Predicting Heart Disease using Machine Learning

Lê Đăng Khoa, Nguyễn Hưng Thịnh

Scholar, FPT University, Thu Duc City, Ho Chi Minh – Viet Nam

Email: khoaldse150847@fpt.edu.vn – thinhnhse151079@fpt.edu.vn

Abstract: Artificial Intelligence, particularly Machine Learning (ML), has made significant strides in various research domains. This paper explores the application of machine learning in detecting heart diseases, a prevalent and life-threatening condition affecting numerous individuals globally. By leveraging attributes such as chest pain, cholesterol levels, and age, machine learning algorithms, specifically supervised learning methods, play a pivotal role in diagnosing cardiovascular diseases. This study employs three prominent algorithms—K-Nearest Neighbor (KNN), Random Forest, and Logistic Regression—to classify individuals with heart diseases and those without. The results indicate a prediction accuracy of 73.82% for KNN, 79.36% for Random Forest, and 78.53% for Logistic Regression, showcasing the effectiveness of these algorithms in disease diagnosis.

1. Keywords

Heart Disease; Machine Learning; K Nearest Neighbor (K-NN); Random Forest; Logistic Regression, Gird Search.

2. Introduction

The human body consists of various organs, each serving specific functions. Among these, the heart plays a crucial role by pumping blood throughout the body; its failure can lead to fatal consequences, making heart disease a leading cause of mortality today [1]. Ensuring the health of the cardiovascular system, and other bodily systems, is essential. Sadly, cardiovascular diseases affect people worldwide, emphasizing the need for early diagnosis to save lives and resources. Utilizing data mining techniques is instrumental in predicting heart diseases by identifying unknown patterns and trends within vast databases [2]. Data mining involves extracting knowledge from extensive data sets [3]. Machine learning, an emerging scientific and technological field, facilitates early heart disease diagnosis by classifying whether an individual might be suffering from this condition, preventing significant harm.

Narain and colleagues (2016) conducted a study aiming to enhance the precision of the commonly used Framingham risk score (FRS) for cardiovascular disease (CVD) prediction. Their innovative approach utilized a quantum neural network and data from 689 symptomatic CVD individuals, along with a validation dataset from Framingham research [5]. The proposed system achieved a remarkable 98.57% accuracy in CVD risk prediction, surpassing the FRS's 19.22% accuracy and other existing methods. This suggests its potential as a valuable tool for doctors, aiding in accurate CVD risk forecasting, improving treatment plans, and enabling early diagnosis [6].

Shah and his team (2020) focused on developing a predictive model for cardiovascular disease using machine learning techniques. They utilized the Cleveland heart disease dataset with 303 instances and 17 attributes. Employing various supervised classification methods such as naive Bayes, decision tree, random forest, and k-nearest neighbor (KKN), the study revealed that the KKN model exhibited the highest accuracy at 90.8% [7]. This study underscored the importance of selecting appropriate models and techniques, showcasing the potential of machine learning in accurate cardiovascular disease prediction.

Drod et al. (2022) aimed to identify significant risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease (MAFLD) using machine learning techniques. Analyzing data from 191 MAFLD patients, they employed multiple ML approaches like multiple logistic regression classifier, univariate feature ranking, and principal component analysis (PCA) [8]. The study identified hypercholesterolemia, plaque scores, and diabetes duration as crucial factors. Their ML model accurately identified high-risk (85.11%) and low-risk (79.17%) patients, indicating the effectiveness of ML in detecting CVD in MAFLD patients based on simple criteria.

Meanwhile, Hasan and Bao (2020) conducted a comparative study with the primary objective of identifying the most efficient feature selection approach for anticipating cardiovascular illness [9]. They considered three well-known feature selection methods (filter, wrapper, and embedding) and retrieved feature subsets using a Boolean process-based common "True" condition. Several models, including random forest, support vector classifier, k-nearest neighbors, naive Bayes, and XGBoost, were evaluated to determine the best predictive analytics. XGBoost coupled with the wrapper technique emerged as the most accurate predictor for cardiovascular illness, achieving an accuracy of 73.74% [10]. This was followed by SVC with 73.18% accuracy and ANN with 73.20%. These findings highlight the effectiveness of the XGBoost classifier in combination with the wrapper technique for accurate cardiovascular illness prediction.

The main limitation of previous research lies in its small dataset, which increases the risk of overfitting and renders the models unsuitable for larger datasets. In contrast, our study utilized a dataset on cardiovascular disease comprising over 30,000 patients and 51 features, but imbalanced. By meticulously cleaning and processing the data, we created a new dataset. This involved identifying key features, removing unnecessary ones, resulting in a comprehensive dataset featuring more than 3000 patients and 8 features. This approach significantly reduces the likelihood of overfitting. Table 1 provides a summary of cardiovascular disease prediction studies conducted on large datasets, underscoring the efficacy of employing substantial data for research.

Table 1

Methods	Best Accuracy	Datasets
KNN	73.82%	3000 patients and 8 features
Random Forest	79.36%	3000 patients and 8 features
Logistic Regression	78.53%	3000 patients and 8 features

3. Literature Review

Research efforts in the field of cardiovascular disease prediction have led to the development of various methods using supervised machine learning algorithms. Numerous research papers explore this area, including a survey paper that evaluates different machine learning models and techniques

[4]. Some studies have focused on innovative approaches, such as creating a Graphical User Interface (GUI) using Weighted Association rule based Classifier to determine heart disease presence [5]. Another novel method introduced involves the coactive neuro-fuzzy interference system (CANFIS) for heart disease prediction [6]. Additionally, research papers have summarized commonly used techniques for heart disease prediction and their complexities [7]. Classifier approaches have been explored, demonstrating the application of Naive Bayes for heart disease detection in specific studies [8]. Furthermore, comprehensive surveys have been conducted, reviewing various papers that utilize data mining algorithms for heart disease prediction [9].

4. Proposed Methods

4.1. K-Nearest Neighbor (K-NN)

K-NN is a classification algorithm where the classification of an unknown data point is based on its k nearest neighbors. The parameter k is defined beforehand. The algorithm selects k neighbors with the lowest Euclidean distance from the unknown data point. The unknown data point is then classified into the category that the majority of its neighbors belong to among the selected k neighbors [11].

4.2. Random Forest

Random Forest operates by constructing multiple decision trees from the training data. Each tree predicts a class, and the final result is determined by the class that the majority of decision trees predict [10]. To use this algorithm, the number of trees to be created must be defined in advance. Random Forest employs bootstrap aggregating (bagging) to decrease result variance.

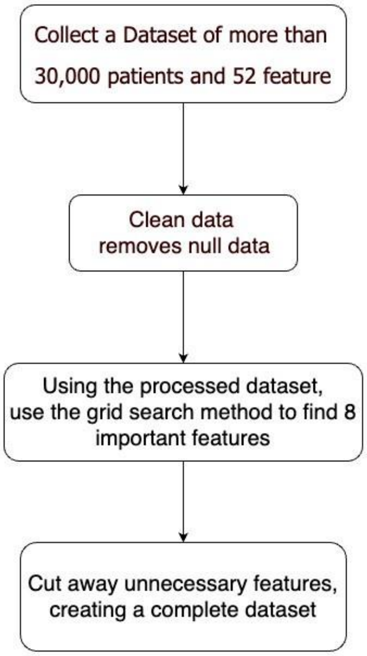
4.3. Logistic Regression

Logistic Regression is a widely utilized classification algorithm in machine learning. Unlike linear regression, which predicts continuous outcomes, logistic regression handles binary classification problems where outcomes fall into two classes [16]. In the context of heart disease prediction, logistic regression determines whether a person is at risk of developing cardiovascular issues [19]. The algorithm calculates probabilities using the sigmoid function and classifies data points based on a predefined threshold (usually 0.5). It is particularly suitable for binary classification tasks.

5. Datasets – Data Mining

5.1 Data cleaning

During our research, we gathered a large dataset comprising over 30,000 patients and 53 related features; however, this dataset was imbalanced. To ensure data reliability, we initiated a data cleaning process. Firstly, we eliminated invalid or missing (null) data points [18]. Subsequently, to fill in the missing values, we used the average values of the corresponding features. Next, we removed data points contributing to the imbalance, making the dataset more balanced. Finally, the processed data was saved for further analysis [20].


Collect a Dataset of more than
30,000 patients and 52 feature
Clean data
removes null data
Using the processed dataset,
use the grid search method to find 8
important features
Cut away unnecessary features,
creating a complete dataset

5.2 Feature importance analysis

To optimize our model, we employed the grid search method on the processed dataset to identify the 8 most important features. Once these features were identified, we proceeded to remove other unnecessary features, creating a complete and optimized dataset for use in predictive and classification models related to heart diseases [18]. This process ensured that we utilized a high-quality and optimized dataset, guaranteeing accurate and reliable results in heart disease prediction and prevention. In summary, the dataset comprises more than 3000 patients and 8 features.

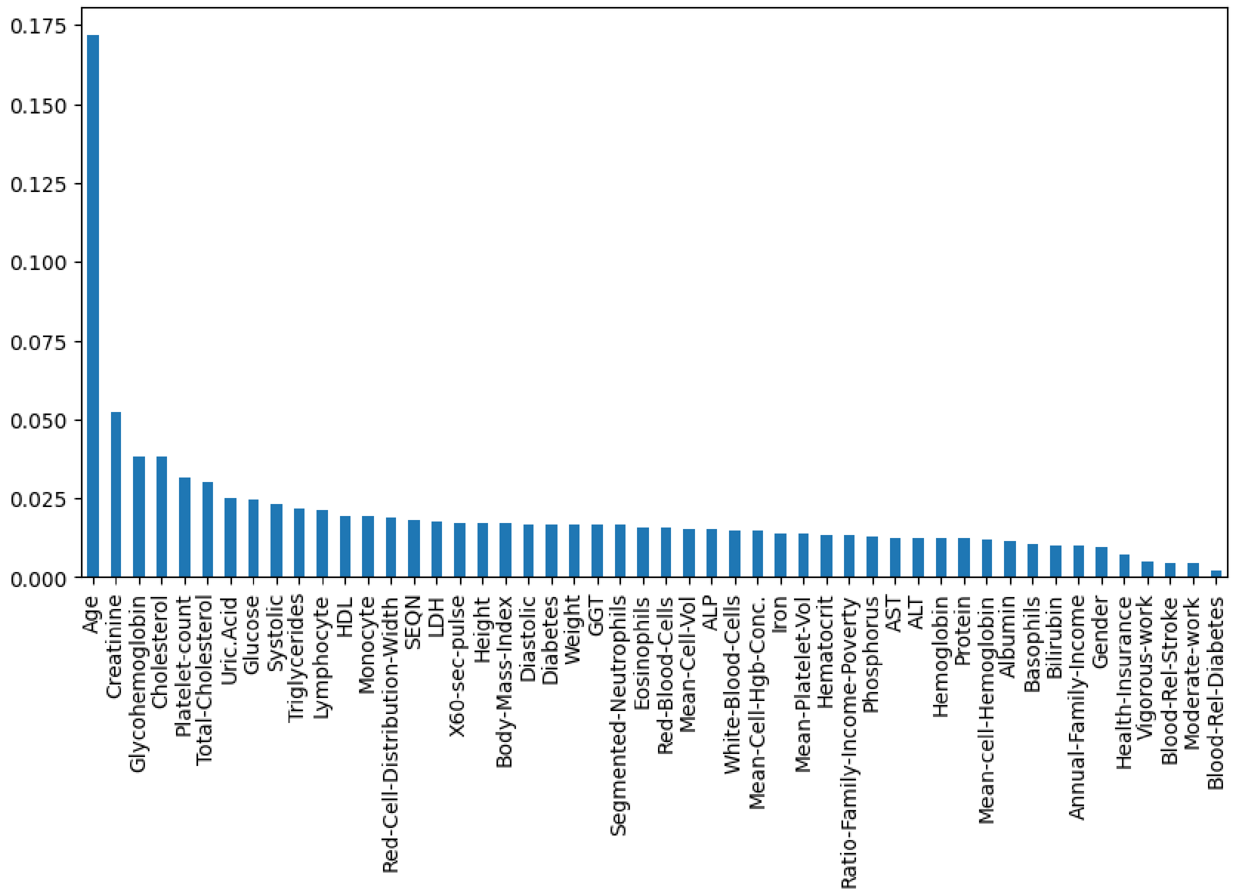


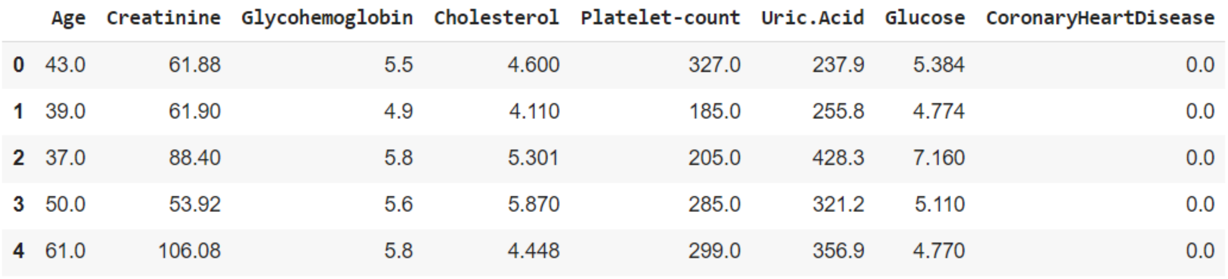
Figure 1 Feature importance analysis

Findings:

*Important features are: *Age, Creatinine, Glycohemoglobin, Cholesterol, Platelet-count, Uric.Acid, Glucose, CoronaryHeartDisease**

5.3 Filter Data

In the process of filtering the data, we focused on key attributes including Age, Creatinine levels, Glycohemoglobin, Cholesterol, Platelet count, Uric Acid, Glucose, and the presence of Coronary heart disease. These specific parameters were carefully selected due to their significant relevance in cardiovascular health. Age plays a crucial role as heart diseases often correlate with aging. Creatinine levels and Glycohemoglobin provide insights into kidney and diabetes-related risks, which are often linked to heart conditions [12]. Cholesterol levels, especially LDL (low-density lipoprotein), are well-known risk factors for heart disease. Platelet count can indicate potential blood clotting issues, while Uric Acid levels are associated with hypertension and heart disease. Monitoring Glucose levels is essential due to the direct connection between diabetes and cardiovascular problems . Lastly, the presence of Coronary heart disease is a direct indicator of existing cardiac issues. By filtering the data based on these parameters, we were able to identify patterns and correlations that are invaluable in understanding the intricate relationship between these factors and heart diseases [14]. This meticulous filtration process ensured that our analysis was focused on the most relevant aspects, enhancing the precision and effectiveness of our predictive models.

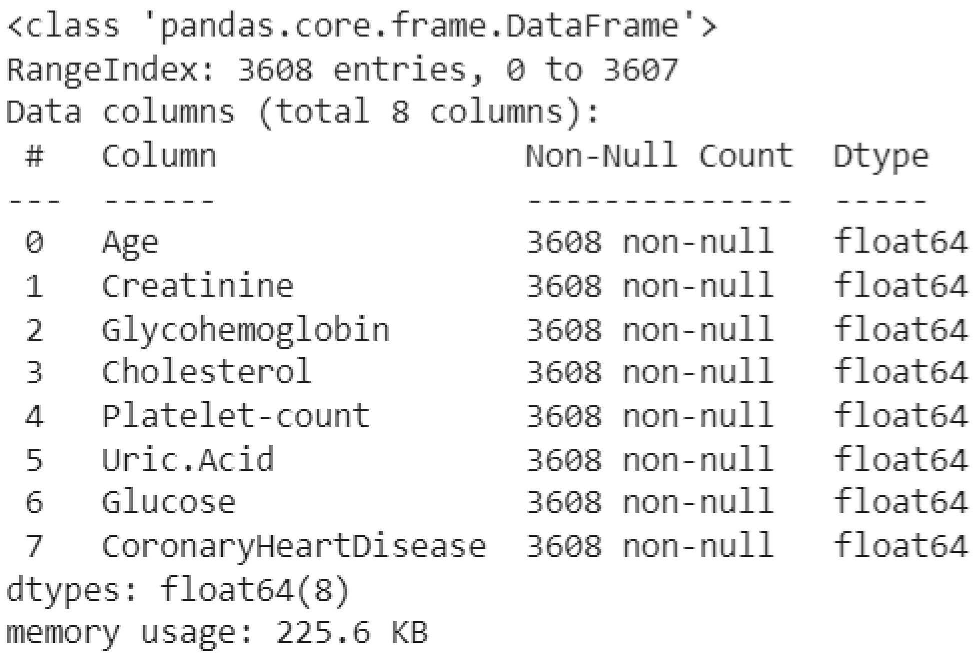


	Age	Creatinine	Glycohemoglobin	Cholesterol	Platelet-count	Uric.Acid	Glucose	CoronaryHeartDisease
0	43.0	61.88	5.5	4.600	327.0	237.9	5.384	0.0
1	39.0	61.90	4.9	4.110	185.0	255.8	4.774	0.0
2	37.0	88.40	5.8	5.301	205.0	428.3	7.160	0.0
3	50.0	53.92	5.6	5.870	285.0	321.2	5.110	0.0
4	61.0	106.08	5.8	4.448	299.0	356.9	4.770	0.0

6. Experimental Setup

The initial setup involves acquiring a dataset containing features of individuals with and without heart disease, along with the corresponding disease status (positive or negative). Python is chosen as the programming language for conducting the experiment. The dataset encompasses eight importants features.

Subsequently, the data analysis process begins. To gain insights into the dataset, it's crucial to understand its structure and content. This is achieved by utilizing the info() function provided by the Pandas library. Employing this function provides a concise summary of the DataFrame, offering essential information about the dataset's columns, data types, and the presence of missing values [13]. This step is fundamental in comprehending the dataset's characteristics before proceeding with further analysis or modeling.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3608 entries, 0 to 3607
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Age                   3608 non-null   float64
 1   Creatinine            3608 non-null   float64
 2   Glycohemoglobin       3608 non-null   float64
 3   Cholesterol           3608 non-null   float64
 4   Platelet-count        3608 non-null   float64
 5   Uric.Acid             3608 non-null   float64
 6   Glucose               3608 non-null   float64
 7   CoronaryHeartDisease  3608 non-null   float64
dtypes: float64(8)
memory usage: 225.6 KB
```

6.1 Data exploration

The *describe()* function from the Pandas library is employed to obtain statistical insights about the dataset, including the mean values of the attributes in use. Specifically, an attribute called " CoronaryHeartDisease " is selected, with a value of 1 indicating the presence of heart disease in the patient and 0 denoting the absence of heart disease. To assess the balance of the dataset, the distribution of the "target" attribute is examined. This evaluation is carried out using a countplot, a visualization tool provided by the Seaborn library [17]. The countplot provides a graphical representation of the distribution of the target values, indicating whether the dataset is balanced or skewed towards a particular class. This step is crucial in understanding the class distribution, which is essential for ensuring unbiased and accurate model predictions [11].

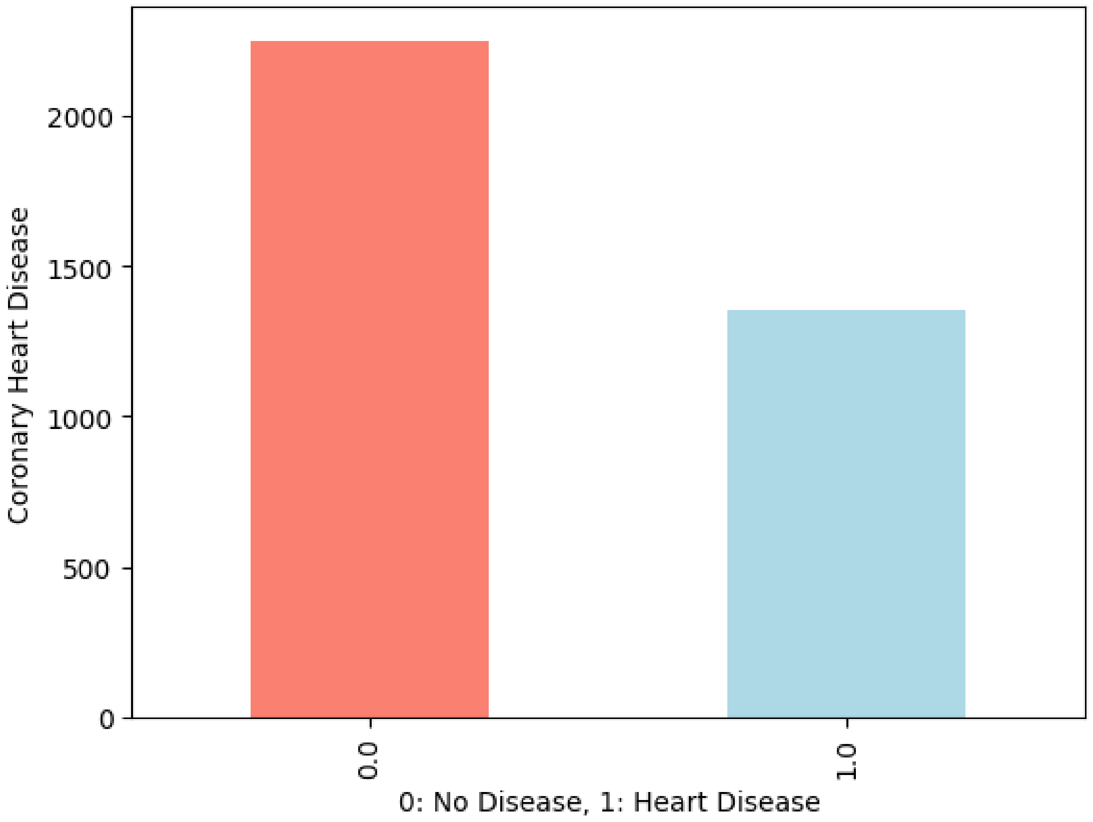


Figure 2 Coronary Heart Disease

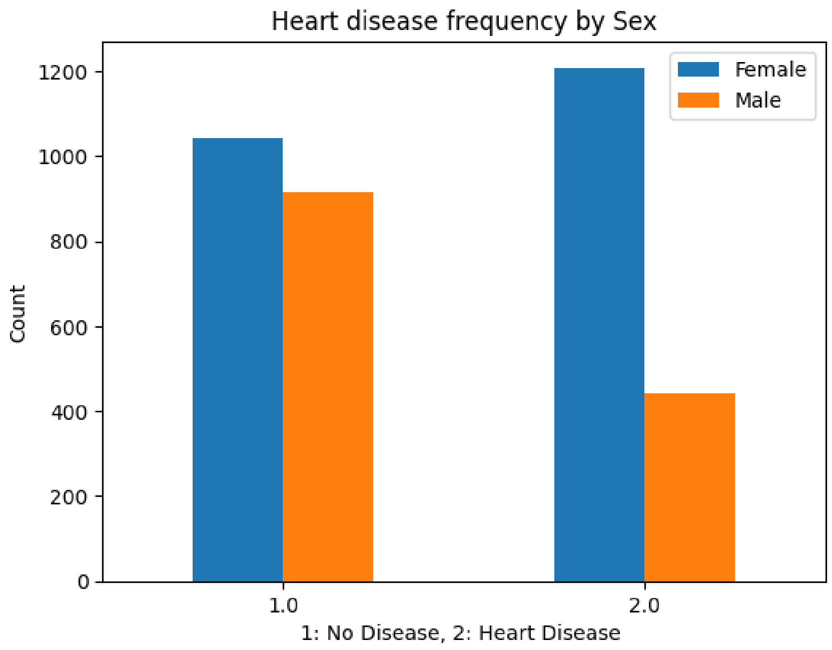


Figure 3 Heart disease frequency by Sex

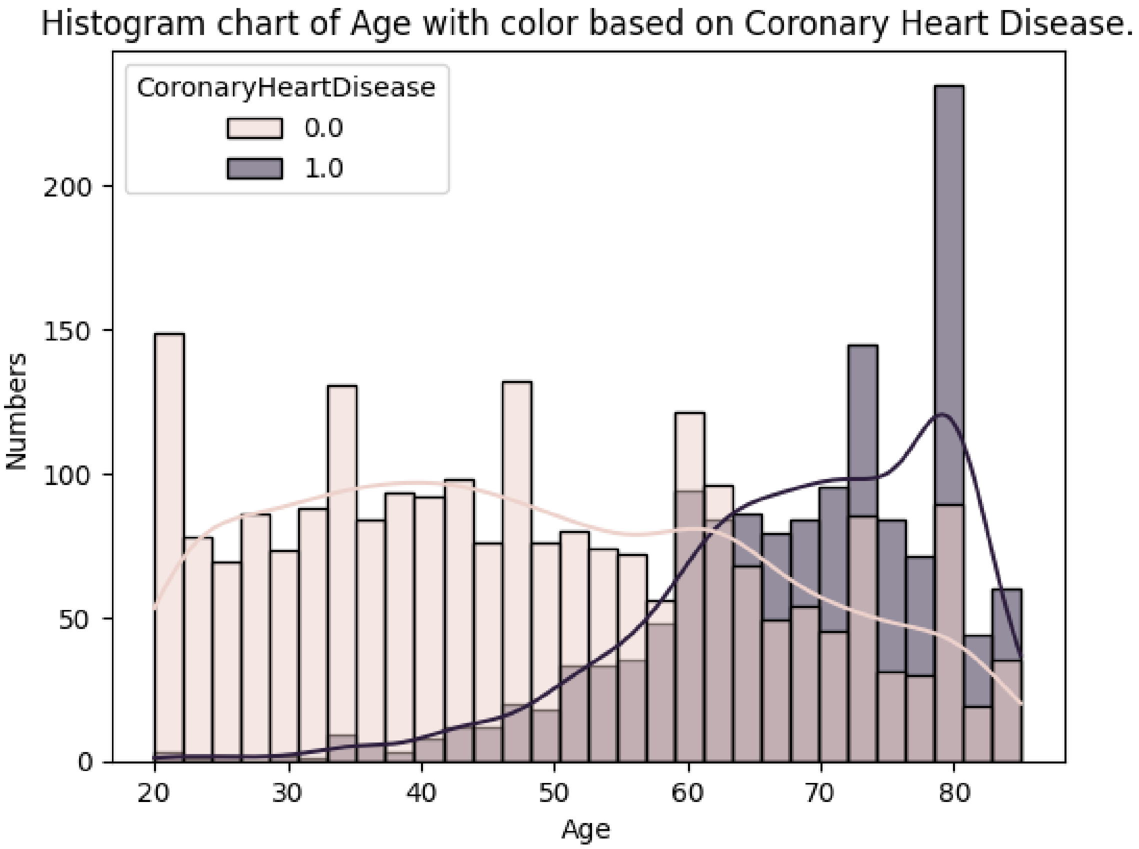


Figure 4 Histogram chart of Age with color based on Coronary Heart Disease

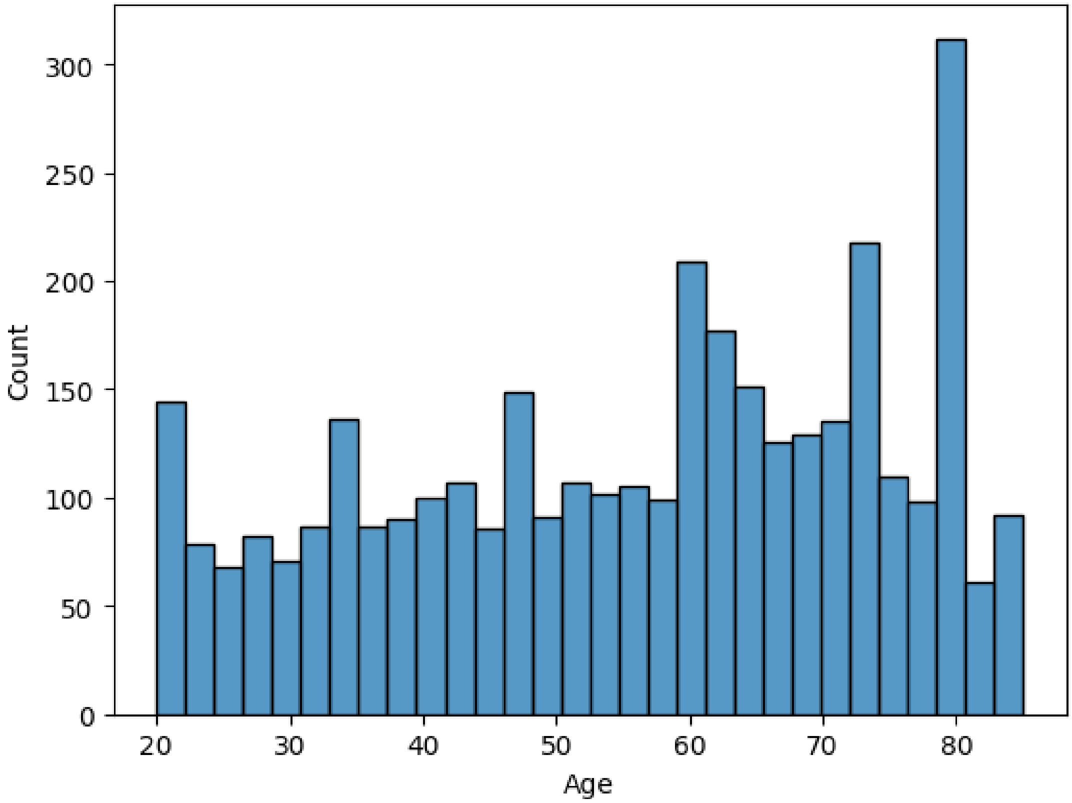


Figure 5 Age

6.2 Clustering

Clustering, a machine learning technique, groups instances based on their similarities. While the k-means algorithm is commonly used, it proves ineffective for categorical data. To address this, the k-modes algorithm was introduced by Huang in 1997. Unlike k-means, k-modes employs dissimilarity measures for categorical data and replaces means with modes for clusters, allowing it to handle categorical data effectively [21].

Given our dataset's categorical nature, we opt for k-modes analysis. To determine the ideal cluster number, we use the elbow curve with Huang initialization. This method involves creating k-modes models with varying cluster numbers, calculating the costs, and plotting them [22]. The "elbow method" identifies the optimal cluster number where adding more clusters no longer significantly improves the model fit [26].

Segmenting the dataset by gender offers advantages in prediction due to biological disparities between men and women affecting disease manifestation and progression. Men tend to develop heart disease earlier, and their symptoms and risk factors may differ from women. Studies show men have higher coronary artery disease (CAD) risks compared to women, and CAD risk factors vary between genders [15]. Analyzing data separately reveals unique risk factors and disease patterns not evident when consolidating data. Moreover, heart disease prevalence rates vary significantly between men and women.

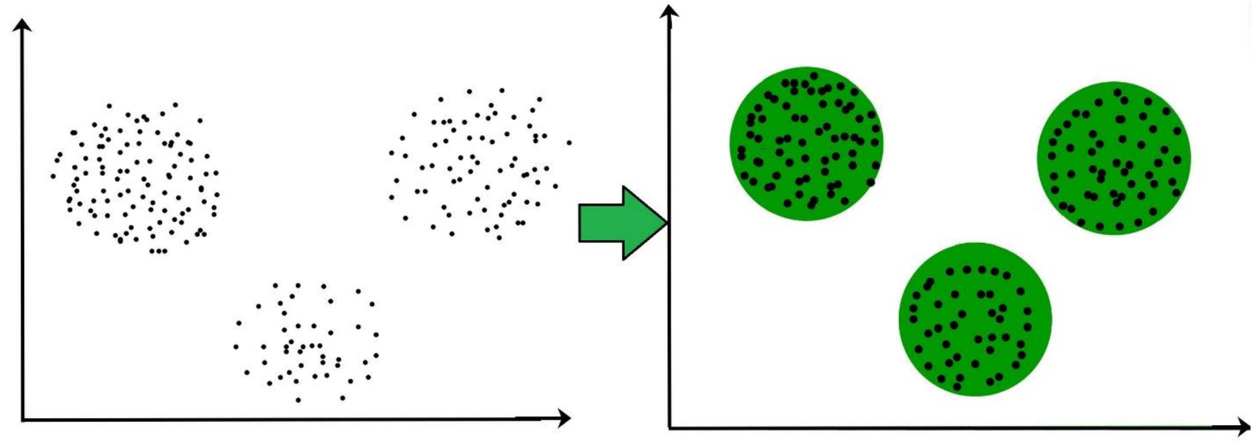


Figure 6 Clustering

6.3 Correlation Table

Additionally, a correlation table is generated to assess the relationships between various categories. Figure 5 illustrates significant correlations between mean arterial pressure and factors such as 'Age', 'Creatinine', and 'Glycohemoglobin'. These highly correlated factors highlight interdependencies among features [23]. This correlation matrix provides valuable insights into intra-feature dependencies, aiding in a comprehensive understanding of the dataset's intricate relationships.

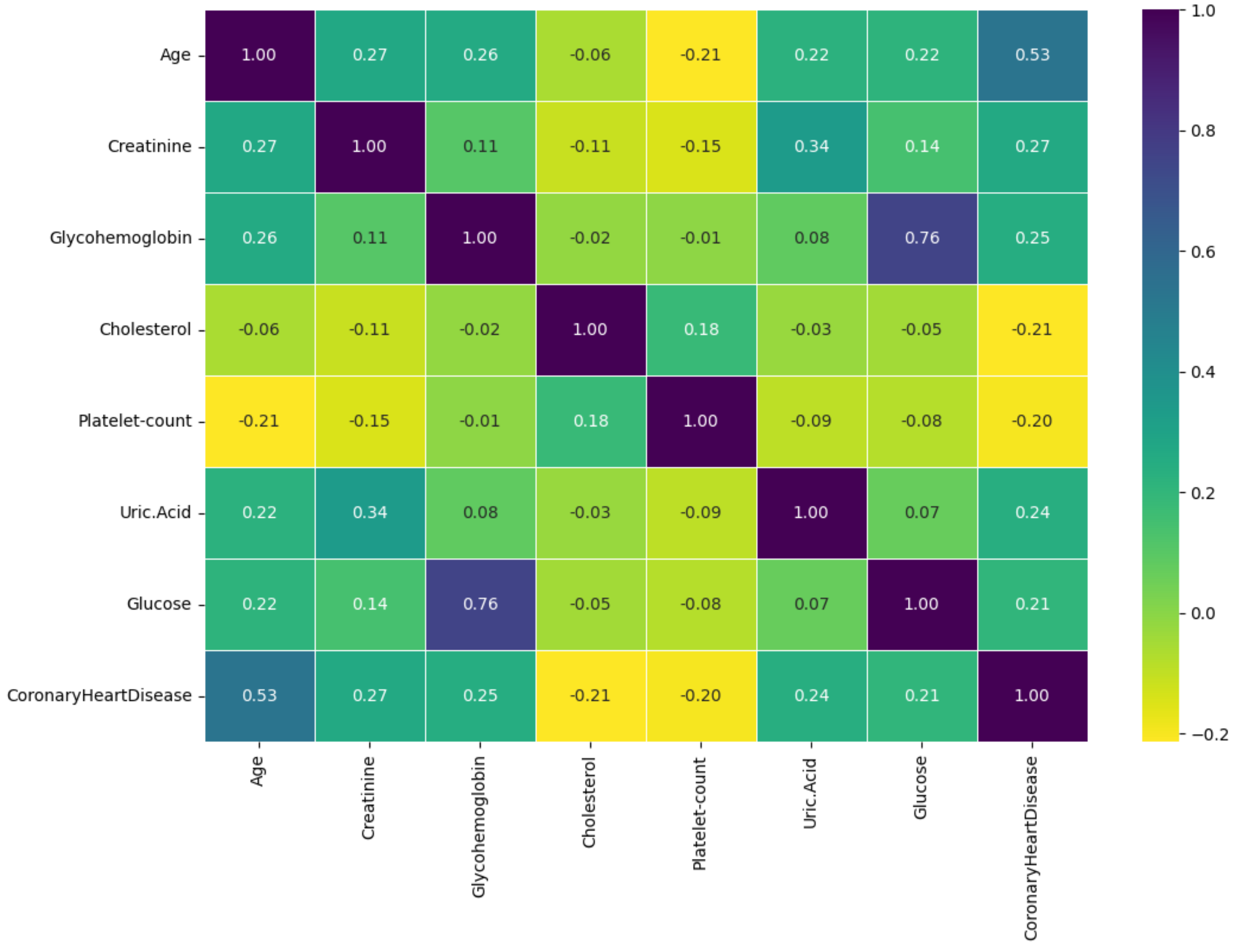


Figure 7 Correlation heatmap

7. Modelling

7.1 Result

After evaluating the correlation using a heatmap, it became evident that attributes like 'cp' (chest pain) and 'thalack' (maximum heart rate achieved) exhibit positive correlations with the target attribute. Following this correlation analysis, the next step involves converting categorical variables such as 'Age', 'Creatinine', 'Glycohemoglobin', 'Cholesterol', 'Platelet-count', 'Uric.Acid', 'Glucose', and 'CoronaryHeartDisease' into binary variables. This transformation is achieved using the get_dummies method from the Pandas library [24].

Subsequently, dummy variables having been created, standardization is applied to columns like 'Age', 'Creatinine', 'Glycohemoglobin', 'Cholesterol' to account for their varied quantities and units. This standardization process is executed utilizing the Scikit-learn library in Python [25].

The dataset is divided into two segments: the training data, constituting 80% of the entire dataset, and the testing data, which encompasses the remaining 20%. Once the data is appropriately prepared, machine learning algorithms are applied, and the confusion matrix is generated. The algorithm's performance is then assessed in terms of accuracy, providing valuable insights into the model's predictive capabilities [27].

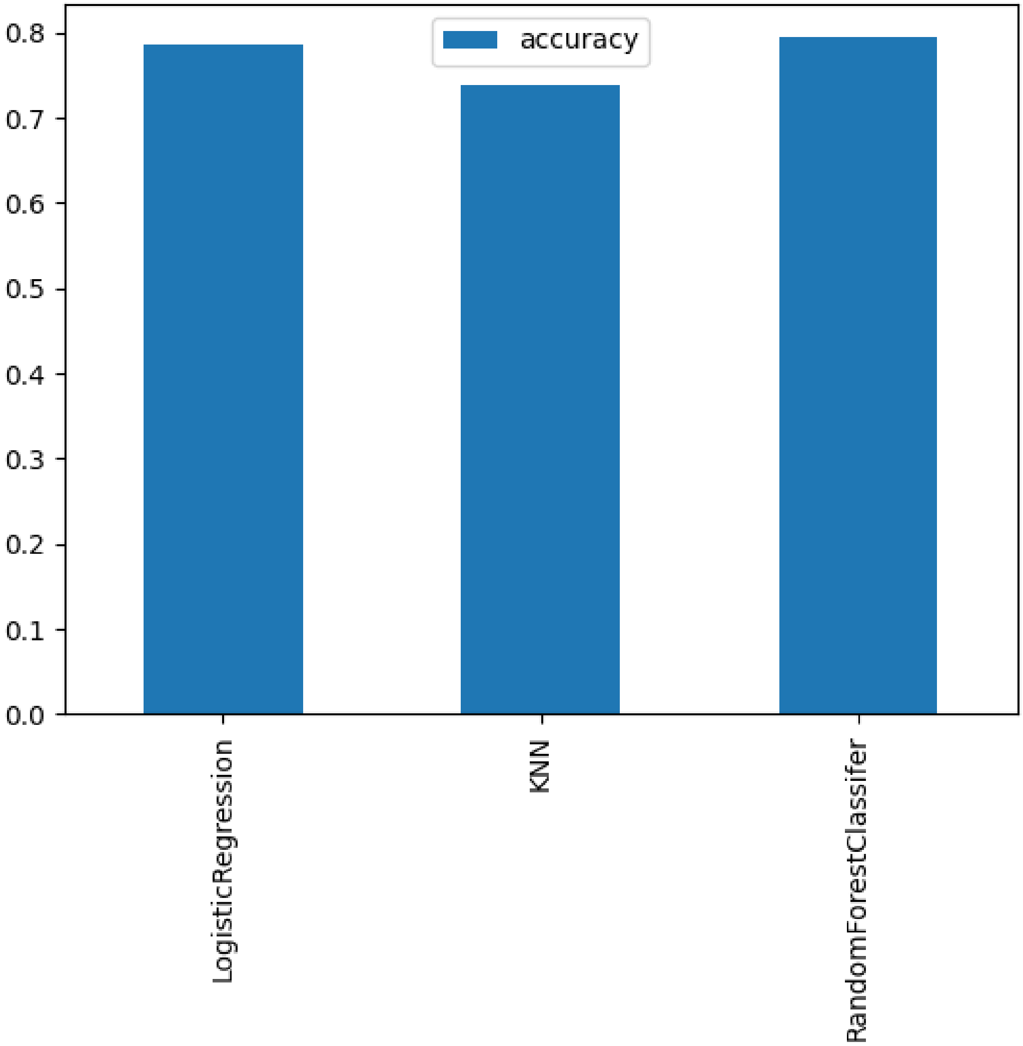


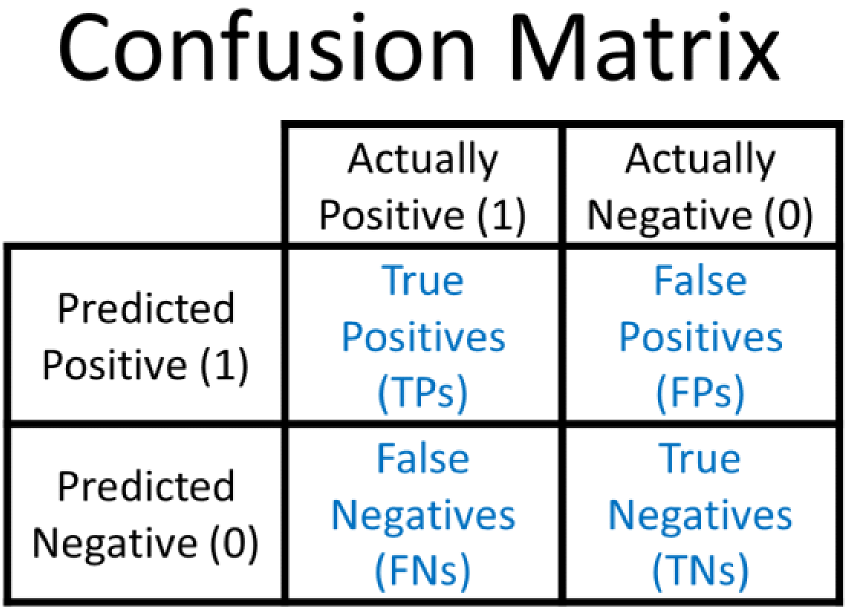
Figure 8 Result Modelling

7.2 Tuning or Improving our models

We utilized Logistic Regression to fine-tune our model and achieved an accuracy of 78.67%. This result indicates that our model can accurately predict 78.67% of the cases tested. It provides a significant insight into the model's predictive capability in determining the risk of heart disease based on the considered features.

7.3 Evaluating Models

The accuracy has been found out with the use of a confusion matrix.



Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 9 Format confusion matrix

[[TP FP
FN TN]]
The accuracy of the algorithm can be calculated using the formula:
Accuracy = {(TP + TN) / TP + FP + TN + FN)} * 100

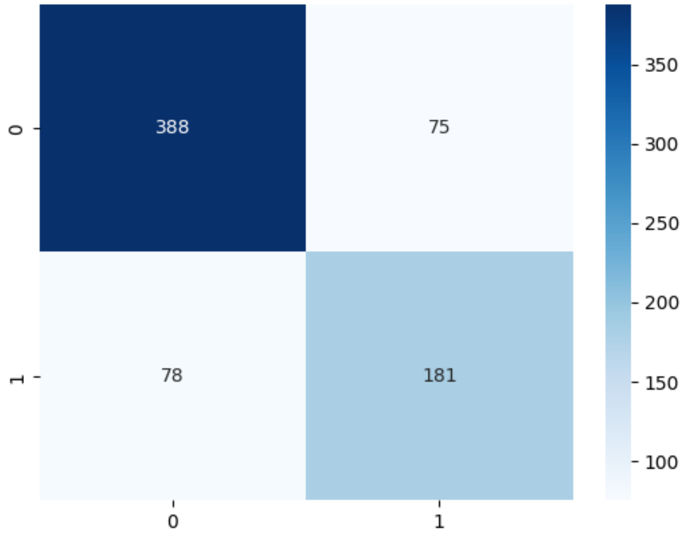


Figure 10 Confusion matrix

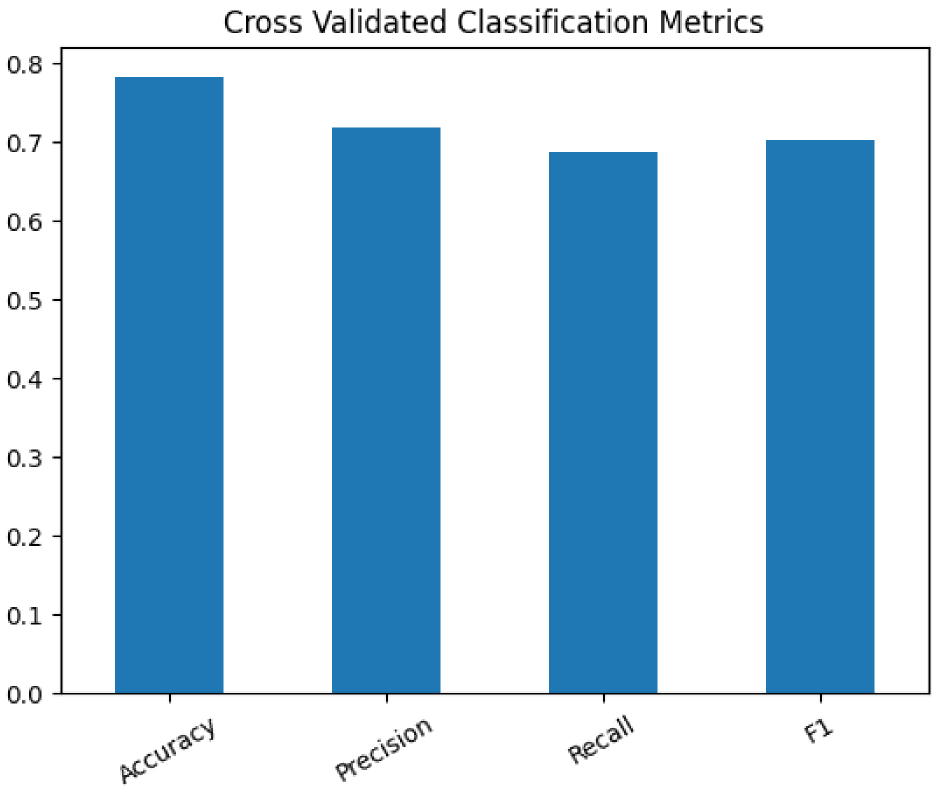


Figure 11 Cross Validated Classification Metrics

7. Conclusions

In summary, the study demonstrates the significant value of machine learning in predicting heart disease, a major societal concern. The algorithms employed in the research performed admirably using the available data, suggesting the potential for machine learning to mitigate both physical and mental damage caused by heart disease [28]. However, it's important to acknowledge certain limitations. The study was confined to a single dataset and specific demographic and clinical variables, limiting its applicability to broader populations and alternative risk factors such as lifestyle choices and genetics. Furthermore, the model's performance on new data was not assessed, and the interpretability of the results was not explored [29]. To address these limitations, future research should consider comparing the k-modes clustering algorithm with other widely used methods, assess the impact of missing data and outliers, and evaluate the model's performance on unseen data [30]. Establishing the robustness of the results and the interpretability of the clusters formed by the algorithm will be crucial for informed decision-making based on the study's findings.

8. References

1. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.
2. Bhatla, N., & Jyoti, K. (2012). An analysis of heart disease prediction using different data mining techniques. International Journal of Engineering, 1(8), 1-4.
3. Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), 129-137.
4. Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687.
5. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Intelligent and effective heart disease prediction system using weighted associative classifiers. International Journal on Computer Science and Engineering, 3(6), 2385-2392.
6. Parthiban, L., & Subramanian, R. (2008). Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3).
7. Chitra, R., & Seenivasagam, V. (2013). Review of heart disease prediction system using data mining and hybrid intelligent techniques. ICTACT Journal on Soft Computing, 3(04), 605-609.
8. Medhekar, D. S., Bote, M. P., & Deshmukh, S. D. (2013). Heart disease prediction system using naive Bayes. International Journal of Enhanced Research in Science, Technology & Engineering, 2(3).
9. Kaur, B., & Singh, W. (2014). Review on heart disease prediction system using data mining techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 2(10), 3003-3008.
10. Li, J., Loerbroks, A., Bosma, H., & Angerer, P. (2016). Work stress and cardiovascular disease: A life course perspective. Journal of Occupational Health, 58(3), 216–219.
11. Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. Procedia Computer Science, 85, 962–969.
12. Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications, 17(8), 43–48.
13. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, 7, 81542–81554.
14. Waigi, R., Choudhary, S., Fulzele, P., & Mishra, G. (2020). Predicting the risk of heart disease using advanced machine learning approach. Eur. J. Mol. Clin. Med., 7, 1638–1645.
15. Breiman, L. (2001). Random forests. Mach. Learn., 45, 5–32.
16. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD ’16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
17. Gietzelt, M., Wolf, K.-H., Marschollek, M., & Haux, R. (2013). Performance comparison of accelerometer calibration algorithms based on 3D-ellipsoid fitting methods. Comput. Methods Programs Biomed., 111, 62–71.
18. K, V., & Singaraju, J. (2011). Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks. Int. J. Comput. Appl., 19, 6–12.
19. Narin, A., Isler, Y., & Ozer, M. (2016). Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO).

20. Shah, D., Patel, S., & Bharti, S. K. (2020). Heart Disease Prediction using Machine Learning Techniques. SN Comput. Sci., 1, 345.
21. Alotaibi, F. S. (2019). Implementation of Machine Learning Model to Predict Heart Failure Disease. Int. J. Adv. Comput. Sci. Appl., 10, 261–268.
22. Hasan, N., & Bao, Y. (2020). Comparing different feature selection algorithms for cardiovascular disease prediction. Health Technol., 11, 49–62.
23. Ouf, S., & ElSeddawy, A. I. B. (2021). A proposed paradigm for intelligent heart disease prediction system using data mining techniques. J. Southwest Jiaotong Univ., 56, 220–240.
24. Khan, I. H., & Mondal, M. R. H. (2020). Data-Driven Diagnosis of Heart Disease. Int. J. Comput. Appl., 176, 46–54.
25. Han, J. A., & Kamber, M. (2011). Data Mining: Concepts and Techniques, 3rd ed.; Morgan Kaufmann Publishers.
26. Yu, D., Zhao, Z., & Simmons, D. (2016). Interaction between Mean Arterial Pressure and HbA1c in Prediction of Cardiovascular Disease Hospitalisation: A Population-Based Case-Control Study. J. Diabetes Res., 2016, 8714745. [PubMed]
27. Huang, Z. (1997). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. DMKD, 3, 34–39.
28. Maas, A. H., & Appelman, Y. E. (2010). Gender differences in coronary heart disease. Neth. Heart J., 18, 598–602.
29. Mohanty, M. D., & Mohanty, M. N. (2022). Verbal sentiment analysis and detection using recurrent neural network. In Advanced Data Mining Tools and Methods for Social Computing (pp. 85–106). Academic Press.
30. Menzies, T., Kocagüneli, E., Minku, L., Peters, F., & Turhan, B. (2015). Using Goals in Model-Based Reasoning. In Sharing Data and Models in Software Engineering (pp. 321–353). Morgan Kaufmann.