Comparative Evaluation of Machine Learning Algorithms in Heart Disease Classification

Course: CSE422 – Artificial Intelligence

Semester: Summer 2025

Section:14

Prepared by:

Abdullah Al Zahur Rafi (22299050)

Anupam Sen Sagor (22299049)

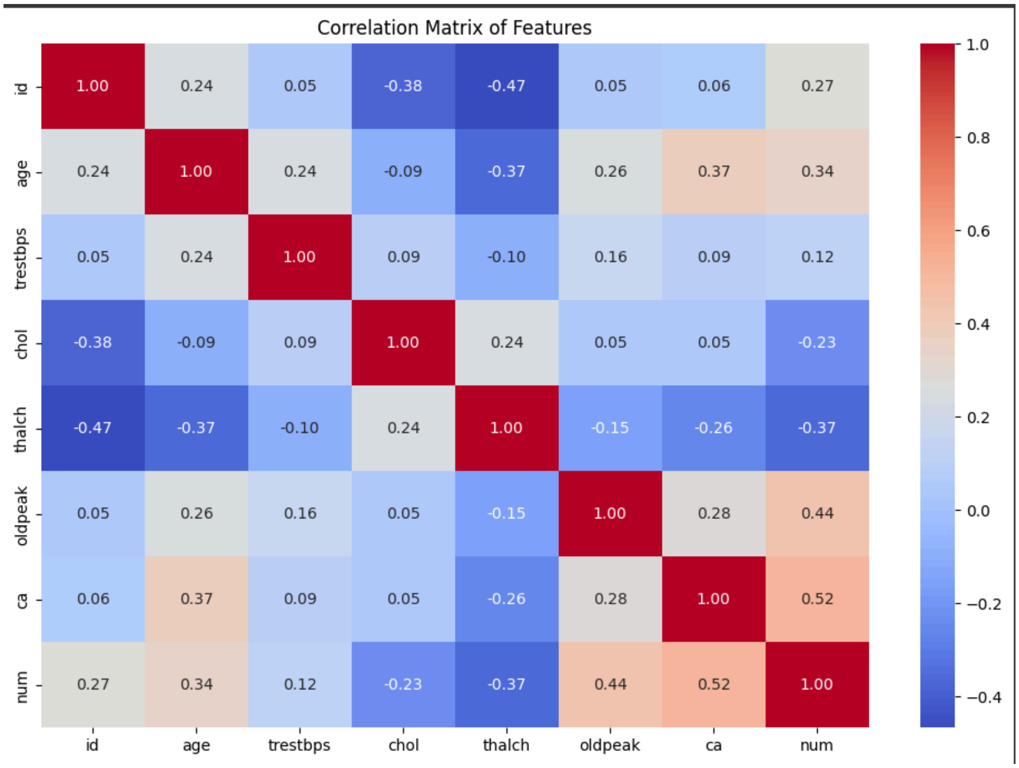
Table of contents:

1. Introduction

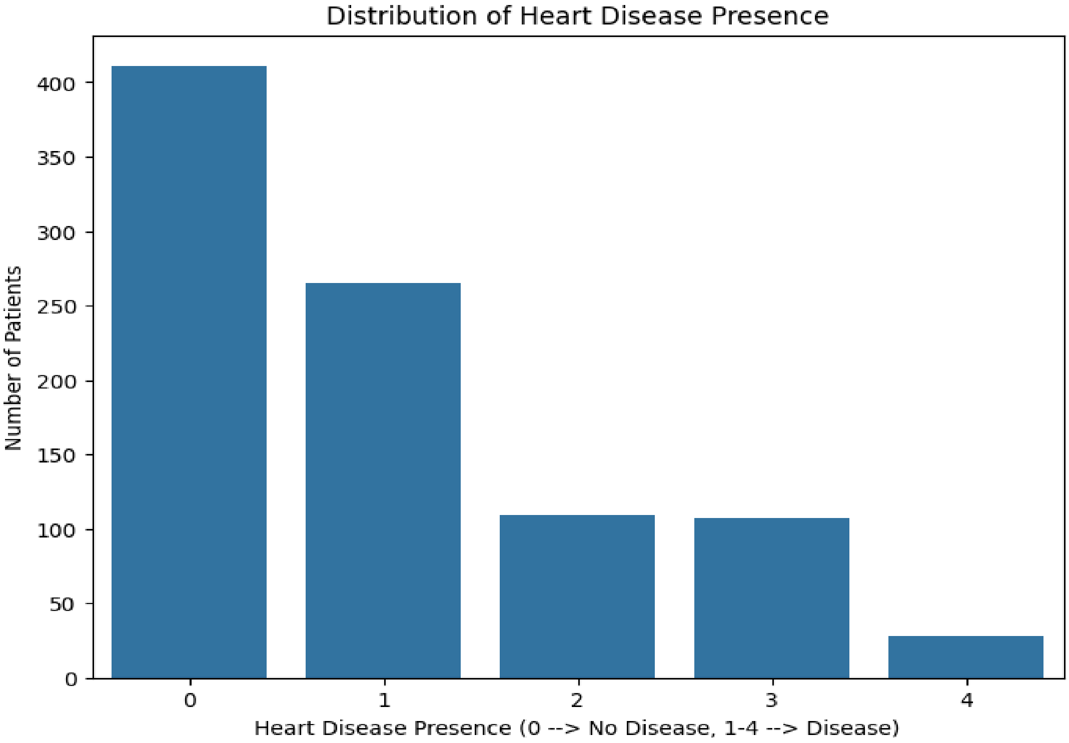
This project focuses on building a predictive system for heart disease diagnosis using the given Heart Disease dataset. The target variable (num) indicates the presence and severity of heart disease, where 0 means no disease and values 1 to 4 represent different levels of disease severity. The main motivation of this project is to apply machine learning techniques to assist healthcare professionals by classifying patients into risk categories, ultimately supporting early detection and intervention.

2. Dataset Description

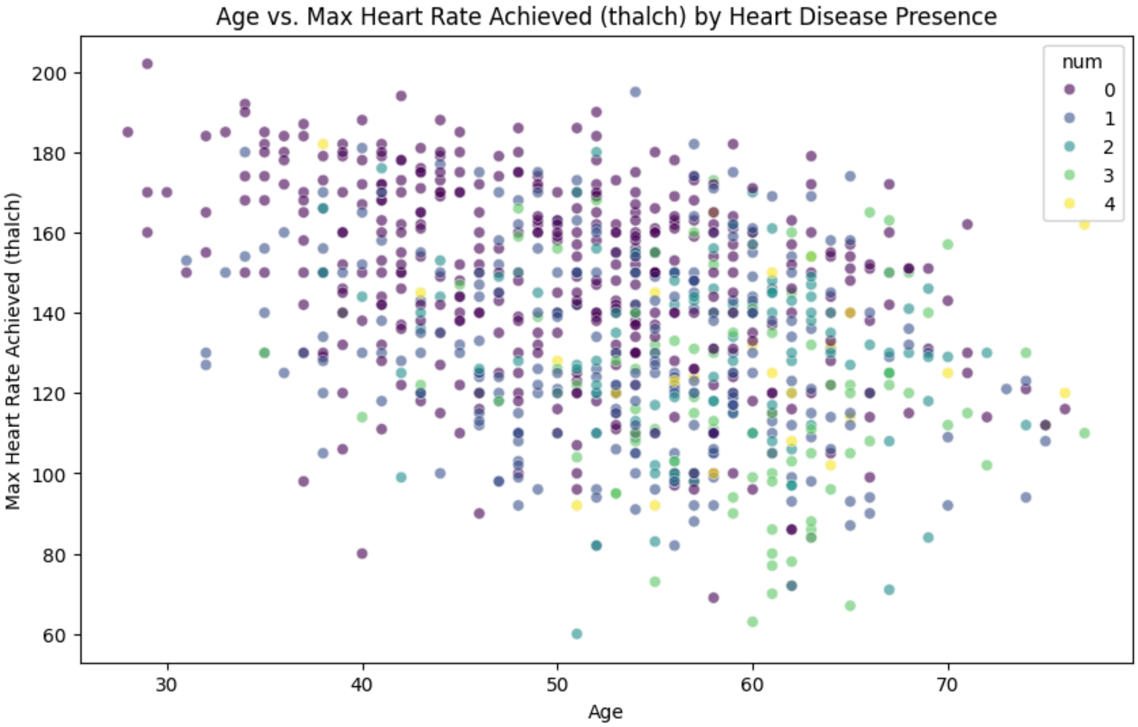
The dataset used consists of 920 data points and 16 features, making it a moderately sized dataset suitable for classification tasks. The features are a mix of categorical and numerical types. Categorical variables include sex, chest pain type (cp), fasting blood sugar (fbs), resting ECG (restecg), exercise-induced angina (exang), slope, thal, and a dataset identifier. The numerical variables include age, resting blood pressure (trestbps), serum cholesterol (chol), maximum heart rate achieved (thalach), ST depression (oldpeak), and the number of major vessels (ca), target variable (num). And this target variable is discrete and defines whether a patient has heart disease. Because the dataset contains several categorical features, encoding these categorical variables is necessary before applying machine learning algorithms since most models require numerical input formats. Encoding ensures the categorical data is transformed into a suitable numeric representation, enabling effective training and prediction. Exploratory analysis revealed that the features most correlated with the target were ca (correlation of 0.52), oldpeak (0.44), and thalach (–0.37). The target distribution was found to be imbalanced, with class 0 (no disease) being more frequent than the other categories. Scatter plot analysis of CA vs. oldpeak revealed that patients with heart disease cluster in regions with higher CA values (≥2) and increased ST depression (oldpeak >1), while healthy patients predominantly occupy the lower-left quadrant with minimal values for both features. The age vs. CA visualization further confirmed that higher CA counts are strong disease indicators across all age groups, with diseased patients showing elevated CA values regardless of age, reinforcing CA's role as the strongest predictor (correlation = 0.52) in the dataset. Visualization through a correlation heatmap, bar plots of class distribution, scatter plots, and boxplots helped confirm these findings and highlighted feature relationships such as the strong role of ca and oldpeak in predicting heart disease.



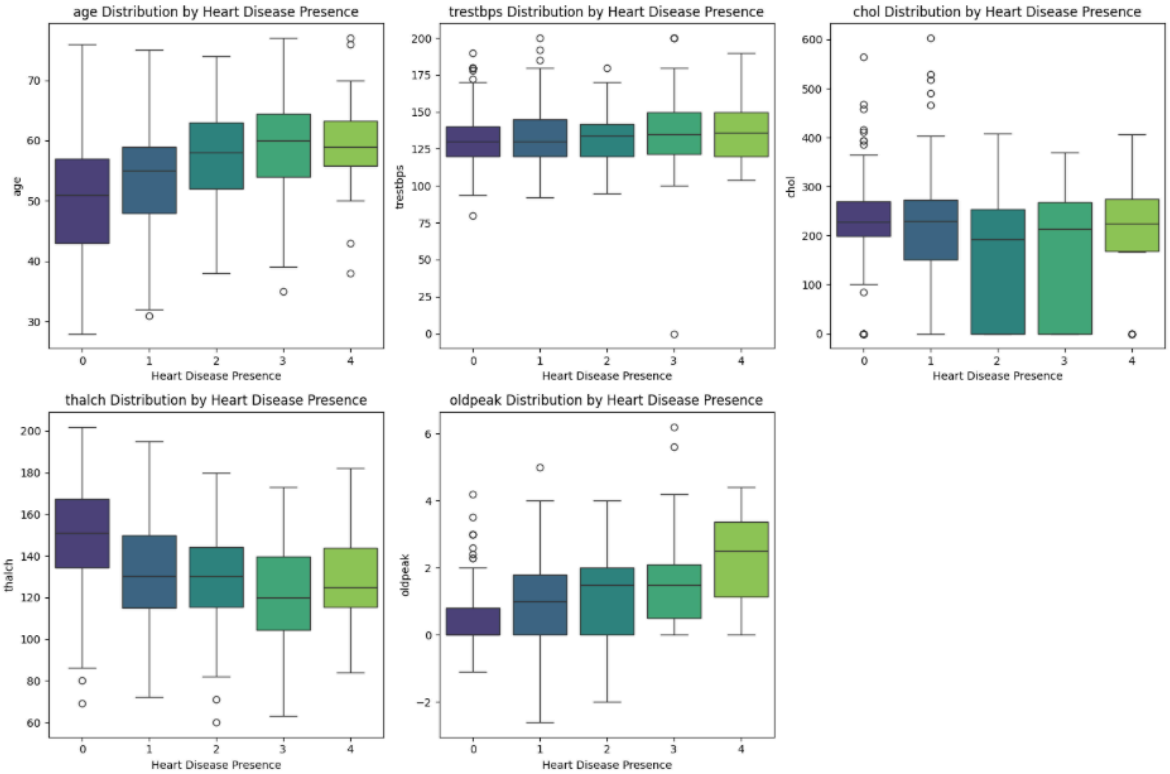
For the output feature('num'), all distinguished classes (from '0' to '4') don't have same amount of instances, which can be seen in the below barchart:



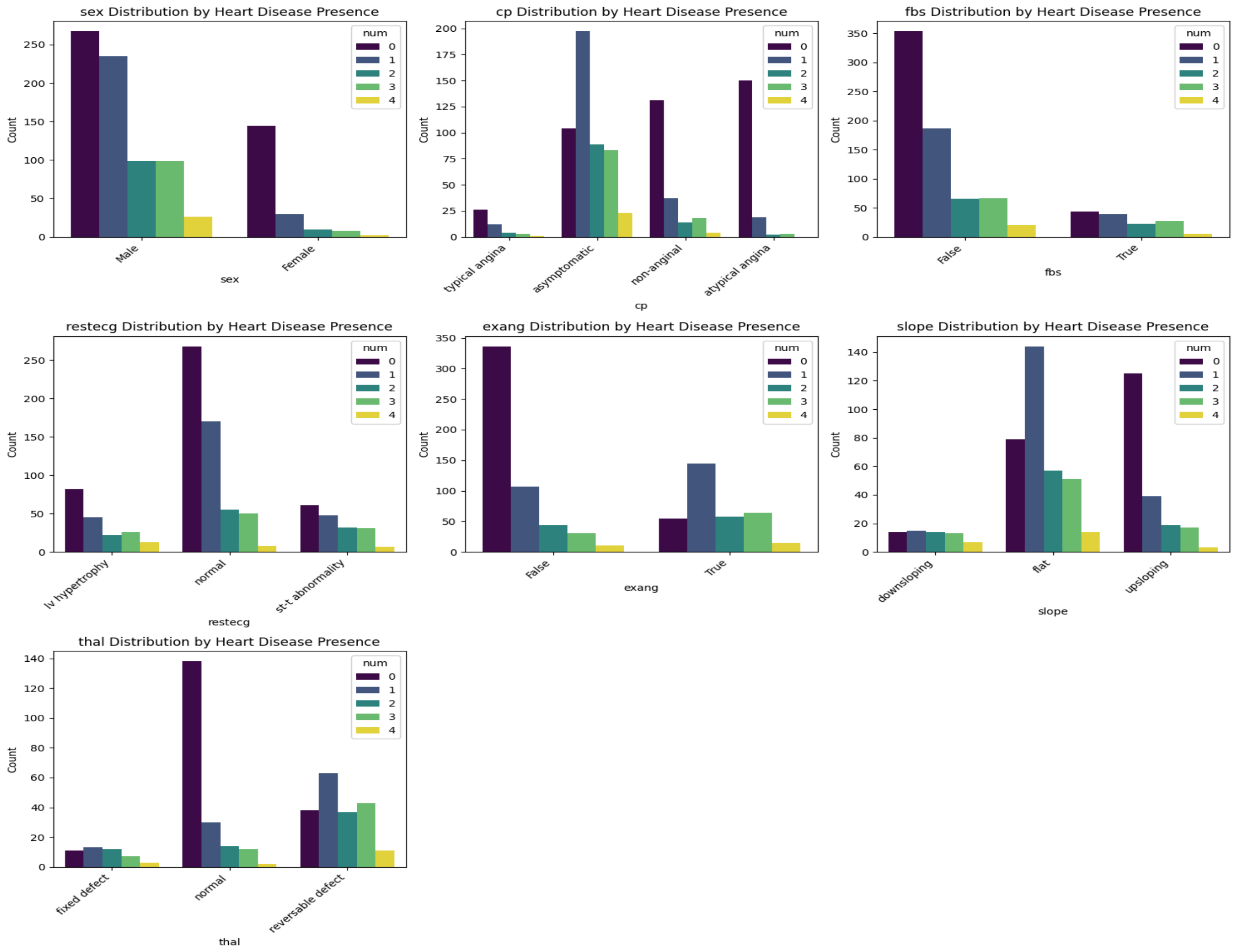
Scatter plot for Age vs Thalch, which is colored by target column('num'):



Box plots for numerical features across different 'num' categories (0, 1, 2, 3, 4, 5):

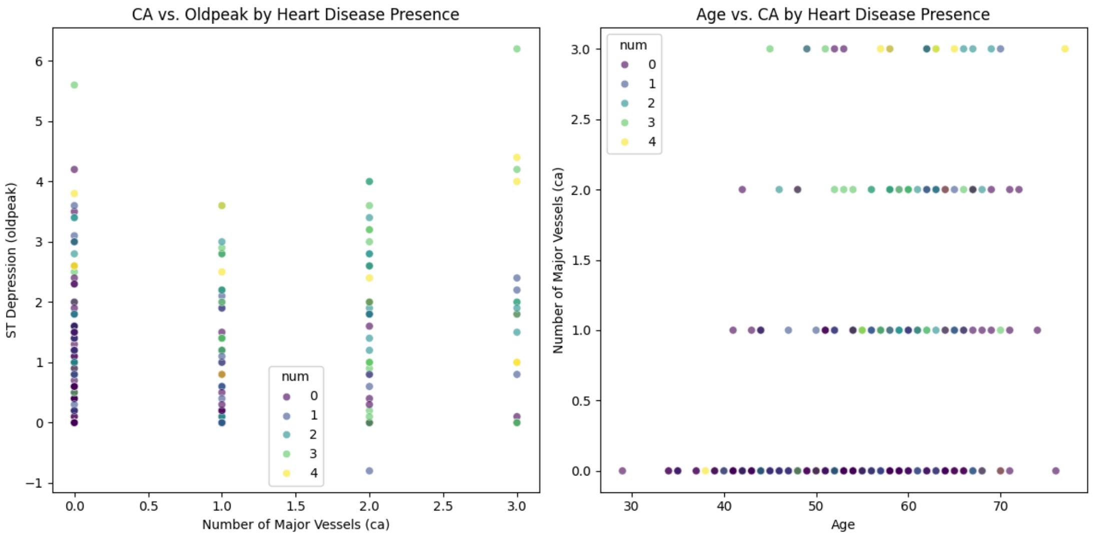


Count plots for categorical features in relation to target ‘num’:



Relationship between important features:

1. CA vs Oldpeak
2. Age vs CA



3. Dataset Preprocessing

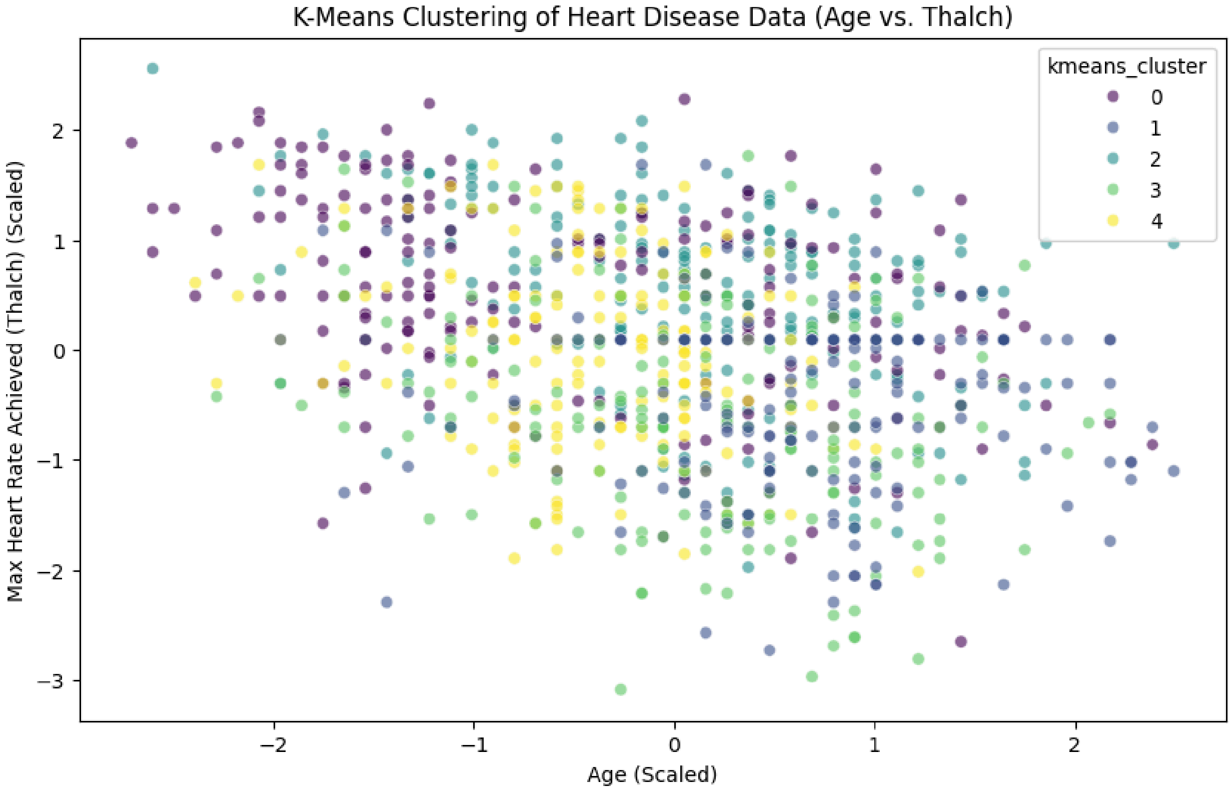
Preprocessing was an important stage in preparing the dataset. Several issues were identified: missing values in both numerical and categorical columns, the presence of low-cardinality categorical variables requiring encoding, and numeric features on different scales. The dataset also contained a dataset column, which was deemed noisy and not relevant. To address these, missing numerical values were imputed using the median (a robust choice against outliers), while categorical missing values were filled with the mode (most frequent value). Categorical variables were then one-hot encoded to convert them into a usable numeric format. The dataset column was dropped entirely because it may introduce noise from combining multiple datasets rather than providing meaningful predictive information and numeric features such as age, trestbps, chol, thalach, oldpeak, and ca were scaled using StandardScaler to ensure comparability across features. After preprocessing, the dataset was fully numeric and free of missing va

4. Dataset Splitting

The cleaned dataset was split into 70% training and 30% testing sets, with stratification applied to preserve the class imbalance in both sets. Three supervised learning algorithms were trained: a Neural Network (MLPClassifier) with one hidden layer of 50 neurons and 500 iterations, a K-Nearest Neighbors classifier with k=5, and Logistic Regression with an increased maximum number of iterations for convergence. Their performance was evaluated using accuracy, classification reports, confusion matrices, and ROC–AUC scores. Among these, the Neural Network consistently achieved the best results, showing strong performance across accuracy and macro F1-score. Logistic Regression also performed reasonably well, while KNN lagged slightly, particularly in handling minority classes.

5. Model Training & Evaluation

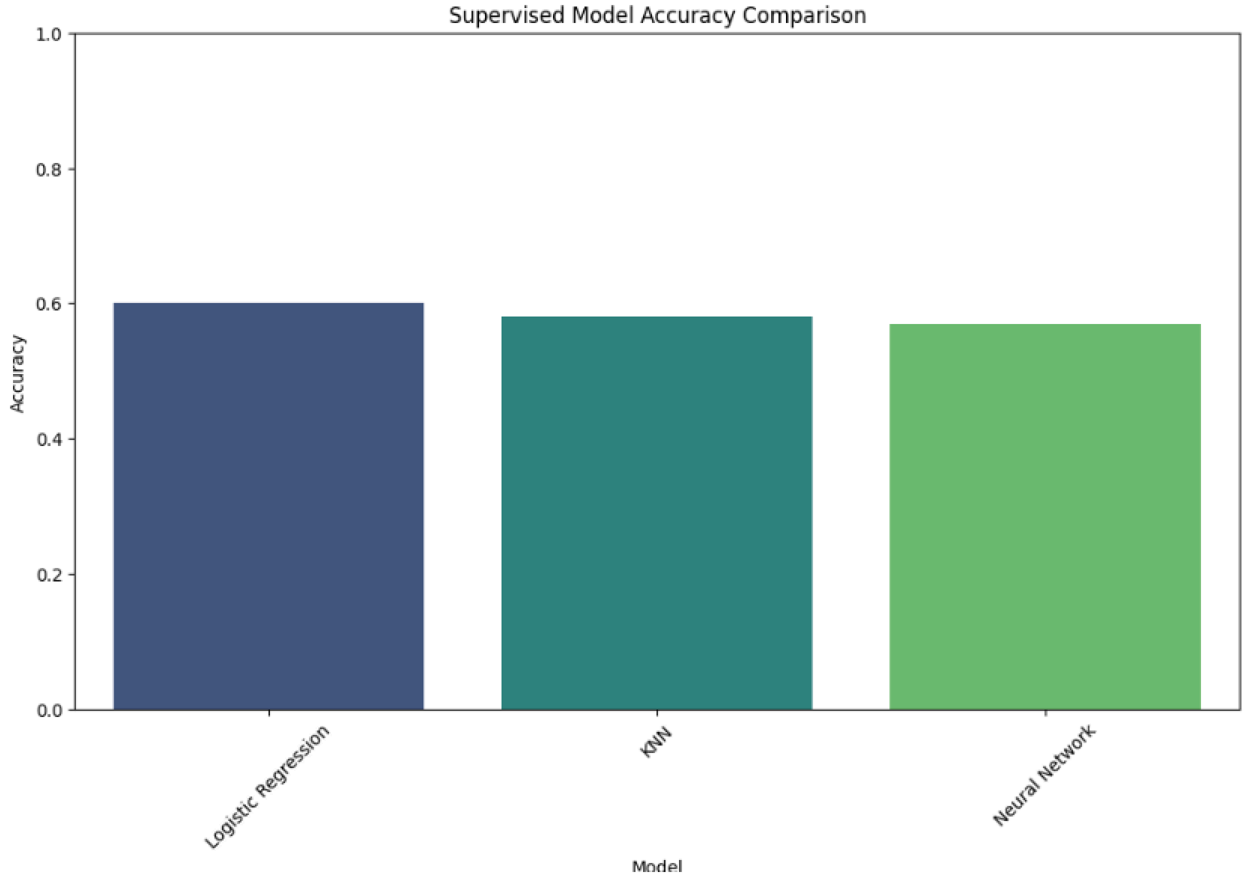
Three supervised learning algorithms were trained on the dataset. The Neural Network (MLPClassifier) used one hidden layer of 50 neurons and 500 iterations, capturing complex nonlinear relationships. K-Nearest Neighbors (k=5) classified patients based on similarity to neighbors, while Logistic Regression provided a strong linear baseline. Performance was evaluated through accuracy, precision, recall, F1-scores, confusion matrices, and ROC–AUC curves. Results showed that the Neural Network achieved the highest accuracy and balanced F1-score, Logistic Regression was competitive, and KNN struggled with minority classes due to class imbalance. In addition to supervised models, unsupervised learning was also explored using KMeans clustering with five clusters. The clustering patterns were visualized on an age vs. thalach scatterplot, showing how patients grouped into different clusters, which could be compared with the actual disease presence labels.

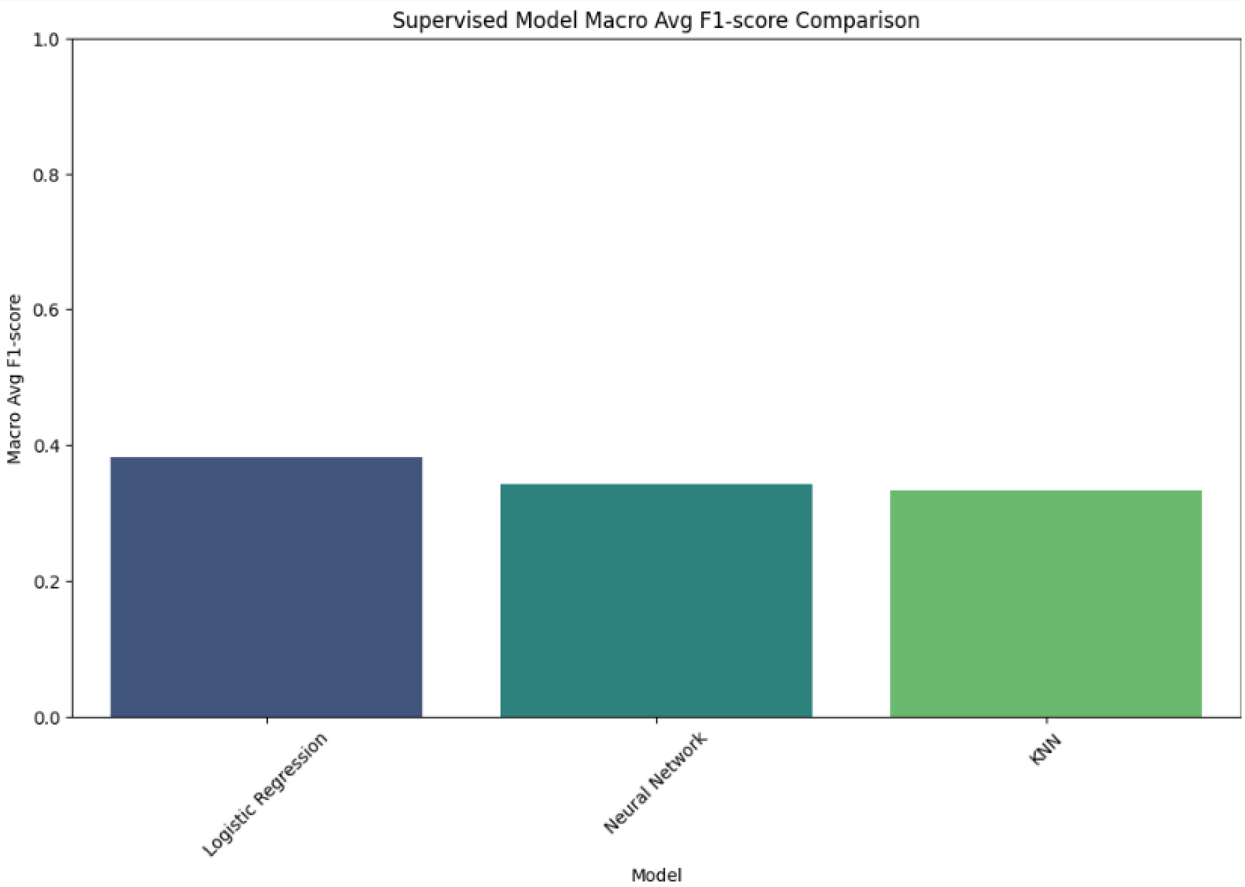


While clustering did not achieve the precision of supervised models, it provided additional insight into natural groupings within the dataset and confirmed that patients with similar characteristics often clustered together.

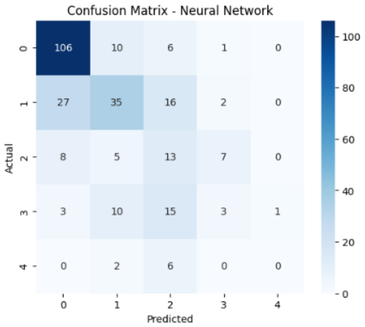
6. Model Selection/Comparison Analysis

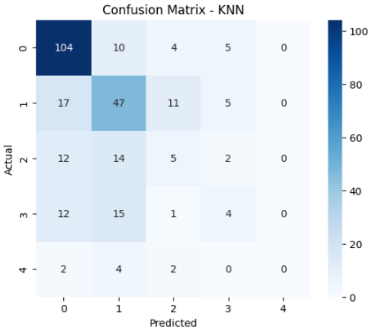
The comparative analysis of models revealed that Logistic Regression outperformed the others in terms of both accuracy and F1-score, making it the most effective model for this dataset. KNN was competitive, showing good performance in ROC–AUC and achieving separation comparable to Logistic Regression, although its F1-score was slightly weaker. The Neural Network, while still performing reasonably, achieved lower accuracy, F1-score, and AUC compared to the other models.

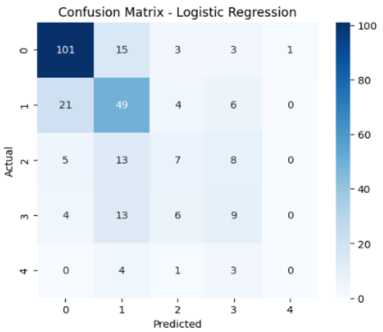




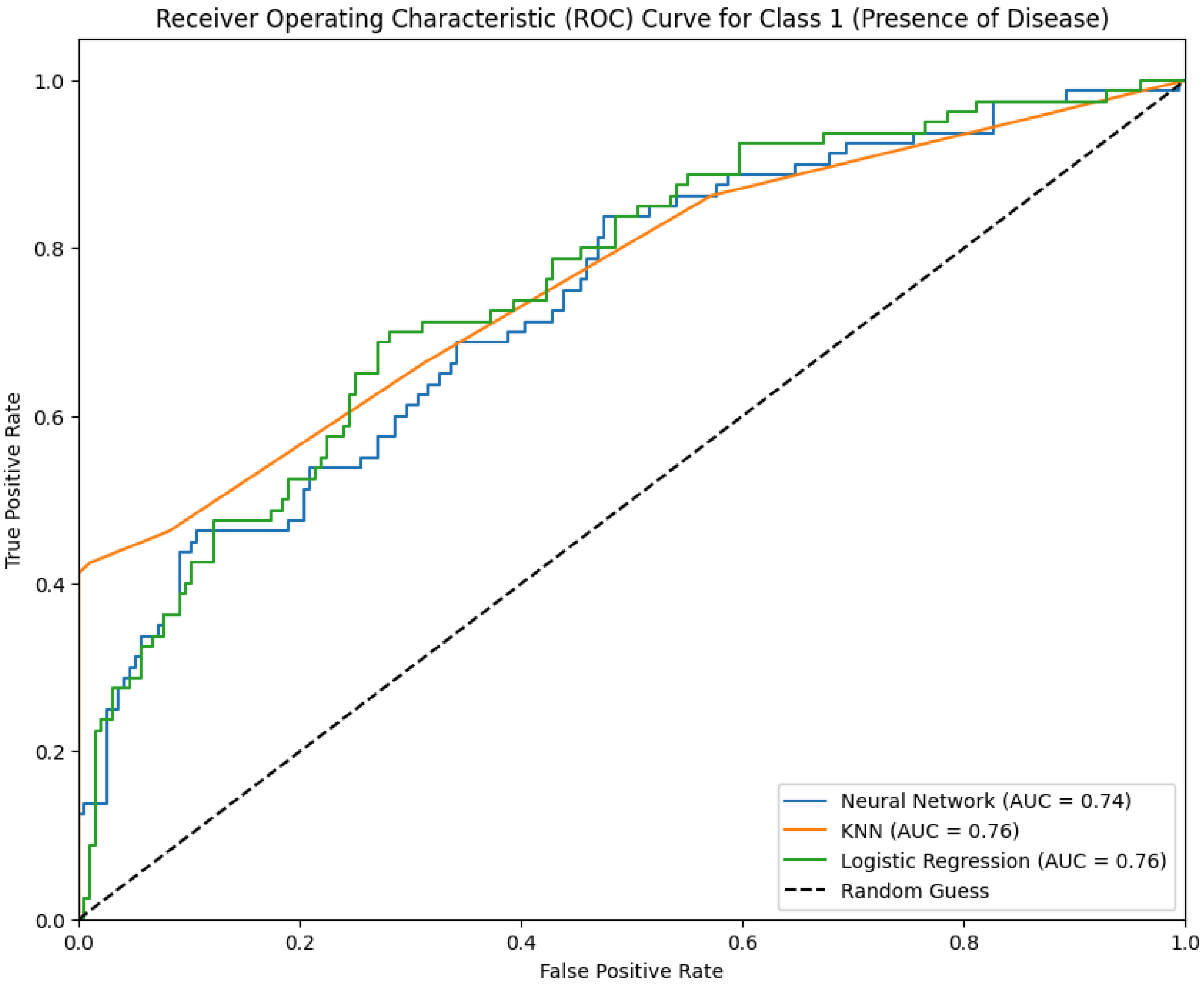
Confusion matrices showed that while all models performed well on predicting patients without disease (class 0), they struggled more with higher disease levels (class 3 and 4), which were underrepresented in the dataset.







ROC–AUC curves confirmed that KNN and Logistic Regression achieved the best separation between classes (AUC ≈ 0.76), outperforming the Neural Network (AUC = 0.74), particularly for class 1 (presence of disease).


Receiver Operating Characteristic (ROC) Curve for Class 1 (Presence of Disease)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Neural Network (AUC = 0.74)
KNN (AUC = 0.76)
Logistic Regression (AUC = 0.76)
Random Guess

7. Conclusion & Future Work

In conclusion, this project demonstrated the complete machine learning pipeline: exploratory data analysis, preprocessing, stratified splitting, supervised classification, unsupervised clustering, and comparative evaluation of models. The results show that Logistic Regression performed best overall, while KNN was competitive and the Neural Network underperformed, likely due to the dataset's small size and imbalance. Models could detect heart disease presence fairly well but struggled with the less frequent severe cases. The main challenges were class imbalance, handling missing values, and ensuring fair evaluation using multiple metrics. These issues explain why simpler models worked better than complex ones. Future improvements include using resampling methods, class-weighted models, or ensemble techniques like Random Forests and XGBoost, as well as expanding the dataset to support more advanced models.