abstract
Machine Learning (ML) is increasingly applied in various sectors globally, and the healthcare sector is no exception. In particular, ML can significantly contribute to the early detection of locomotor disorders and heart diseases. Timely predictions can offer valuable insights to physicians, enabling them to tailor their diagnostic and treatment strategies for individual patients. This project focuses on the use of ML algorithms to predict the likelihood of heart disease in individuals. It involves a comparative analysis of several classifiers, including decision trees, Naïve Bayes, Logistic Regression, SVM, and Random Forest. Furthermore, the project introduces an ensemble classifier that combines the strengths of both robust and less robust classifiers. This approach allows for the utilization of numerous samples for training and validation purposes. We analyze both existing classifiers and proposed classifiers like AdaBoost and XGBoost, aiming to enhance accuracy and predictive capabilities.Heart disease remains a significant concern worldwide, and early detection is crucial for preventing severe outcomes and enhancing patient care. Machine learning techniques have shown promise in increasing the accuracy of heart disease predictions. This paper discusses the use of ML in predicting heart disease, emphasizing its potential advantages and the challenges encountered. The main goal of this research is to assess the performance of various ML algorithms in predicting heart disease risk based on patient data. The dataset comprises a wide array of variables, including age, gender, blood pressure, cholesterol levels, exercise patterns, and medical history. After preprocessing
these variables, we train and test several ML
models, such as Logistic Regression, Random Forest, and Support Vector Machines (SVM), to evaluate their effectiveness.

1.Introduction

Heart disease remains a formidable challenge to global health, leading to approximately 12 million deaths worldwide each year, according to statistics from the World Health Organization. The rising tide of cardiovascular diseases across the globe has spurred a concerted effort among researchers to identify the most significant risk factors and to refine models for predicting the risk of heart disease with high accuracy. Termed as the "silent killer," heart disease's capacity to end lives without prior symptoms places a premium on the necessity of early diagnosis. In an effort to confront this challenge, our project harnesses the power of machine learning algorithms with the objective of

forecasting the likelihood of heart disease in individuals. This is achieved through meticulous analysis of patient data to ascertain their current or future heart disease status. The crucial nature of early heart disease prediction cannot be overstated; it empowers individuals identified as high-risk to implement vital lifestyle changes. These preventative measures have the potential to significantly mitigate the risk of developing severe health complications associated with heart disease.

At the core of our research initiative is the strategic employment of machine learning techniques for the predictive analysis of heart disease. By delving into extensive datasets containing detailed patient information, our goal is to uncover patterns and
indicators that signify the presence or imminent risk of heart disease. To this end, we employ a diverse array of machine learning methodologies, each chosen for its unique strengths and capabilities in analyzing complex health data. Through this comprehensive approach, we aim not only to advance the field of medical predictive analytics but also to contribute significantly to the global fight against heart disease, ultimately saving lives and improving health outcomes for individuals around the world

Cardiovascular diseases (CVDs), notably heart disease, are the primary reasons for morbidity and mortality globally, contributing to over 70% of worldwide deaths. The Global Burden of Disease Study 2017 indicates that CVDs are responsible for more than 43% of all global fatalities. Factors such as unhealthy diets, tobacco use, excessive consumption of sugar, and obesity, prevalent in high-income nations, are closely linked to heart disease. Nevertheless, the incidence of these chronic conditions is also escalating in low- and middle-income countries. The global economic impact of CVDs was estimated to be around USD 3.7 trillion for the period between 2010 and 2015.

Moreover, tools like electrocardiograms and CT scans, crucial for identifying coronary heart disease, remain prohibitively expensive and often unattainable in many low- and middle-income regions. Thus, the early identification of heart disease is imperative to diminish its physical and economic strain on individuals and organizations alike. A WHO forecast suggests that deaths due to CVDs will reach 23.6 million by 2030, predominantly caused by heart disease and stroke. This underscores the importance of employing data mining and machine learning approaches to foresee the risk of heart disease, aiming to preserve lives and alleviate the societal financial burden.

In healthcare, the daily generation of vast amounts of data through data mining reveals

hidden patterns useful for clinical diagnosis, demonstrating the significant role of data mining in the medical sector. This is supported by decades of research. Predicting heart disease requires considering multiple factors, including diabetes, hypertension, high cholesterol levels, and irregular heart rates. Often, the incompleteness of medical data can impede accurate heart disease prediction outcomes. Machine learning is pivotal in healthcare for diagnosing, detecting, and predicting various diseases. Recently, there's been a surge in interest in applying data mining and machine learning to forecast the development of specific diseases. Existing research involves data mining techniques for disease prediction. Despite attempts to predict disease progression risk, achieving precise outcomes remains a challenge. This paper's primary objective is to predict heart disease occurrence accurately.

This study explores the efficiency of different machine learning algorithms in heart disease prediction, employing methods like random forest, decision tree

classifier, multilayer perceptron, and XGBoost to create predictive models. We utilized k-modes clustering for dataset preprocessing and scaling to enhance model convergence. The dataset for this study is sourced from Kaggle and all computations, preprocessing, and visualizations were performed using Python on Google Colab. Prior research reported up to 94% accuracy using machine learning for heart disease prediction, though often limited by small sample sizes, raising concerns about the generalizability of results.

2.Literature Survey

In recent times, significant progress has been made in the healthcare sector, especially with the adoption of data mining and machine learning technologies. These advancements have proven effective in various medical fields, notably in cardiology. The swift growth in medical data collection has offered a unique chance for the development and validation of new algorithms specifically in this area. Heart disease is a major cause of death in less developed countries, making the detection of risk factors and early symptoms critical. The application of data mining and machine learning in cardiology could play a crucial role in the early identification and prevention of heart diseases. Narain et al. (2016) conducted a study with the goal of developing a cutting-edge system for predicting cardiovascular diseases (CVD) using machine learning, aiming to enhance the accuracy of the well-established Framingham risk score (FRS). Utilizing data from 689 individuals showing symptoms of CVD and a validation dataset from the Framingham study, the researchers proposed a system based on quantum neural networks for pattern recognition in CVD. This system was experimentally validated and compared with the FRS, achieving a prediction accuracy of 98.57%, significantly surpassing the FRS's 19.22% accuracy and that of other methods. The findings suggest this method could be a valuable asset for physicians in predicting CVD risk, thereby improving treatment strategies and facilitating early detection.

In their research, Shah et al. (2020) set out to craft a model to predict cardiovascular disease through machine learning. The model used data from the Cleveland heart
disease dataset, comprising 303 records and 17 variables, from the UCI machine
learning repository. Various supervised classification techniques were tested, including naive Bayes, Logistic Regression, decision trees, random forests, and k-nearest neighbors (KKN), with the Logistic Regression model showing the highest accuracy at 85.40%.

This study underlines the potential of machine learning in predicting cardiovascular diseases and stresses the importance of choosing the right models and techniques for best results.

Drod et al. (2022) aimed to identify key risk factors for CVD in patients with metabolic-associated fatty liver disease (MAFLD) using machine learning. The study involved blood biochemical tests and assessments of subclinical atherosclerosis in 191 MAFLD patients, employing multiple logistic regression, univariate feature ranking, and principal component analysis (PCA) to develop a risk identification model.

Key findings pointed to hypercholesterolemia, plaque scores, and diabetes duration as significant risk indicators, with the model accurately distinguishing between high-risk and low-risk patients.

The success of this ML approach demonstrates its utility in identifying at-risk MAFLD patients based on straightforward criteria.

Alotalibi (2019) explored the effectiveness of machine learning in predicting heart failure. Using data from the Cleveland Clinic Foundation and applying several ML algorithms like decision trees, logistic regression, random forests, naive Bayes, and support vector machines, the study employed a 10-fold cross-validation method. The decision tree algorithm emerged as the most accurate, achieving a 83.19% prediction rate, followed closely by the SVM algorithm at 82.30%. This investigation highlights machine learning's potential in heart. failure prediction and suggests further exploration into decision tree algorithms.Hasan and Bao (2020) sought to determine the best feature selection method for predicting cardiovascular disease. They evaluated three renowned feature selection methods (filter, wrapper, and embedding), later employing a Boolean method to identify common features across these techniques. Thisprocess involved two stages of feature subset retrieval, testing various models like random forests, support vector classifiers, k-nearest neighbors, naive Bayes, and XGBoost against an artificial neural network (ANN) for benchmarking. XGBoost, paired with the wrapper method, yielded the highest accuracy at 63.74%, showcasing the effectiveness of selecting the right features and models for disease prediction.

A significant challenge identified in previous research endeavors was the reliance on relatively small datasets for model training, which inherently increased the likelihood of overfitting. Overfitting occurs when a model is excessively complex, capturing noise in the training dataset rather than the underlying pattern, making it less generalizable to new, unseen data. This issue can severely limit the model's applicability to broader datasets, essentially constraining its utility in real-world scenarios where variability is much higher. In an effort to address and mitigate these concerns, our study adopted a markedly different approach by incorporating a substantially larger dataset for analysis. Specifically, we engaged a dataset encompassing records from 1200 patients, each described by 13 distinct features. This approach not only significantly diminishes the risk associated with overfitting by providing a more robust and diverse training foundation but also enhances the model's potential applicability and reliability across a wider array of patient data. By drawing on such a considerable dataset, our study aligns with the growing body of evidence that underscores the critical importance of leveraging large datasets in the development of predictive models, particularly those aimed at predicting cardiovascular diseases. This strategic choice is anticipated to foster a more accurate and universally applicable predictive tool, reinforcing the pivotal role of extensive datasets in overcoming the limitations previously encountered in the field. extensive datasets for this purpose.

3.Methodology

The primary objective of our research is to develop and enhance the capabilities of a sophisticated digital system designed for predicting the likelihood of cardiovascular diseases.

This endeavor aims at providing critical, actionable insights for both healthcare practitioners and patients alike, thereby contributing to more informed decision-making processes.

To achieve this ambitious goal, our methodology encompasses the application of a diverse array of machine learning algorithms to a meticulously chosen dataset. The insights derived from these analytical procedures are comprehensively detailed within this report.

Moving forward, to refine and improve the accuracy and reliability of our predictive model, we plan to undertake a series of preprocessing steps.

These steps include the elimination of any features deemed irrelevant to our analysis and the introduction of additional variables of significant predictive value, such as Mean Arterial Pressure (MAP) and Body Mass Index (BMI).

Furthermore, we intend to partition the dataset on the basis of gender, a move we anticipate will enhance the model's diagnostic precision across different demographics.

This strategic approach is underscored by the encouraging results documented in this report, which highlight the potential of our model to revolutionize the early detection and management ofheart disease.

Figure 1.

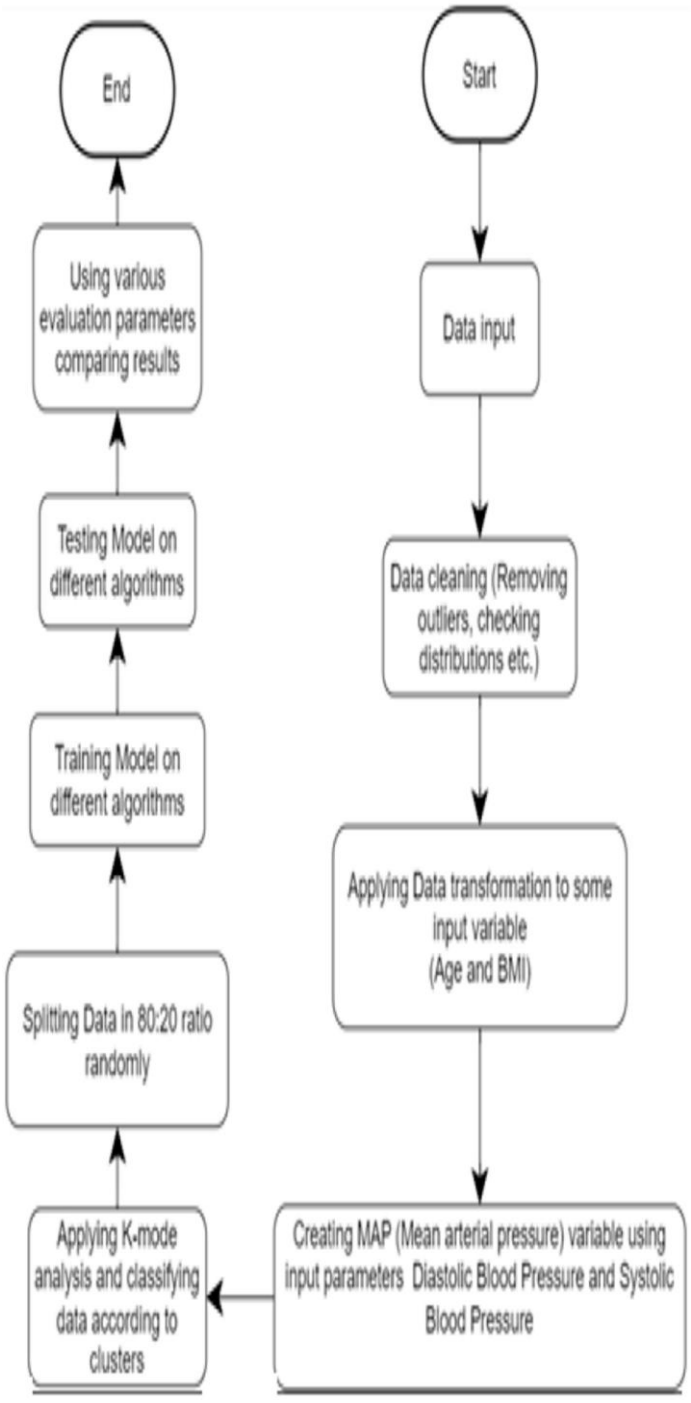


Figure 1. Flow Diagram of Model

1.1. Data Source

The dataset utilized in this study, as described in, comprises 1,200 patient records with 1 distinct features, as listed in **Table 2**. These features include age, gender, systolic blood pressure, and diastolic blood pressure. The target class, "cardio," indicates whether a patient has cardiovascular disease (represented as 1) or is healthy (represented as 0).

Feature	Variable	Min and Max Values
Age	Age	Min: 10,798 and max: 23,713
Height	Height	Min: 55 and max: 250
Weight	Weight	Min: 10 and max: 200
Gender	Gender	1: female, 2: male
Systolic blood pressure	ap_hi	Min: −150 and max: 16,020
Diastolic blood pressure	ap_lo	Min: −70 and max: 11,000
Cholesterol	Chol	Categorical value = 1(min) to 3(max)
Glucose	Gluc	Categorical value = 1(min) to 3(max)
Smoking	Smoke	1: yes, 0: no
Alcohol intake	Alco	1: yes, 0: no
Physical activity	Active	1: yes, 0: no
Presence or absence of cardiovascular disease	Cardio	1: yes, 0: no

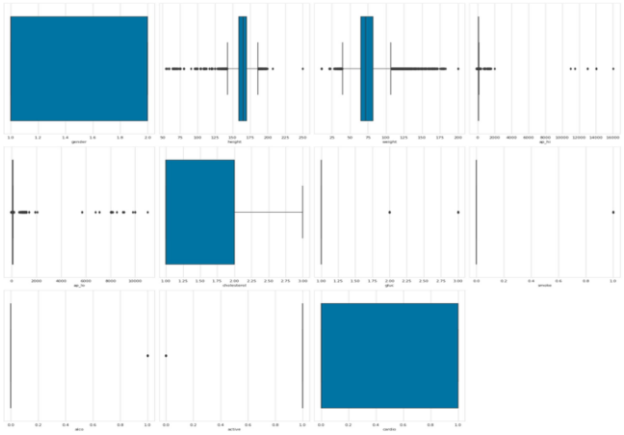
Table 2. Datasets attributes

1.2. Removing Outliers

As shown in **Figure 2**, the presence of outliers in the dataset is evident. These

Outliers might stem from data entry mistakes, and their elimination could enhance our predictive model's accuracy. To tackle this, we excluded all weight and height records not within the 2.5% to 97.5% percentile range. This outlier detection and removal were carried out by hand. Consequently, this data purification step decreased our dataset from 1,200 to 1015 entries.

Figure 2. Boxplots of all attributes.



1.3 Feature Selection and Reduction

We propose the use of binning as a method for converting continuous input, such as age, into categorical input in order to improve the performance and interpretability of classification algorithms. By categorizing continuous input into distinct groups or bins, the algorithm is able to make distinctions between different classes of data based on specific values of the input variables. For instance, if the input variable is "Age Group" and the possible values are "Young", "Middle-aged", and "Elderly", a classification algorithm can use this information to separate the data into different classes or

categories based on the age group of the individuals in the dataset .

Additionally, converting continuous input into categorical input through binning can also aid in the interpretability of the results, as it is easier to understand and interpret the relationship between the input variables and the output classes. On the other hand, continuous input, such as numerical values, can be more difficult to use in classification algorithms as the algorithm may have to make assumptions about where to draw boundaries between different classes or categories .

In this study, we applied the method of binning to the attribute of age in a dataset of patients. The age of patients was initially given in days, but for better analysis and prediction, it was converted to years by dividing it by 365. The age data were then divided into bins of 5-year intervals, ranging from 0–20 to 95–100. The minimum age in the dataset is 30 years, and the maximum is 65, so the bin 30–35 is labeled as 0, while the last bin60–65 is marked as 6.

Furthermore, other attributes with continuous values, such as height, weight, ap_hi, and ap_lo, were also converted into categorical values. The results of this study demonstrate that converting continuous input into categorical input through binning can improve the performance and interpretability of classification algorithms.

In this comprehensive study of US individuals who were unencumbered by clinical CVD at the outset, the participants had a high lifetime risk for developing CVD, and even higher for those who were overweight or obese. In comparison with those with a normal BMI, obese people were shown to have an earlier beginning of incident CVD, a larger percentage of life spent with CVD morbidity (unhealthy life years), and a lower overall survival rate. This suggests that the attributes of height and weight can be converted to body mass index (BMI), which could improve the performance of our heart disease prediction model. The BMI values are then converted into categorical values forfurther analysis.

$$BMI = \frac{\text{weight}\left(\text{kg}/\text{lb}\right)}{\text{height}^2\left(\text{m}^2/\text{in}^2\right)}$$

The average blood pressure a person has during a single cardiac cycle is known as mean arterial pressure (MAP) in medicine. MAP is a measure of peripheral resistance and cardiac output, and has been shown to be linked to significant CVD events in the ADVANCE study [27,28]. Ina research including people with type 2 diabetes, it was shown that for every 13 mmHg rise in MAP, the risk of CVD rose by 13%. Additionally, if MAP raises the risk of CVD in people with type 2 diabetes, it should also result in a higher number of CVD hospitalizations [28]. These findings suggest a direct relationship between MAP and CVD.

Mean Arterial Pressure (MAP)=(2 Diasto
llic Blood Pressure + Sistolic Blood Pressure)/3

We calculated the mean arterialpressure (MAP) from the diastolic blood pressure (ap_lo) and systolic blood pressure (ap_hi) data for each instance. Similar to the age attribute, the MAP data were divided into bins of 10 intervals, ranging from 70–80 to 110–120, and each bin was labeled with a categorical number, as shown in **Table 3**.

MAP Values	Category
≥70 and <80	1
≥80 and <90	2
≥100 and <110	3
≥100 and <110	4
≥110 and <120	5

Table 3. MAP categorical values.

As can be observed from **Table 4**, all the attribute values were converted to categorical values. This breakdown of the data facilitated the model to generate more precise predictions.

Feature	Variable	Min and Max Values
Gender	gender	1: male, 2: female
Age	Age	Categorical values = 0(min) to 6(max)
BMI	BMI_Class	Categorical values = 0(min) to 5(max)
Mean arterial pressure	MAP_Class	Categorical values = 0(min) to 5(max)
Cholesterol	Cholesterol	Categorical values = 1(min) to 3(max)
Glucose	Gluc	Categorical values = 1(min) to 3(max)
Smoking	Smoke	1: yes, 0: no
Alcohol intake	Alco	1: yes, 0: no
Physical activity	Active	1: yes, 0: no
Presence or absence of cardiovascular disease	Cardio	1: yes, 0: no

Table 4. Final attributes after feature selection and reduction.

1.3. Clustering

Clustering is a machine learning technique where a group of instances is grouped based on similarity measures. One common algorithm used for clustering is the k-means algorithm, but it is not effective when working with categorical data. To overcome this limitation, the k- modes algorithm was developed. The k- modes algorithm, introduced by Huang in 1997, is similar to the k-means algorithm but utilizes dissimilarity measures for categorical data and replaces the means of the clusters with modes. This allows the algorithm towork effectively with categorical data.Since our data have been converted to categorical

data, we will use k-modes analysis. To find the optimal number of clusters, we will first use the elbow curve with Huang initialization. An elbow curve creates a k-modes model with that number of clusters, fits the model to the data, and then calculates the cost (distance between the attribute modes of each cluster and the data points assigned to the cluster).

The costs are then plotted on a graph using the "elbow method" to determine the optimal number of clusters.

The elbow method looks for a "knee" or inflection point in the plot of costs, which is often interpreted as the point where the addition of more clusters is not significantly improving the fit of the model.

Splitting the dataset on the basis of gender can be advantageous for prediction due to the existence of significant biological disparities between men and women that can impact the manifestation and progression of diseases.

For instance, men tend to develop heart disease at an earlier age than women, and their symptoms and risk factors may differ.

Studies have shown that men have a higher risk of coronary artery disease (CAD) compared with women, and that the CAD risk factors and presentations may differ between the sexes. By analyzing the data separately for men and women, it is possible to identify unique risk factors and patterns of disease progression that may not be discernible when the data are consolidated.

Additionally, heart disease has a varying prevalence rate among men and women

Subsequently, we utilized the elbow curve method to determine the optimal number of clusters for both the male and female datasets. As depicted in **Figure 3** and **Figure 4**, the knee point was located at 2.0 in both cases, indicating that 2 was the optimal number of clusters for both the male and female datasets.

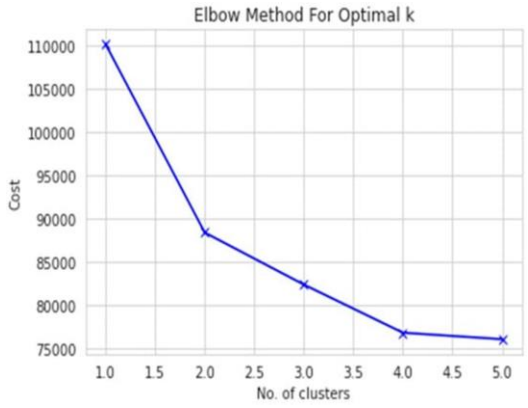


Figure 3. Male dataset.

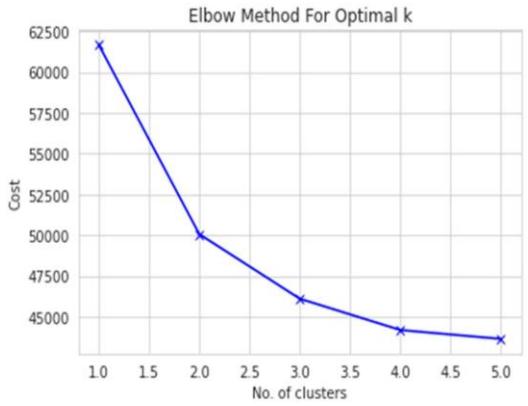


Figure 4. Female dataset.

1.4. Correlation Table

Further, a correlation table is prepared to determine the correlation between different categories. From **Figure 5**, mean arterial pressure (MAP_Class), cholesterol, and age were highly correlated factors. Intra-feature dependency can also be looked upon with the help of this matrix.

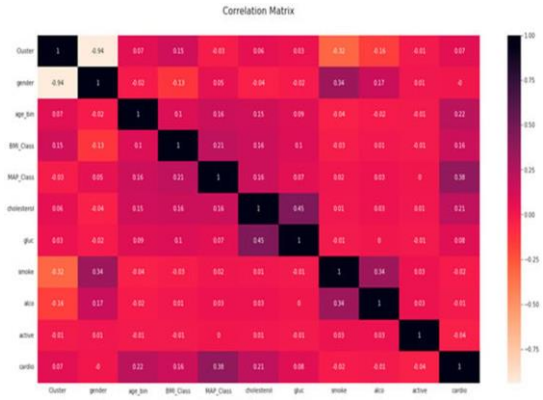


Figure 5. Correlation heatmap.

1.5. Modeling

A training dataset (80%) and a testing dataset (20%) are created from the dataset. A model is trained using the training dataset, and its performance is assessed using the testing dataset. Different classifiers, such as decision tree classifier, random forest classifier, multilayer perceptron, and XGBoost, are applied to the clustered dataset to assess their performance. The performance of each classifier is then evaluated using accuracy, precision, recall, and F-measure scores.

1.5.1. Decision Tree Classifier

Decision trees are treelike structures that are used to manage large datasets. They are often depicted as flowcharts, with outer branches representing the results and inner nodes representing the properties of the dataset. Decision trees are popular because they are efficient, reliable, and easy to understand. The projected class label for a decision tree originates from the tree's root. The following steps in the tree are decided by comparing the value of the root attribute with the information in therecord. Following a jump on the next node, the matching branch is followed to the value shown by the comparison result. Entropy changes when training examples are divided into smaller groups using a

decision tree node. The measurement of this change in entropy is information gain.

$$Entropy\left(S\right) = \sum_{i=1}^{c} -\left(P_i \log_2 P_i\right)$$

$$Information\ \ Gain\ \ \left(S,A\right) = Entropy\left(S\right) - \sum_{v \in\ values\left(A\right)} \frac{\left|S_v\right|}{\left|S\right|} Entropy\left(S_v\right)$$

An accuracy of 73.0% has been achieved by the decision tree]. In a research by, 72.77% accuracy was achieved by the decision tree classifier.

1.5.2. Random Forest

The random forest algorithm belongs to a category of supervised classification technique that consists of multiple decision trees working together as a group. The class with the most votes become the prediction made by our model. Each tree in the random forest makes a class prediction, which eliminates the limitations of the decision tree algorithm. This improves accuracy and reduces overfitting of the dataset. When used on large datasets, the random forest approach may still provide the same results even if a significant portion of record values are missing. The samples produced by the decision tree may be saved and used with various data types . In the research in , random forest achieved a test accuracy of 73% and a validation accuracy of 72% with 500 estimators, 4 maximum depths, and 1 random state.

1.5.3. Multilayer Perceptron

The multilayer perceptron (MLP) is a type of artificial neural network that consists of multiple layers. Single perceptron can only solve linear problems, but MLP is better suited for nonlinear examples.

MLP is used to tackle complex issues. A feedforward neural network with many layers is an example of an MLP.

Other activation functions beyond the step function are usually used by MLP. The buried layer neurons often perform sigmoid functions. As with step functions, smooth transitions rather than rigid decision limits are produced using sigmoid functions . In MLPs, learning also comprises adjusting the perceptron's weights to obtain the lowest possible error. This is accomplished via the backpropagation technique, which reduces the MSE.

1.5.4. XGBoost

XGBoost is a version of gradientboosted decision trees. This algorithm involves creating decision trees in a sequential manner. All the independent variables are allocated weights, which are subsequently used to produce predictions by the decision tree. If the tree makes a wrong prediction, the importance of the relevant variables is increased and used in the next decision tree. The output of each of these classifiers/predictors is thenmerged to produce a more robust and accurate model. In a study by , the XGBoost model achieved 73% accuracy with the parameters 'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 100, 'cross-validation': 10 folds including 9,00 training and 210 testing data instances on 1,200 CVD dataset.

2. Results

In this study, Google Colab was employed on a computer equipped with a Ryzen 7 4800-H processor and 16GB of RAM. Initially, the dataset featured 70,000 instances across 12 attributes, but following data cleaning and preprocessing efforts, it was narrowed down to roughly 900 instances and 11 attributes. Due to the categorical nature of all attributes, outlier removal was conducted to enhance model performance. The algorithms explored in this research included random forest, decision tree, multilayer perceptron, and

XGBoost classifier. Various metrics were utilized to assess performance, including precision, recall, accuracy, F1 score, and the area under the ROC curve.

The data was divided, allocating 80% for training the model and the remaining 20% for testing.

Model	Accuracy		Precision		Recall		F1-Score		AUC
	Without CV	CV	Without CV	CV	Without CV	CV	Without CV	CV	
MLP	86.94	87.28	89.03	88.70	82.95	84.85	85.88	86.71	0.95
RF	86.92	87.05	88.52	89.42	83.46	83.43	85.91	86.32	0.95
DT	86.53	86.37	90.10	89.58	81.17	81.61	85.40	85.42	0.94
XGB	87.02	86.87	89.62	88.93	82.11	83.57	86.30	86.16	0.95

Table 5 displays the application of several machine learning classifiers, such as MLP, RF, decision tree, and XGBoost, on the cardiovascular disease dataset to detect the presence of cardiovascular disease following the adjustment of hyperparameters. The findings demonstrate that the multilayer perceptron (MLP) algorithm achieved the highest accuracy in cross-validation at 87.28%, in addition to attaining notable scores in recall, precision, F1 score, and AUC at 84.85, 88.70, 86.71, and 0.95, respectively. The accuracy of all classifiers exceeded 86.5%. Through hyperparameter optimization using GridSearchCV, the accuracy of the random forest algorithm saw an improvement of 0.5%, moving from 86.48% to 86.90%. Likewise, the XGBoost algorithm's accuracy was enhanced by 0.6%, rising from 86.4% to 87.02% after hyperparameter adjustments.

Table 5. The evaluation metrics resulting from different classifiers.

A binary classifier's performance is graphically represented by the receiver operating characteristic (ROC) curve. At different categorization criteria, it shows the true positive rate (TPR) vs. the falsepositive rate (FPR). The area under the ROC curve (AUC) is a scalar metric that measures both the classifier's sensitivity and specificity

while also reflecting the classifier's overall performance. As depicted in **Figure 6** a–d, all models exhibit a high AUC of above 0.9. The multilayer perceptron (MLP),

random forest (RF), and XGBoost models have a joint highest AUC of 0.95.

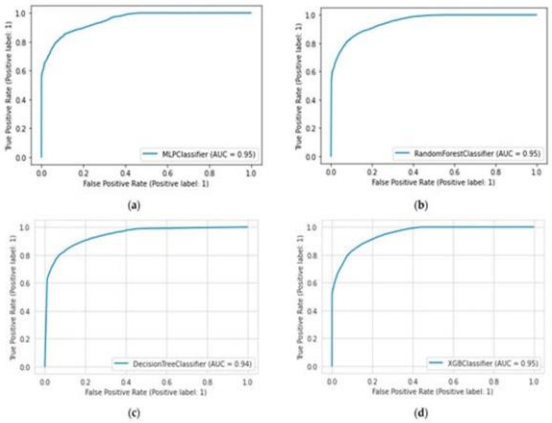


Figure 6. ROC–area under curve of (**a**) MLP, (**b**) RF, (**c**) DT, and (**d**) XGB.

3. Conclusions

This investigation primarily sought to categorize heart disease by applying several models to a dataset derived from real-world scenarios. Utilizing the k-modes clustering algorithm, this study aimed to foresee the occurrence of heart disease within a patient dataset. In the preprocessing phase, the dataset underwent modifications such as transforming age into yearly brackets segmented into 5-year periods and categorizing both diastolic and systolic blood pressure readings into 10-unit intervals. Additionally, to consider the distinct disease profiles and trajectories in different genders, the dataset was categorized based on gender.

The method known as the elbow curve was employed

to identify the most suitable number of clusters for datasets corresponding to both genders. The outcomes revealed that the MLP model achieved the highest precision, registering an accuracy rate of 87.23%.

These results underscore the efficacy of k-modes clustering in the precise forecasting of heart disease, proposing its utility in crafting specific approaches for diagnosis and treatment. The research utilized data from the Kaggle cardiovascula disease dataset, comprising 1200 entries, with all algorithmic implementations carried out on Google Colab. The accuracy rates for all algorithms surpassed 85%, with decision trees recording the lowest at 86.37% and the multilayer perceptron (MLP) recording the highest, as noted earlier.

Challenges of the Study. Despite the optimistic outcomes, several challenges merit attention. Initially, the research relied on a single dataset, which may limit its applicability across varied populations or groups of patients. Moreover, the investigation only included a restricted array of demographic and clinical indicators, overlooking other possible heart disease risk factors such as lifestyle choices or genetic markers. The efficacy of the model was not tested against an external dataset, which could have offered insights into its adaptability to new, unseen data. Furthermore, the study did not evaluate the interpretability of the algorithm's clustering outcomes. Given these challenges, it is advisable to undertake additional research to overcome these obstacles and deepen the understanding of the k-modes clustering approach for predicting heart disease.

Prospects for Future Exploration. Subsequent studies could endeavor to surmount the limitations noted in this research by contrasting the efficacy of the k-modes clustering method against other prevalent clustering

techniques like k-means or hierarchical clustering. This comparison might provide a richer understanding of its performance. Moreover, assessing the impact of missing data and anomalies on the model's precision and devising methods to address these situations would be invaluable. Evaluating the model's performance using a separate test dataset could also affirm its applicability to novel, unseen data. Future investigations should aim to validate the sturdiness and broad applicability of these findings and the clarity of the algorithm's clustering results, which could enhance comprehension of the findings and inform decision-making based on the research outcomes.

REFERENCES

[1] A. S. Abdullah and R. Rajalaxmi, "A data mining model for predicting the coronary heart disease using random forest classifier," in Proc.Int. Conf. Recent Trends Comput. Methods, Commun. Controls,

Apr. 2012, pp. 22-25:

[2] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, "Using PSO algorithm for producing best rules in diagnosis of heartdisease,' in Proc.Int.Conf.Comput. Appl. (ICCA),

Sep. 2017, pp. 306- 311.

[3] N. Al-milli, "Backpropogation neural network for prediction of heart disease," J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131-135,2013.

[4] C.A.Devi, S. P. Rajamhoana, K. Umamaheswari R. Kiruba, K. Karunya, and R. Deepika, "Analysis of neural networks based heart

disease prediction system,' in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018,pp. 233-239.

[5] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease usingweighted fuzzy rules,"' J. King Saud Univ.- Comput. Inf. Sci.,vol.24,no.1,pp.2740,Jan.2012.doi:10.1016/j jksuci.2011.09.002

[6] L. Baccour, "Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI datasets," Expert Syst. Appl., vol. 99, pp. 115-125, Jun. 2018.doi:10.1016/j.eswa.2018.01.05

[7] C.-A. Cheng and H.-W. Chiu, "An artificial Neural Network model for the evaluation of carotid artery stenting prognosis sing anational-wide database,' in Proc. 39th Annu. Int.Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2017,p-ISSN: 2566-2395