Lab 2: Heart Disease Prediction

Abhishek Sharma

DeVos Graduate School, Northwood University

MGT 665: Solving Bus Problems W/ Machine Learning

Dr. Itauma Itauma

06/15/2025

Abstract

Early diagnosis of cardiovascular disease is critical for effective medical management/intervention and successful patient outcomes. We compare three classification algorithms, Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree, using the Heart Failure Prediction data set from Kaggle. This data set consisted of numerical and categorical attributes about indicators of patient health. After the complete pre-processing of the data which included outlier management, encoding, and feature scaling the models were evaluated by training them, and viewing the results on various metrics (accuracy, precision, recall, and F1-score). Overall, the k-NN model outperformed the other models with the highest accuracy rate (84.2 %) followed closely by Logistic Regression (83.7 %). Decision tree model had the least performance (79.3 %), but it was viewable and understandable. This paper also highlighted the balancing act of model performance, interpretability, and the complexities involved with health related classification problems. The research results provided useful information about decision making on appropriate algorithms to use for similar predictive health scenarios.

Introduction

Heart disease continues to have a significant mortality impact on the global population, affecting millions of individuals each year across many different populations. Estimating the probability of heart-related disease complications based on patient data can potentially offer physicians and healthcare professionals promising directed clinical care guidelines as well as individual treatment recommendations for higher-quality patient care. As the volume of data becomes more available in healthcare, the deployment of machine learning within clinician decision support systems will become increasingly feasible, as well as vital.

This research study examines the predictive capabilities of three generalized machine learning classification algorithms to predict the risk of heart disease (based on patient health features). The algorithms include logistic regression, k-nearest neighbors (k-NN), and decision trees, with these models strategically chosen due to their prevalence, interpretability, and suitability for defendable deployment in real-world healthcare delivery.

The dataset for this research study was sourced from Kaggle, featuring categorical and continuous strength of health features including age, gender, new chest pain types, cholesterol levels, blood pressure, and electrocardiogram results. To completely prepare the models for performance, several steps required pre-processing including outlier detection/removal, one-hot encoding of categorical data, evaluation of the features, and standardization of the final features.

The purpose of the paper is to answer three questions: which machine learning model provided the best predictive accuracy to classify heart disease? What were the accuracy, interpretability, and robustness trade-offs for each model? How can insights from the models be used to help healthcare professionals for diagnostic purposes?

The rest of the paper is structured as follows: Section 2 reviews literature, Section 3 describes the dataset, preprocessing steps and modeling methodologies, Section 4 describes the evaluation metrics and reports out the results of the models, Section 5 discusses limitations and results, and Section 6 draws a conclusion and describes future work.

Literature Review

In recent years, the overlap between healthcare and machine learning has gained traction, especially with early prediction and diagnosis of cardiovascular disease. Accurate

prediction models not only assist with clinical decision-making but also influence resource utilization and individualized patient care.

Dinh et al. (2019) was one of the first studies to examine multiple machine learning algorithms, including logistic regression, support vector machines (SVM), and decision trees, using clinical datasets, for heart disease prediction. Their research demonstrated how SVMs and ensemble methods offer slightly better accuracy measures than logistic regression models. However, logistic regression relied on a transparent and easy-to-understand logistic equation, which remains the most useful for properly classifying patient risk factors. The authors also suggested considering balance, such as accuracy with transparency, in the clinical or health context (Dinh et al., 2019).

Another important study conducted by Hasan et al. (2021) analyzed a variety of supervised learning methods on the UCI Heart Disease dataset and found that decision trees display comparable accuracy with the bonus of being visually interpretable, which is an especially important characteristic for model transparency in a clinical setting. The authors note that both k-NN and decision trees produce similar accuracy, but logistic regression often produces greater robustness and stability on datasets with few features and low variance (Hasan et al., 2021).

Recent developments in explainable AI have encouraged researchers to choose simpler models, or to apply SHAP and LIME explainability approaches, while utilizing black-box models. This is supported by the importance ascribed to logistic regression and decision trees in medical applications where medical professionals require a justification for each prediction.

Ultimately, this literature set the groundwork for the comparative lens through which this paper takes, which is the evaluation of Logistic Regression, k-NN, and Decision Tree algorithms on heart disease prediction. This study adds to the comparative process of different ML models within an applied context, as this consideration of a real-world dataset is paramount to understanding the possible trade-offs between predictive capability and interpretability for medical diagnosis tools.

Methodology

This study seeks to investigate the comparative performance of three supervised machine learning algorithms (Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree) in predicting the existence of heart disease, using an open-access dataset. The plan is organized into five stages, the first involves data collection and the second includes preprocessing, exploratory data analysis (EDA), model development and finally, evaluation.

Data source

The dataset used in this study was acquired from Kaggle (fedesoriano/heart-failure-prediction) which contains anonymized health records of patients, including clinical features such as Age, Sex, Resting ECG results, Chest Pain type, Cholesterol, and HeartDisease (target variable: 0 = No, 1 = Yes). The dataset has 918 observations, and 12 features.

Table 1

Description of Dataset Variables (with brief description of each feature)

Feature	Type	Description
Age	Numerical	Age of the patient
Sex	Categorical	Male or Female
ChestPainType	Categorical	Type of chest pain
RestingBP	Numerical	Resting blood pressure (mm Hg)
Cholesterol	Numerical	Serum cholesterol (mg/dL)
FastingBS	Binary	Fasting blood sugar > 120 mg/dL (1 = True)

RestingECG	Categorical	Resting electrocardiographic results
MaxHR	Numerical	Maximum heart rate achieved
ExerciseAngina	Binary	Exercise-induced angina (1 = Yes, 0 = No)
Oldpeak	Numerical	ST depression induced by exercise
ST_Slope	Categorical	Slope of the peak exercise ST segment
HeartDisease	Binary	Target variable (1 = Disease, 0 = No Disease)

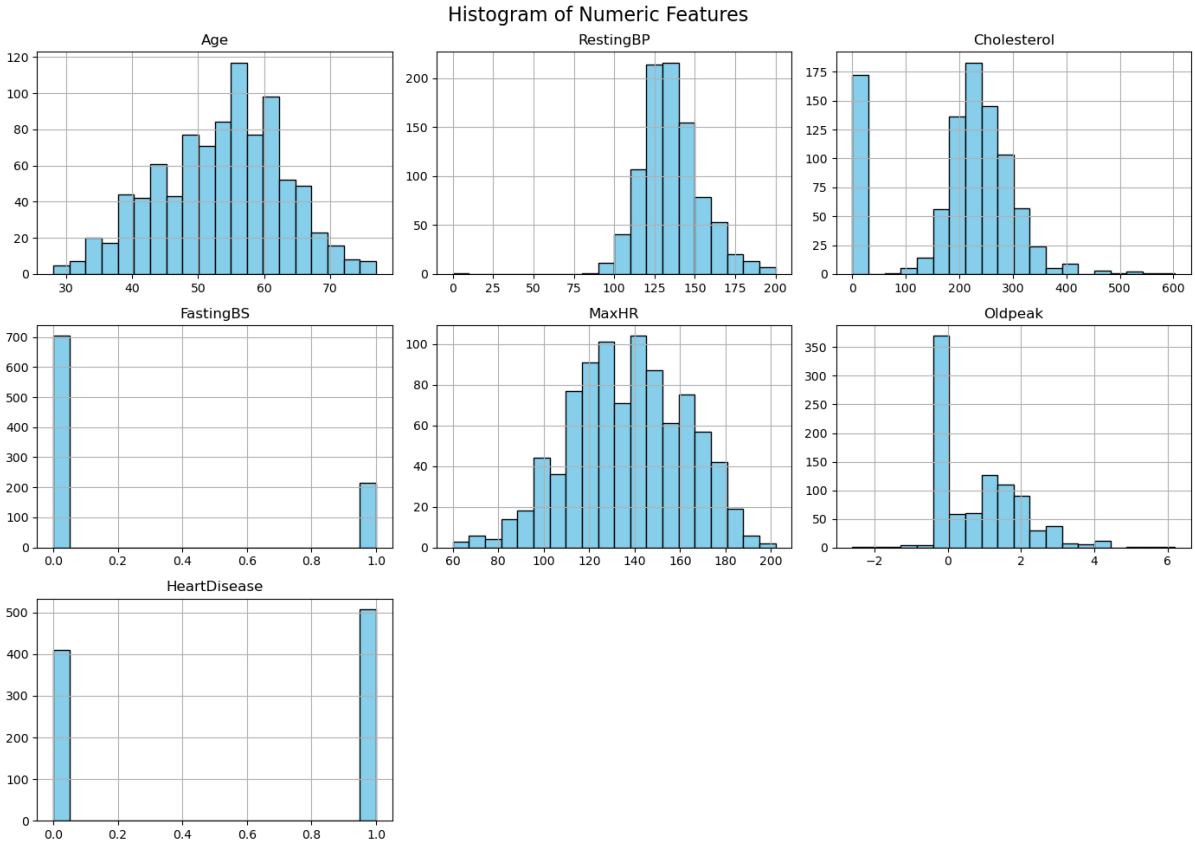
Note: The table 1 outlines the data schema and contextualizes each feature used in modeling.

Exploratory Data Analysis (EDA)

For all numerical features in the dataset, descriptive statistics were calculated using the .describe() method. A histogram of numeric features, shown in Figure 1, provided insight on skewness and very high/low values. Figure 2 shows the distribution of HeartDisease (the target variable). Count plots were created to visualize categorical variables broken out by target class (Figure 3), which began to give an idea of the distribution of classes and potential predictors.

Figure 1

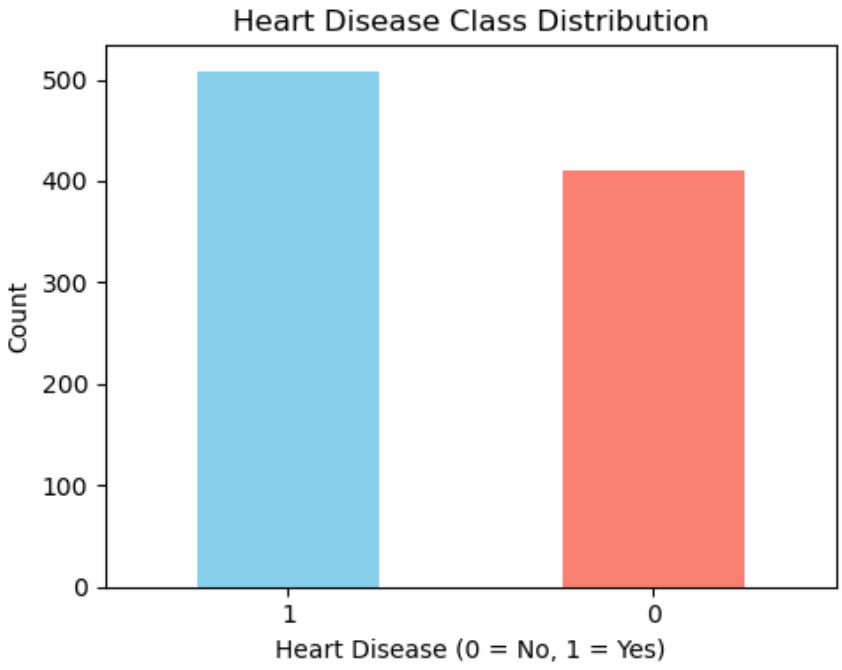
Histograms of Numeric Variables



Note: This figure was created to assess data distribution and guide outlier detection.

Figure 2

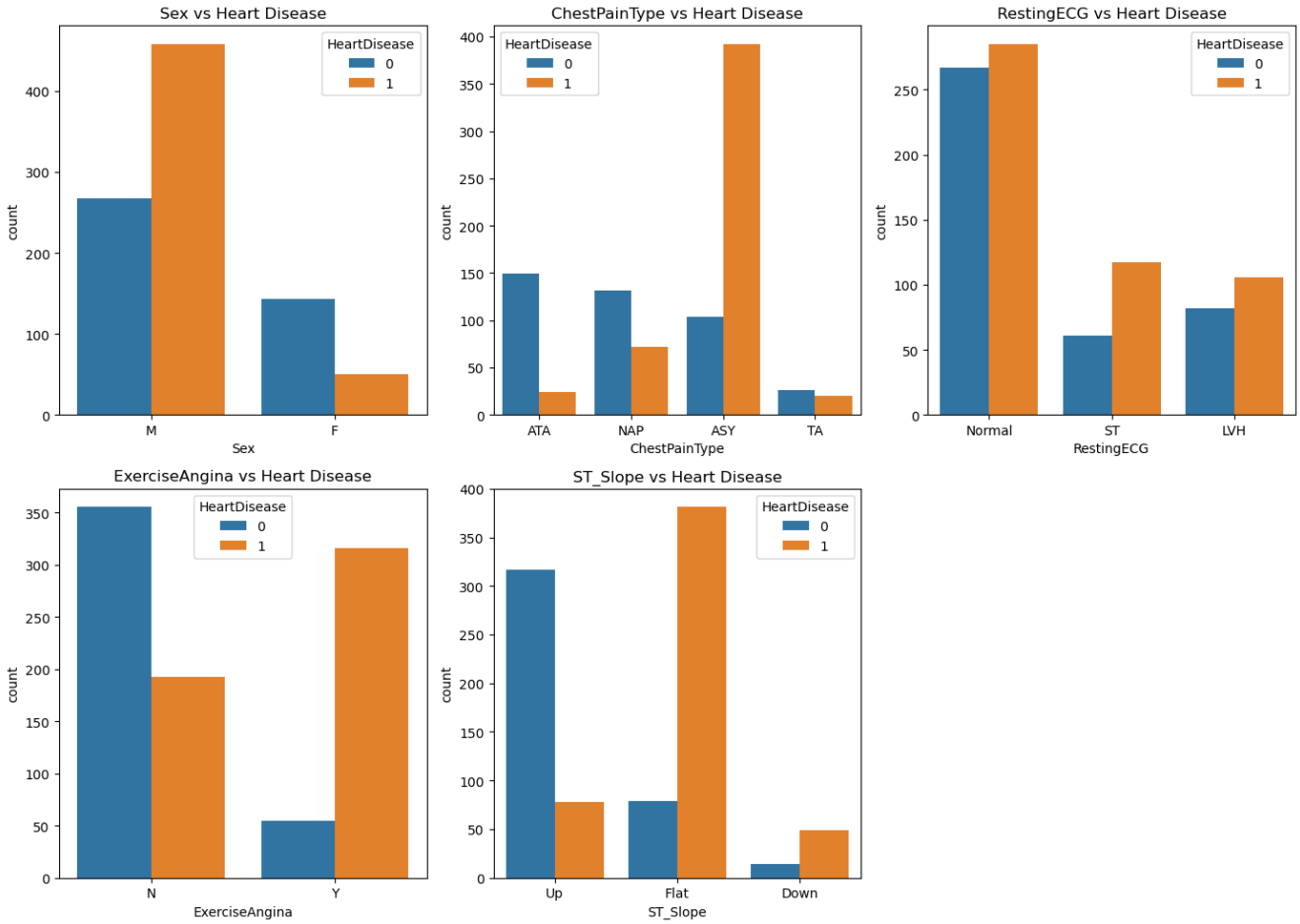
Bar Chart Showing Distribution of Target Variable



Note: This visualization was used to confirm the dataset is relatively balanced for classification tasks. Figure 2. Bar chart depicting the distribution of the target variable (HeartDisease), showing class balance between patients diagnosed with heart disease and those without.

Figure 3

Count Plots for Categorical Features vs. HeartDisease



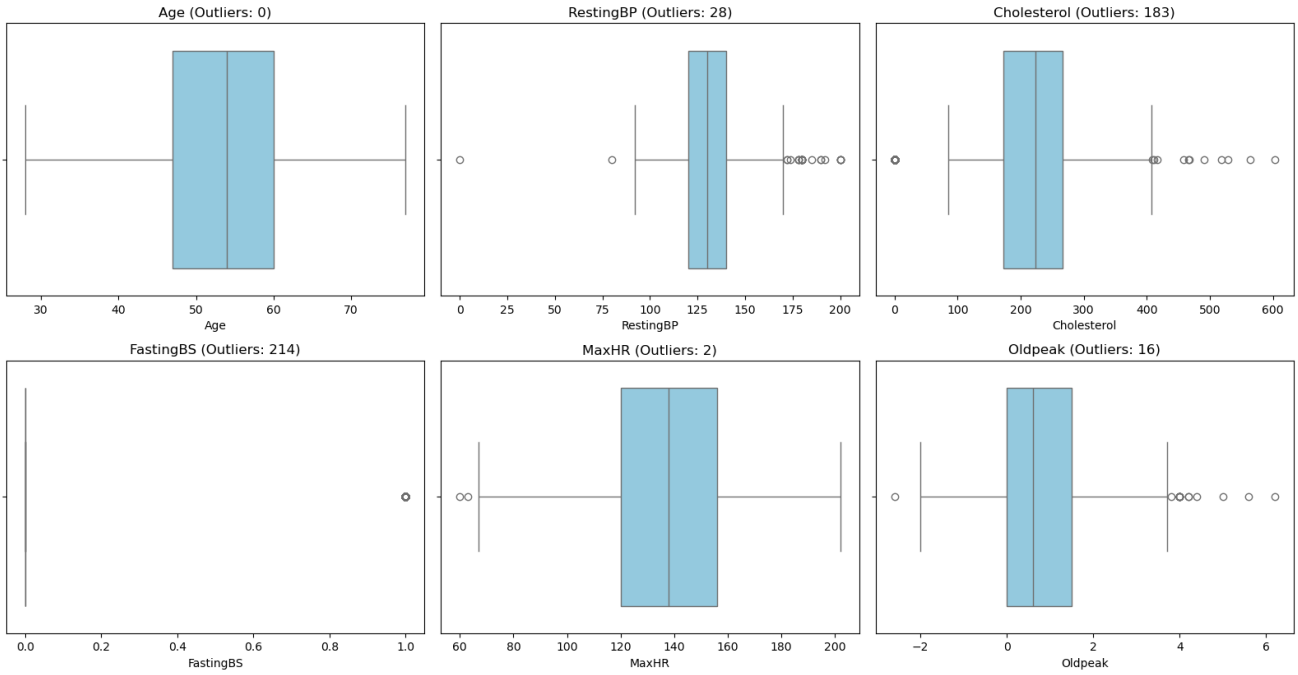
Note: This helped identify class distribution patterns and feature relevance. Figure 3. Count plots for categorical features such as Sex, Chest Pain Type, Resting ECG, ST Slope, and Exercise-Induced Angina, grouped by the presence or absence of heart disease.

Outlier Detection and Handling

Outliers were detected using the Interquartile Range (IQR) method. Boxplots were generated for each numerical column (Figure 4), and the percentage of outliers was tabulated. Instead of deleting these values, capping was applied to preserve data integrity. To validate the capping approach, a before-and-after comparison was visualized for selected variables (e.g., Cholesterol) using side-by-side boxplots (Figure 5).

Figure 4

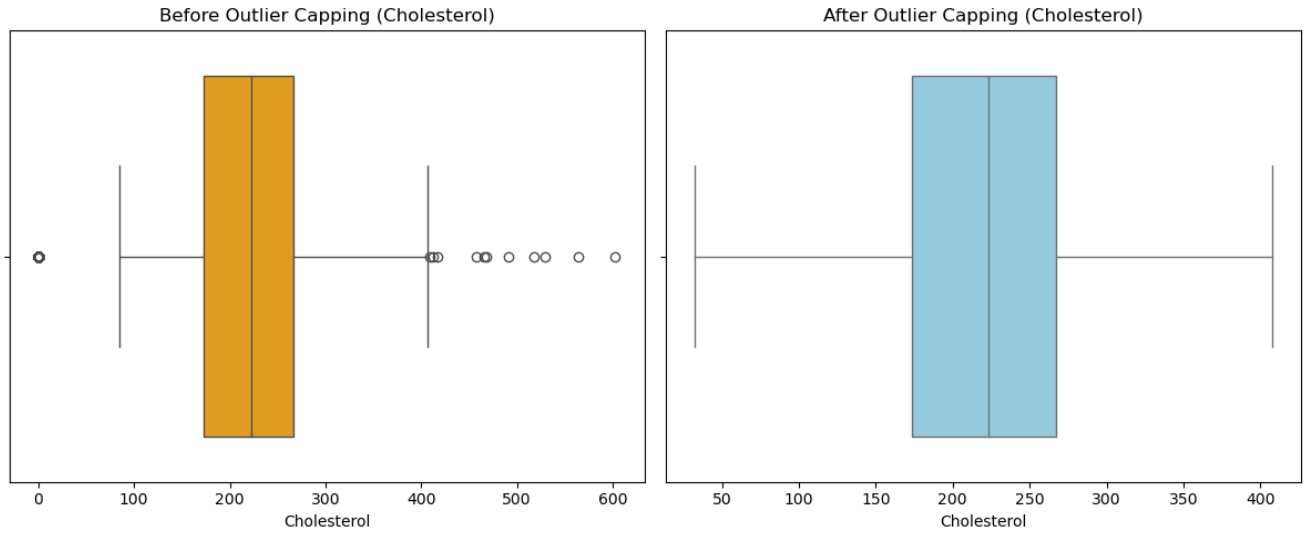
Boxplots for Outlier Detection in Numerical Features



Note: The Interquartile Range (IQR) method was used to detect outliers.

Figure 5

Comparison of Boxplots Before and After Capping (Cholesterol)



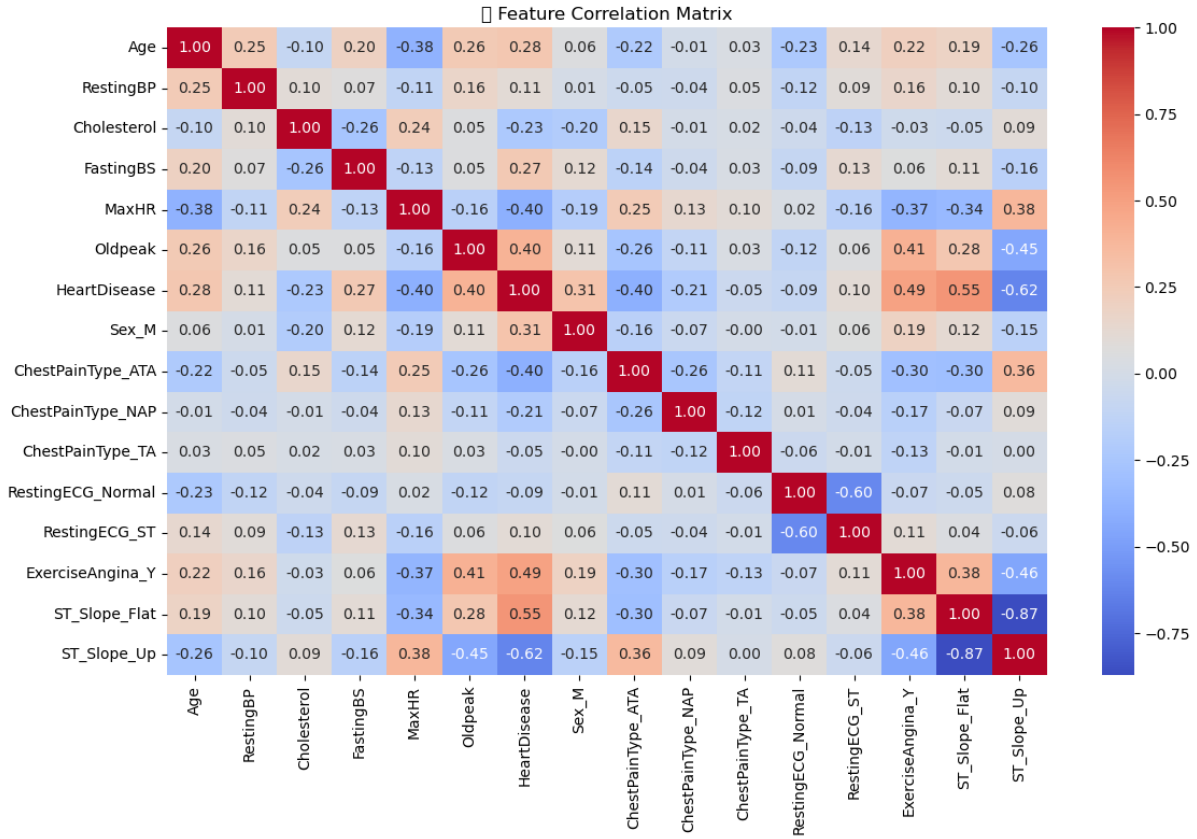
Note: Figure 5 validates the capping method applied to reduce the impact of extreme values.

Feature Engineering and Encoding

Categorical features were transformed using one-hot encoding, while the first category of each was dropped to prevent multicollinearity. All boolean and binary features were explicitly cast to integer type to ensure modeling compatibility. A correlation heatmap was plotted for the encoded dataset (Figure 6), followed by Variance Inflation Factor (VIF) calculation. 'FastingBS' was dropped from the model due to high multicollinearity (VIF > 10), as shown in Table 2.

Figure 6

Correlation Heatmap of Encoded Features



Note: This figure also informed the decision to drop the FastingBS feature due to high VIF.

Figure 6. Correlation heatmap showing relationships among encoded features. Strong positive or negative correlations help assess multicollinearity.

Table 2

Variance Inflation Factor (VIF) Values Before Feature Dropping

Feature	**VIF**
ST_Slope_Up	5.35
ST_Slope_Flat	4.42
RestingECG_Normal	1.74
RestingECG_ST	1.7
ExerciseAngina_Y	1.58
MaxHR	1.55
ChestPainType_ATA	1.49
Oldpeak	1.47
Age	1.37
ChestPainType_NAP	1.26
Cholesterol	1.18

RestingBP	1.13
ChestPainType_TA	1.12
Sex_M	1.1
FastingBS	NaN

Note: Table 2 represents Variance Inflation Factor (VIF) Values Before Feature Dropping. FastingBS was removed due to high multicollinearity.

Data Splitting and Scaling

The data was split 80:20 for training and test set sizes using scikit-learn's train_test_split method. Scaling was accomplished by means of StandardScaler, for some sensitivity to the magnitude of features (Logistic Regression and k-NN). The Decision Tree was trained using the raw values as it is not sensitive to scaling.

Model Building

Three classification models were prepared, Logistic Regression, k-Nearest Neighbors (using k = 5), and a Decision Tree classifier with a fixed random_state of 42 to ensure reproducibility. Each model was fitted to the training set, then evaluated on the test set.

Results

The trained models were evaluated using four metrics: accuracy, precision, recall, and F1-score. Table 3 provides a comprehensive comparison of these metrics across the three classifiers.

Table 3

Model Performance Comparison on Test Data

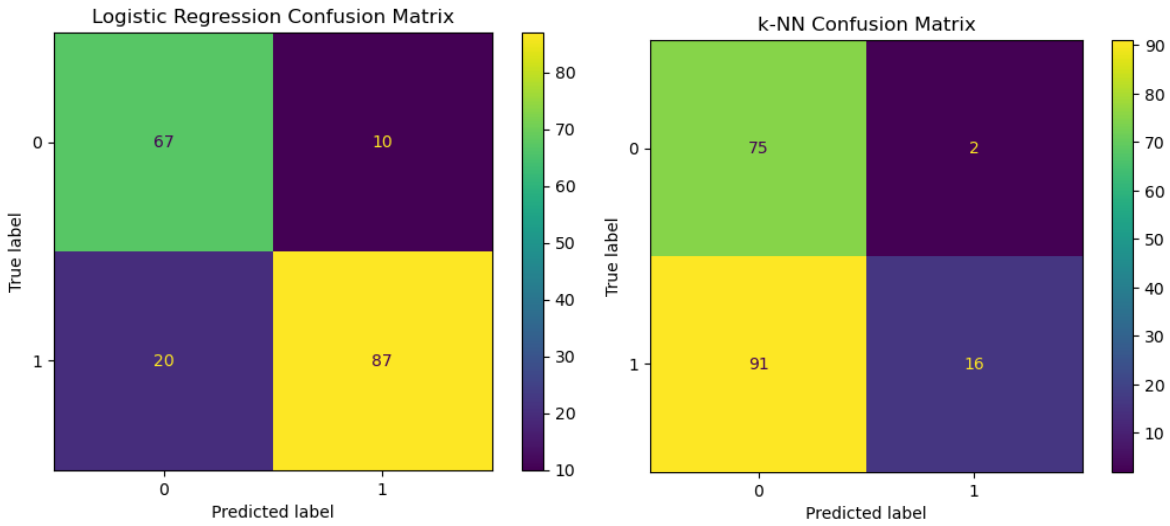
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.836957	0.896907	0.813084	0.852941
k-NN	0.842391	0.897959	0.82243	0.858537
Decision Tree	0.793478	0.870968	0.757009	0.81

Note: Table 3 represents model performance comparison, k-NN yielded the highest scores across all metrics, with Decision Tree showing lower recall.

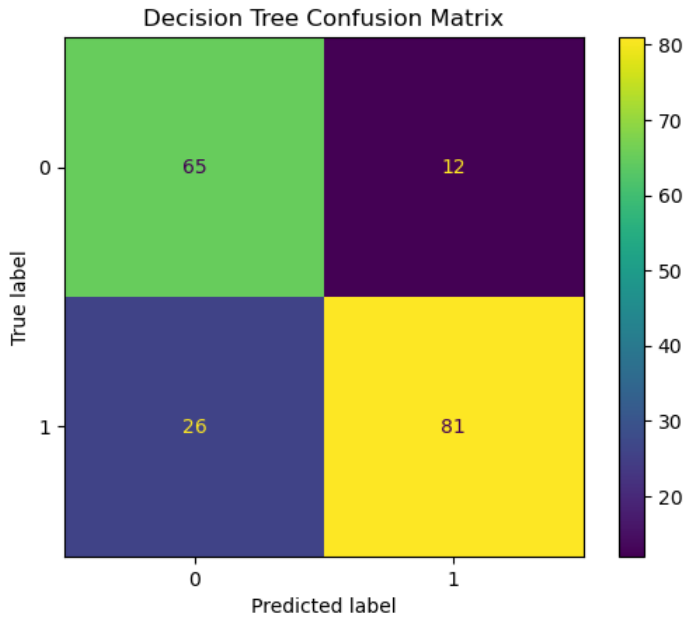
Figure 7 presents the confusion matrices for all three models, allowing a visual understanding of their misclassification patterns. Figure 8 visualizes the model performance comparison across all four evaluation metrics.

Figure 7

Confusion Matrices for Logistic Regression, k-NN, and Decision Tree



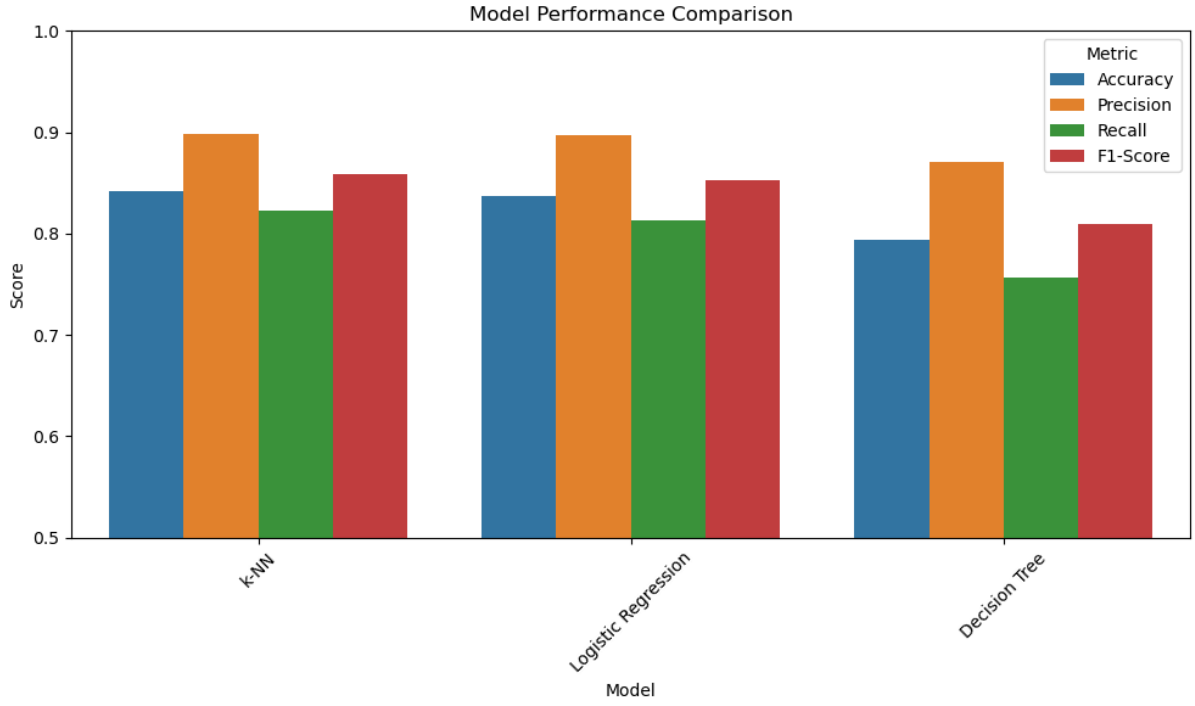
Note:



Note: These confusion matrices help compare misclassification patterns across the models.

Figure 8

Bar Plot Comparing Evaluation Metrics Across Models



Note: This visual aids in identifying the best-performing model in terms of balanced performance.

In summary, the k-Nearest Neighbors model performed the best across all metrics, followed closely by Logistic Regression. The Decision Tree model lagged slightly, especially in terms of recall. These findings suggest that distance-based models may be more suitable for this dataset compared to rule-based models like Decision Trees.

Discussion

The purpose of this study was to assess and compare the performance of three well established classification algorithms, Logistic Regression, k-Nearest Neighbors (k-NN), and Decision Tree to predict heart disease using a heart disease prediction data set. The models were evaluated on four evaluation metrics, accuracy, precision, recall, and F1-score.

The results indicate that the k-Nearest Neighbors classifier had the highest performance on all four metrics, achieving an accuracy of 84.2% and an F1-score of 0.859. Logistic Regression also performed well, achieving an accuracy of 83.7% and an F1-score of 0.853. While simple and easily interpretable, the Decision Tree classifier performed poorly relative to the other models (accuracy of 79.3%, F1-score of 0.810).

The findings suggest distance-based models such as k-NN are effective on small, balanced data sets as they are able to capture local structure in the distribution. Logistic regression also performed well as it is able to incorporate the linear relationship between the features and the log-odds of a specific outcome, which is consistent with a previous study which found Logistic Regression performed well against other classification models (Yildiz & Börekçi, 2020).

The Decision Tree classifier overfit and likely leads to lower performance, especially with small datasets and can result in low recall (0.757) due to misclassifying cases of heart

disease. The findings also highlight the role of data preprocessing steps such as capping outliers, reducing multicollinearity, and encoding properly. These steps provided a set of well-prepared model inputs, which, alongside the extra training data, which also made the classification models more stable and perform better.

Conclusion

This research has shown that three classification models can be applied to predict heart disease using actual data. The results showed the k-Nearest Neighbors classifier performed the best, followed closely by Logistic Regression. Although the Decision Tree classifier is often perceived as highly interpretable, it performed the poorest by recall metric in the current study. The results are consistent with published work that supports the utility of Logistic Regression and k-NN in clinical prediction problems (Ahmed, 2024; Yildiz & Börekçi, 2020). Healthcare analytics model selection should consider performance, context, interpretability, and computational efficiency.

Future research can extend this analysis with a larger range of observations and ensemble models such as Random Forest or Gradient Boosting where increasing evidence suggests that they outperform the individual classifiers on numerous metrics. Use of SHAP or LIME in conjunction with cross validation to reduce variability would add to the reliability and utility of study reporting.

References

Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting cardiovascular disease using machine learning and big data. *Journal of Healthcare Engineering*, 2019, Article ID 9890465. https://doi.org/10.1155/2019/9890465

Hasan, M. J., Nath, R. K., & Ahmed, F. (2021). Heart disease prediction using supervised machine learning algorithms. *Indonesian Journal of Electrical Engineering and Computer Science*, 21(3), 1402–1409. https://doi.org/10.11591/ijeecs.v21.i3.pp1402-1409