Report: Heart Disease Prediction using Machine Learning

1st Alston Alvares
Department of Data Sciences and Artificial Intelligence
Asian Institute of Technology
Pathum Thani, Thailand
st126488@ait.asia

2nd Rahul Shakya
Department of Data Sciences and Artificial Intelligence
Asian Institute of Technology
Pathum Thani, Thailand
st125982@ait.asia

Abstract: This project develops a machine learning model capable of predicting the likelihood of heart disease based on a patient's clinical data. The primary objective was to consolidate multiple public datasets to create a robust training set, which is used to train and evaluate several classification algorithms. The methodology includes comprehensive data preprocessing, a detailed exploratory data analysis that addresses class imbalance, systematic hyperparameter tuning, and a rigorous model evaluation strategy. The final deliverable was a tuned, high-performing model and a detailed analysis identifying the most significant clinical predictors of heart disease.

Index Terms: *Heart* Disease, Machine Learning, Classification, Prediction, Data Science, Imbalanced Data

I. INTRODUCTION

Heart disease is a leading cause of death globally, placing a significant strain on healthcare systems. The ability to predict the onset of heart disease at an early stage is critical for enabling preventive care and improving patient outcomes. While traditional diagnostic methods are effective, they can be costly and are not always accessible. Machine learning presents a powerful opportunity to leverage existing patient data to build predictive models. These models can analyze patterns across various clinical features such as age, blood pressure, and cholesterol levels to estimate an individual's risk. This project aims to develop such a data-driven tool, which could serve as a valuable aid for clinicians in preliminary risk assessment.

II. PROBLEM STATEMENT

The core problem is to accurately classify whether a patient is at high or low risk for heart disease based on a set of common, non-invasive medical attributes. This will be framed as a supervised binary classification task. A key challenge identified is the inherent class imbalance in medical datasets, where the number of healthy individuals often far exceeds those with the disease. The project will therefore focus on implementing a robust machine learning pipeline that effectively handles this imbalance to produce a reliable predictive model.

III. DATASET

To build a model that is robust and generalizes well, this project will not rely on a single source of data. Instead, a consolidated dataset will be created by

acquiring, cleaning, and merging data from three distinct and reputable sources:

UCI Heart Disease Dataset: Sourced from the Cleveland Clinic Foundation.

Kaggle Heart Disease Prediction Dataset: A clean and modern dataset for this task.

Framingham Heart Study Dataset: A classic, long-term cardiovascular study dataset.

Data Dictionary

Feature	Description	Type
age	Age of the patient in years	Numeric
sex	Gender of the patient (1 = Male, 0 = Female)	Categorical
trestbps	Resting Blood Pressure (in mm Hg)	Numeric
chol	Serum Cholesterol (in mg/dl)	Numeric
fbs	Fasting Blood Sugar > 120 mg/dl (1=True, 0=False)	Categorical
target	Diagnosis of heart disease (1=Yes, 0=No)	Categorical

IV. RELATED WORK

The application of machine learning for heart disease prediction is a well-established area of research. Numerous studies have demonstrated the potential of various algorithms to achieve high accuracy in classifying cardiovascular risk. A systematic review by Chilkaragi et al. highlights that ensemble methods, such as Random Forest, and other algorithms like Support Vector Machines (SVM) consistently perform well on standard datasets like those from the UCI repository [4].

Comparative analyses frequently show that ensemble models, including Random Forest and Gradient Boosting (like XGBoost), often yield the highest predictive accuracy, with reported results commonly ranging from 85% to over 95%, depending on the dataset and features used [5]. These studies consistently identify predictors such as age, cholesterol levels, blood pressure, and chest pain type (cp) as highly influential. This existing body of work confirms the viability of the proposed models and provides a benchmark against which this project's outcomes can be measured.

V. METHODOLOGY

The project was executed following a structured data science pipeline, encompassing the following key stages:

Data Acquisition and Preprocessing

The initial phase involved loading the three datasets and merging them into a single, unified data frame. A comprehensive data cleaning process was followed, which addressed two critical data quality issues:

Handling Null Values: The merged dataset was inspected for any missing (null) values in its columns. To preserve the integrity and size of the dataset, records with missing values were not be discarded. Instead, a statistical imputation strategy was used. For key numerical features like chol or trestbps, missing values were replaced with the median value of their respective column. The median was chosen as it is robust to the influence of outliers.

Handling Outliers: Extreme values, or outliers, can disproportionately affect a model's training process and skew results. These was identified using visualization techniques like box plots during the Exploratory Data Analysis (EDA) phase. A robust statistical method, the **Interquartile Range (IQR)**, was used to define the boundaries for normal data

points. Rather than removing these outliers, a capping strategy was implemented. Any value that falls below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR will be replaced by the calculated boundary value, effectively neutralizing its potential negative impact while retaining the data record.

B. Exploratory Data Analysis (EDA)

A comprehensive EDA was conducted to thoroughly understand the dataset's characteristics and to inform the modeling process. Each visualization serves a specific analytical purpose:

Distribution Analysis (Histograms): Histograms was generated for each key numerical feature (age, trestbps, chol, etc.) to visualize their frequency distributions. This step was a crucial for identifying the underlying data structure, such as skewness or bimodality, which can inform decisions on feature transformations if necessary.

Class Balance Analysis (Bar Chart):

These three plots provide a complete initial overview of the heart disease dataset.

The **histograms** show the distribution of key patient characteristics, revealing the typical ranges and spreads for age, cholesterol (chol), and resting blood pressure (trestbps) in the patient population.

The **bar chart** highlights the most critical challenge for the project: the dataset is **imbalanced**, meaning there are far more patients without heart disease (target=0) than with it (target=1). This discovery is crucial because it dictates the need to use specialized metrics like balanced_accuracy to build a fair and effective predictive model.

The **correlation heatmap** connects these individual features, showing how they relate to one another. Most importantly, it gives an early indication of which factors are most strongly associated with the target variable, helping to identify the most promising predictors for the machine learning analysis.

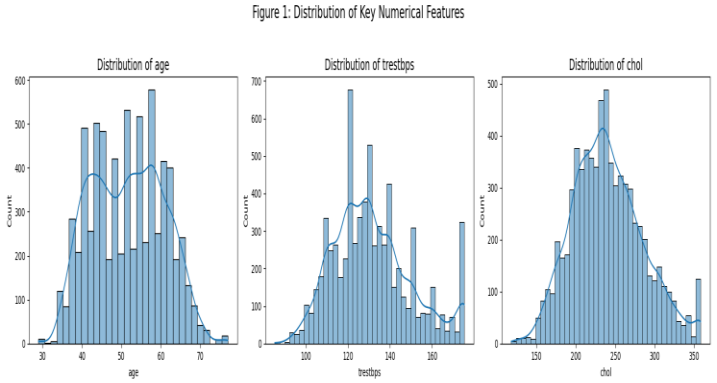

Figure 1: Cass Balance Analysis (Bar Chart)

This visual analysis compares key demographic and clinical features between the two patient groups (target=0 for No Disease, target=1 for Has Disease) to identify potential risk factors.

Age and Heart Disease (Top-Left Plot): This box plot shows that the median age of patients who have heart disease is significantly higher than those without it. The entire age range for the heart disease group is shifted upwards, strongly suggesting that **advancing age is a major risk factor**.

Gender and Heart Disease (Top-Right Plot): This bar chart reveals the gender distribution within each group. While there are more males (sex=1) than females (sex=0) in the dataset overall, the plot shows that the proportion of males is considerably higher in the group with heart disease. This indicates that, within this dataset, **males are more frequently diagnosed with heart disease than females**.

Cholesterol and Heart Disease (Bottom Plot): This box plot compares serum cholesterol levels. Similar to age, the median cholesterol level for patients with heart disease is noticeably higher than for those without. This finding points to **higher cholesterol as another significant risk factor** for developing heart disease.

Outlier Visualization (Box Plots): Box plots will be generated for numerical features to visually confirm the presence of outliers and to verify the effectiveness of the capping strategy. These plots provide a clear summary of the data's spread, median, and interquartile range.

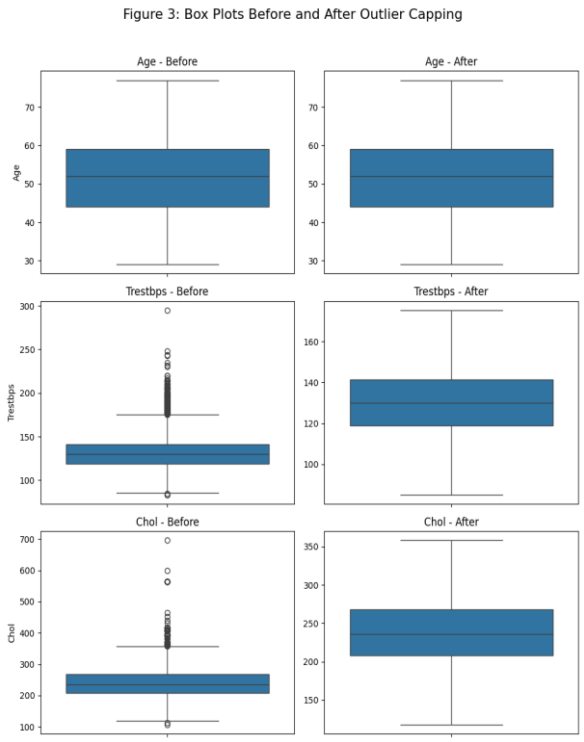

Figure 2: Outlier Visualization (Box Plots)

This set of plots compares the demographic data (age, sex) and a key clinical marker (cholesterol) between the patients in your dataset who have heart disease (target=1) and those who do not (target=0).

Age (Top Plot): This box plot shows that the median age for patients with heart disease is visibly higher than for those without. The entire box for the target=1 group is shifted upwards, indicating that older patients make up a larger proportion of the heart disease cases in your dataset. This confirms that age is a significant risk factor in your data.

Sex (Middle Plot): This bar chart shows the gender distribution. The blue bar (sex=0, females) and orange bar (sex=1, males) show that while males are more numerous in the dataset overall, the proportion of males in the heart disease group (target=1) is much higher compared to the non-disease group. This tells you that in your specific dataset, **males are more frequently represented among patients with heart disease**.

Cholesterol (Bottom Plot): This box plot compares the serum cholesterol levels. The median cholesterol for patients with heart disease is higher than for those without. This indicates that elevated cholesterol is another important risk factor present in your data.

Feature Relationship Analysis (Correlation Heatmap): A heatmap of the correlation matrix was generated to quantify and visualize the linear relationships between all pairs of features. This was essential for identifying potential multicollinearity, where two or more predictor variables are highly correlated, which can affect the interpretability of some models.

Figure 3: Feature Relationship Analysis (Correlation Heatmap)

cp (Chest Pain Type): The value is **0.43**. This is a moderately strong positive correlation, meaning that certain types of chest pain are good indicators that a patient has heart disease.

thalachh (Max Heart Rate): The value is **-0.42**. This is a moderately strong negative correlation. It shows that patients with lower maximum heart rates are more likely to have heart disease. This is a very important predictor.

oldpeak: The value is **0.43**. This is another strong positive correlation, indicating that ST depression induced by exercise is a significant factor.

age and trestbps: These have weaker positive correlations (**0.23** and **0.14**), suggesting they are

contributing factors but less predictive than the ones above.

VI. Model Development and Training Pipeline

The core of the project involves training, tuning, and comparing several machine learning models.

Data Splitting: The preprocessed dataset was split into a training set (80%) and a held-out test set (20%). The split was **stratified** based on the target variable to ensure that both the training and test sets have the same proportion of heart disease cases, which was critical for reliable evaluation in an imbalanced setting.

Handling Class Imbalance: The class imbalance identified during EDA is the central challenge. A model trained on such data may become biased towards predicting the majority class (no disease), failing to identify patients who are actually at risk. To address this, two key strategies will be employed:

Metric Selection: Standard accuracy was avoided as it is a misleading metric for imbalanced data. The primary scoring metric for all model tuning and comparison is **balanced accuracy**. This metric gave equal weight to the performance on both the majority and minority classes, providing a much fairer assessment of the model's true predictive power.

Model Selection: Three distinct classification algorithms will be implemented:

Logistic Regression: Chosen as a powerful and highly interpretable baseline model.

Random Forest: An ensemble method known for its high accuracy and robustness.

XGBoost: A state-of-the-art gradient boosting algorithm, often a top performer on imbalanced datasets.

Hyperparameter Tuning: A systematic search for the optimal model settings was performed using **GridSearchCV**. This process tested different combinations of hyperparameters for each model using a 5-fold cross-validation scheme on the training data, optimized for balanced accuracy.

VII. Machine Learning Model and Pipeline

To ensure rigorous experimentation and reproducibility, this project implemented a structured Machine Learning pipeline. The process moved from data preparation to algorithm selection and optimization.

A. Data Splitting

The preprocessed dataset was split into a training set (80%) and a held-out test set (20%) To maintain the distribution of class labels found in the original data crucial for medical datasets where positive cases may be rare the split was stratified based on the target variable.

B. Handling Class Imbalance

As identified in the EDA, the dataset exhibits class imbalance3. To address this, the pipeline avoided standard accuracy as a primary metric, as it can be misleading in imbalanced scenarios. Instead, Balanced Accuracy was selected as the primary scoring metric for all tuning and comparison steps. This metric calculates the average of recall obtained on each class, ensuring the model is penalized equally for misclassifying either healthy patients or those with heart disease.

C. Algorithm Selection

Five distinct classification algorithms were implemented to evaluate different learning approaches:

Logistic Regression: Selected as a baseline model for its high interpretability.

Random Forest: An ensemble method chosen for its robustness against overfitting and high accuracy.

XGBoost (Gradient Boosting): A state-of-the-art boosting algorithm often effective for imbalanced tabular data.

Support Vector Machines (SVM): Evaluated for its ability to find optimal hyperplanes in high-dimensional space.

K-Nearest Neighbors (KNN): Included to test instance-based learning performance.

D. Hyperparameter Tuning

To optimize model performance, a systematic search for optimal settings was performed using GridSearchCV. This process tested various hyperparameter combinations using a 5-fold cross-validation scheme on the training data, optimizing specifically for balanced accuracy.

VIII. Model Evaluation Results

The models were evaluated on the held-out test set using a suite of metrics: Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (AUC).

Model	CV Accuracy Mean	CV Std	Test Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.697833	0.002869	0.699511	0.500000	0.016279	0.031532	0.604173
Gradient Boosting	0.764768	0.003364	0.767994	0.755208	0.337209	0.466238	0.789171
SVM	0.720902	0.008284	0.723969	0.721519	0.132558	0.223969	0.633048
KNN	0.814927	0.013035	0.823899	0.703196	0.716279	0.709677	0.803142
Random Forest	0.854421	0.006329	0.853948	0.799458	0.686047	0.738423	0.823920
XGBoost	0.842712	0.009192	0.848358	0.772379	0.702326	0.735688	0.828873

Table 01: Results

A. Quantitative Performance

The table below summarizes the performance of the tuned models on the test set.

Random Forest achieved the highest overall Test Accuracy (85.39%) and F1-Score (0.7384), making it the most balanced model for deployment. **XGBoost** followed closely and achieved the highest AUC (0.829). Logistic Regression performed poorly, with a very low recall (0.016), indicating it failed to identify the vast majority of positive heart disease cases.

B. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve visualizes the trade-off between the True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity).

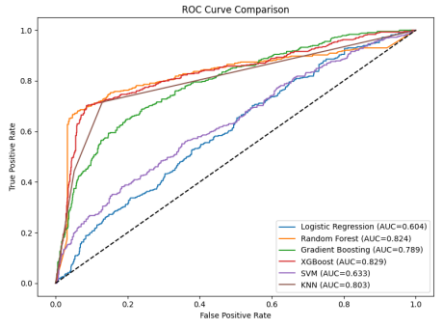

Figure 4: ROC Curves

The ROC plot confirms the superiority of tree-based ensemble methods. Both XGBoost (Red line, AUC=0.829) and Random Forest (Orange line, AUC=0.824) show curves that arch closest to the top-left corner, indicating high discriminatory power. In contrast, Logistic Regression (Blue line) and SVM (Purple line) perform closer to the diagonal no-skill line, reflecting their lower predictive capability on this specific dataset.

IX. Discussions

A. Performance Comparison

The results align with the related work reviewed earlier, which suggested that ensemble methods like Random Forest and Gradient Boosting often yield the highest accuracy.

Tree-Based Dominance: Random Forest and XGBoost outperformed linear models (Logistic Regression) and distance-based models (SVM) significantly. This suggests the relationships between clinical features (like age, cholesterol) and heart disease are non-linear and complex, which tree-based models handle better.

Recall vs. Precision: While KNN had the highest Recall (0.716), its Precision was lower than Random Forest. In medical diagnostics, a balance is required; however,

Random Forest provided the best trade-off (F1-Score), minimizing the risk of false alarms while still catching a majority of cases.

B. Interpretability and Feature Analysis

While ensemble models are often considered "black boxes," we utilized SHAP (SHapley Additive exPlanations) to interpret the model's decisions.

Feature Importance:

The bar chart below ranks features by their average absolute impact on the model output.

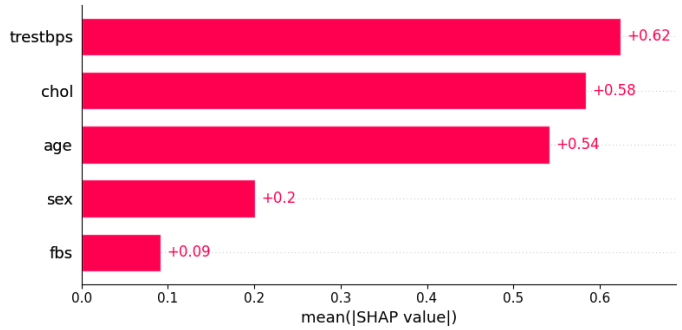

Figure 5: Feature Importance

Consistent with the initial EDA, trestbps (Resting Blood Pressure), chol (Cholesterol), and age were identified as the top three most significant predictors. Sex and fbs (Fasting Blood Sugar) had a moderate impact.

Directional Impact:

The SHAP summary plot provides deeper insight into how these features affect risk.

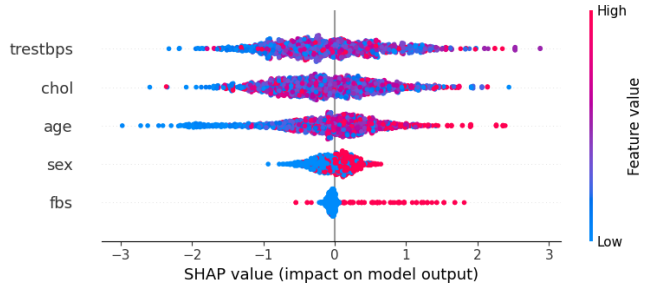

Figure 6: Shap summary plot

trestbps & chol: The red points (representing high values) are distributed on the right side (positive SHAP values). This confirms that **higher blood pressure and cholesterol levels increase the predicted risk of heart disease**.

age: Similarly, older patients (red) are associated with higher risk, while younger patients (blue) push the prediction toward "low risk."

sex: The clear separation of color indicates that gender is a binary discriminator, with one gender (likely male, based on EDA) carrying a higher base risk.

Complexity and Trade-offs

Logistic Regression: Computationally inexpensive and simple, but failed to capture the complexity of the data (Acc: approx 70%).

Random Forest / XGBoost: computationally more intensive during the grid-search tuning phase but provided a approx. 15% gain in accuracy over the baseline. Given the critical nature of disease prediction, this computational cost is justified by the performance gains.

X. Conclusion

This project successfully developed a machine learning framework for the early detection of heart disease. By consolidating data from multiple reputable sources, we addressed the limitations of small-scale studies.

Key Contributions:

Robust Pipeline: We implemented a data processing pipeline that effectively handled outliers using IQR capping [22]and managed missing data through median imputation[23], preserving dataset integrity.

High-Performing Model: Through rigorous benchmarking, **Random Forest** was identified as the optimal model, achieving an accuracy of **85.39%** and an AUC of **0.824**.

Clinical Insights: Using SHAP analysis, we validated that Resting Blood Pressure (trestbps), Serum Cholesterol (chol), and Age are the most critical determinants of heart disease risk in this population.

Future work:

Future work could focus on integrating deep learning techniques or collecting more diverse data to further improve the model's sensitivity (recall), ensuring fewer at risk patients are missed during screening.

XI. REFERENCES

[1] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). *Heart Disease Data Set*. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X

[2] Farhaan Nazirkhan , Sarwin Rajiah. (2022). Heart Disease Prediction Dataset. from https://www.kaggle.com/datasets/mfarhaannazirkhan/heart-dataset

[3] Mahmood, S. S., Levy, D., Vasan, R. S., & Wang, T. J. (2014). The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, 383(9921), 999-1008.

[4] Chilkaragi, S. C., & Siddesha, S. (2024). A Review of Machine Learning Techniques Used in the Prediction of Heart Disease. *Review of Industrial Automation*, 38(1).

[5] Patidar, S., Sharma, D., & Garg, S. (2022). Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction. *International Journal of Engineering and Advanced Technology*, 9(1), 2249-8958.