Machine Learning Techniques for Precise Heart Disease Prediction

Mukesh Kumar Saini

PhD, MBA, MCA, Technologist and Consulting Data Engineer
Email: *drmsaini78[at]gmail.com*

Abstract: ***Diagnosing and forecasting cardiovascular disease represents a pivotal task in medicine, crucial for accurately categorizing and effectively treating patients under the care of cardiologists. Within the medical domain, the integration of machine learning has grown, offering the capability to identify patterns from extensive datasets. Employing machine learning for the classification of cardiovascular disease occurrences holds promise in reducing diagnostic errors. This study introduces a novel method using k-modes clustering with Huang initialization to enhance the precision of classification. Various models, including random forest (RF), decision tree (DT), multilayer perceptron (MLP), and XGBoost (XGB), were employed and their parameters optimized using GridSearchCV. Evaluation was conducted on a practical dataset comprising 70,000 instances sourced from Kaggle, yielding the following accuracies: decision tree: 86.37% (with cross-validation) and 86.53% (without), XGBoost: 86.87% (with) and 87.02% (without), random forest: 87.05% (with) and 86.92% (without), multilayer perceptron: 87.28% (with) and 86.94% (without). Additionally, these models demonstrated robust AUC values: decision tree: 0.94, XGBoost: 0.95, random forest: 0.95, multilayer perceptron: 0.95. The study concludes that the multilayer perceptron model, particularly with cross-validation, exhibited superior performance with the highest accuracy of 87.28%.***

Keywords: heart disease; machine learning; k-modes; classification; multilayer perceptron; model evaluation

1. Introduction

Globally, cardiovascular disease (CVDs) is the primary cause of morbidity and mortality, responsible for more than 70% of all deaths. According to the 2017 Global Burden of Disease study, CVDs account for over 43% of fatalities [1,2]. Risk factors such as poor diet, smoking, high sugar intake, and obesity are prevalent in high-income countries [3,4], while chronic disease rates are rising in low- and middle-income nations [5]. The economic burden of CVDs from 2010 to 2015 was estimated at approximately USD 3.7 trillion [6,7].

Diagnostic technologies like electrocardiograms and CT scans, crucial for detecting coronary heart disease, are often prohibitively expensive in many regions, contributing to significant mortality rates [5]. Employers also face substantial healthcare costs, with 25-30% of annual medical expenditures linked to employees with cardiovascular disease [8]. Early detection is crucial to mitigate both the physical and financial impacts of heart disease. The World Health Organization projects that CVD-related deaths could reach 23.6 million by 2030, primarily from heart disease and stroke [9]. To reduce these numbers and alleviate economic burdens, leveraging data mining and machine learning techniques to predict heart disease risk is imperative.

In the medical field, data mining uncovers hidden patterns crucial for clinical diagnosis [10], essential given factors like diabetes, hypertension, high cholesterol, and irregular pulse rates when predicting heart disease [11]. Machine learning, pivotal in healthcare, aids in disease diagnosis, detection, and prognosis. Recent interest has focused on using these technologies to forecast disease likelihood, though accurate predictions remain challenging [12]. This study aims to precisely predict heart disease probabilities, evaluating various machine learning algorithms including random forest [13], decision trees, multilayer perceptrons, and XGBoost [14].

To enhance model performance, k-modes clustering was applied for dataset preprocessing and scaling. Using a publicly available dataset from Kaggle, computations, preprocessing, and visualization were conducted using Python in Google Colab. Previous studies reported up to 94% accuracy in heart disease prediction with machine learning [15], but these results often suffer from limited sample sizes and may lack generalizability. Our research addresses this gap by employing a larger and more diverse dataset, enhancing the robustness and applicability of our findings.

2. Literature Survey

In recent years, the healthcare industry has made significant strides in leveraging data mining and machine learning techniques, particularly within medical cardiology. These technologies have proven effective across various healthcare applications, aiming to identify risk factors and early signs of heart disease, a leading cause of mortality in developing nations [12–16].

Narain et al. (2016) [17] conducted a study focusing on enhancing the accuracy of cardiovascular disease (CVD) prediction using a novel machine-learning-based system. Their approach, utilizing a quantum neural network, demonstrated an impressive forecasting accuracy of 98.57%, far surpassing the 19.22% accuracy of the widely used Framingham risk score (FRS) and other existing methods. This advancement suggests promising potential for doctors in improving treatment plans and enabling early diagnosis.

Shah et al. (2020) [18] aimed to develop a predictive model for cardiovascular disease using machine learning techniques. They employed various supervised classification methods on the Cleveland heart disease dataset, achieving the highest accuracy of 90.8% with the k-nearest neighbor (KNN) model. This underscores the effectiveness of machine learning in

Licensed Under Creative Commons Attribution CC BY
DOI: https://dx.doi.org/10.21275/SR23121209074

cardiovascular disease prediction and stresses the importance of selecting appropriate models for optimal outcomes.
Drod et al. (2022) [2] focused on identifying significant risk variables for cardiovascular disease in patients with metabolic-associated fatty liver disease (MAFLD) using machine learning techniques. Their study highlighted the success of multiple logistic regression, univariate feature ranking, and principal component analysis (PCA) in identifying critical clinical characteristics. The model achieved an AUC of 0.87, demonstrating its utility in detecting high-risk CVD patients.

Alotalibi (2019) [19] explored machine learning techniques for predicting heart failure using data from the Cleveland Clinic Foundation. Their study found that the decision tree algorithm exhibited the highest accuracy of 93.19%, followed closely by support vector machines (SVM) at 92.30%. This research underscores the potential of machine learning as a robust tool for heart disease prediction.

Hasan and Bao (2020) [20] compared various feature selection methods and machine learning algorithms for predicting cardiovascular disease. Their study identified the XGBoost classifier coupled with the wrapper technique as the most accurate, achieving 73.74% accuracy. This highlights the importance of effective feature selection in enhancing predictive performance.

One common limitation across these studies is the use of relatively small datasets, which may lead to overfitting and limit generalizability. In contrast, our study utilized a dataset comprising 70,000 patients and 11 features, aiming to mitigate overfitting and enhance applicability. Table 1 summarizes these cardiovascular disease prediction studies conducted on large datasets, reinforcing the efficacy and relevance of utilizing substantial data in advancing predictive analytics.

Table 1: Related work on heart disease prediction using large datasets

Authors	Approach	Best Accuracy	Dataset
Shorewall, 2021 [5]	Stacking of KNN, random forest, and SVM outputs with logistic regression as the metaclassifier	75.1% (stacked model)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Maiga et al., 2019 [7]	- Random forest -Naive Bayes -Logistic regression -KNN	70%	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Waigi at el., 2020 [12]	Decision tree	72.77% (decision tree)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
Our and ElSeddawy, 2021 [21]	Repeated random with random forest	89.01% (random forest classifier)	UCI cardiovascular dataset (303 patients, 14 attributes)
Khan and Mondal, 2020 [22]	Holdout cross-validation with the neural network for Kaggle dataset	71.82% (neural networks)	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)
	Cross-validation method with logistic regression (solver: lbfgs) where k = 30	72.72%	Kaggle cardiovascular disease dataset 1 (462 patients, 12 attributes)
	Cross-validation method with logistic linear SVM where k = 10	72.22%	Kaggle cardiovascular disease dataset (70,000 patients, 12 attributes)

3. Methodology

This study aims to predict the probability of heart disease through computerized heart disease prediction, which can be beneficial for medical professionals and patients. To achieve this objective, we employed various machine learning algorithms on a dataset and present the results in this study report. To enhance the methodology, we plan to clean the data, eliminate irrelevant information, and incorporate additional features such as MAP and BMI. Next, we will separate the dataset based on gender and implement k-modes clustering. Finally, we will train the model with the processed data. The improved methodology will produce more accurate results and superior model performance, as demonstrated in Figure 1.

3.1 Data Source

The dataset utilized in this study, as described in [23], comprises 70,000 patient records with 12 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).

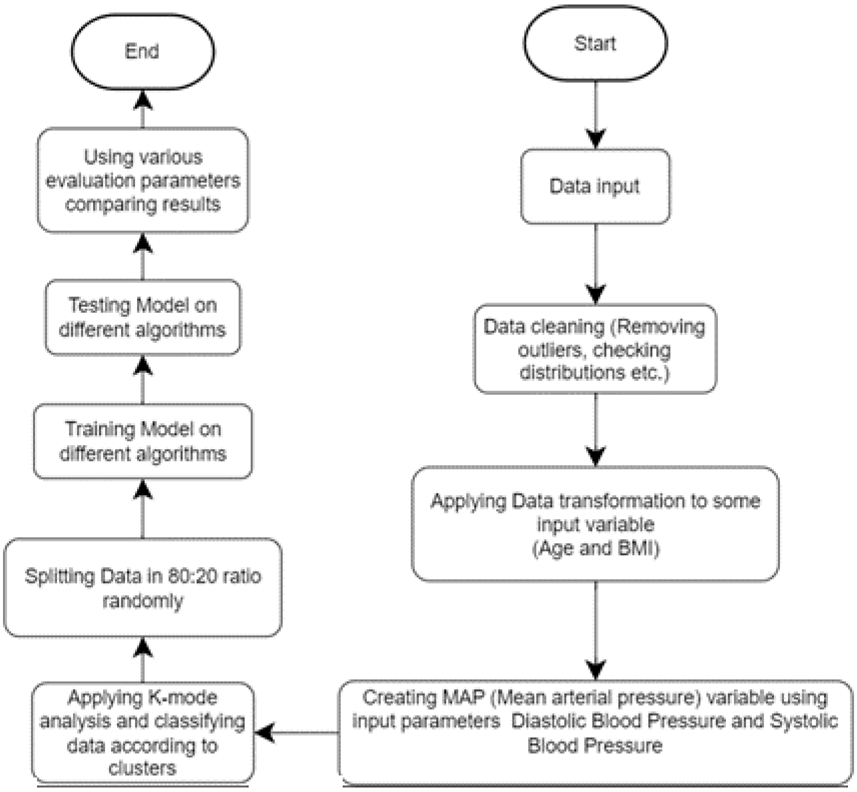

Start
Data input
Data cleaning (Removing outliers, checking distributions etc.)
Applying Data transformation to some input variable (Age and BMI)
Creating MAP (Mean arterial pressure) variable using input parameters Diastolic Blood Pressure and Systolic Blood Pressure
Applying K-mode analysis and classifying data according to clusters
Splitting Data in 80:20 ratio randomly
Training Model on different algorithms
Testing Model on different algorithms
Using various evaluation parameters comparing results
End

Figure 1: Flow diagram of Model

Table 2: Datasets attributes

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

3.2 Removing Outliers

Figure 2 illustrates the clear presence of outliers within the dataset. These outliers are likely attributable to errors in data entry. Removing these outliers could enhance the predictive model's performance. To tackle this issue, we systematically excluded instances where ap_hi, ap_lo, weight, and height values fell outside the 2.5% to 97.5% range. This outlier identification and elimination process was conducted manually. Consequently, the dataset size decreased from 70,000 to 57,155 rows.

Licensed Under Creative Commons Attribution CC BY

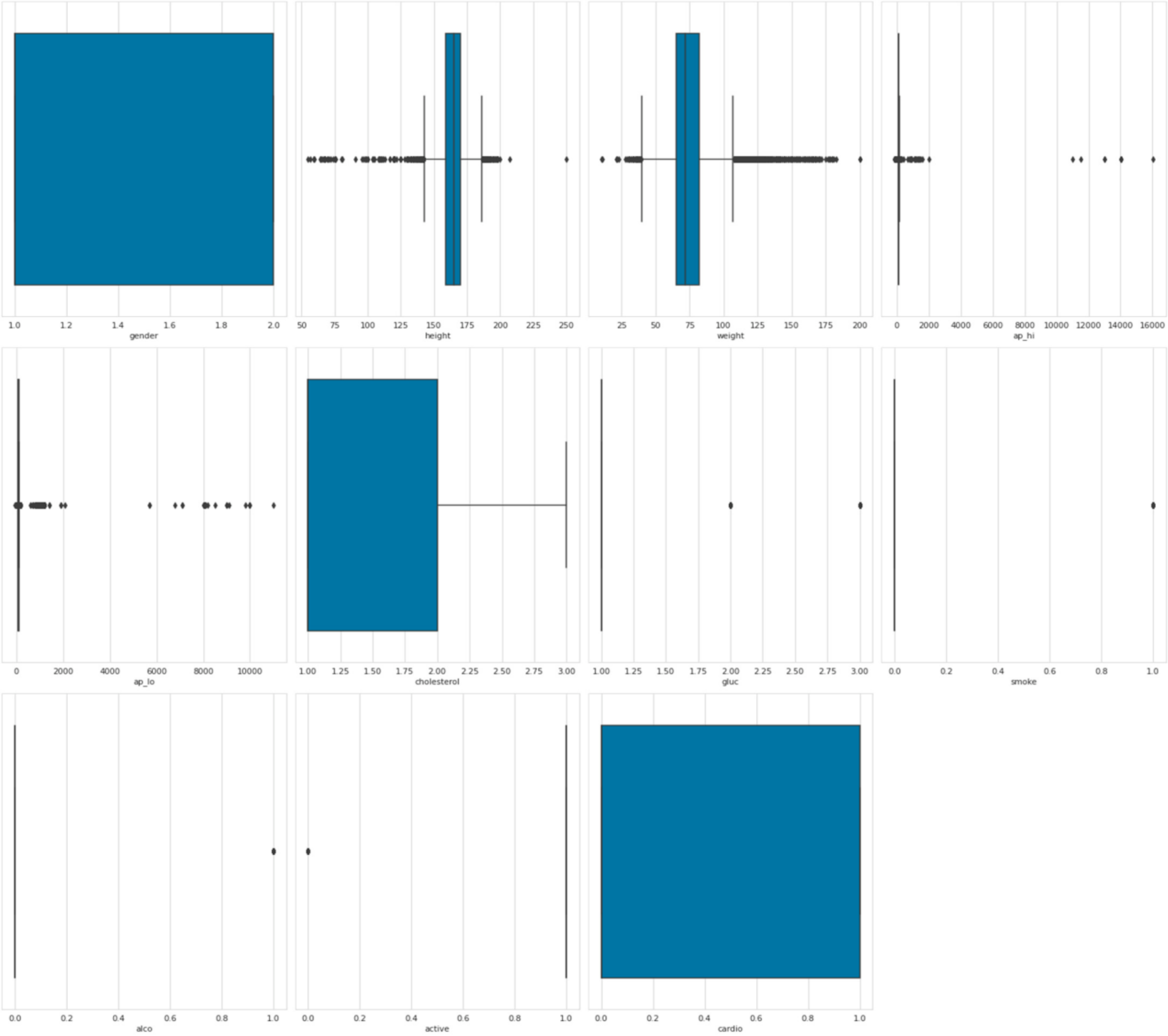


Figure 2: Boxplots of all attributes

Figure Selection and Reduction

In our research, we advocate for the utilization of binning as a method to enhance the efficacy and interpretability of classification algorithms when dealing with continuous input variables, such as age. By categorizing continuous data into distinct groups or bins, algorithms can better differentiate between different classes based on specific values of input variables. For instance, converting age into "Age Group" categories like "Young", "Middle-aged", and "Elderly" enables classifiers to effectively segregate data based on age demographics [24]. Moreover, binning continuous input aids in result interpretability by simplifying the relationship between input variables and output classes. Continuous numerical values can pose challenges for classification algorithms, which must infer boundaries between classes [25].

In our study, we applied binning to age data sourced from a patient dataset initially recorded in days, which we converted into years for analysis purposes by dividing by 365. Age values were subsequently grouped into 5-year intervals, ranging from 0–20 to 95–100. Given our dataset's age range (30 to 65 years), bins such as 30–35 were categorized as 0, while 60–65 were labeled as 6. Additionally, we applied similar binning techniques to other continuous attributes like height, weight, ap_hi, and ap_lo. Our findings indicate that converting these attributes into categorical values through binning enhances the performance and interpretability of classification algorithms.

In a comprehensive study of individuals in the US initially free from clinical cardiovascular disease (CVD), participants demonstrated a significant lifetime risk for developing CVD, particularly among those who were overweight or obese. Compared to individuals with a normal BMI, obese participants showed earlier onset of incident CVD, a higher percentage of life affected by CVD morbidity (unhealthy life years), and lower overall survival rates [26].

This highlights the potential benefit of converting continuous attributes like height and weight into categorical variables such as body mass index (BMI), thereby potentially improving the predictive accuracy of our heart disease models.

$$BMI = weight\ (kg/lb)/\ height2\ (m2/in2)$$

Mean arterial pressure (MAP) is a crucial measure in medicine, representing the average blood pressure during a single cardiac cycle. It serves as an indicator of both peripheral resistance and cardiac output, factors that are significantly linked to cardiovascular disease (CVD) events, as evidenced by studies such as ADVANCE [27,28]. Research involving individuals with type 2 diabetes has highlighted a direct correlation between MAP and CVD risk. Specifically, for every 13 mmHg increase in MAP, there is a corresponding 13% rise in the risk of experiencing cardiovascular events [28].

Moreover, the elevated MAP levels associated with type 2 diabetes are also linked to an increased likelihood of hospitalizations due to cardiovascular disease [28]. These findings underscore the critical role of MAP as a predictor of cardiovascular health outcomes, emphasizing its importance in assessing and managing CVD risk, particularly in populations with conditions such as type 2 diabetes.

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

We calculated the mean arterial pressure (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.

Table 3: MAP categorical values.

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

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.

Table 4: Final attributes after feature selection and reduction

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

3.3 Clustering

Clustering is a machine learning technique used to group instances based on similarity measures. While the k-means algorithm is widely used for clustering, it isn't effective with categorical data. To address this limitation, the k-modes algorithm was introduced by Huang [29] in 1997. Unlike k-means, k-modes uses dissimilarity measures suited for categorical data and replaces cluster means with modes, making it suitable for such datasets. Given that data has been converted to categorical format, we opted to employ k-modes analysis. To determine the optimal number of clusters, we utilized the elbow curve method with Huang initialization. This approach involves creating multiple k-modes models with varying numbers of clusters, fitting them to the data, and then evaluating the cost (distance between attribute modes of each cluster and assigned data points). The resulting costs are plotted on a graph using the elbow method to identify the optimal number of clusters, typically where adding more clusters does not significantly improve model fit.

In our study, we also recognized the potential benefits of splitting the dataset by gender for predictive analysis. Biological differences between men and women can influence disease manifestation and progression. For example, men often develop heart disease at an earlier age than women, with distinct risk factors and symptom presentations [30]. Studies indicate varying prevalence rates and risk profiles for conditions like coronary artery disease (CAD) between genders. Analyzing data separately by gender allows for the identification of unique risk factors and disease progression patterns that may not be evident when analyzing aggregated data. To this end, we applied the elbow curve method separately to datasets for males and females. Figures 3 and 4 illustrate our findings, with both curves showing a clear "knee" point at 2.0 clusters, indicating that 2 clusters were optimal for both male and female datasets. This approach helps tailor clustering analysis to account for gender-specific variations in disease characteristics and risk factors.

3.4 Correction 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.

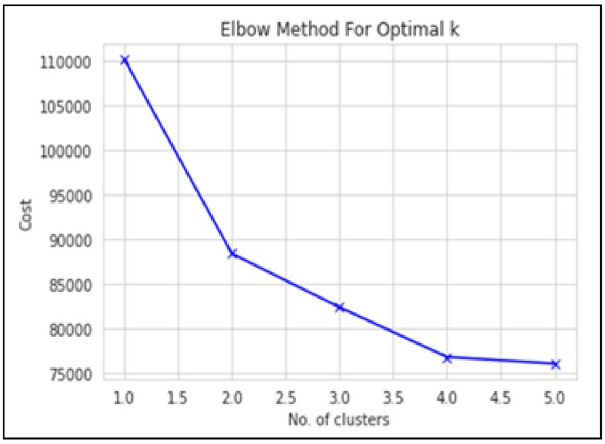


Figure 3: Male dataset

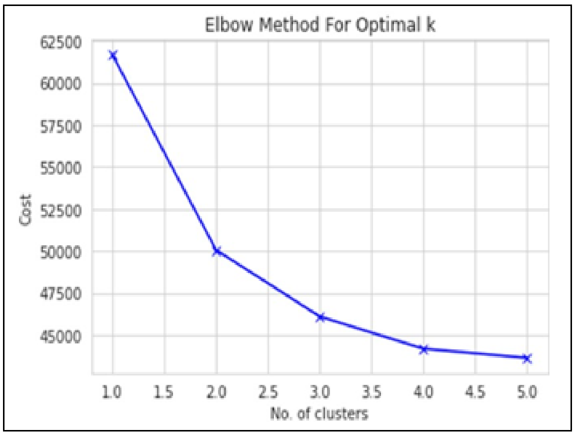


Figure 4: Female dataset

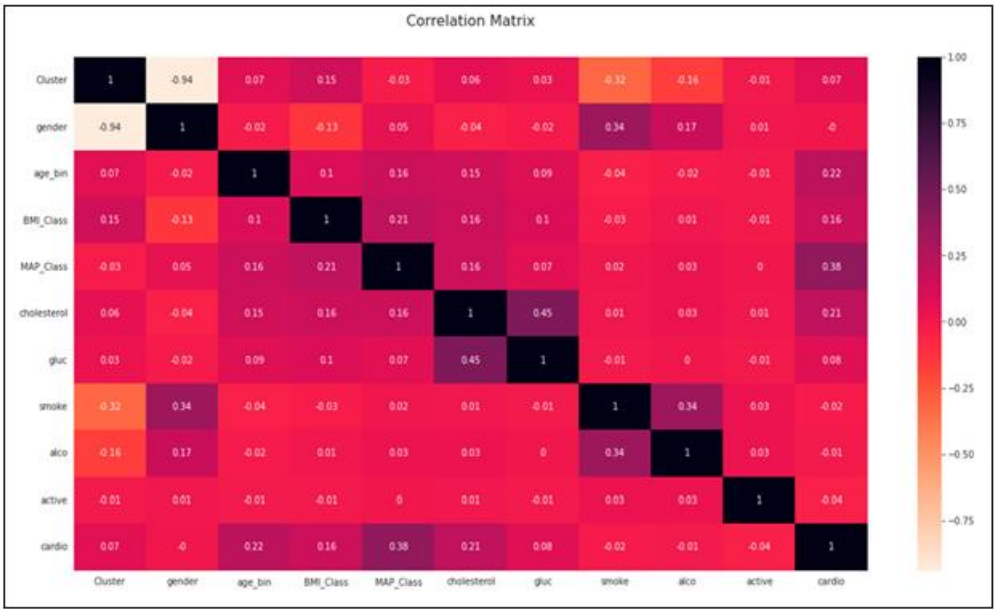


Figure 5: Correlation heatmap

3.6 Modeling

A training dataset consisting of 80% of the data and a testing dataset containing the remaining 20% are derived from the original dataset. The training dataset is used to train various classifiers such as decision tree classifier, random forest classifier, multilayer perceptron, and XGBoost. After training, each classifier's performance is evaluated using the testing dataset. This evaluation includes assessing metrics such as accuracy, precision, recall, and F-measure scores to gauge the effectiveness of each classifier in making predictions on the dataset.

3.6.1 Decision Tree Classifier

Decision trees are hierarchical structures used to manage large datasets, often depicted as flowcharts where outer branches signify outcomes and inner nodes represent dataset properties. They are favored for their efficiency, reliability, and simplicity in interpretation. Starting from the root of a decision tree, the projected class label is determined. Each subsequent step in the tree involves comparing the attribute values with the data record, directing the path through branches based on these comparisons. When a decision tree node splits training examples into smaller groups, entropy, a measure of disorder, changes. The quantification of this entropy change is termed information gain.

$$Entropy(S) = \sum c\ i=1\ -(Pi\ log2\ Pi)$$
$$Information\ Gain\ (S,\ A) = Entropy(S) - \sum v \in values(A)\ |Sv|$$
$$|S|\ Entropy(Sv)$$

An accuracy of 73.0% has been achieved by the decision tree [5]. In a research by [12], 72.77% accuracy was achieved by the decision tree classifier.

3.6.2 Random Forest

The random forest [13] algorithm is a supervised classification technique that utilizes an ensemble of multiple decision trees working collaboratively. Predictions are made by aggregating the class with the most votes across all trees. Unlike individual decision trees, each tree in a random forest independently predicts classes, which helps mitigate the shortcomings of single-tree models, improving overall accuracy and reducing overfitting to the dataset. Moreover, random forests can handle large datasets effectively, even in the presence of missing values, by leveraging samples generated by decision trees that can accommodate various data types [31]. In a study detailed in [7], a random forest model achieved a test accuracy of 73% and a validation accuracy of 72% using 500 estimators, a maximum depth of 4, and a specified random state.

3.6.3 Multilayer Perceptron

The multilayer perceptron (MLP) is an artificial neural network composed of multiple layers, making it capable of handling nonlinear problems, unlike single-layer perceptrons. MLPs are specifically designed to address complex issues. An example of an MLP is a feedforward neural network with multiple hidden layers [32].MLPs typically employ activation functions other than the step function. Neurons in the hidden layers often use sigmoid functions, which facilitate smooth transitions rather than rigid decision boundaries [33]. In MLPs, learning involves adjusting the weights of perceptrons to minimize errors. This adjustment process is achieved through backpropagation, which aims to reduce the mean squared error (MSE).

3.6.4 XGBoost

XGBoost [14] represents a variant of gradient boosted decision trees, where trees are sequentially constructed. Each independent variable is assigned weights, used for making predictions within the decision trees. Incorrect predictions lead to increased importance of relevant variables in subsequent trees. The outputs from these predictors are combined to form a robust and precise model. In a study by [34], the XGBoost model achieved 73% accuracy using parameters such as 'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 100, and employing 10-fold cross-validation with 49,000 training and 21,000 testing instances from a 70,000 CVD dataset

4. Results

This research utilized Google Colab running on a Ryzen 7 4800-H processor with 16 GB of RAM to conduct analysis on a dataset initially comprising 70,000 rows and 12 attributes. After rigorous cleaning and preprocessing, the dataset was refined to approximately 59,000 rows and 11 attributes, all of which were categorical. Outliers were systematically removed to enhance model efficiency and accuracy.

The study employed several machine learning algorithms tailored for classification tasks, namely random forest, decision tree, multilayer perceptron (MLP), and XGBoost classifiers. These algorithms were evaluated using various performance metrics including precision, recall, accuracy, F1 score, and area under the ROC curve (AUC). The dataset was split into training (80%) and testing (20%) subsets to assess model performance and generalization.

Hyperparameter tuning was automated using GridSearchCV, a method from the scikit-learn library. GridSearchCV optimizes hyperparameters by performing an exhaustive search over specified parameter values, guided by a chosen scoring method, typically utilizing k-fold cross-validation to ensure robust evaluation.

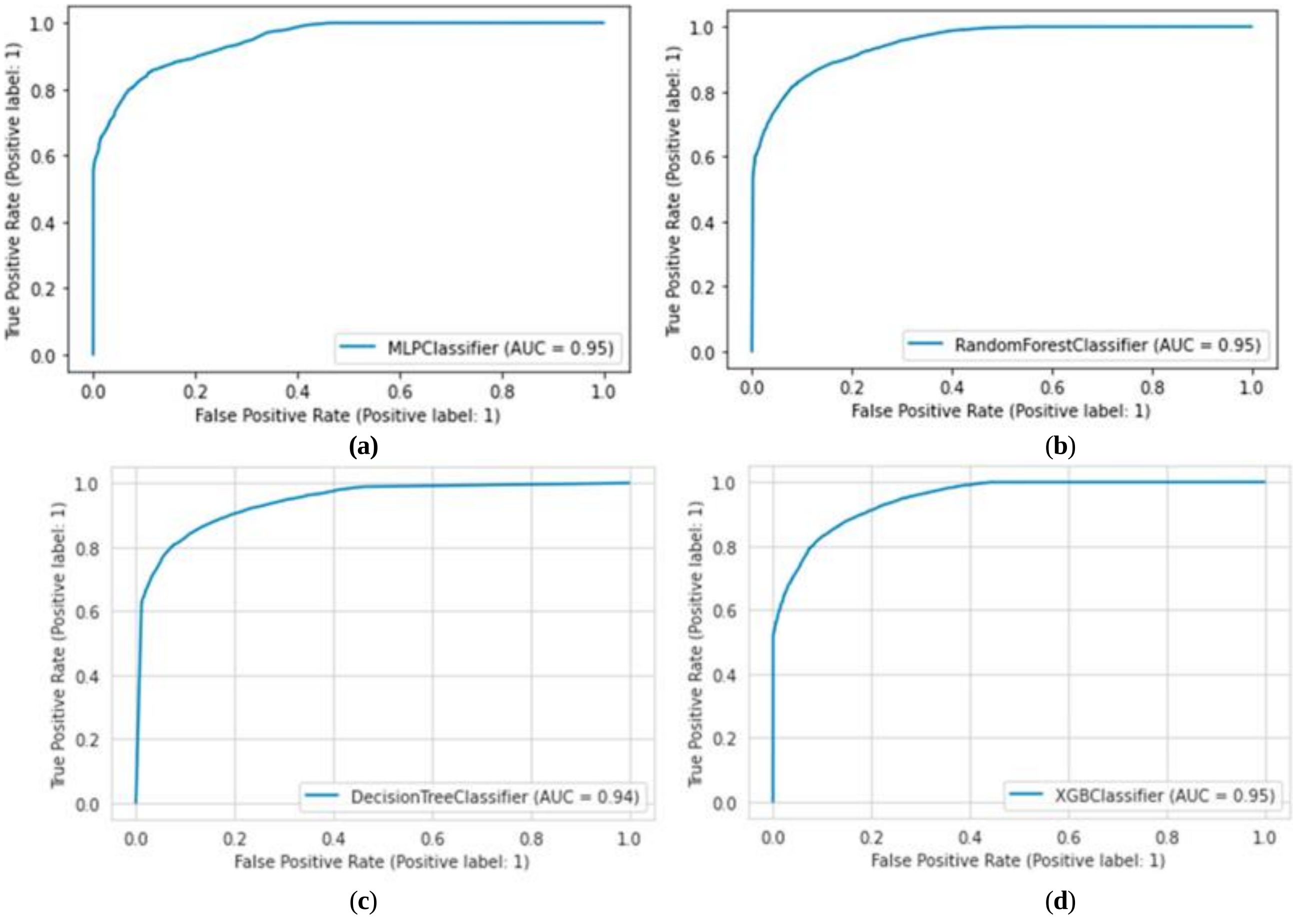
Results from the study, as summarized in Table 5, demonstrated strong performance across all classifiers post hyperparameter tuning. The MLP algorithm emerged with the highest cross-validation accuracy of 87.28%, accompanied by notable precision (88.70%), recall (84.85%), F1 score (86.71%), and AUC (0.95) metrics. Other classifiers also achieved high accuracies, with the random forest algorithm improving from 86.48% to 86.90% accuracy following hyperparameter tuning, and XGBoost achieving a significant increase from 86.4% to 87.02%.

This research showcases the effectiveness of employing advanced machine learning techniques and rigorous hyperparameter tuning to achieve robust predictive models for identifying cardiovascular disease. The utilization of GridSearchCV facilitated the optimization of model performance, ensuring that each algorithm was fine-tuned to yield optimal results on the dataset.

Table 5: The evaluation metrics resulting from different classifiers.

	Accuracy		Precision		Recall		F1-Score		
Model	Without CV	CV	Without CV	CV	Without CV	CV	Without CV	CV	AUC
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

The performance of a binary classifier is visually represented by the receiver operating characteristic (ROC) curve. This curve plots the true positive rate (TPR) against the false positive rate (FPR) at various decision thresholds. The area under the ROC curve (AUC) is a single numerical metric that assesses the classifier's sensitivity and specificity across all thresholds, providing an overall measure of its effectiveness. In Figure 6a–d, all models demonstrate excellent performance with AUC values exceeding 0.9. Specifically, the multilayer perceptron (MLP), random forest (RF), and XGBoost models achieve the highest AUC of 0.95 collectively.



(a) **(b)**

(**c**) (**d**)

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

5. Conclusions

The primary aim of this study was to utilize different models to classify heart disease using a real-world dataset. Specifically, the k-modes clustering algorithm was employed on a dataset of patients with heart disease to predict its presence. The dataset underwent preprocessing steps such as converting age into years and binning it into 5-year intervals, as well as dividing systolic and diastolic blood pressure data into 10 intervals. Additionally, the dataset was stratified by gender to account for the unique characteristics and progression of heart disease between men and women.

To determine the optimal number of clusters for both male and female datasets, the elbow curve method was applied. Results indicated that the MLP model achieved the highest accuracy at 87.23%. These findings suggest that k-modes clustering holds promise in accurately predicting heart disease, potentially aiding in the development of targeted diagnostic and treatment strategies.

The study utilized the Kaggle cardiovascular disease dataset comprising 70,000 instances, with all computations performed on Google Colab. Accuracy rates for all algorithms exceeded 86%, with decision trees achieving the lowest at 86.37% and multilayer perceptron achieving the highest accuracy, as previously mentioned.

6. Limitations

Despite the promising outcomes, several limitations should be considered. Firstly, the study relied on a single dataset, which may limit its generalizability to other populations or patient groups. Moreover, the study focused on a restricted set of demographic and clinical variables and did not account for other potential heart disease risk factors such as lifestyle choices or genetic predispositions. Additionally, the model's performance on a separate test dataset, which would provide insights into its generalizability, was not evaluated. Lastly, the study did not assess the interpretability of the clusters formed by the algorithm.

7. Future Research

Future investigations could address these limitations by comparing the performance of the k-modes clustering algorithm with other commonly used clustering techniques like k-means or hierarchical clustering. Evaluating the impact of missing data and outliers on model accuracy and developing robust strategies to handle these issues would also be beneficial. Furthermore, assessing the model's performance on an independent test dataset could establish its reliability with unseen data. Finally, efforts should focus on enhancing the interpretability of cluster results, which would facilitate understanding and decision-making based on study findings. These steps would contribute to advancing the application of clustering algorithms in heart disease prediction and treatment strategies.

Author Contributions: Contributed to conceptualization, data curation, methodology, software, validation, visualization, and writing of the original draft.

Funding: This research did not receive any external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] Estes, C.; Anstee, Q.M.; Arias-Loste, M.T.; Bantel, H.; Bellentani, S.; Caballeria, J.; Colombo, M.; Craxi, A.; Crespo, J.; Day, C.P.; et al. Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030. *J. Hepatol.* **2018**, *69*, 896–904

[2] Droz˙dz˙, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. *Cardiovasc. Diabetol.* **2022**, *21*, 240.

[3] Murthy, H.S.N.; Meenakshi, M. Dimensionality reduction using neuro-genetic approach for early prediction of coronary heart disease. In Proceedings of the International Conference on Circuits, Communication, Control and Computing, Bangalore, India, 21–22 November 2014; pp. 329–332.

[4] Benjamin, E.J.; Muntner, P.; Alonso, A.; Bittencourt, M.S.; Callaway, C.W.; Carson, A.P.; Chamberlain, A.M.; Chang, A.R.; Cheng, S.; Das, S.R.; et al. Heart disease and stroke statistics—2019 update: A report from the American heart association. *Circulation* **2019**, *139*, e56–e528.

[5] Shorewala, V. Early detection of coronary heart disease using ensemble techniques. *Inform. Med. Unlocked* **2021**, *26*, 100655.

[6] Mozaffarian, D.; Benjamin, E.J.; Go, A.S.; Arnett, D.K.; Blaha, M.J.; Cushman, M.; de Ferranti, S.; Després, J.-P.; Fullerton, H.J.; Howard, V.J.; et al. Heart disease and stroke statistics—2015 update: A report from the American Heart Association. *Circulation* **2015**, *131*, e29–e322.

[7] Maiga, J.; Hungilo, G.G.; Pranowo. Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data. In Proceedings of the 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia, 24–25 October 2019; pp. 45–48.

[8] Li, J.; Loerbroks, A.; Bosma, H.; Angerer, P. Work stress and cardiovascular disease: A life course perspective. *J. Occup. Health* **2016**, *58*, 216–219.

[9] Purushottam; Saxena, K.; Sharma, R. Efficient Heart Disease Prediction System. *Procedia Comput. Sci.* **2016**, *85*, 962–969.

[10] Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *Int. J. Comput. Appl.* **2011**, *17*, 43–48.

[11] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* **2019**, *7*, 81542–81554.

[12] Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. *Eur. J. Mol. Clin. Med.* **2020**, *7*, 1638–1645.

[13] Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.

[14] Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.

[15] Gietzelt, M.; Wolf, K.-H.; Marschollek, M.; Haux, R. Performance comparison of accelerometer calibration algorithms based on 3D-ellipsoid fitting methods. *Comput. Methods Programs Biomed.* **2013**, *111*, 62–71.

[16] K, V.; Singaraju, J. Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. *Int. J. Comput. Appl.* **2011**, *19*, 6–12.

[17] Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.

[18] Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN Comput. Sci.* **2020**, *1*, 345.

[19] Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 261–268.

[20] Hasan, N.; Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. *Health Technol.* **2020**, *11*, 49–62.

[21] Ouf, S.; ElSeddawy, A.I.B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. *J. Southwest Jiaotong Univ.* **2021**, *56*, 220–240.

[22] Khan, I.H.; Mondal, M.R.H. Data-Driven Diagnosis of Heart Disease. *Int. J. Comput. Appl.* **2020**, *176*, 46–54.

[23] Kaggle Cardiovascular Disease Dataset. Available online: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease- dataset Han, J.A.; Kamber, M. *Data Mining: Concepts and Techniques*, 3rd ed.; Morgan Kaufmann Publishers: San Francisco, CA, USA, 2011.

[24] Rivero, R.; Garcia, P. A Comparative Study of Discretization Techniques for Naive Bayes Classifiers. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 674–688.

[25] Khan, S.S.; Ning, H.; Wilkins, J.T.; Allen, N.; Carnethon, M.; Berry, J.D.; Sweis, R.N.; Lloyd-Jones, D.M. Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity. *JAMA Cardiol.* **2018**, *3*, 280–287.

[26] Kengne, A.-P.; Czernichow, S.; Huxley, R.; Grobbee, D.; Woodward, M.; Neal, B.; Zoungas, S.; Cooper, M.; Glasziou, P.; Hamet, P.; et al. Blood Pressure Variables and Cardiovascular Risk. *Hypertension*

2009, *54*, 399–404.

[27] Yu, D.; Zhao, Z.; Simmons, D. Interaction between Mean Arterial Pressure and HbA1c in Prediction of Cardiovascular Disease Hospitalisation: A Population-Based Case-Control Study. *J. Diabetes Res.* **2016**, *2016*, 8714745.

[28] Huang, Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. *DMKD* **1997**, *3*, 34–39.

[29] Maas, A.H.; Appelman, Y.E. Gender differences in coronary heart disease. *Neth. Heart J.* **2010**, *18*, 598–602.

[30] Bhunia, P.K.; Debnath, A.; Mondal, P.; D E, M.; Ganguly, K.; Rakshit, P. Heart Disease Prediction using Machine Learning. *Int. J. Eng. Res. Technol.* **2021**, *9*.

[31] Mohanty, M.D.; Mohanty, M.N. Verbal sentiment analysis and detection using recurrent neural network. In *Advanced Data Mining Tools and Methods for Social Computing*; Academic Press: Cambridge, MA, USA, 2022; pp. 85–106.

[32] Menzies, T.; Kocagüneli, E.; Minku, L.; Peters, F.; Turhan, B. Using goals in model-based reasoning. In *Sharing Data and Models in Software Engineering*; Morgan Kaufmann: San Francisco, CA, USA, 2015; pp. 321–353.

[33] Fayez, M.; Kurnaz, S. Novel method for diagnosis diseases using advanced high-performance machine learning system. *Appl. Nanosci.* **2021**.

[34] Hassan, C.A.U.; Iqbal, J.; Irfan, R.; Hussain, S.; Algarni, A.D.; Bukhari, S.S.H.; Alturki, N.; Ullah, S.S. Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. *Sensors* **2022**, *22*, 7227.

[35] Subahi, A.F.; Khalaf, O.I.; Alotaibi, Y.; Natarajan, R.; Mahadev, N.; Ramesh, T. Modified Self-Adaptive Bayesian Algorithm for Smart Heart Disease Prediction in IoT System. *Sustainability* **2022**, *14*, 14208.

[36] *Saini, Mukesh Kumar; Singh, Jaibir. "Big data analytics and machine learning: Personalized, predictive health and boost exactitude medicine research" In Big data analytics and machine learning, 2021 International Journal of Management IT and Engineering, pp. 47-56. International Journal of Management IT and Engineering 2021.*

Author Profile

Dr. Mukesh Kumar Saini holds PhD, MBA, and MCA degrees in Computer Science and Information Technology, and brings over two decades of industry experience. He specializes in Artificial Intelligence, Machine Learning, Data Engineering and their applications in clinical settings. His research focuses on personalized healthcare, with an emphasis on clinical applications designed to optimize healthcare provider costs.