Heart Health Insights: Analysing Data and Visualising Predictive Factors for Cardiovascular Disease

Udayini Vedantham , Nithish Chouti , Karan Naik and Sunil C K *

Department of Computer Science and Engineering, Indian Institute of Information Technology Dharwad, 580009, India

ARTICLE INFO

Keywords:

Cardiovascular Disease (CVDs)

Machine Learning

Machine Learning Algorithms

Data Analytics

Visualizations

ABSTRACT

Cardiovascular diseases (CVDs) are a leading cause of mortality worldwide, posing a significant public health challenge. This study aims to contribute to the existing research on CVD prediction by exploring the application of data analysis and visualization techniques. The researchers employed a range of machine learning algorithms, such as decision trees, support vector machines, random forest, and logistic regression, to analyze a comprehensive dataset of 1,190 observations with 11 independent variables and the presence or absence of heart disease. The study focused on identifying the most significant factors contributing to the risk of heart disease using Principal Component Analysis (PCA) and evaluating the accuracy of different machine learning models in predicting the likelihood of CVDs. The findings offer important insights into how data analytics and visualization can be applied to detect and prevent cardiovascular diseases early on.

1. Introduction

Cardiovascular diseases (CVDs) represent a pose a significant global health dilemma, accounting for a substantial portion of morbidity and mortality worldwide [1]. According to the World Health Organization (WHO), CVDs are responsible for 17.7 million fatalities annually, making up approximately 31% of all deaths globally [1]. In India, heart-related illnesses have emerged as the major cause of death, with 1.7 million lives lost to cardiac ailments in 2016 alone [1]. The economic impact of cardiovascular diseases in India has been staggering, with estimates suggesting costs of up to $237 billion between 2005 and 2015 [2]. Beyond the mortality rates, CVDs also impose a significant burden on personal productivity and healthcare expenditure.

Machine learning algorithms have become a potent tool in combating cardiovascular diseases, allowing for analyzing sizeable clinical data sets to identify hidden patterns and correlations that traditional methods might miss [2]. By harnessing the potential of machine learning, researchers can identify the critical risk factors associated with heart disease, develop targeted interventions, and implement preventive strategies to enhance public health outcomes and alleviate the burden of CVDs.

This research paper aims to add to the current knowledge on predicting cardiovascular diseases by examining how data analysis and visualization techniques can be applied. The study will utilize various machine learning algorithms—such as decision trees, support vector machines, random forest, and logistic regression—to analyze a comprehensive dataset of 1,190 observations, each with 11 independent variables related to heart disease. The goal is to pinpoint the key factors contributing to heart disease risk using Principal Component Analysis (PCA) [3] and to evaluate the accuracy of different machine learning models in predicting cardiovascular diseases [4].

The prime contributions of this work are:

- Employing Principal Component Analysis (PCA) [3] to identify pivotal risk factors for heart disease offers valuable insights into leveraging data analytics and visualization for the early detection and prevention of cardiovascular diseases.

*Corresponding author

21bcs130@iiitdwd.ac.in (U.V.); 21bcs074@iiitdwd.ac.in (N.C.); 21bcs051@iiitdwd.ac.in (K.N.); sunilck@iiitdwd.ac.in (S.C.K.)

- This paper employs a variety of machine learning algorithms such as decision trees, support vector machines, random forest, and logistic regression to analyze a detailed dataset of 1,190 observations, each containing 11 independent variables along with the presence or absence of heart disease.
- Assessing the performance of diverse machine learning models in predicting the probability of cardiovascular diseases can guide the development of targeted interventions and preventive measures to improve public health outcomes and reduce the burden of CVDs.

2. Literature Review

Cardiovascular diseases (CVDs) are a significant cause of morbidity and mortality worldwide, with heart disease being the primary cause of death globally [5]. Cardiovascular diseases, particularly heart disease, stand out as a leading cause of mortality globally, emphasizing the pressing necessity for accurate prediction and early detection strategies [3]. Machine learning has emerged as a powerful tool in this domain, showcasing notable achievements in predicting heart disease.

Addressing the challenge of imbalanced data, Ishaq et al. [6] adopted a pragmatic approach by utilizing random forest feature importance scores to rank and select pertinent features. To mitigate class imbalance, they incorporated the Synthetic Minority Over-sampling Technique (SMOTE) [5]. Furthermore, techniques such as the mean value approach for handling missing data and feature importance methods for selection have been crucial in improving prediction accuracy [7].

Feature selection and extraction play pivotal roles in ML-based heart disease prediction. Principal Component Analysis (PCA) is a widely used linear transformation technique for dimensionality reduction, optimizing variance and identifying orthogonal feature space directions [3]. Studies have demonstrated the effectiveness of PCA in feature extraction, enhancing prediction accuracy [3] [4]. Moreover, feature fusion techniques have been instrumental in refining datasets derived from medical records and sensor data, further improving prediction accuracy [8].

Various machine learning algorithms, including Decision Trees (DT), Random Forest (RF) [5], Support Vector Machines (SVM) [9], Neural Networks (NN), and K-Nearest Neighbors (KNN) [9], have been employed for heart disease prediction [10] [4]. Hybrid models combining DT and RF have shown superior performance, with a significantly improving accuracy [10]. In a paper, Abu Sufian et. al [5] performed a comparative analysis of various ML algorithms and got the best accuracy of 95% with the XGBoost model. Ensemble learning models, such as combinations of KNN and Logistic Regression (LR), have also proven effective in predicting heart disease [11].

Evaluation metrics such as accuracy, sensitivity, specificity, and F1-score [12] have been crucial in assessing the performance of ML models for heart disease prediction.It has been recommended that AUC-ROC curves and accuracy scores be utilized for thorough model assessment [5].

In summary, the application of machine learning in predicting heart disease shows significant promise. The selection of appropriate ML algorithms, feature extraction techniques, and data balancing methods, along with the utilization of ensemble learning models and dimensionality reduction techniques like PCA [3], are key factors in achieving high accuracy in heart disease prediction.

3. Methodology

This section includes the dataset description, data pre-processing techniques and feature extraction using Principal Component Analysis for the train test and split. Python programming language and its various libraries are used for analyzing the data.

Figure 1 illustrates the sequential implementation of the model. It begins with Data Preprocessing, where the raw data is cleaned and prepared. Next, Feature Extraction using PCA (Principal Component Analysis) [3] is performed to reduce dimensionality. The data is divided into training and testing sets through a Train Test Split process. Machine Learning Algorithms are then employed on the training set, and the data is fitted into the model.The analysis of the dataset (Experimental Analysis) is conducted to understand the relation between the variables of the dataset. The predictions made by the models are evaluated (Evaluation), and finally, the finalized model is exported for application (Export Model for Application).

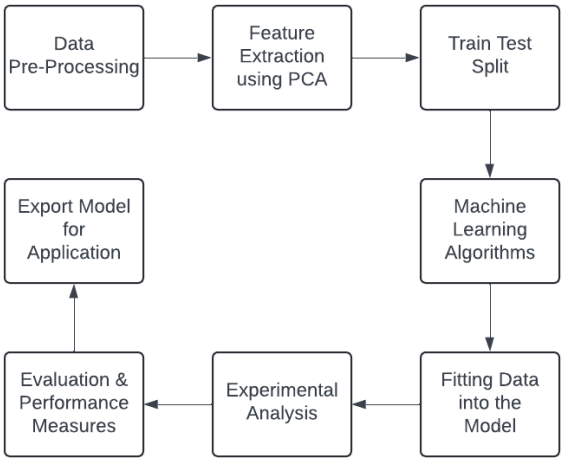

Figure 1: Flowchart of the Complete Procedure.

3.1. Dataset Description

The Heart Disease Dataset (Comprehensive) is a curated dataset that combines five popular heart disease datasets into a single comprehensive dataset [13]. The dataset contains medical data from patients with and without heart disease, with a total of 1,190 instances and 12 features. The 5 datasets used to create this comprehensive dataset are:

- Cleveland: 303 instances
- Hungarian: 294 instances
- Switzerland: 123 instances
- Long Beach VA: 200 instances
- Stalog (Heart) Data Set: 270 instances

The variables of the dataset [13] and description of nominal attributes[13] are depicted in the Table 1 and Table 2 respectively below.

Table 1: Description of Heart Disease Dataset (Comprehensive) [13] Attributes

S.No.	Attribute	Code given	Unit	Data type
1	age	age	in years	Numeric
2	sex	sex	1,0	Binary
3	chest pain type	chest_pain_type	1,2,3,4	Nominal
4	resting blood pressure	resting_bp_s	in mm Hg	Numeric
5	serum cholesterol	cholesterol	in mg/dl	Numeric
6	fasting blood sugar	fasting_blood_sugar	1,0 > 120 mg/dl	Binary
7	resting electrocardiogram results	resting_ecg	0,1,2	Nominal
8	maximum heart rate achieved	max_heart_rate	60-202	Numeric
9	exercise induced angina	exercise_angina	0,1	Binary
10	oldpeak =ST	oldpeak	0.0-6.2	Numeric
11	the slope of the peak exercise ST segment	ST_slope	0,1,2	Nominal
12	class	target	0,1	Binary

Table 2: Medical Attributes and Descriptions of Nominal Attributes

Attribute	Description
Sex	1 = male, 0 = female;
Chest Pain Type	– Value 1: typical angina – Value 2: atypical angina – Value 3: non-anginal pain – Value 4: asymptomatic
Fasting Blood Sugar	(fasting blood sugar > 120 mg/dl) 1 = true; 0 = false
Resting Electrocardiogram Results	– Value 0: normal – Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) – Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Exercise Induced Angina	1 = yes; 0 = no
The Slope of Peak Exercise ST Segment	– Value 1: upsloping – Value 2: flat – Value 3: downsloping
Target	1 = heart disease, 0 = Normal

3.2. Data Pre - Processing

The Heart Disease Dataset (Comprehensive) [13] required some preprocessing to make the data usable for analysis and model development. The initial step involved eliminating duplicate instances from the dataset [13] containeing 1,190 instances, but after removing the duplicates, the dataset was reduced to 918 instances.

The dataset was then scrutinized for missing values[3]. Although the dataset had no missing values, it contained some 0 values in the "RestingBP" and "Cholesterol" columns, which are not biologically feasible. To tackle this, we employed hot-deck imputation. Here, the missing values were filled by randomly selecting values from the respective columns.

The dataset also contained some negative values in the "Oldpeak" column, which were removed, resulting in a final dataset of 905 instances. Finally, the analysis of the raw data revealed that the "FastingBS" column had about 77% of its values as 0, indicating that this feature may not significantly impact the classification task[3]. Therefore, the "FastingBS" column was dropped from the dataset.

After these preprocessing steps, the dataset was ready for further analysis and model development.

3.3. Feature extraction using PCA

In the research study, Principal Component Analysis (PCA) [3] was employed as a feature extraction technique to reduce the dimensionality of the feature data[3]. . The PCA algorithm was applied to the feature data (x_data) to retain the most significant information while reducing the number of features. A total of 9 principal components were selected to effectively capture the variance in the data.

The explained variance ratio of the principal components was calculated to understand the amount of variance each principal component contributes to the overall dataset. This information is crucial for determining the optimal number of principal components to retain for the analysis. The explained variance ratio provides insights into the proportion of variance explained by each principal component.

A scree plot was generated to represent the explained variance ratio of the principal components visually. The scree plot helps identify the point of diminishing returns in terms of explained variance, aiding in the decision making process regarding the number of principal components to retain for the analysis.

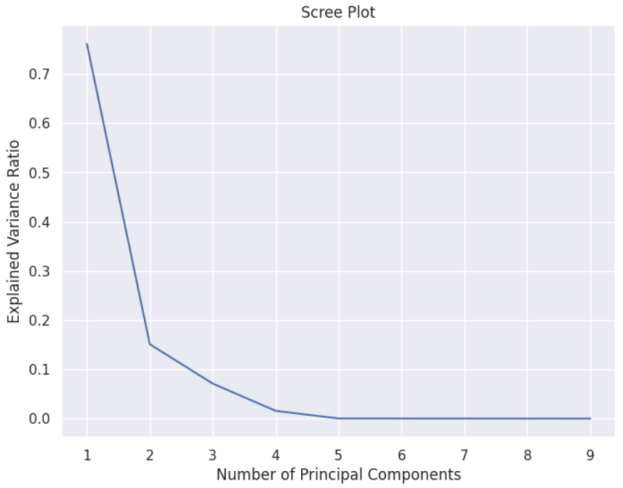

Figure 2: Scree Plot to Visually represent the Explained Variance Ratio

3.4. Train Test Split

Following the feature extraction using PCA, the transformed data (x_pca) was split into training and testing sets to facilitate the model training and evaluation process. The dataset was split into training and testing sets, with a test size of 20% and a random state of 0 to ensure reproducibility. Additionally, the data was stratified based on the target variable (y) to ensure a proportional distribution of classes in both the training and testing sets. This step is essential to preserve the dataset's integrity and ensure unbiased evaluation of the models.

4. Machine Learning Algorithms

Machine learning encompasses the automated generation of models, representing a type of data analysis [9]. With minimal human intervention,they can detect patterns and make predictions based on the information provided. Eight machine learning (ML) algorithms are briefly discussed in this section.

4.1. Logistic Regression

Logistic Regression (LR)[3, 10, 4, 9] is a statistical technique employed for binary classification tasks, such as predicting whether a patient has cardiovascular disease or not. LR uses a sigmoid function [3] to map the linear combination of input features (e.g., age, blood pressure, cholesterol levels) to a probability value between 0 and 1, showing the probability of the positive class (cardiovascular disease). The model is trained to maximize the likelihood of observing the actual class labels given the input features, and a decision threshold (usually 0.5) is applied to assign instances to the positive or negative class. Logistic regression proves to be a valuable tool for predicting the probability of heart disease by considering patient characteristics. This aids in early intervention and risk assessment.

4.2. Random Forest Classifier

Random Forest (RF)[5, 3] is a flexible and reliable ensemble learning algorithm[5] that utilizes decision trees [5] to achieve accurate predictions and model resilience. RF builds numerous decision trees, each trained on a random subset of the training data. It then merges these individual tree predictions using ensemble techniques. The ensemble approach helps to reduce overfitting and improve generalization performance. RF incorporates randomization techniques, such as bootstrap sampling and feature subset selection, to further enhance the diversity and robustness of the individual trees. When making predictions, Random Forest aggregates the outputs of individual trees, typically using voting for classification tasks. Random Forest is a highly effective and popular ensemble learning method recognized for its capability to manage complex datasets and achieve exceptional predictive accuracy in a vast range of applications, including cardiovascular disease prediction.

4.3. XGBoost

XGBoost (Extreme Gradient Boosting)[5, 3] is a powerful ensemble learning method that utilizes a gradient boosting framework to greatly enhance predictive accuracy, particularly when employing decision tree models. XGBoost sequentially builds a series of weak learners (typically decision trees) and combines them to create a strong predictive model. It uses a boosting technique, where each new model in the ensemble corrects errors made by the previous models, and minimizes the loss function [5] (e.g., log loss for classification) by optimizing the gradient of the loss function. XGBoost is known for its efficiency, scalability, and ability to handle large datasets, making it a powerful choice for cardiovascular disease prediction tasks.

4.4. Naive Bayes

Naive Bayes [5] is a probabilistic machine learning algorithm, relying on Bayes' theorem and assumes independence between features. It is frequently utilized for classification tasks, especially in text classification and spam filtering. This classifier calculates the posterior probability of a class given input features, assuming feature independence. Despite its simplicity, Naive Bayes can perform remarkably in diverse classification tasks, including predicting cardiovascular disease, particularly when the independence assumption holds reasonably well.

4.5. Neural Networks

Neural Networks (NN), or Artificial Neural Networks (ANN) [5], are a type of machine learning model inspired by the biological architecture of the human brain. They comprise interconnected nodes (neurons) arranged in layers, enabling them to capture intricate patterns and dependencies within datasets. Neural networks can be successfully used for predicting cardiovascular disease by capturing intricate relationships between input features (e.g., patient characteristics) and the target variable (presence of cardiovascular disease). The network is trained using a process of forward propagation and backpropagation, where the error between the predicted output and the actual target is minimized by adjusting the network's weights and biases.

4.6. Support Vector Machines

Support Vector Machines (SVM) [3] is a robust supervised machine learning technique utilized for classification and regression tasks. SVM strives to identify the optimal hyperplane that effectively divides data points into distinct classes while maximizing the margin between these classes. SVM can efficiently handle non-linear decision boundaries by using kernel functions (e.g., polynomial, radial basis function) to map input features into higher-dimensional space. SVM's ability to handle complex data patterns and deliver high accuracy makes it a suitable choice for cardiovascular disease prediction.

4.7. Decision Trees

Decision Trees are adaptable and explainable supervised machine learning methods employed for classification and regression purposes. They acquire straightforward decision rules from training data to forecast the target variable through iterative feature space partitioning. Decision Trees aim to minimize impurity or maximize information gain at each node, using measures such as Gini Impurity or Entropy [3, 14]. The decision rules acquired through Decision Trees offer valuable insights into the factors influencing cardiovascular disease prediction.

4.8. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a straightforward supervised learning approach for classification and regression tasks. It predicts by identifying the k nearest data points (neighbors) in the training set and utilizing their labels (for classification) or averaging their values (for regression). Selecting the appropriate value for k is vital to prevent overfitting or underfitting. KNN may pose computational challenges with large datasets, and feature scaling is typically applied to improve distance calculations. Despite its limitations, KNN's simplicity and capability to manage non-linear data patterns render it widely used in diverse machine learning applications.

5. Experimental Analysis

5.1. Data Exploration

The dataset contains crucial features ranging from 0 or 1 to a range of values,like age spanning from 20s to 70s, sex represented by 0 or 1 for females or males, and chest pain type categorized into four intensities, among others.

The target variable can assume values of 0 or 1, indicating the presence of heart disease. This information is vital for recognizing potential class imbalance, a common issue in medical datasets where one class may be significantly less represented. Addressing class imbalance through appropriate sampling techniques or class weighting strategies can enhance the model's performance and ensure that predictions are not biased towards the majority class [5]. Exploratory data analysis revealed diverse distributions among the features, highlighting the need for careful preprocessing and feature engineering to enhance the predictive capabilities of the AI models.

5.2. Distribution of Target Variable

It is observed that people who have a possibility of Heart Disease and who do not are approximately 500 and 400 respectively. This shows that more than half of the people in the dataset exhibit signs or have the possibility to be diagnosed with cardiac conditions.

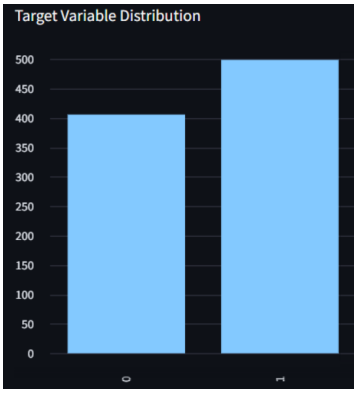

Figure 3: Distribution of Target Variable

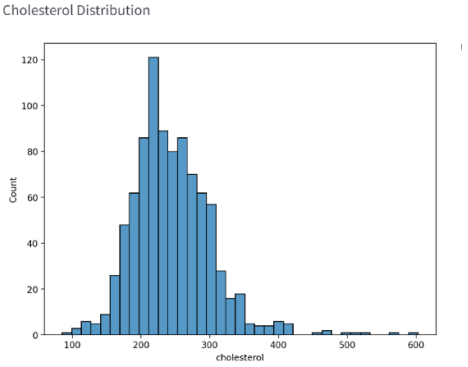

Figure 4: Distribution of Cholesterol Levels

5.3. Distribution of Cholesterol Levels

By examining the dataset, we can note that cholesterol levels vary widely, ranging from as low as 100 to as high as the 500s. Additionally, it's apparent that around 70.7 percent of individuals fall within the cholesterol range of 200 to 300. These insights are valuable for understanding specific health patterns and conducting more targeted research, particularly concerning individuals with low and high cholesterol levels.

5.4. Distribution of Max Heart Rate

Analyzing the dataset reveals that the maximum heart rate is pivotal in determining the likelihood of Heart Disease (Figure 5). The graph's horizontal axis represents heart rate in beats per minute (bpm), while the vertical axis displays the number of individuals in the dataset with each heart rate. It is evident that most heart rates in the dataset cluster between 120 and 160 bpm, accounting for 72.5 percent within this range.

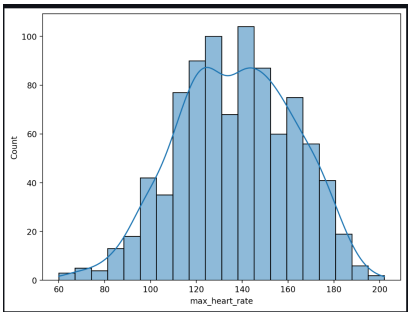

Figure 5: Distribution of Max Heart Rate

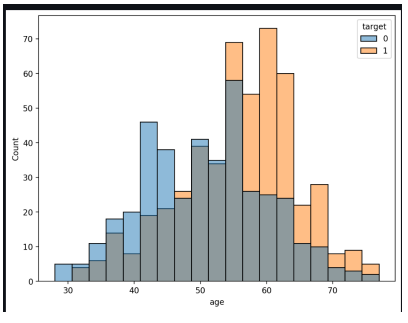

Figure 6: Age vs Target Variable Distribution

5.5. Age vs Target Variable Distribution

The dataset has different patterns in heart diseases based on age(Figure 6) [5]. We can observe the age of range 30 to 50, around 150 individuals i.e., only 36 percent have the possibility of having heart disease while 265 individuals i.e., 64 percent are healthy and safe from the risk of having any heart issues. If we go to the middle range of age 55 to 70, around 315 individuals i.e., 68 percent have the possibility of heart disease while 150 individuals i.e., 32 percent have no heart diseases. From this analysis, we can infer that as the age increases from 40 to 50, there is more possibility of getting a heart disease, so this information can be beneficial while addressing heart diseases.

5.6. Chest Pain type vs Target Variable Distribution

The bar chart (Figure 7) below shows the chest pain type and its relation to the target variable (heart disease possibility). By considering the chest pain type 2 then, it can be observed that only around 20 individuals have heart issues. At the same time, 150 members are healthy, so in this pain type, less than 15 percent tend to have heart issues, and in the chest pain type 4, around 80 percent of the individuals have heart diseases. As the chest pain intensity increases, there is more chance of getting heart issues.

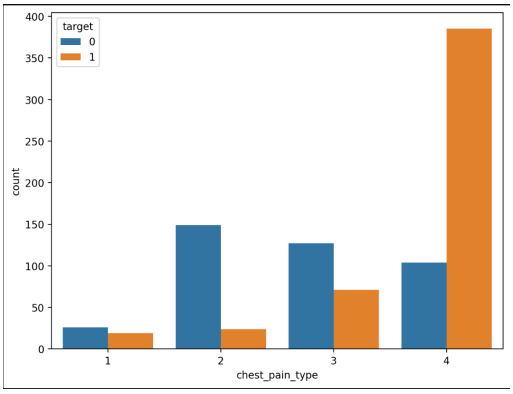

Figure 7: Chest Pain Type vs Target Variable Distribution

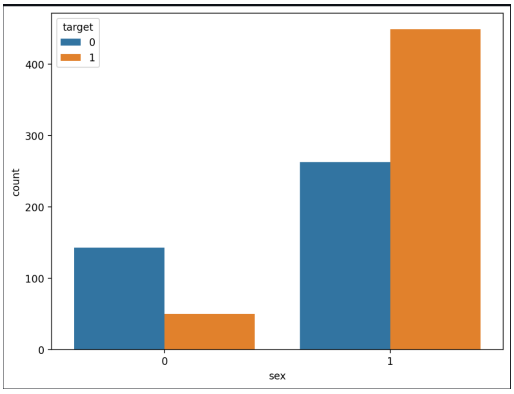

Figure 8: Gender vs Target Variable Distribution

5.7. Gender vs Target Variable Distribution

The dataset analysis also shows that among the people in our study, gender significantly influences the prevalence of heart disease, where 0 indicates female and 1 indicates male. Among the male members, more than 60 percent are prone to having heart issues, while in females, just around 22 percent have the possibility of having heart disease. The same can be observed from Figure 8.

5.8. Correlation between all variables (HeatMap)

A correlation heatmap [5] is a visual tool to analyze the relationships between multiple numerical variables in a dataset. It is a color-coded matrix that helps you quickly identify and understand how strongly variables are correlated. It resembles a grid where each variable in the data occupies a row and a column. The cells within the grid represent the correlation between each pair of variables. Darker colors represent stronger correlations, and lighter colors indicate weak correlations. Red or orange shades usually mean positive correlations, while blue or green shades indicate negative ones.

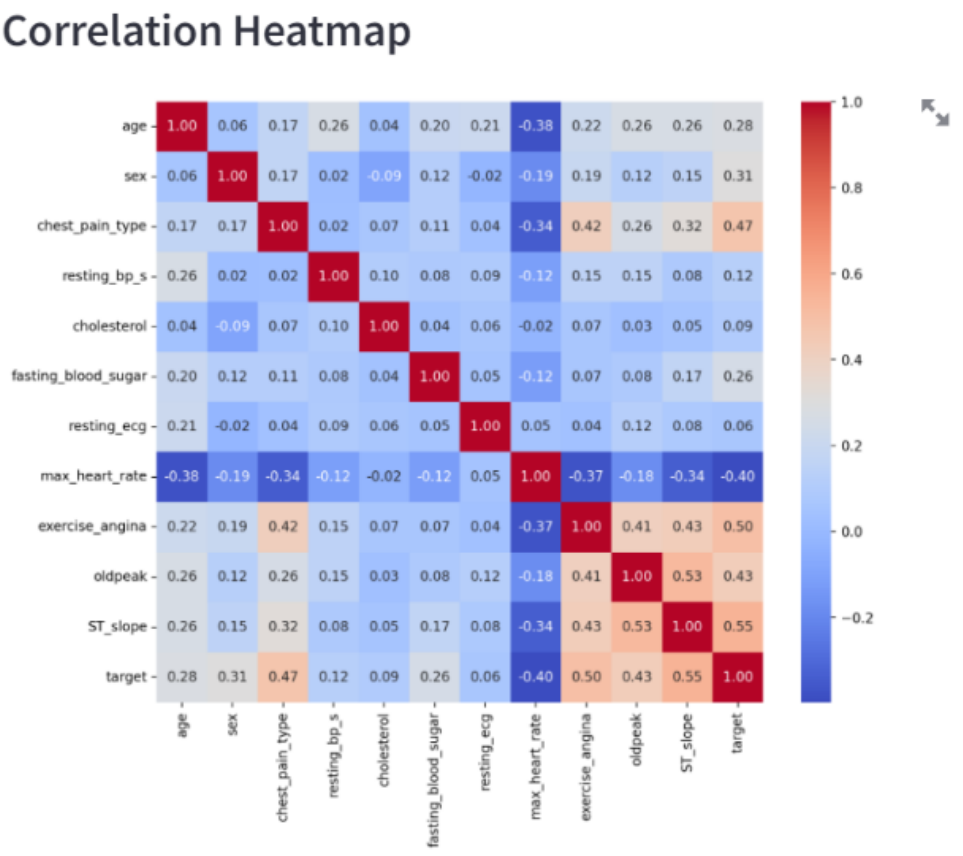

Figure 9: Correlation Heatmap for the Features of the Dataset

6. Performance Metrics

The performance of different machine learning models [3] in predicting the probability of cardiovascular diseases was assessed using the following metrics [14]. The Accuracy Score gauges the overall percentage of accurate predictions made by the model. Precision assesses the proportion of optimistic predictions that are genuinely correct, prioritizing the quality of positive results. Meanwhile, Recall measures the percentage of actual positive cases correctly identified by the model, highlighting the comprehensiveness of positive result detection. F1-Score combines Precision [3] and Recall [3] to provide a balanced view of model performance.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

- TP (True Positives) are people with the disease identified as having the disease
- TN (True Negatives) are healthy people identified as not having the disease
- FP (False Positives) are healthy people identified as having the disease
- FN (False Negatives) are people with the disease identified as not having the disease

7. Model Performance Comparison

7.1. AUC - ROC Curve

The curve of AUC-ROC (Area Under the Receiver Operational Characteristic Curve) visually represents how efficiently a binary classification model carries out regarding various classification thresholds [5]. It is often used in machine learning to assess a model's ability to distinguish between two classes, usually the positive class (e.g. presence of heart disease) and the negative class (e.g. absence of heart disease). It is recommended to be used for balanced datasets. The AUC - ROC curve for the Random Forest Classifier was the best among all eight algorithms used for the dataset.

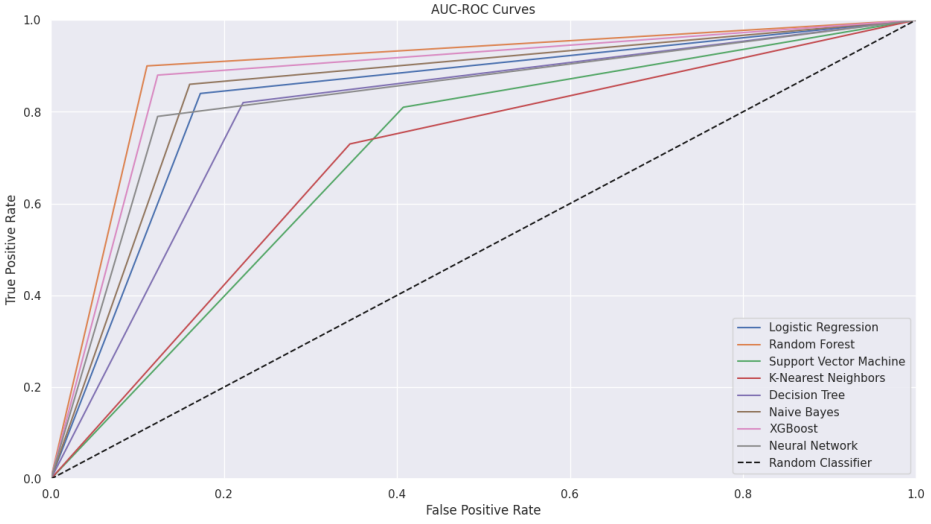

Figure 10: AUC - ROC Curve for Comparison of Machine Learning Algorithms

7.2. Performance Metrics Comparison Table

Our tests and evaluations have found that the Random Forest Classifier is the most effective algorithm for predicting the likelihood of someone having a heart problem, achieving an accuracy rate of 89.5%. The precision of this algorithm is 90.9%, with a recall of 90.0% and an F-1 score of 90.4%. In comparison, the study by S. J. Anirban Ghosh et al. [3] reported an accuracy score of 86.4%. Thus, our approach improved the accuracy to 89.5% for the Random Forest classifier and outperformed other machine learning algorithms. Below is a comparison table (Table 3) of the various algorithms used.

Table 3: Comparison of Machine Learning (ML) Algorithms

Performance of Machine Learning (ML) Models				
ML Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.834	0.857	0.840	0.849
Random Forest	0.895	0.909	0.900	0.904
SVM	0.713	0.711	0.810	0.757
Decision Tree	0.801	0.820	0.820	0.820
Naive Bayes	0.851	0.869	0.860	0.864
Neural Network	0.829	0.888	0.790	0.836
XGBoost	0.879	0.898	0.880	0.889
K-Nearest Neighbor	0.697	0.723	0.730	0.726

8. Streamlit UI application

8.1. About Streamlit

Streamlit [15] is an open-source Python framework for creating interactive, web-based applications for data science and machine learning projects. It allows researchers to quickly build and share custom dashboards and visualizations, facilitating real-time data exploration and enhancing the accessibility and impact of their research to everyone.

8.2. Exporting Our Model into the UI Application

We have crafted a user-friendly interface utilizing the Streamlit Python library, enhancing the viewing experience for both Data Visualizations and Predictions. We have employed a Python library called Pickle to store the model in a file and integrate the application with our machine-learning models. When constructing the application, we import the same library and load the model, enabling seamless predictions of new data within our environment. Additionally, we have dedicated a separate page for visualizations, presenting a comprehensive dashboard showcasing all dataset plots.

9. Conclusion

This research examines the application of data analytics, visualization techniques, and machine learning (ML) methods in forecasting cardiovascular diseases (CVDs). By processing a collected dataset comprising 1,190 instances and 11 variables indicative of heart disease, this study applies several ML algorithms, including decision trees, support vector machines, random forests, and logistic regression. Utilizing Principal Component Analysis (PCA), significant predictors of heart disease were pinpointed, and the predictive accuracies of various ML algorithms in determining the risk of CVDs were assessed. The results highlight the effectiveness of data-driven strategies in the early diagnosis and prevention of cardiovascular disorders, providing crucial insights for enhancing health care outcomes. The quantitative evaluation revealed that the random forest model achieved an 89.5% accuracy rate in CVD prediction. Moreover, PCA helped in delineating key risk factors, underscoring the pivotal role of data analytics and visualization in propelling public health efforts forward.

Conflicts of Interest

The authors report no conflicts of interest.

Declaration of Generative AI and AI-assisted technologies in the writing process

Authors declare that no generative AI tools were used in the creation of this manuscript.

References

[1] A. Alkahtani, A. Albarakati, M. Alshehri, A. Alshehri, A. Alshehri, and V. Lozano, "Early prediction in classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods," *NCBI*, no. 1, 2023.

[2] M. Ekta, M. S. Manikandan, C. M. Sujatha, and S. Ramakrishnan, "Machine learning–based heart disease prediction system for indian population," *ScienceDirect*, 2020.

[3] S. J. Anirban Ghosh, "A study on heart disease prediction using different classification models based on cross validation method," *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)*, vol. 11, 2022.

[4] N. Absar, E. K. Das, S. N. Shoma, M. U. Khandaker, M. H. Miraz, M. R. I. Faruque, N. Tamam, A. Sulieman, and R. K. Pathan, "The efficacy of machine-learning-supported smart system for heart disease prediction," *Healthcare*, vol. 10, no. 6, p. 1137, 2022.

[5] M. A. Sufian, "Ai models for early detection and mortality prediction in cardiovascular diseases," *TechRxiv*, 2023.

[6] A. Ishaq, S. Sadiq, M. Umer, D. S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using smote and effective data mining techniques," *IEEE Access*, vol. PP, pp. 1–1, 03 2021.

[7] O. Taylan, A. S. Alkabaa, H. S. Alqabbaa, E. Pamukçu, and V. Leiva, "Early prediction in classification of cardiovascular diseases with machine learning, neuro-fuzzy and statistical methods," *Biology*, vol. 12, no. 1, p. 117, 2023.

[8] N. Almansouri, M. Awe, S. Rajavelu, K. Jahnavi, R. Shastry, A. Hasan, H. Hasan, M. Lakkimsetti, R. AlAbbasi, B. Gutiérrez, and A. Haider, "Early diagnosis of cardiovascular diseases in the era of artificial intelligence: An in-depth review," *Cureus*, vol. 16, no. 3, p. e55869, 2024.

[9] R. Gandham, K. R. Manambakam, S. V. N. Madala, N. S. Nannapaneni, S. Tokala, and M. K. Enduri, "Predictive modeling for heart disease detection with machine learning," in *2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 325–329, IEEE, 2023.

[10] S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning machine (elm) technique for heart disease diagnosis," in *2015 IEEE Canada International Humanitarian Technology Conference (IHTC2015)*, pp. 1–3, IEEE, 2015.

[11] K. V. V. Reddy, I. Elamvazuthi, A. Aziz, S. Paramasivam, H. Chua, and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," *Applied Sciences*, vol. 11, no. 18, p. 8352, 2021.

[12] M. Limbitote, D. Mahajan, K. Damkondwar, and P. Patil, "A survey on prediction techniques of heart disease using machine learning," *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)*, vol. 09, June 2020.

[13] "Heart disease dataset (comprehensive)." https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive, 04 2024.

[14] J. Srivastava, L. Bhargva, D. J. P. Yadav, Monica, P. Saxena, and P. K. Aggarwal, "Automated heart disease prediction system using machine learning approaches," in *2023 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI)*, pp. 1–6, IEEE, 2023.

[15] S. Shukla, A. Maheshwari, and P. Johri, "Comparative analysis of ml algorithms & streamlit web application," in *Proceedings of the 2021 International Conference on Advanced Computing and Communication Networks (ICAC3N)*, pp. 175–180, 12 2021.