Heartbeat Sentinel - Decoding and Predicting Heart Failure

Ayushman Anupam
Email: ayushmantutu@gmail.com
29 November 2024

Abstract

Heart disease remains one of the leading causes of mortality worldwide, necessitating advancements in early diagnosis and prediction. This project leverages a comprehensive dataset, compiled by integrating five prominent heart disease datasets (Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog), to analyze and predict heart disease using machine learning techniques.

Firstly through Exploratory Data Analysis (EDA), the project investigates the distribution, correlations, and trends within the dataset, highlighting critical patterns linked to heart disease. Then machine learning models are applied to predict the target variable, aiming to identify high-risk individuals based on their medical features. This study not only provides insights into the factors contributing to heart disease but also demonstrates the potential of predictive analytics in healthcare to improve diagnostic accuracy and patient outcomes.

Contents

1 Introduction

Cardiovascular diseases (CVDs) simply Heart disease remain one of the leading causes of mortality globally, underscoring the importance of early diagnosis and effective risk prediction strategies.

This project aims on predicting wheather a person is havig a heart failure risk or not. Using historical data, the project employs machine learning techniques, specifically two models are used (a) Random Forest Model and (b) XGBoost Model trained on a diverse dataset, incorporating features such as age, gender, cholesterol levels, blood pressure, glucose levels, and other clinical variables. The feature set is prepared by excluding the target variables, and the model is trained on the remaining features.

The performance of the models are presented as confusion matrix and are evaluated on the metrices like Accuracy, precision, Recall, F1 score and ROC/AUC for both models. This approach provides a robust way to predict heart risk, which in turn provides valuable insights for medical professionals, facilitating early intervention and personalized treatment strategies to mitigate heart disease risks.

2 Dataset Description

The dataset provides detailed information about heart disease diagnosis based on several medical features. The data was curated by combining five different heart disease datasets, making it the largest dataset available for heart disease research. These datasets include records from the Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog (Heart) datasets, which have been integrated to create a comprehensive collection of 918 observations after removing duplicates and can be accessed on Kaggle.

The features in the dataset include attributes such as age, sex, chest pain type, blood pressure, cholesterol, fasting blood sugar, ectrocardiogram results, maximum heart rate, exercise-induced angina, oldpeak (measured depression), and the slope of the peak exercise ST segment. The target variable HeartDiseast indicates whether a patient has heart disease (1) or not (0).

All these features are discussed in detail below:

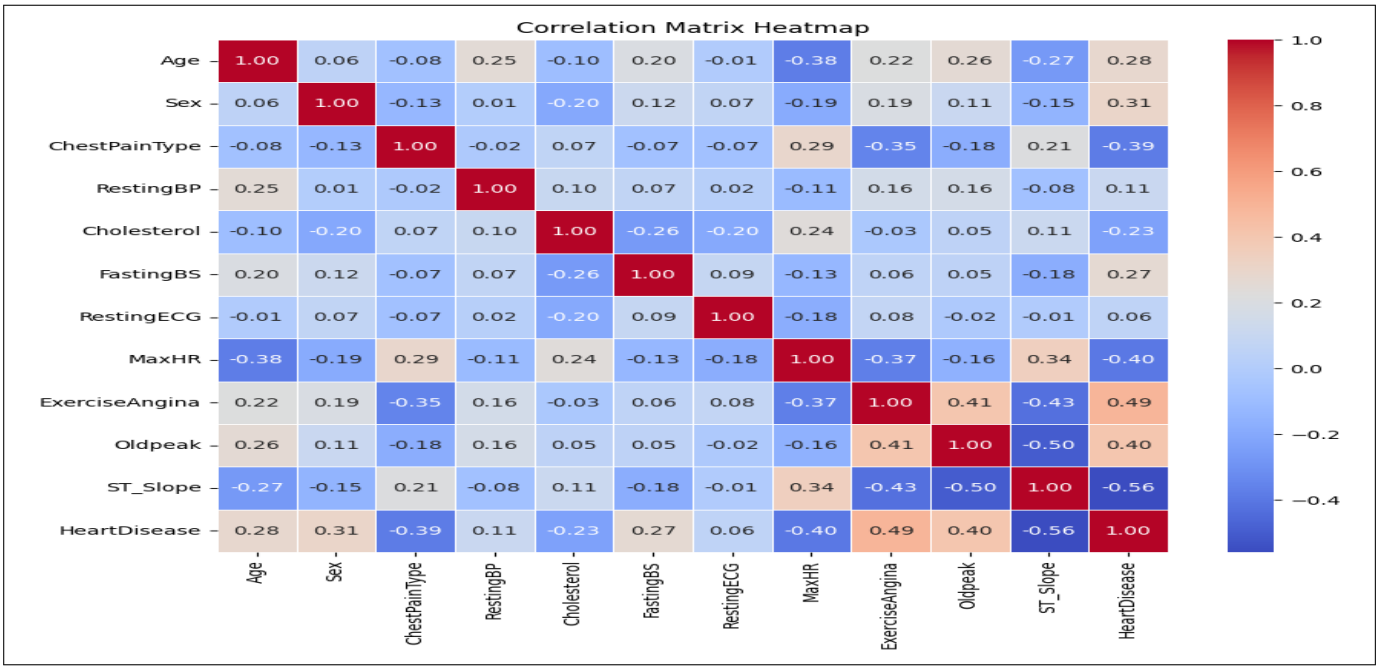
1. **Age**: Age of the individual.
2. **Sex**: Gender (M = Male, F = Female).
3. **ChestPainType**: Type of chest pain experienced (e.g., ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic).
4. **RestingBP**: Resting blood pressure (in mm Hg).
5. **Cholesterol**: Serum cholesterol in mg/dL.
6. **FastingBS**: Fasting blood sugar ¿ 120 mg/dL (1 = Yes, 0 = No).
7. **RestingECG**: Results of resting electrocardiogram (eg. Normal, ST, etc.).
8. **MaxHR**: Maximum heart rate achieved.
9. **ExerciseAngina**: Exercise-induced angina (Y = Yes, N = No).
10. **Oldpeak**: ST depression induced by exercise relative to rest.
11. **ST_Slope**: Slope of the peak exercise ST segment (eg Up, Flat, Down).
12. **HeartDisease**: Target variable (1 = Heart Disease, 0 = No Heart Disease).

3 Exploratory Data Analysis

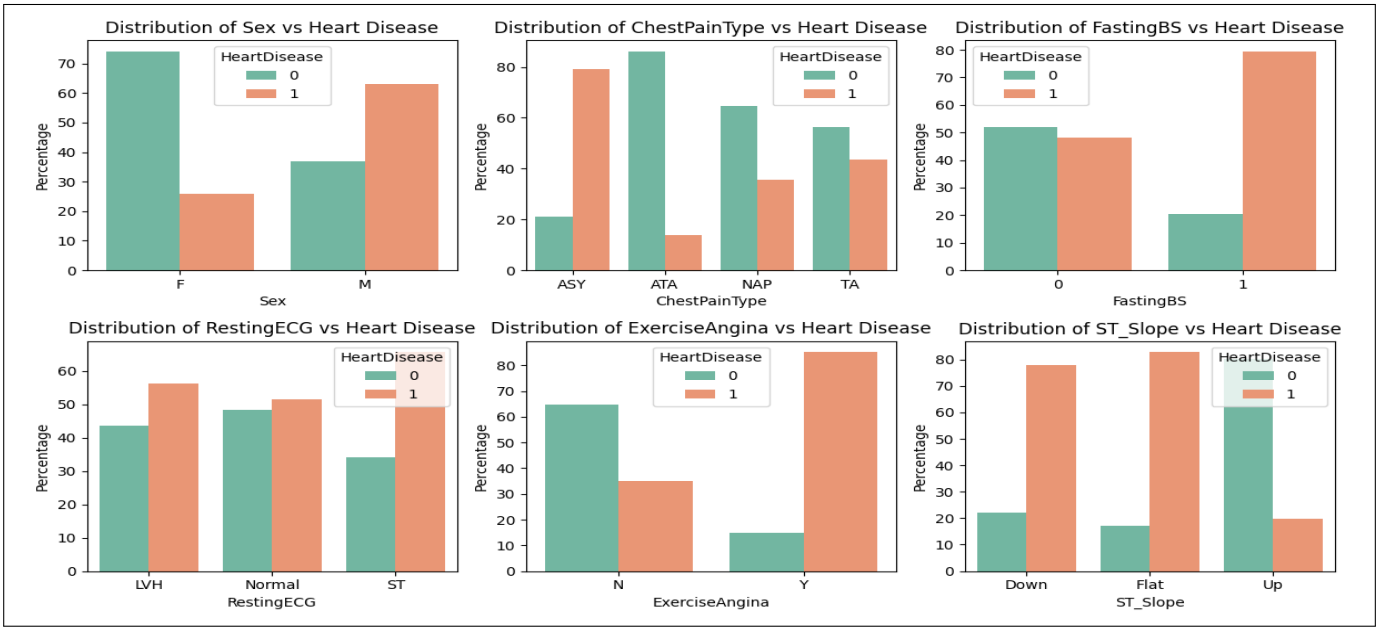
Here, we identify trends of our dataset by conducting univariate and bivariate data analyses. Univariate analysis helps to summarize individual variables, bivariate analysis explores relationships between two variables and examines interactions between multiple variables to uncover deeper patterns and insights influencing housing prices.

3.1 Data Visualization

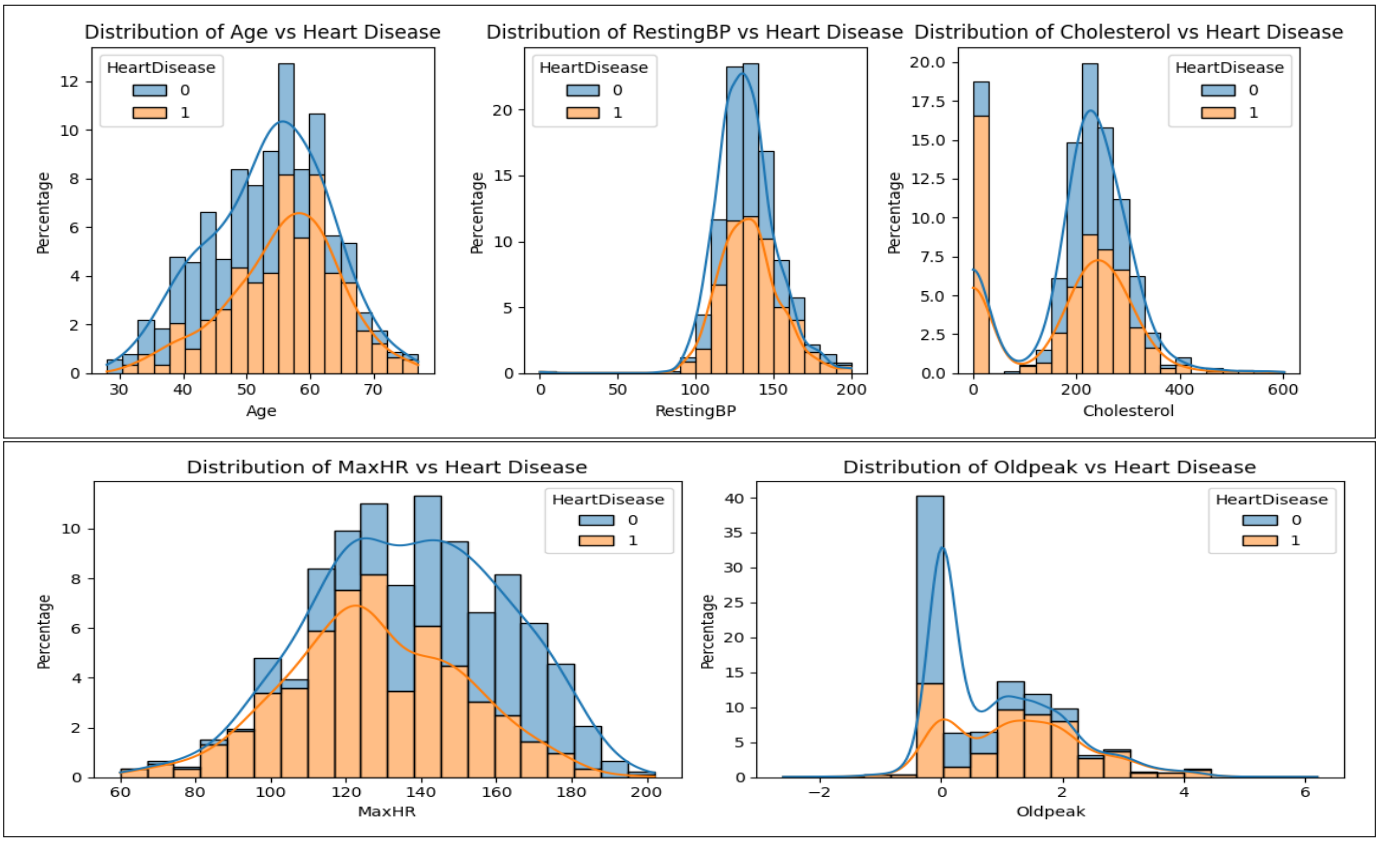
3.1.1 Figure 01. Correlation Matrix Heatmap



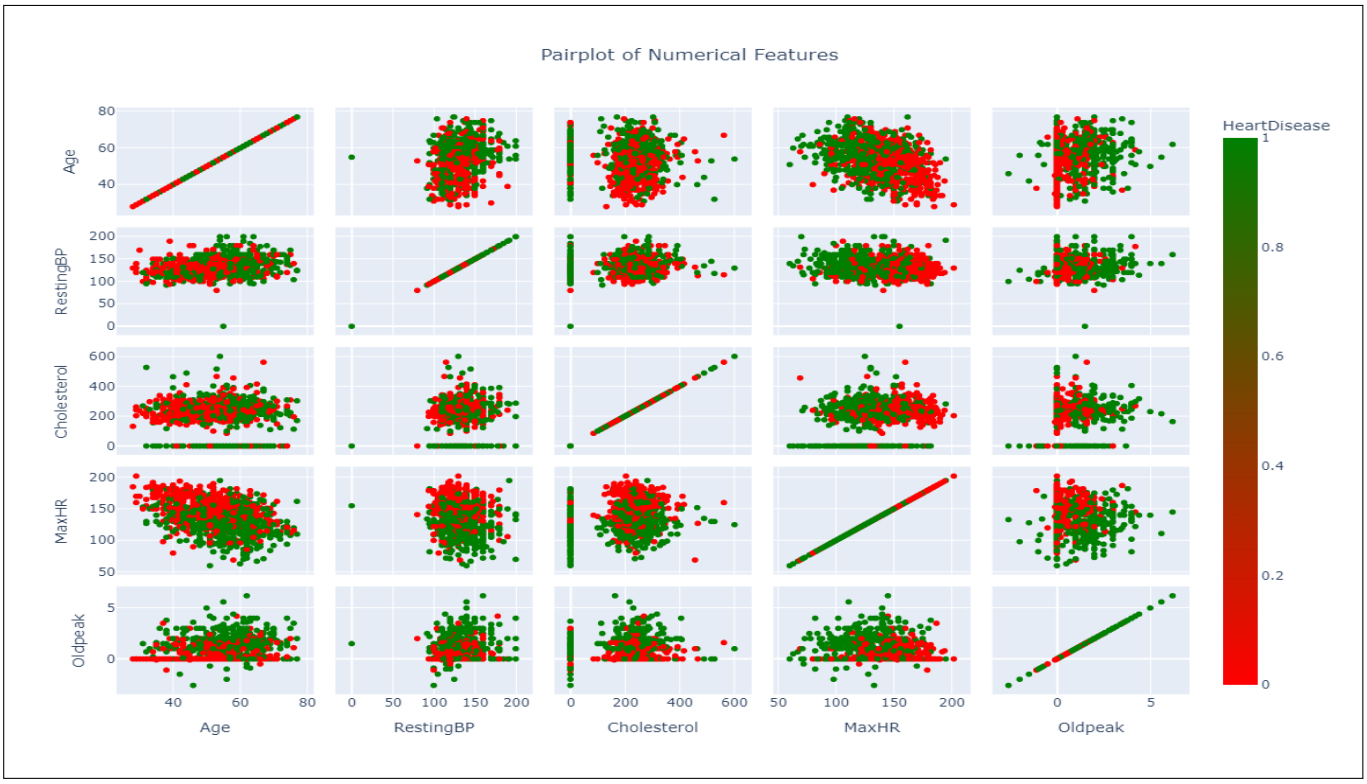
3.1.2 Figure 02. Histogram of Categorical Data



3.1.3 Figure 03. Histogram of Numerical Data



3.1.4 Figure 04. Scatter matrix of Numerical Data



3.2 Result and conclusion from EDA

3.2.1 Histogram Analysis

The histograms for categorical features such as ChestPainType and FastingBS reveal distinct distributions: For ChestPainType: most observations fall into categories such as asymptomatic pain, indicating common patterns among patients with heart disease. Whereas, for FastingBS: Binary distributions (1 for sugar level is greater than 120mg/dL, 0 otherwise) highlight key differences between healthy and at-risk individuals.
Additionally, histograms for numerical variables like Age and Cholesterol display skewness or clustering, suggesting that certain age groups or cholesterol levels are more prone to heart disease.

3.2.2 Scatter Plot Analysis

From scatter plots of numerical variablesshows that Older individuals generally exhibit lower maximum heart rates (MaxHR) and clustering of higher cholesterol levels (Cholesterol) with elevated resting blood pressure (RestingBP) is observed, indicating potential shared risk factors.

3.2.3 Correlation Map Analysis

The correlation heatmap reveals important relationships:

- Strong positive correlations exist among features like MaxHR and ExerciseAngina, which could be significant predictors of heart disease.
- Weak correlations are observed between FastingBS and other variables, suggesting it has limited predictive power on its own.
- Features such as Age, Cholesterol, and Oldpeak (ST depression) show strong links with the target variable, HeartDisease.

These findings emphasize distinct distributions and relationships among features, guiding feature selection and model development for heart disease prediction.

4 Feature Selection and Training for model

For this project, feature selection focused on identifying key clinical variables, such as age, cholesterol levels, blood pressure, and glucose levels, while excluding the target variable. Data preprocessing ensured clean, normalized inputs to support robust training. Both Random Forest and XGBoost models were trained on these features, with hyperparameter optimization to enhance accuracy and reliability. This approach improved early risk prediction consistency, aiding in reliable heart failure risk assessment.

4.1 Feature Selection and Training for Random Forest Model

The Random Forest model is a powerful ensemble machine learning algorithm that builds multiple decision trees during training by taking combination of feature at a time and merges their outputs for more accurate and stable predictions. Its ability to handle both classification and regression tasks, along with robustness against overfitting, makes it ideal for medical risk prediction. For this project, the

Random Forest model was trained using clinically significant features such as age, cholesterol levels, blood pressure, and glucose levels, identified through feature importance analysis. The model leverages its inherent feature ranking capability to prioritize impactful variables. Data preprocessing, including normalization and handling missing values, ensured quality inputs. Hyperparameter tuning, such as optimizing the number of trees and maximum depth, enhanced model accuracy and reliability.

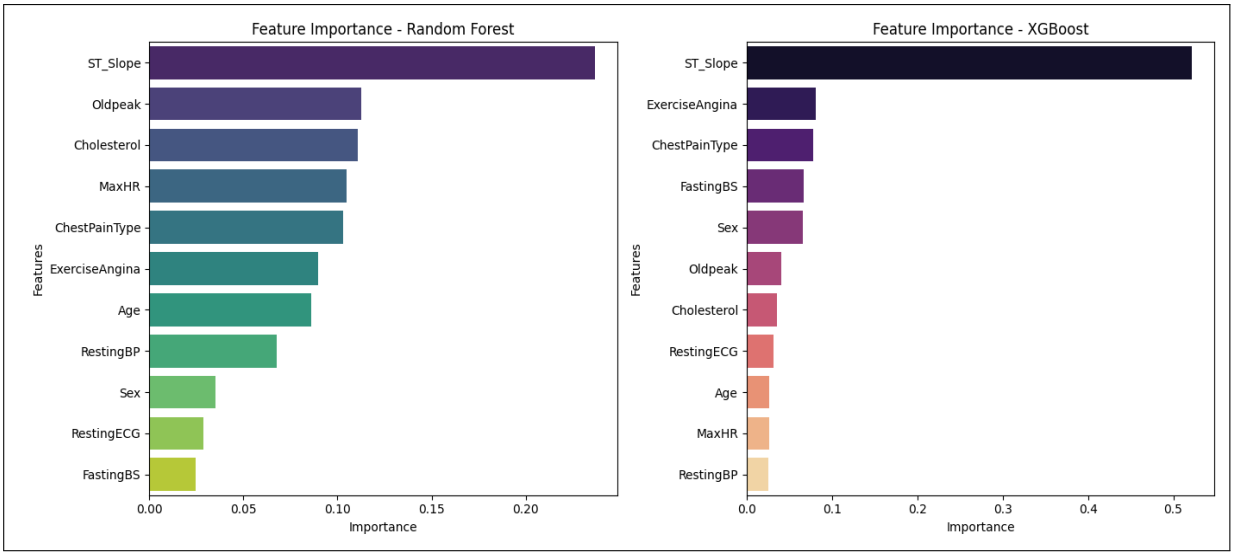
```
from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier(n_estimators=100,
                          random_state=42)
rf_model.fit(X_train, y_train)
importances_rf = rf_model.feature_importances_
rf_feature_importances = pd.Series(importances_rf, index=X.columns)
                          .sort_values(ascending=False)
```

4.2 Feature Selection and Training for XGBoost Model

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable machine learning algorithm designed for structured data. It utilizes gradient boosting frameworks to create strong predictive models through iterative improvements, offering excellent performance in classification tasks with fast computation and effective handling of missing data. For this project, XGBoost was trained on selected features, including age, cholesterol levels, blood pressure, and glucose levels, chosen for their clinical relevance. The model's built-in feature importance metric helped refine the feature set further. Preprocessing steps, such as normalization and addressing missing values, ensured consistent data quality. Hyperparameter tuning, including adjustments to learning rate, tree depth, and boosting rounds, maximized accuracy and robustness in predicting heart failure risk.

```
import xgboost as xgb
xgb_model = xgb.XGBClassifier(n_estimators=100,
                              random_state=42)
xgb_model.fit(X_train, y_train)
importances_xgb = xgb_model.feature_importances_
xgb_feature_importances = pd.Series(importances_xgb, index=X.columns)
                              .sort_values(ascending=False)
```

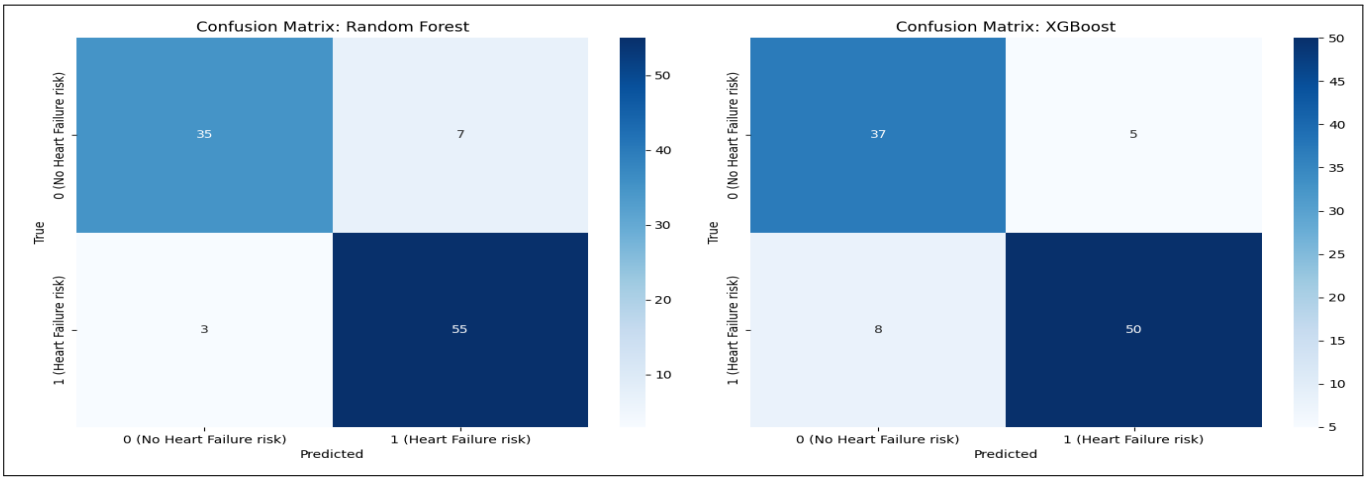
4.2.1 Figure 05. Feature importnace for both model



5 Prediction and Result

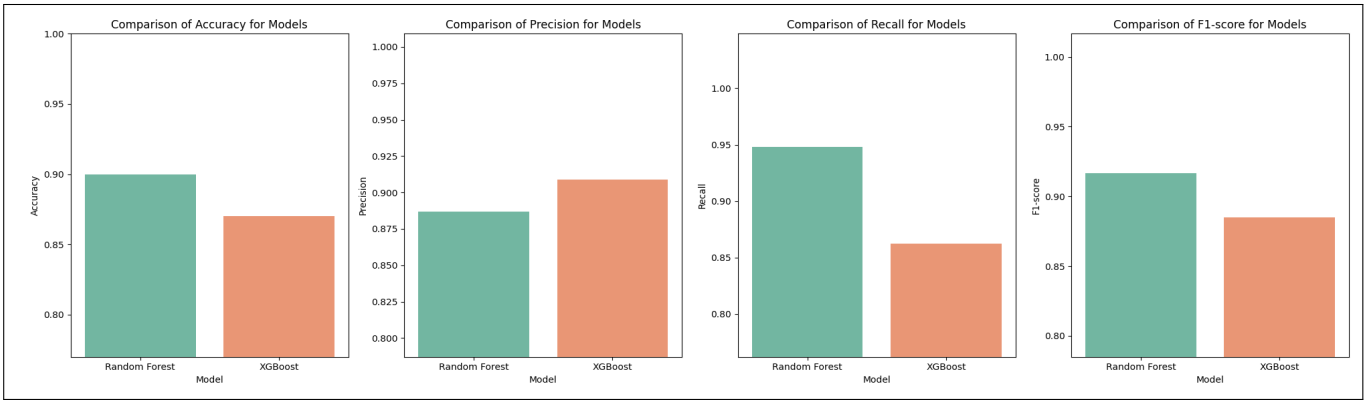
The prediction phase utilized the trained Random Forest and XGBoost models to classify individuals into risk categories for heart failure. Each model was evaluated using metrics such as Accuracy, Precision, Recall, F1 Score, and ROC/AUC to ensure a comprehensive assessment of performance.

5.0.1 Figure 06. Confusion Matrix (a) Random Forest Model, (b) XGBoost Model



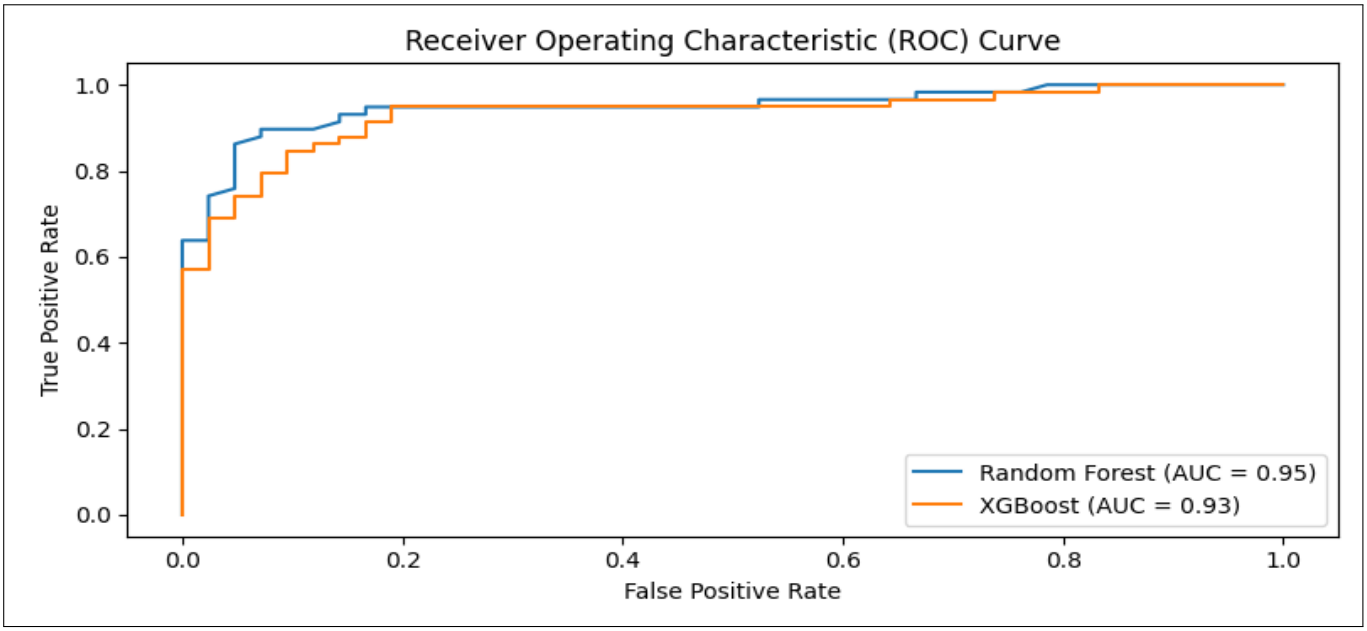
Above confusion matrix for each model revealed a low rate of false negatives, which is critical in the medical context, as false negatives could result in missed diagnoses of individuals at high risk of heart failure.

5.0.2 Figure 07. Comaprison of (a) Accuracy, (b) precision, (c) Recall, (d) F1 score for both model



Above graph from results shows that both models achieved high predictive accuracy, but the Random Forest model outperformed the XGBoost model in terms of accuracy and classification performance. The Random Forest model demonstrated a higher overall accuracy and better precision, which is crucial in medical applications where false positives can be costly.

5.0.3 Figure 08. Comaprison of (a) Receiver Operating Characteristic (ROC), (b) Area Under curve (AUC)



From above, ROC/AUC scores for both models were high, indicating that both models had good discriminatory power, meaning they could effectively distinguish between high-risk and low-risk individuals.

Both models performed well, but Random Forest proved to be more robust in this particular case, highlighting its efficiency in handling medical datasets with potentially noisy or imbalanced features.

These findings confirm the reliability and efficacy of both models in predicting heart failure risk, and underscore the potential of machine learning in supporting clinical decision-making. With further fine-tuning, these models could serve as valuable tools for early intervention and personalized treatment strategies, ultimately improving patient outcomes in heart disease management.

6 Conclusion and disussion

Cardiovascular diseases (CVDs) remain a significant global health challenge, necessitating innovative approaches for early diagnosis and risk prediction. In this project, we successfully developed and implemented machine learning models to predict the risk of heart failure using historical data and clinical variables. By employing Random Forest and XGBoost models, we leveraged advanced algorithms to provide accurate and reliable predictions, contributing to the growing field of data-driven healthcare.

Strengths

- Machine learning models like Random Forest and XGBoost are capable of capturing complex non-linear relationships in data, which is critical in medical applications.
- The inclusion of clinically relevant variables ensures the models remain interpretable and useful in real-world healthcare settings.

Limitations

- The dataset size and quality directly impact model performance. Any bias or missing data in the training set can lead to skewed predictions.
- Although the models performed well, their predictive power depends on the availability of accurate input features during deployment.
- The models are not a substitute for clinical diagnosis but should be used as a decision-support tool alongside clinical judgment.

This project underscores the potential of machine learning in healthcare, particularly in predicting heart failure risk. By leveraging models like Random Forest and XGBoost, we can offer a reliable approach for early risk identification in the future, paving the way for personalized medicine and better management of cardiovascular diseases. Despite the limitations, the outcomes of this study provide a strong foundation for further research and application, contributing to the broader goal of reducing CVD-related morbidity and mortality.

7 Future Work

- Further refinement of the models is recommended, including exploring other advanced machine learning techniques, such as neural networks or ensemble learning approaches.
- Expanding the dataset to include more diverse populations and additional clinical parameters could enhance model generalizability.
- Integration of these models into real-time clinical systems could provide immediate risk assessments, enabling proactive healthcare interventions.

8 References

- **Random Forest** Details about Random Forest model
- **XGBoostl** Details about XGBoost Model
- **matplotlib Library**
- **Seaborn Library**
- **plotly library** for interactive plots

8.1 Resources for the Project

The complete project, including code (python and latex), dataset and images and everything is available on my GitHub page:
Heartbeat Sentinel: Decoding and Predicting Heart Failure