Predictive Modelling for Cardiovascular Disease Prediction Using Machine Learning Algorithms

I. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for approximately 17.9 million deaths annually, representing 31% of all global mortality (1). These conditions often progress undetected until they reach a critically severe stage, underscoring the importance of early detection and timely medical intervention to improve survival rates and quality of life.

Recent advancements in machine learning (ML) offer promising opportunities for the early detection and intervention of cardiovascular conditions. By analysing extensive medical datasets that include complex patient histories and clinical parameters, these technologies can predict health outcomes with remarkable precision (2). The primary goal of this research is to utilize advanced predictive modelling to help healthcare providers identify cardiovascular conditions early, potentially saving lives and reducing healthcare costs.

This study will conduct a comparative analysis across various algorithms to determine which best suits the early detection of cardiovascular conditions. The aim is to demonstrate the significant impact of machine learning techniques in addressing real-world healthcare challenges, particularly in reducing healthcare costs and improving patient outcomes.

II. DATASET OVERVIEW

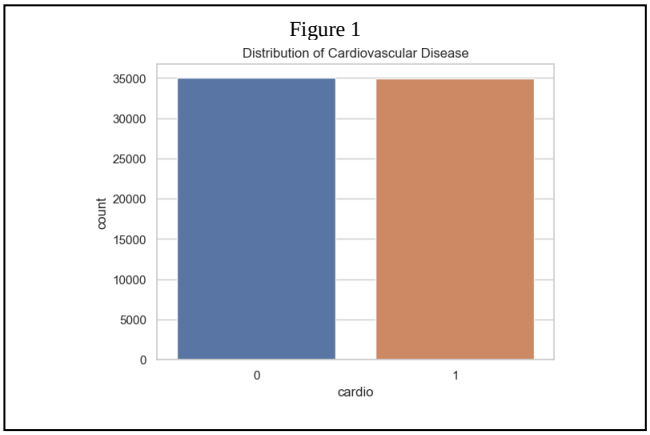
This study utilizes the Cardiovascular Disease Dataset from the UCI Machine Learning Repository (3), available on Kaggle (4), encompassing 70,000 patient records. The dataset comprises 12 attributes, including both continuous and categorical data types (Table 1). Each record provides demographic information, health status indicators, and lifestyle choices, all essential for analyzing cardiovascular health risks. The dataset's binary target variable, 'cardio', indicates the presence (1) or absence (0) of cardiovascular disease, making it vital for developing predictive models.

Table 1

Attribute	Type	Description
id	Integer	Unique Identifier
age	Continuous	Age of the patient (in days)
gender	Categorical	Gender (1: Female, 2: Male)
height	Continuous	Height in centimeters
weight	Continuous	Weight in kilograms
ap_hi	Continuous	Systolic blood pressure
ap_lo	Continuous	Diastolic blood pressure
cholesterol	Categorical	Cholesterol level (1 = Normal, 2 = Above Normal, 3 = High)
gluc	Categorical	Glucose level (1 = Normal, 2 = Above Normal, 3 = High)
smoke	Categorical	Smoking status (1: Yes, 0: No)
alco	Categorical	Alcohol intake (1: Yes, 0: No)
active	Categorical	Physical activity (1: Yes, 0: No)
cardio	Binary	Cardiovascular disease (1: Yes, 0: No)

As illustrated in Figure 1, the dataset is evenly balanced between the two classes, with each class comprising approximately 35,000 instances. This balanced distribution is crucial for training unbiased machine learning models, ensuring that performance metrics accurately reflect the model's predictive capabilities. The binary nature of the target variable, 'cardio', simplifies the application of various machine learning models, allowing predictions based on a comprehensive set of health indicators and lifestyle choices.

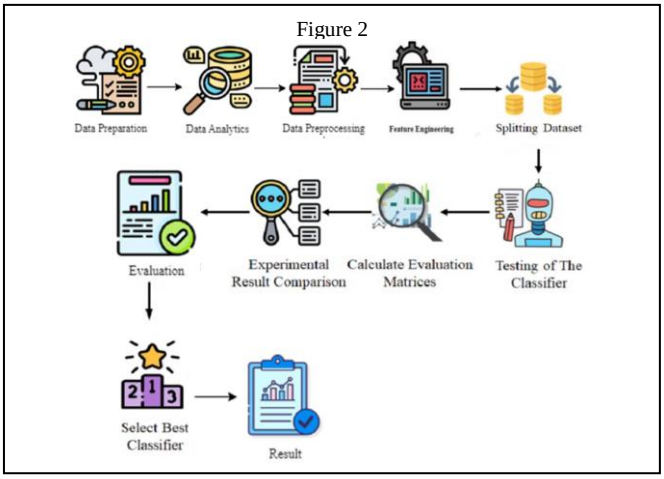
Figure 1



III. Exploratory Data Analysis and Preprocessing

The following workflow outlines the steps taken in this study (Figure 2):

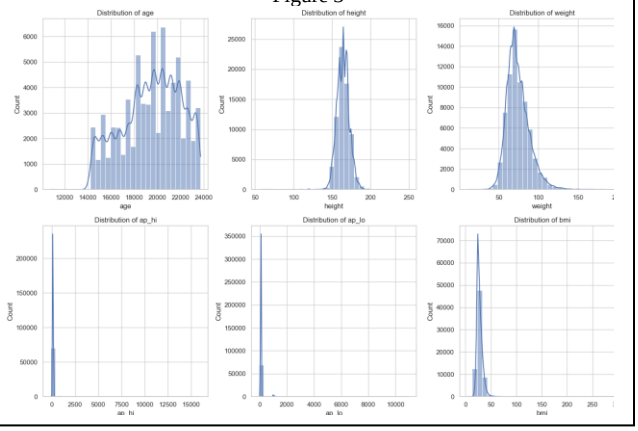
Figure 2



A. Initial Observations and Outlier Detection and Handling

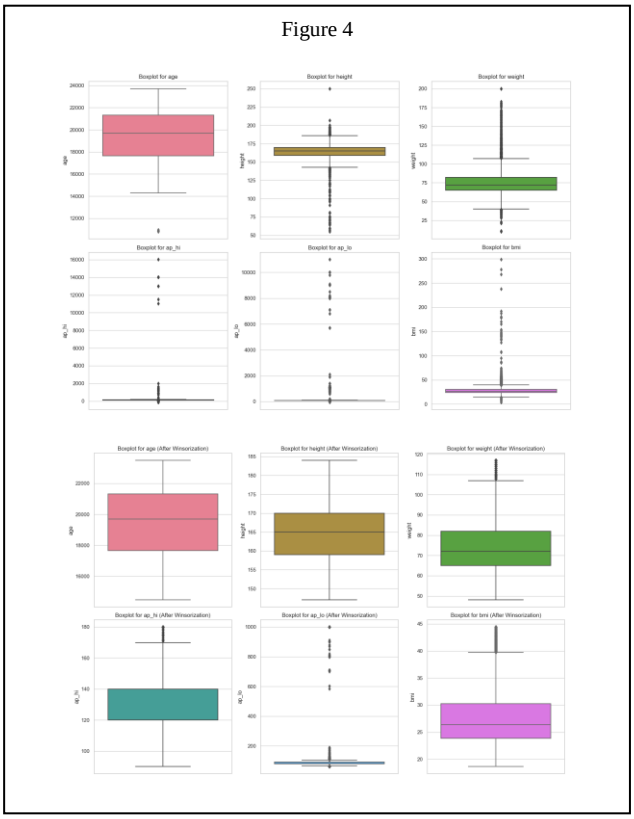
The initial exploratory data analysis (EDA) was conducted using statistical summaries and graphical representations to identify patterns, anomalies, and distributions across the features. Descriptive statistics were computed to understand the central tendencies and variability within the data. This analysis highlighted the need for normalization of skewed data distributions, particularly in physiological measurements, as shown in Figure 3.

Figure 3



Extreme care was taken in handling outliers, particularly in blood pressure measurements, to ensure that data points were not discarded without thorough examination. Outliers such as blood pressure (ap_hi and ap_lo) were managed through Winsorization, a method that caps extreme values at specified percentiles to minimize the impact of potential data entry errors

Figure 4



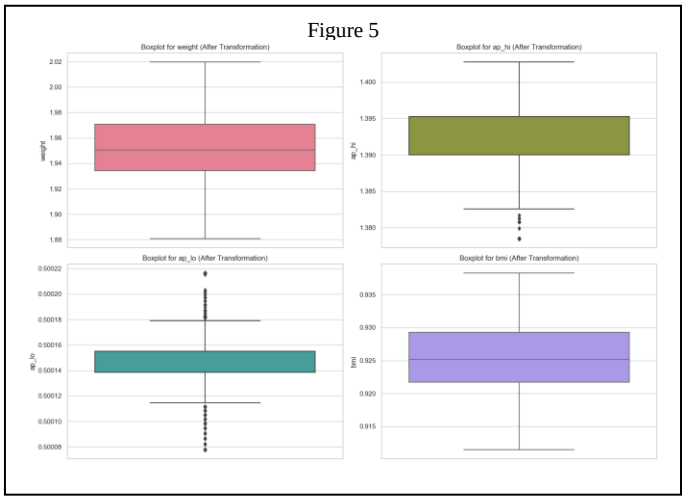
while preserving as much data integrity as possible (5). For this dataset, values beyond the 95th percentile and below the 5th percentile were adjusted to these thresholds, as depicted in Figure 4.

To address skewness in variables such as weight and BMI, a Box-Cox transformation was applied. This transformation is defined by the formula

$$y(\lambda) = \frac{(y^{\lambda} - 1)}{\lambda}, \quad \text{for } \lambda \neq 0 \qquad (1)$$

It optimizes the transformation parameter λ to stabilize variance and enhance the symmetry of distributions, which is crucial for the effectiveness of linear models (6). This approach is preferred over simpler methods like logarithmic transformations because it adapts to various distribution shapes, thus enhancing the suitability of the data for linear models, as demonstrated in Figure 5.

Figure 5



B. Feature Engineering

Feature engineering plays a pivotal role in enhancing the predictive power of machine learning models. By transforming and selecting relevant features, model accuracy can be significantly improved. Below, the clinical significance of key features and the rationale behind their engineering are detailed, which are critical for predicting cardiovascular disease (Table 2).

Cholesterol Levels:

Cholesterol levels are categorized into normal, above normal, and high to reflect their well-documented impact on cardiovascular health. Elevated cholesterol is a major risk factor for coronary artery disease as it contributes to arterial blockage (7). In this study, cholesterol levels were treated as categorical variables to simplify their interpretation and utilization in the predictive models.

Blood Pressure:

Both systolic and diastolic blood pressures are critical indicators of cardiovascular risk. High values are associated with an increased risk of heart disease and stroke (7).

Furthermore, a derived feature is introduced, 'pulse pressure,' which is calculated as:

$$Pulse\ Pressure = Systolic\ Blood\ Pressure - Diastolic\ Blood\ Pressure \quad (2)$$

This measure assesses arterial stiffness and is a predictor of cardiovascular events, especially in the elderly (7).

Body Mass Index (BMI):

BMI was calculated from height and weight data using the formula:

$$BMI = \frac{weight\ (kg)}{(height\ (m))^2} \quad (3)$$

This index was categorized into underweight (BMI < 18.5), normal (BMI 18.5 to 24.9), overweight (BMI 25 to 29.9), and obese (BMI ≥ 30) based on WHO guidelines, providing a quantifiable measure of obesity levels which are strongly correlated with cardiovascular risk (7).

Physical Activity and Lifestyle Choices:

The levels of physical activity, smoking status, and alcohol intake were encoded as binary features. Regular physical activity is known to reduce the risk of heart failure, supporting its inclusion as a significant predictive factor (7). Similarly, smoking and alcohol consumption data were utilized due to their established effects on heart health.

Glucose Levels:

Given the strong link between diabetes and cardiovascular diseases, glucose levels were categorized into normal, above normal, and high. This feature helps in identifying individuals at an increased risk of cardiovascular issues due to elevated glucose levels (8).

Table 2

Feature	Description
age_years	Age in years
gender	Gender
height	Height
weight	Weight
ap_hi	Systolic Blood Pressure
ap_lo	Diastolic Blood Pressure
smoke	Smoker
alco	Alcohol Consumer
active	Physically Active
bmi	Body Mass Index
age_x_cholesterol	Age x Cholesterol Interaction
bmi_x_bp	BMI x BP Interaction
pulse_pressure	Pulse Pressure

V. Problem Requirements and Metrics

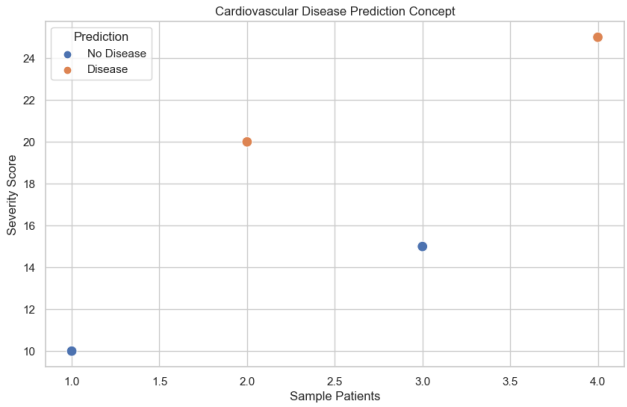
A. Problem Definition and Requirements

Cardiovascular disease prediction is a binary classification problem where the goal is to predict whether a patient will have a cardiovascular disease based on various health indicators. The requirements for the predictive model include:

- *Accuracy*: High accuracy is crucial as the outcome influences clinical decisions.
- *Timeliness:* The model must quickly provide predictions to be feasible in clinical settings.
- *Interpretability:* Medical staff must be able to understand how decisions are made by the model to trust and effectively use it.

Figure 6 illustrates the classification challenge addressed by our predictive models. Each point represents a patient sample, categorized based on the predicted presence or absence of disease. This visual aids in understanding the classification problem our models aim to solve.

Figure 6



B. Evaluation Metrics

Table 3 outlines the essential metrics for evaluating the performance of the models in this study, connecting each metric with its corresponding requirement for the predictive model. These metrics will assist in assessing each model's effectiveness and ensure that the chosen model meets the necessary clinical standards and requirements.

Table 3

Requirement	Metric	Description	Formula
Effective Classification	Accuracy	Measures the overall correctness of the model.	$\frac{TP+TN}{TP+TN+FP+FN}$ (4)
Minimizing False Positives	Precision	Important for conditions where false positives carry high risk.	$\frac{TP}{TP+FP}$ (5)
Minimizing False Negatives	Recall	Critical in medical diagnostics where missing a positive case could be dangerous.	$\frac{TP}{TP+FN}$ (6)
Balance Between Precision & Recall	F1-Score	Provides a balance between Precision and Recall, useful in situations with uneven class distribution.	$2 \times \frac{Precision \times Recall}{Precision + Recall}$ (7)
Detailed Error Analysis	Confusion Matrix	Visualizes the performance of a classification model, offering insight into the nature of errors made.	Displaying True Positives, False Positives, True Negatives, and False Negatives

VI. Model Descriptions and Categorization

This section describes the selected machine learning models for this study and explains the exclusion of certain others. All chosen models are supervised learning methods suitable for the labeled dataset.

A. Selected Model Descriptions and Justifications

- *Random Forest*: An ensemble method combining multiple decision trees to enhance robustness and accuracy. Effective for unbalanced datasets and reducing overfitting, ideal for clinical predictions (9).
- *XGBoost:* Utilizes gradient-boosted decision trees with regularization and tree pruning for superior performance and overfitting mitigation. It excels in handling structured data and is known for its speed and accuracy in predictive modelling (10).
- *Logistic Regression:* Models the probability of binary outcomes, offering straightforward and effective binary classification for predicting cardiovascular disease. Its simplicity and efficiency make it suitable for high-dimensional datasets.
- *Gaussian Naive Bayes:* Uses Bayes' theorem with the assumption of predictor independence, performing well with large categorical or Gaussian-distributed datasets (11).
- *Support Vector Machines (SVM):* Creates hyperplanes in high-dimensional spaces for data classification, handling complex decision boundaries and non-linear data effectively (12).
- *K-Nearest Neighbors (KNN):* Classifies based on the closest training examples, adapting well to dynamic healthcare datasets (12).
- *Decision Tree Classifier:* Builds a tree-like model to split the dataset, easily interpretable for clinical settings (12).
- *Linear Discriminant Analysis (LDA):* Finds a linear combination of features to separate classes, efficient for linear decision boundaries and computationally straightforward. LDA is particularly useful for reducing dimensionality before subsequent classification (12).

B. Exclusions of Other Models

Certain algorithms were excluded based on project needs:

- *Deep Learning Models:* Excluded due to their need for large datasets, high computational costs, and lack of transparency (13).
- *Complex Ensemble Methods like Stacking:* Complexity and computational demands outweigh benefits for this dataset (14).
- *Reinforcement Learning:* Inapplicable as it is suited for dynamic decision-making tasks, not static, supervised projects.

VII. Analysis and Evaluation

This section discusses the performance of various predictive models used in the study, based on the metrics outlined in Section V. Each model was evaluated on its accuracy, precision, recall, F1-score, and the insights gleaned from confusion matrices. Additionally, cross-validation was implemented to ensure the robustness and generalizability of the results. An 80-20 split ensured representativeness and reproducibility. Hyperparameters for all models were optimized using grid search, enhancing accuracy and robustness. 5-fold cross-validation ensured robustness and minimized overfitting, balancing efficiency and reliability.

A. Model Performance Comparison

The overview of model performance based on key metrics such as accuracy, precision, recall, and F1-score are illustrated in Table 4.

- *Accuracy:* XGBoost, Logistic Regression, and Linear Discriminant Analysis (LDA) show the highest accuracy scores (0.73), making them strong candidates for predicting cardiovascular disease. These models demonstrate a good balance between detecting positive and negative

Table 4

Model	Accuracy	Precision (No Disease/Disease)	Recall (No Disease/Disease)	F1-Score (No Disease/Disease)	Support (No Disease)	Support (Disease)	Cross-Validation Accuracy
RandomForestClassifier	0.72	0.71/0.72	0.73/0.70	0.72/0.71	6988	7012	0.716 ± 0.002
XGBClassifier	0.73	0.72/0.75	0.77/0.70	0.74/0.72	6988	7012	0.733 ± 0.001
Logistic Regression	0.73	0.72/0.75	0.76/0.70	0.74/0.72	6988	7012	0.728 ± 0.003
GaussianNB	0.61	0.58/0.67	0.79/0.44	0.67/0.53	6988	7012	0.613 ± 0.005
Support Vector Machine	0.60	0.59/0.61	0.64/0.56	0.62/0.59	6988	7012	0.601 ± 0.008
K-Nearest Neighbors	0.58	0.58/0.58	0.58/0.58	0.58/0.58	6988	7012	0.581 ± 0.005
Decision Tree Classifier	0.63	0.63/0.63	0.63/0.63	0.63/0.63	6988	7012	0.635 ± 0.003
Linear Discriminant Analysis	0.73	0.72/0.74	0.76/0.70	0.74/0.72	6988	7012	0.727 ± 0.002

classes, evidenced by their performance in cross-validation, where XGBoost slightly leads with a cross-validation accuracy of 0.733 ± 0.001.

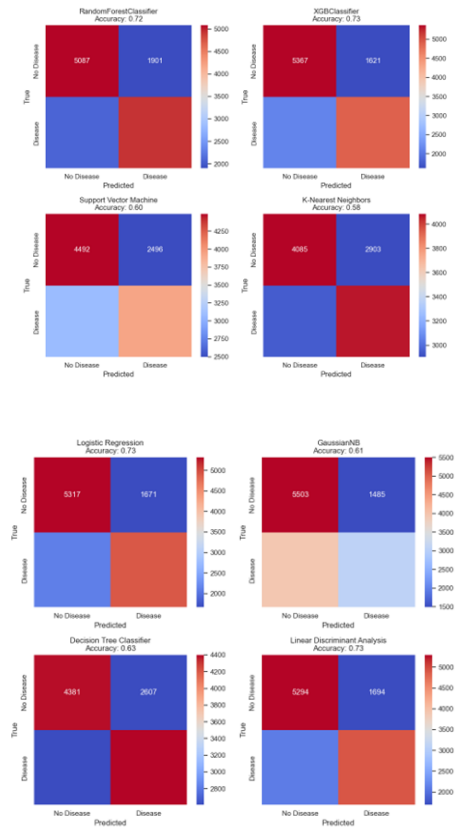
- *Precision and Recall:* Logistic Regression and LDA perform consistently well. Logistic Regression has precision (0.72 for No Disease, 0.75 for Disease) and recall (0.76 for No Disease, 0.70 for Disease), while LDA shows precision (0.72 for No Disease, 0.74 for Disease) and recall (0.76 for No Disease, 0.70 for Disease). Both models indicate their effectiveness in minimizing false positives and capturing true disease cases. Conversely, models like Gaussian Naive Bayes and Support Vector Machine show significant trade-offs between precision and recall, which could lead to higher misclassification rates in a clinical setting.
- *F1-Score:* Logistic Regression, XGBoost, and LDA have comparable F1-scores of approximately 0.74 for the disease class, suggesting a balanced detection capability for both the presence and absence of cardiovascular disease.

B. Confusion Matrix Analysis

The confusion matrices (Figure 7) and detailed counts (Table 5) provide insights into each model's performance in classifying cardiovascular disease.

- *Random Forest and XGBoost:* Both models robustly identify true negatives and true positives but have notable false positives and negatives. XGBoost slightly outperforms Random Forest with fewer false classifications.
- *Logistic Regression and Linear Discriminant Analysis:* These models perform consistently well, with balanced detection of true positives and true negatives. Logistic Regression has 4911 true positives and 1671 false positives, while LDA has 4914 true positives and 1694 false positives, indicating reliable performance.

Figure 7



- *Gaussian Naive Bayes:* This model has a high rate of false negatives (3952), posing risks of missed disease cases, which is critical in clinical settings.
- *Support Vector Machine and K-Nearest Neighbors:* These models exhibit weaker performance with higher false negatives, such as 3079 for Support Vector Machine and 2948 for K-Nearest Neighbors, potentially leading to missed diagnoses.
- *Decision Tree Classifier:* Shows moderate performance with 2607 false positives and 2611 false negatives, indicating it is less reliable compared to XGBoost and Logistic Regression.

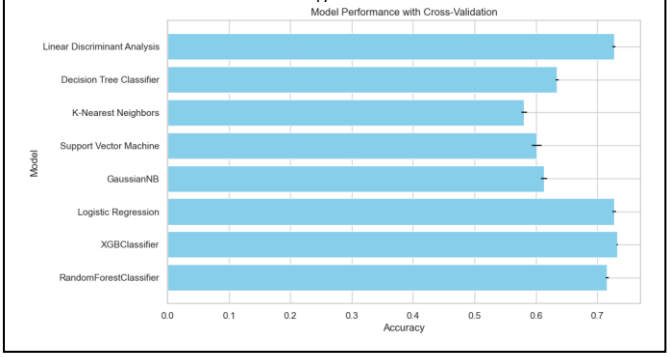
Table 5

Model	TN (True No Disease)	FP (False Disease)	FN (False No Disease)	TP (True Disease)
RandomForestClassifier	5087	1901	2083	4929
XGBClassifier	5367	1621	2116	4896
Logistic Regression	5317	1671	2101	4911
GaussianNB	5503	1485	3952	3060
Support Vector Machine	4492	2496	3079	3933
K-Nearest Neighbors	4085	2903	2948	4064
Decision Tree Classifier	4381	2607	2611	4401
Linear Discriminant Analysis	5294	1694	2098	4914

C. Cross-Validation Performance

The bar graph (Figure 8) illustrates the cross-validation accuracy for each model, comparing their stability across multiple dataset partitions. Models like Logistic Regression, XGBoost, and Linear Discriminant Analysis display high accuracy and less variability in performance, suggesting better generalization across various patient data.

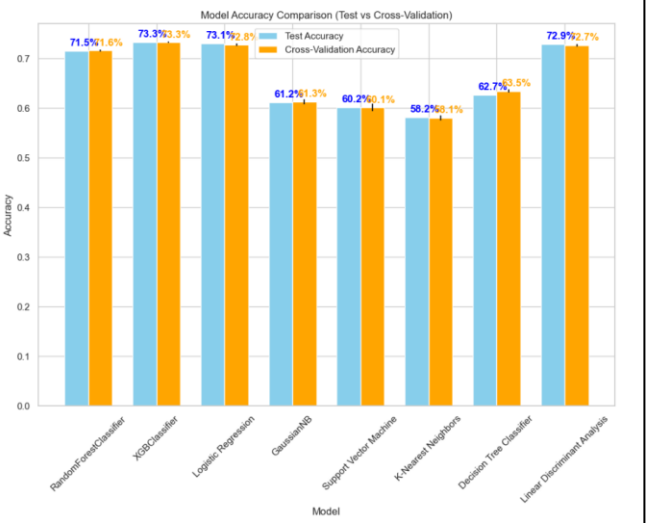
Figure 8



D. Comparative Analysis of Test vs. Cross-Validation Accuracy

Figure 9 shows the test accuracy and cross-validation accuracy for various models, highlighting potential overfitting or underfitting. Logistic Regression, XGBoost, and Linear Discriminant Analysis exhibit minimal differences between test and cross-validation scores. Logistic Regression has scores of 73.3% and 73.1%, XGBoost has 73.3% and 73.3%, and Linear Discriminant Analysis has 72.9% and 72.7%. This indicates consistent performance and minimal overfitting for these models.

Figure 9



VIII. Conclusion

This study explored the use of various machine learning models to predict cardiovascular disease (CVD) using a dataset of 70,000 patient records. Through careful feature engineering, data analysis, and outlier management, the dataset was prepared for predictive modeling.

XGBoost, Logistic Regression, and Linear Discriminant Analysis emerged as the top performers, demonstrating high accuracy, precision, and recall. Their robustness in cross-validation suggests they are reliable for healthcare applications requiring accurate and timely diagnosis. Random Forest also performed well, while Gaussian Naive Bayes and Support Vector Machine showed limitations in precision and recall, raising concerns about false negatives.

Detailed confusion matrices and performance metrics highlighted each model's strengths and weaknesses, emphasizing their practical utility in healthcare. Logistic Regression, XGBoost, and Linear Discriminant Analysis exhibited minimal differences between test and cross-validation scores, indicating consistent performance and minimal overfitting.

Future research could explore hyperparameter tuning for optimization. Additionally, deeper ensemble techniques to combine the strengths of multiple models and more granular patient data could be considered to improve predictive accuracy, especially as computational resources become more accessible. Testing these models in real-world clinical settings will be crucial for assessing their efficacy This study underscores the potential of machine learning in CVD diagnosis and provides a comparative framework for healthcare professionals to select the most suitable models based on clinical criteria.

References

[1] World Health Organization: WHO, "cardiovascular diseases,". https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

[2] Rajkomar, J. M. Dean, and I. S. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/nejmra1814259.

[3] "UCI Machine Learning Repository." Available: https://archive.ics.uci.edu/dataset/45/heart+disease

[4] "Kaggle." Available: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset.

[5] H. A. Miot, "Valores anômalos e dados faltantes em estudos clínicos e experimentais," Jornal Vascular Brasileiro, vol. 18, Jan. 2019, doi: 10.1590/1677-5449.190004.

[6] Z. Kułaga et al., "Polish 2010 growth references for school-aged children and adolescents," European Journal of Pediatrics, vol. 170, no. 5, pp. 599–609, Oct. 2010, doi: 10.1007/s00431-010-1329-x.

[7] D. M. Lloyd-Jones et al., "Heart Disease and Stroke Statistics—2010 update," Circulation, vol. 121, no. 7, Feb. 2010, doi: 10.1161/circulationaha.109.192667.

[8] Djupsjö, J. Kuhl, T. Andersson, M. Lundbäck, M. J. Holzmann, and T. Nyström, "Admission glucose as a prognostic marker for all-cause mortality and cardiovascular disease," Cardiovascular Diabetology, vol. 21, no. 1, Nov. 2022, doi: 10.1186/s12933-022-01699-y.

[9] S. Kusuma and K. R. Jothi, "Cardiovascular Disease Prediction and Comparative Analysis of Varied Classifier Techniques," *2021 2nd Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2021, pp. 1-7, doi: 10.1109/GCAT52182.2021.9587734.

[10] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University. Computer and Information Sciences/Mağalaï Ğam'aï Al-malīk Saud : Ùlm Al-ḥasib Wa Al-ma'lumat*, vol. 34, no. 7, pp. 4514–4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013.

[11] S. S A, "Comparative Study of Naive Bayes, Gaussian Naive Bayes Classifier and Decision Tree Algorithms for Prediction of Heart Diseases," Int J Res Appl Sci Eng Technol, vol. 9, no. 3, pp. 475–486, Mar. 2021, doi: 10.22214/IJRASET.2021.33228.

[12] K. Dissanayake and G. Johar, "Comparative study on heart disease prediction using feature selection techniques on classification algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, pp. 1–17, Nov. 2021, doi: 10.1155/2021/5581806.

[13] R. R. Sarra, A. M. Dinar, M. A. Mohammed, M. K. A. Ghani, and M. A. Albahar, "A robust framework for data generative and heart disease prediction based on efficient deep learning models," *Diagnostics*, vol. 12, no. 12, p. 2899, Nov. 2022, doi: 10.3390/diagnostics12122899.

[14] A. E. Sheikh, N. Mahmoud, and A. Keshk, "Heart disease classification based on hybrid ensemble stacking technique," *IJCI International Journal of Computers and Information*, vol. 8, no. 2, pp. 1–8, Dec. 2021, doi: 10.21608/ijci.2021.207732.