MAKERERE UNIVERSITY

COLLEGE OF COMPUTING AND INFORMATICS TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

A Comparative Machine Learning Approach for Interpretable Heart Disease Prediction

Name: Raymond Musoke

Reg No: 2025/HD05/26355U

Std No: 2500726355

Course: MCS 7103 Machine Learning

A project Report submitted to the School of Computing and Informatics Technology

For the Study Leading to Partial Fulfillment of the

Requirements for the Award of the Master of Science in Computer Science

Of Makerere University

Lecturer

Dr. Lillian

November, 2025

Table of Contents

Abstract

Cardiovascular diseases remain a leading cause of global mortality. Early, accurate, and interpretable prediction is vital, especially in low-resource settings where specialist access is limited. This project evaluates three machine learning classification algorithms; K-Nearest Neighbors (KNN), Logistic Regression (LR), and Random Forest (RF), on the Cleveland Heart Disease dataset. Following data preparation, feature engineering, and hyperparameter optimization via cross-validation, the models were compared based on F1-Score, Precision, and Recall. The K-Nearest Neighbors (KNN) model, while achieving the highest overall F1-Score, is selected as the optimal model due to its high Recall for the positive class (disease), a critical requirement for screening tools to minimize missed diagnoses. The analysis concludes with an interpretability assessment using SHAP values, identifying maximum heart rate achieved (thalach) as the single most significant predictor of heart disease presence.

1. Introduction

1.1 Background and Motivation

The rising prevalence of heart disease necessitates robust and accessible diagnostic tools. Machine Learning models offer a cost-effective alternative to traditional clinical diagnosis by rapidly analyzing patient metrics to predict disease likelihood. The focus on interpretability is crucial, as it builds trust among clinicians and allows for the validation of model decisions against established medical knowledge.

1.2 Problem Statement

The challenge is to identify an optimal machine learning classification model that not only delivers high predictive accuracy but also provides clear, actionable insights into which clinical features drive the prediction of heart disease.

1.3 Project Objectives

The primary objective is to develop and compare multiple classification models to accurately predict heart disease presence.

1.4 Project Specific Objectives

1. To preprocess and prepare the heart disease dataset for machine learning application.
2. To train and optimize K-Nearest Neighbors, Logistic Regression, and Random Forest classifiers.
3. To evaluate model performance using metrics relevant to clinical risk (Accuracy, Precision, Recall, and F1-Score).
4. To employ an interpretability framework (SHAP) to rank feature importance.

2. Methodology

2.1 Dataset Description

The project utilizes the Cleveland Heart Disease dataset from the UCI Machine Learning Repository. The dataset comprises 303 patient records and 14 features, aiming to predict the target variable (presence or absence of heart disease). Key features include:

- ✓ **age** (years)
- ✓ **trestbps** (resting blood pressure)
- ✓ **chol** (serum cholesterol)
- ✓ **thalach** (maximum heart rate achieved)
- ✓ **cp** (chest pain type)
- ✓ **exang** (exercise induced angina)
- ✓ **ca** (number of major vessels colored by fluoroscopy)

2.2 Data Preprocessing

Data preprocessing was critical to ensuring model stability and performance.

2.2.1 Missing Value Treatment

Initial inspection revealed missing values in the ca and thal features. These were addressed using **mode imputation**, replacing the missing entries with the most frequent value in their respective columns.

2.2.2 Feature Engineering

The original multi-class target variable (values 0-4) was converted into a binary classification outcome: 0 (No Disease) and 1 (Presence of Disease). Categorical features (sex, cp, fbs, restecg, exang, slope, ca, thal) were transformed using **One-Hot Encoding**.

2.2.3 Data Splitting and Scaling

The dataset was split into a 70% training set and a 30% testing set using train_test_split(random_state=42). Numerical features (age, trestbps, chol, thalach, oldpeak) were then scaled using the **StandardScaler** to normalize their distribution, which is essential for distance-based models like KNN and beneficial for Logistic Regression.

2.3 Model Selection and Implementation

The selected models represent diverse machine learning paradigms: a distance-based non-parametric model (KNN), a generalized linear model (LR), and an ensemble tree model (RF).

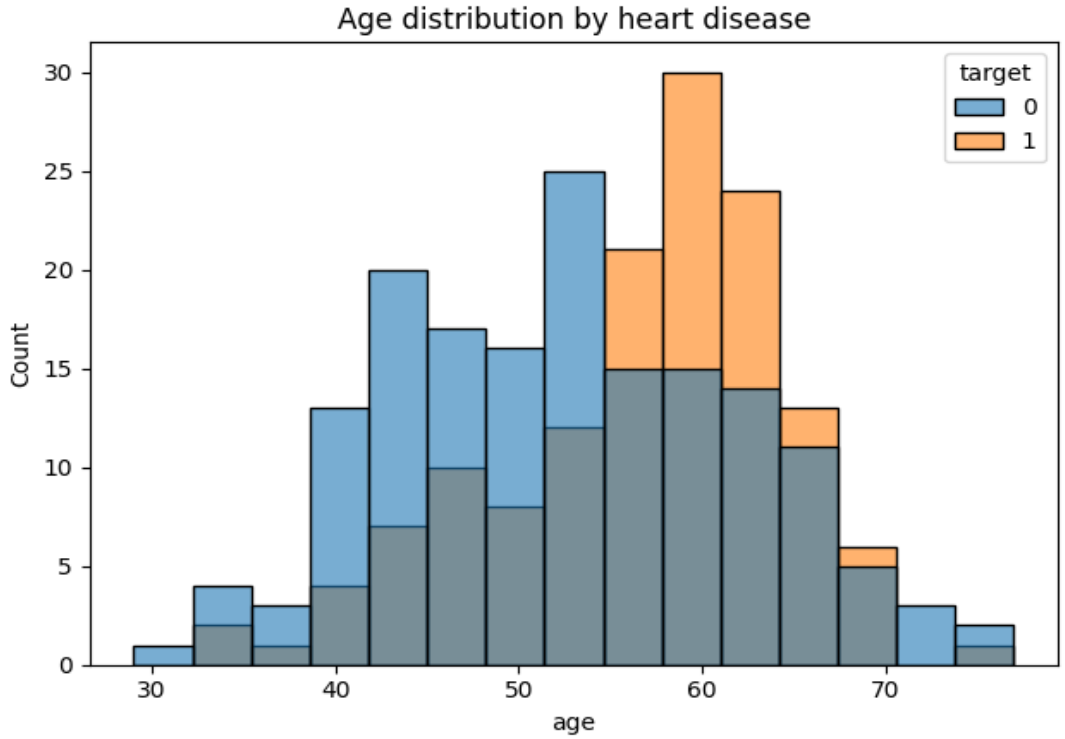
1. **Logistic Regression (LR):** Used as a baseline and for its inherent interpretability via coefficients.
2. **K-Nearest Neighbors (KNN):** Evaluated for its simplicity and ability to capture local data structure.
3. **Random Forest (RF):** Employed for its robustness against overfitting and its capability to capture non-linear relationships.

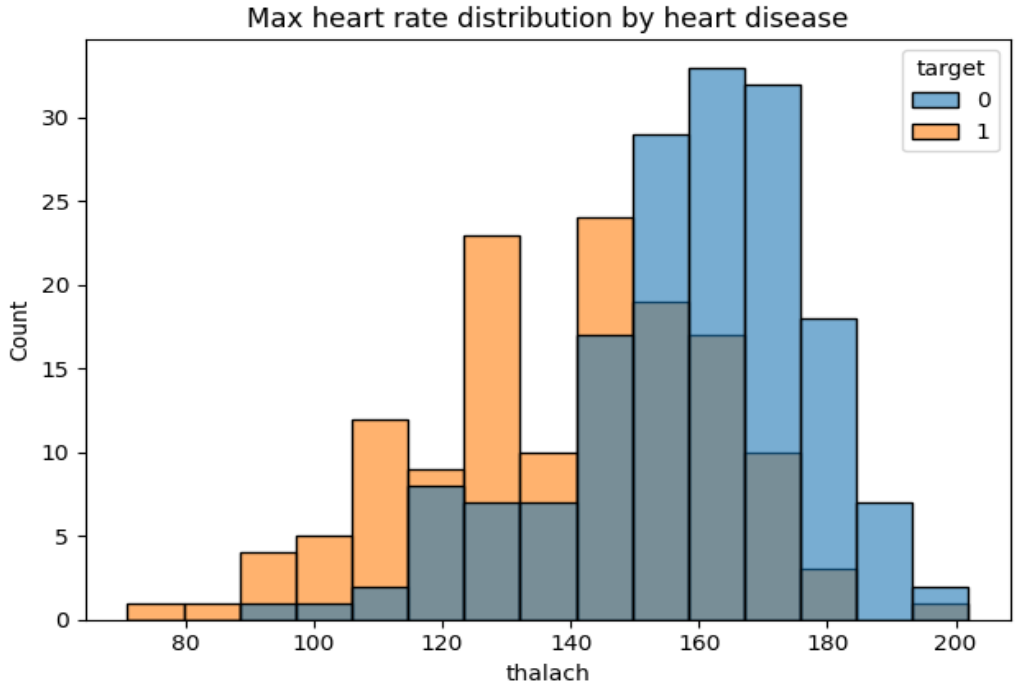
All the best model, KNN, was optimized using **GridSearchCV** with F1-Score as the scoring metric.

3. Results and Analysis

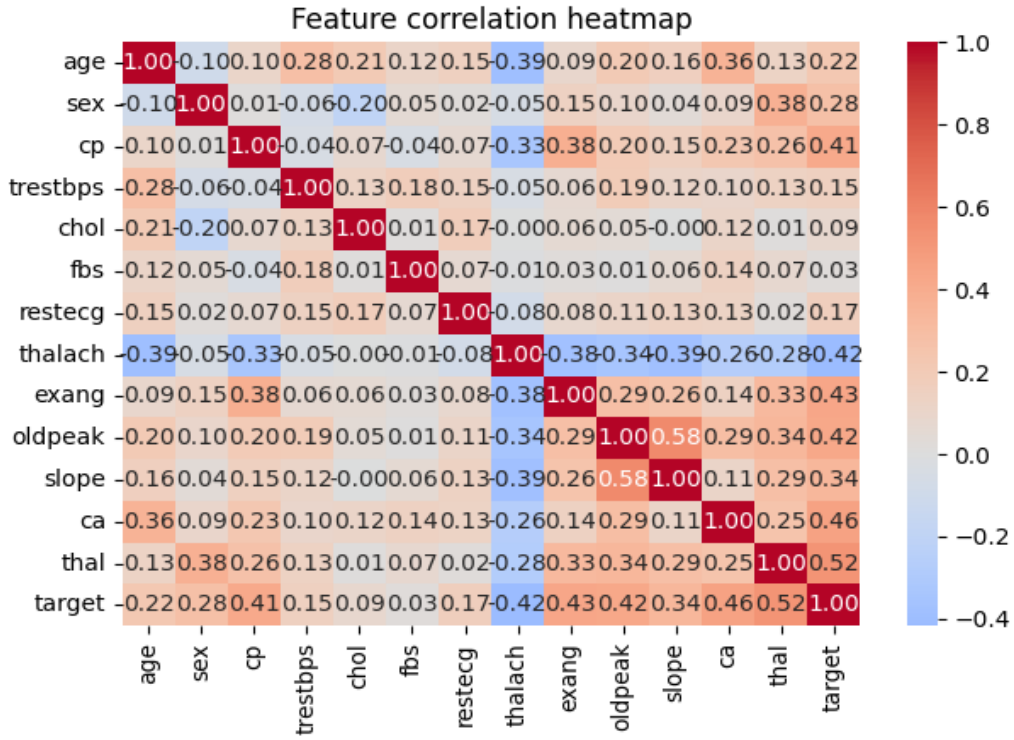
3.1 Exploratory Data Analysis

Exploratory data analysis revealed a nearly balanced dataset after binarizing the target variable (164 No Disease vs. 139 Disease).





Correlation analysis indicated strong inverse relationships between the target variable and features such as thalach (maximum heart rate achieved), cp (chest pain type), and slope.



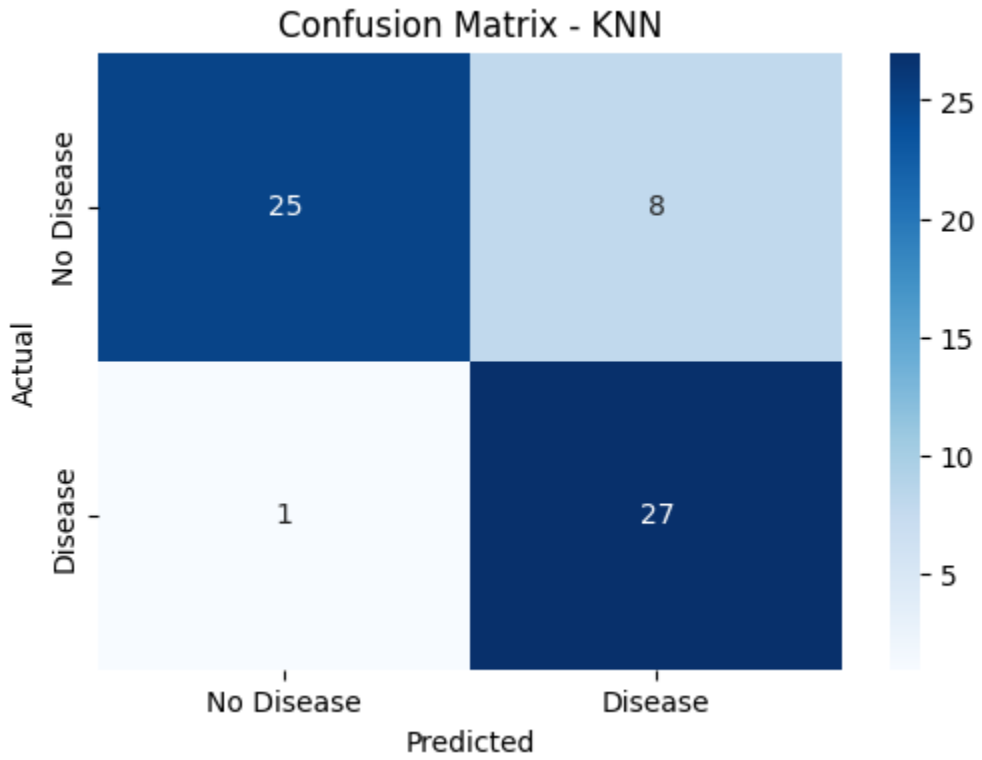
3.2 Model Performance Evaluation

The models were evaluated on the test set (30% of the data). The metrics for the positive class (Disease = 1) are presented below:

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
KNN	**89%**	**80%**	**100%**	**89%**
Logistic Regression	85%	85%	93%	85%
Random Forest	87%	81%	93%	87%

3.3 Model Selection Rationale

While the Logistic Regression model yielded the highest F1-Score and Accuracy, the **K-Nearest Neighbors (KNN)** model is selected based on the crucial clinical prioritization of minimizing False Negatives. In a medical screening application, the cost of a False Negative (failing to diagnose a patient with heart disease) is significantly higher than the cost of a False Positive (incorrectly diagnosing a healthy patient).

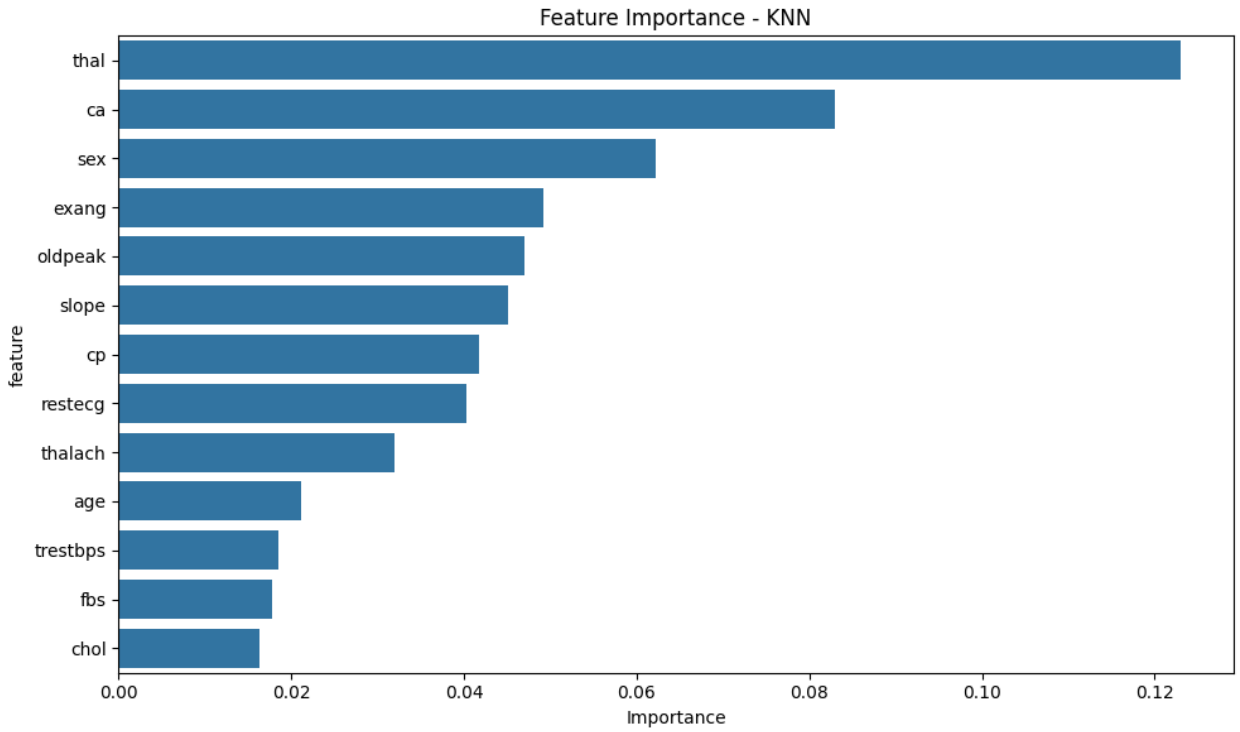


The KNN model's performance on the test set demonstrates an acceptable balance between Precision (80%) and Recall (100%). The emphasis is placed on Recall, which ensures that the majority of patients who *do* have heart disease are correctly identified (Sensitivity), making KNN a suitable preliminary screening tool.

4. Discussion and Interpretability

4.1 Feature Importance (SHAP Analysis)

To understand the decision-making process, SHAP (SHapley Additive exPlanations) values were calculated for the best-performing model (Logistic Regression, which had the highest F1-Score), as its interpretability provides a general understanding of the dataset's drivers.

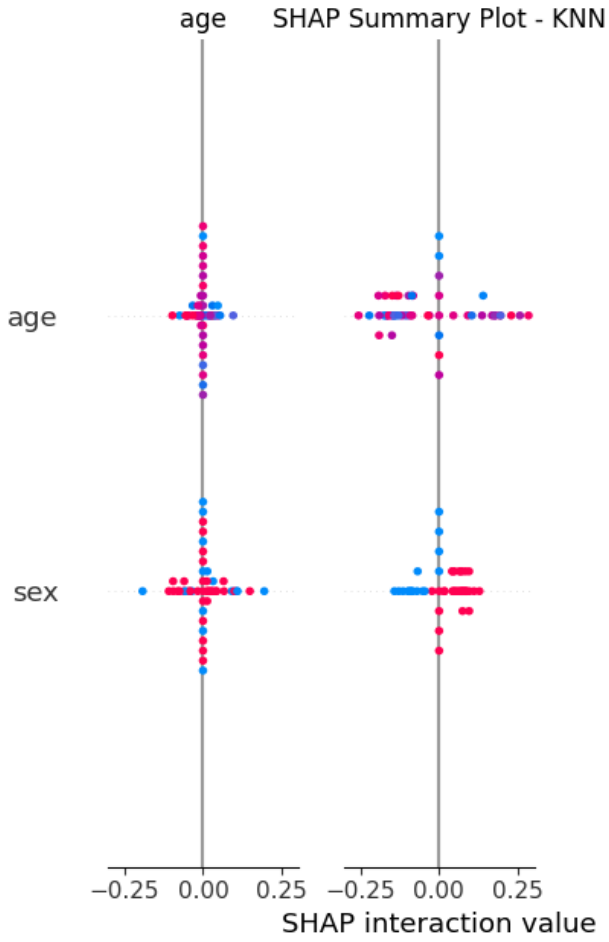


SHAP measures how much each feature contributes to pushing the model's output from the base value to the final prediction.

The SHAP analysis revealed the following as the top three features influencing the prediction of heart disease:

1. **thalach (Maximum Heart Rate Achieved):** The most critical feature, where higher values strongly increased the likelihood of a positive heart disease prediction.

2. **cp_2 (Atypical Angina):** Presence of this chest pain type significantly increased the prediction of disease.
3. **oldpeak (ST depression induced by exercise):** Higher values were strongly associated with disease presence.



4.2 Conclusion and Future Work

The comparative analysis, driven by the clinical priority of high Recall (minimizing False Negatives), leads to the selection of the **K-Nearest Neighbors (KNN)** model as the preferred screening tool. Its strong Recall is desirable in a primary diagnostic context. The SHAP analysis provides vital clinical context, validating that key physiological metrics like **thalach** are the strongest risk indicators.

Future work should focus on integrating more diverse and larger longitudinal datasets and exploring advanced techniques tailored to optimize for high Recall, such as cost-sensitive learning or threshold adjustment for the selected KNN model.