Heart Disease Prediction Using Machine Learning Classification Models

***Abstract*—** According to the Centres for Disease Control and Prevention (CDC), heart disease is one of the leading causes of death for people in the United Sates, posing a heavy burden on society, families, and patients requiring early detection and intervention to mitigate its impact. This study explores the application of machine learning techniques to predict heart disease using data from the Behavioural Risk Factor Surveillance System (BRFSS) and employing seven classification models such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Gradient Boosting, XGBoost, and AdaBoost. The project investigates the relationship between various health and lifestyle factors and heart disease. Data preprocessing, variable encoding, and resampling were performed to address class imbalance and improve model performance. The models were evaluated using metrics such as ROC- AUC, F1-score, accuracy, precision, and recall.

***Keywords*—*Machine learning, Heart Disease Prediction, Gradient Boosting, Classification Models, Healthcare Analytics, Model building, Model Evolution*)**

Introduction

This project aims to develop a predictive model capable of identifying individuals at high risk of developing heart disease by employing advanced machine learning technique utilising data about individuals such as smoking habits, alcohol consumption, body mass index (BMI), and exercise habits collected by BRFSS. BRFSS is a collaborative project between all the states in the United States and participating US territories and CDC, which conducts annual telephone surveys to collect data on the health status of U.S residents. BRFSS completes more than 400,000 adult interviews each year making it the largest continuously conducted health survey system in the world . Heart disease prediction is crucial for healthcare professionals to take preventive actions at early stage. This project will explore the dataset answering the below questions, train and evaluate seven classification models, and determine the most effective algorithm for heart disease prediction.

- Does age distribution relate to heart disease?
- Does heart disease percentage differ between genders?
- Does alcohol consumption influence heart disease?
- Does smoking status impact heart disease?
- Does physical activity correlate with heart disease?
- Does General Health status influence heart disease?
- Does diabetic status influence heart disease?
- Does heart disease percentage differ between racial groups?
- How do machine learning models compare in predicting heart disease?

The data goes through a pre-processing stage which entails cleaning, normalising and splitting the dataset into sets for training and testing. The cleaned data will be explored and visualised to draw insights. Tableau dashboards will be employed to uncover connections and trends within the data. Seven machine learning models (XGBoost, Gradient Boosting, Random Forest, Decision Trees, K-Nearest Neighbours, and AdaBoost) will be employed for this project, each of these models offers different strengths and the project will focus on finding the best preforming models. Each model will be evaluated using performance measures such as ROC- AUC, F1-score, accuracy, precision, and recall. Python libraries (Pandas, Matplotlib, Seaborn and Scikit-learn) will be employed throughout this project.

Research Methodology

A. Data Description:

The dataset originally comes from the Behavioural Risk Factor Surveillance System and the dataset consists of 319,795 entries and 18 columns. It contains information related to individuals' health and lifestyle factors, including.

- **HeartDisease**: Has heart disease (Yes/No).

- **BMI**: Body Mass Index.
- **Smoking**: Whether the individual smokes (Yes/No).
- **AlcoholDrinking**: Whether the individual drinks alcohol excessively (Yes/No).
- **Stroke**: History of stroke (Yes/No).
- **PhysicalHealth**: Number of physically unhealthy days
- **MentalHealth**: Number of mentally unhealthy days
- **DiffWalking**: Difficulty walking or climbing stairs
- **Sex**: Gender of the individual.
- **AgeCategory**: Age group.
- **Race**: Race or ethnicity of the individual.
- **Diabetic**: Diabetes status.
- **PhysicalActivity**: Engagement in physical activity
- **GenHealth**: General health status
- **SleepTime**: Hours of sleep per day.
- **Asthma**: History of asthma (Yes/No).
- **KidneyDisease**: History of kidney disease (Yes/No).
- **SkinCancer**: History of skin cancer (Yes/No).

B. Data Preprocessing:

Prior to conducting the data analysis, the dataset went through several cleaning and preparation steps. No missing values were detected in the dataset, but 18078 duplicated records were found and removed to ensure data integrity.

Box plots were used to identify the spread of data and the presence of outliers for numeric variables. While the four plots are skewed especially Physical Health and Mental health, outliers in BMI, Physical Health and Mental health were retained as they are realistic values that could hold valuable insights. Outliers in SleepTime were removed as they were unrealistic values.

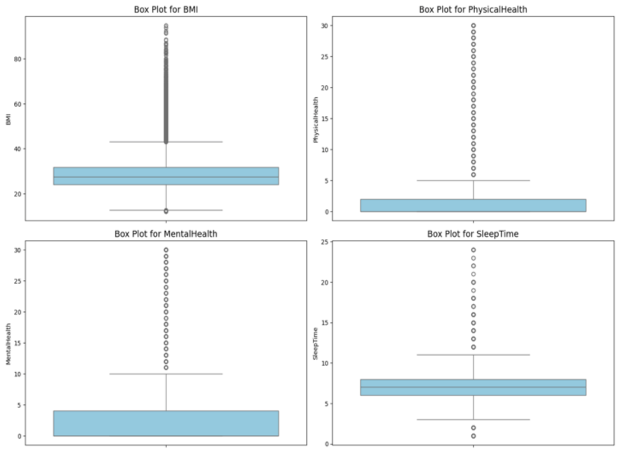


Figure 1. box plots to check for outliers

C. Exploratory Data Analysis:

Heart Disease Percentage in the Dataset

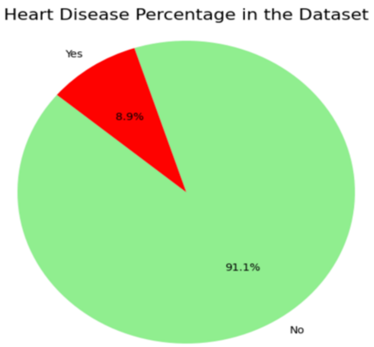


Figure 2. Heart Disease Percentage in the Dataset

This pie chart represents the proportion of individuals in the dataset who have heart disease versus those who don't. 8.9% of individuals with heart disease while 91.1% of individuals without heart disease. The low percentage of heart disease cases suggests that the dataset will require handling in case of machine learning model building as the target variable is imbalanced in the dataset.

Gender Percentage in the Dataset

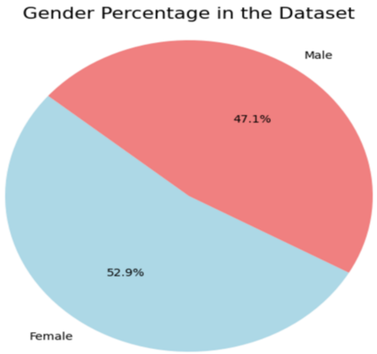


Figure 3. Gender Percentage in the Dataset

This pie chart shows the gender distribution in the dataset. The dataset has a nearly balanced distribution of genders with a slight majority of females 52.9% compared to males 47.1%. The difference between the two genders is relatively small indicating that the dataset has an equal representation of both categories.

Distribution of BMI

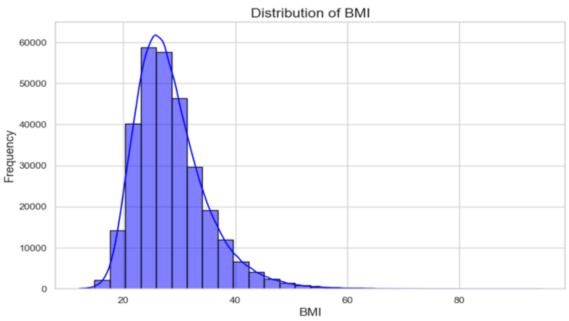


Figure 4. Distribution of BMI.

The BMI range appears to start around 10 and goes beyond 60 with most values falling in the middle with a clear peak around the BMI range of 20 to 30. The histogram is right skewed indication some individuals with higher BMI values.

Distribution of Sleep Time

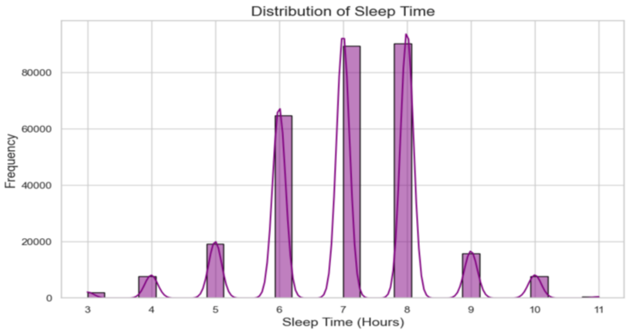


Figure 5. Distribution of Sleep Time

The number of sleeping hours in the dataset is ranging from approximately 3 to 11 hours specially after removing the unrealistic hours from the sleep time variable. This distribution of sleep time shows that most individuals sleep between 6 and 8 hours

Distribution of Mental Health

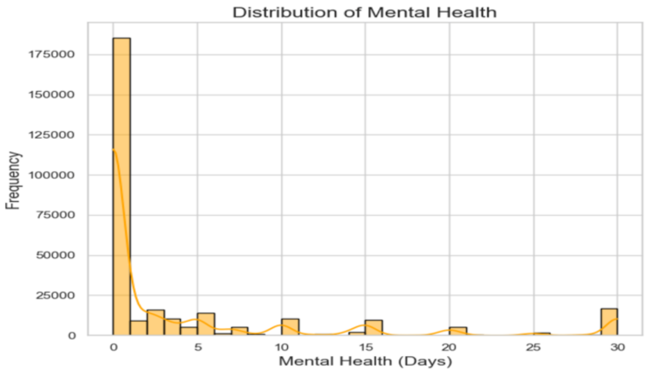


Figure 6. Distribution of Mental Health

The number of days individuals experienced poor mental health ranges from 0 to 30 days. Most individuals reporting 0 days of poor mental health and the frequency decreases massively as the number of days increases. There are some individuals experiencing significant mental health challenges especially those reporting more than 10 days of poor mental health.

Distribution of Physical Health

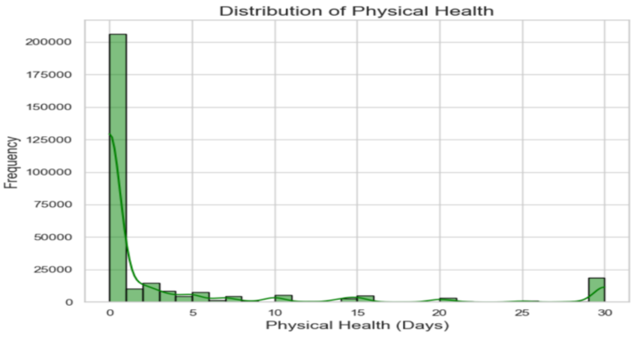


Figure 7. Distribution of Physical Health

The number of days individuals experienced poor physical health ranges from 0 to 30 days. Most individuals in the dataset appear to have good physical health with the majority reporting no poor physical health days. There are some individuals experiencing extended periods of poor physical health.

Heart Disease Count by Age Category

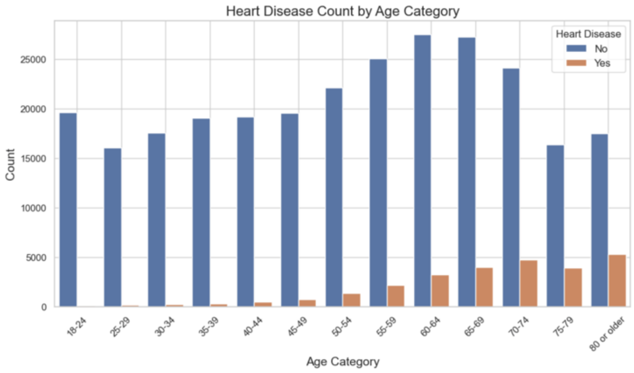


Figure 8. Heart Disease Count by Age Category

This bar chart shows the relationship between age categories and heart disease. It appears that the number of individuals with heart disease increases with age indicating that older individuals are more likely to have heart disease. Heart disease is noticeable starting from 50 to 54 age group and continues to go higher with older age groups.

Heart Disease Percentage by Sex

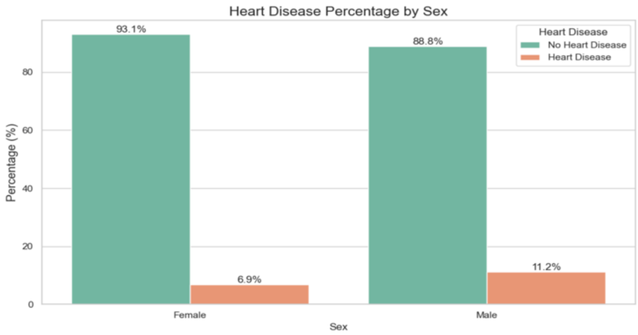


Figure 9. Heart Disease Percentage by Sex

It appears that 93.1% of females without heart disease and 6.9% with heart disease while for males 88.8% without heart disease and 11.2% with heart disease. Males have a higher percentage of heart disease cases 11.2% compared to females 6.9%.

Heart Disease Percentage by Smoking Status

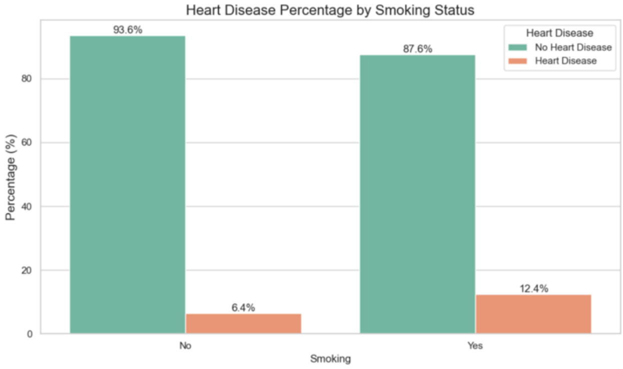


Figure 10. Heart Disease Percentage by Smoking Status

The chart clearly shows the increased risk of heart disease between smokers compared to non-smokers as smokers have almost double the percentage of heart disease cases 12.4% compared to non-smokers 6.4%.

Heart Disease Percentage by Alcohol Drinking Status

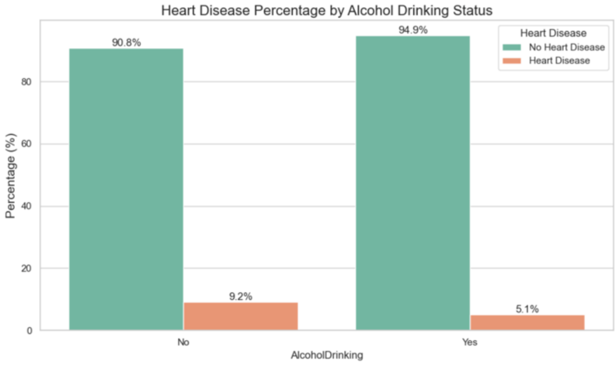


Figure 11. Heart Disease Percentage by Alcohol Drinking Status

A smaller percentage of drinkers have heart disease 5.1% compared to non-drinkers 9.2%. These results agree with some hypotheses about moderate alcohol consumption and type of Alcohol can improve heart health, but this should be interpreted cautiously. [10] [11]

Heart Disease Percentage by Physical Activity

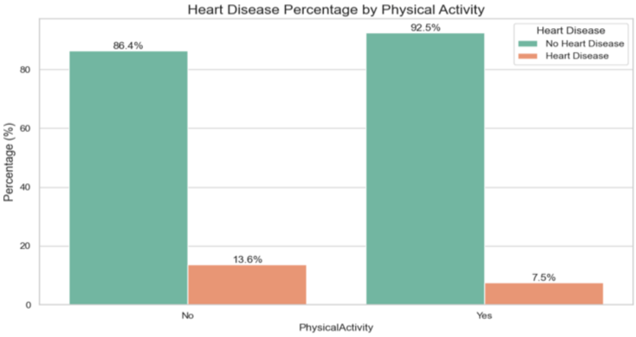


Figure 12. Heart Disease Percentage by Physical Activity

Physically active individuals have a lower prevalence of heart disease 7.5% compared to those who are not physically active 13.6%. The difference in heart disease percentage between physically active and inactive groups is 6.1% which strongly supports the idea that regular physical activity is associated with a reduced risk of heart disease

Heart Disease Percentage by General Health

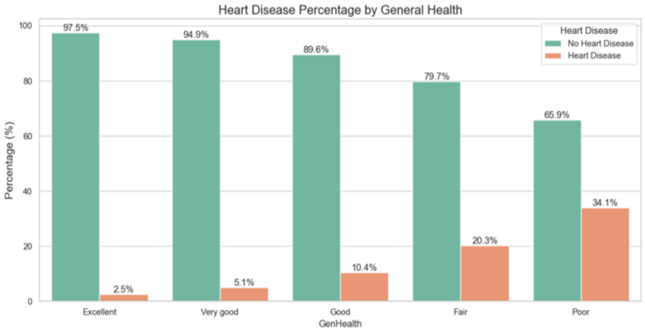


Figure 13. Heart Disease Percentage by General Health

As general health declines from Excellent to Poor, the percentage of individuals with heart disease increases significantly from 2.5% to 34.1%. The chart reflects the importance of maintaining good general health to lower heart disease risk.

Heart Disease Percentage by Race

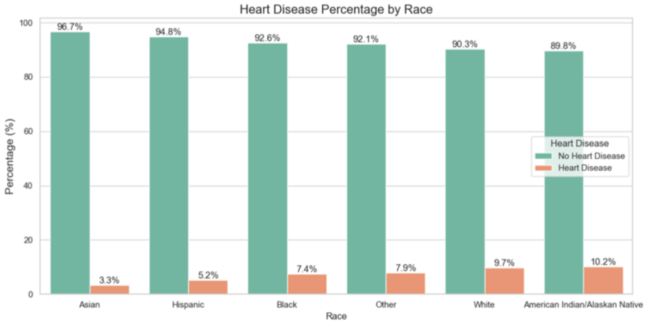


Figure 14. Heart Disease Percentage by Race

Asians have the lowest percentage of individuals with heart disease 3.3% then Hispanics 5.2%. American Indian/Alaskan Native group has the highest prevalence of heart disease 10.2% then Whites 9.7%. The Black 7.4% and Other 7.9% groups come in the middle between the highest and lowest groups.

Heart Disease Percentage by Diabetic Status

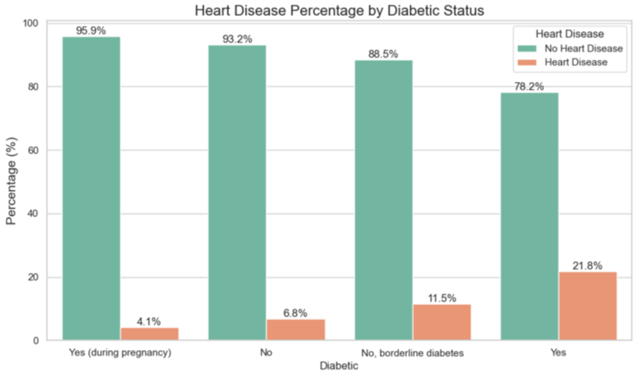


Figure 15. Heart Disease Percentage by Diabetic Status

Individuals with diabetes only during pregnancy have the lowest prevalence of heart disease 4.1%. Individuals with diabetes have the highest heart disease prevalence 21.8% followed by those with borderline diabetes 11.5%. The data strongly suggest that diabetes is a big risk factor for heart disease.

Tableau Dashboard

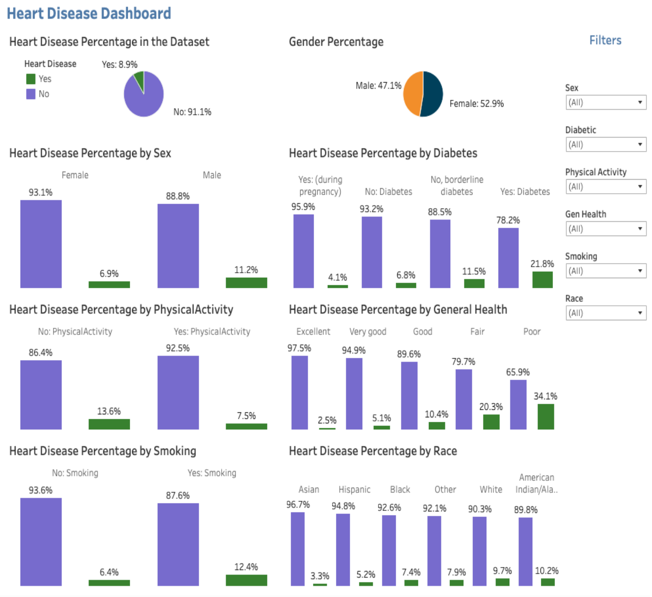


Figure 16. Tableau Dashboard

This is an interactive dashboard that provides an overview of heart disease percentages based on different demographic and lifestyle factors. It highlights key factors like gender, diabetes status, physical activity, general health, smoking, and race showing how heart disease correlates with these factors.

D. Data Preparation and Modelling Building:

D.1 Variable Encoding:

Label Encoding was used to encode binary columns ('HeartDisease', 'Smoking', 'AlcoholDrinking', 'Stroke', 'DiffWalking','Sex', 'PhysicalActivity', 'Asthma', 'KidneyDisease', 'SkinCancer') to 0 and 1. Label encoding is used here as it provides a simple way to map categorical variables to numeric variables. Ordinal Encoding was employed to encode ordinal columns ('GenHealth', 'AgeCategory') with numeric values that preserve the oder of ordinal values. OneHot encoding was employed to encode nominal columns ('Race', 'Diabetic') creating separate binary columns for each category. This method is used to prevent the model from interpreting nominal data as ordinal data. The downside of OneHot encoding is it increase dataset dimensionality which can impose computational challenges.

D.2 Correlation Analysis:

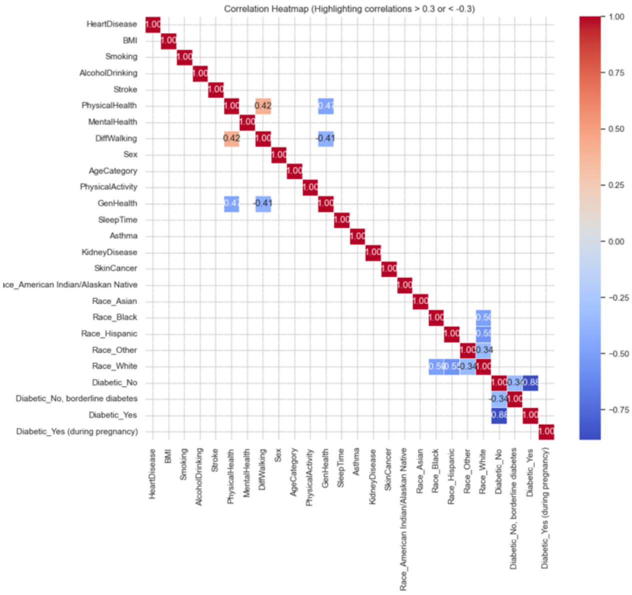


Figure 17. Correlation Analysis

The correlation of each independent variable was checked in relation to dependent variable 'HeartDisease'. Variables 'AgeCategory' and 'GenHealth' showed the strongest relationship to the dependent variable with 0.239 and -0.235 respectively. Most variable showed very weak correlation to the dependent variable suggesting limited direct influence on the target variable.

Some variable showed a notable positive correlation like PhysicalHealth and DiffWalking 0.42, MentalHealth and PhysicalHealth 0.42, GenHealth and AgeCategory 0.41 while other variable showed a notable negative correlation like GenHealth and PhysicalActivity -0.41 but overall, they are all weak correlations. The remaining correlations are either > 0.3 or < -0.3.

D.3 Data Scaling:

As the range for numeric variables varied widely, StandardScaler was used to scale the data. StandardScaler ensures that the features have a mean of 0 and a standard deviation of 1, this improves Model speed and performance. The data was split first to training set and Test set. The Scaler was fit and transformed on the training dataset and then the test dataset was only transformed so there is no data leakage between the training set and the test set to maintain the integrity of evolution process.

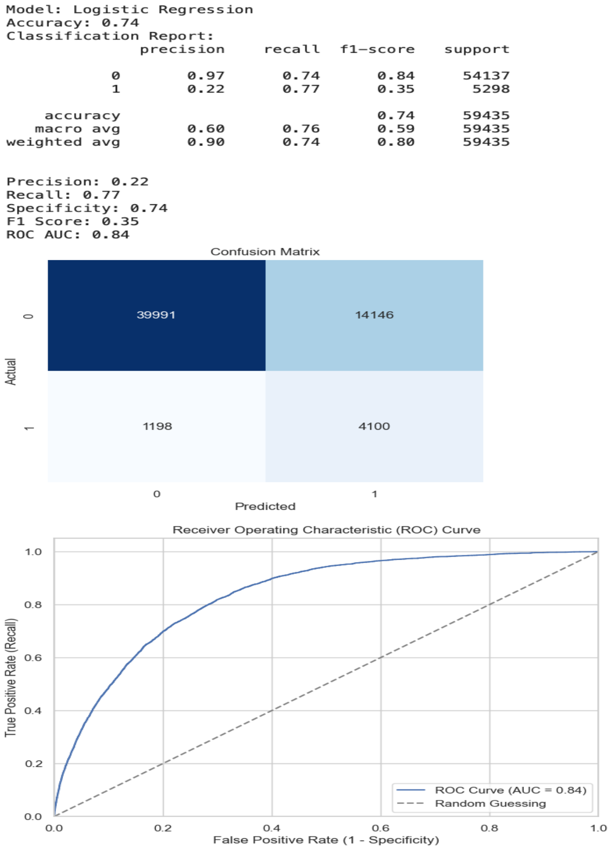
D.4 Data Resampling:

As the dataset was highly imbalanced, resampling methods were used on the training set. Over Sampling, Under Sampling and SMOTE methods were used to resample the training set. Gradient Boosting Classifier Model was trained on the three different samples and evaluated to decide on best resampling methods. Over Sampling and Under Sampling methods rendered best results and Under Sampling method was selected to reduce the training time.

E. Model Evolution:

These evaluation metrics Accuracy, precision, recall, f1-score, Specificity and ROC AUC are used to evaluate each classification model.

E.1 Logistic Regression:



```
Model: Logistic Regression
Accuracy: 0.74
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.74      0.84     54137
           1       0.22      0.77      0.35      5298

    accuracy                           0.74     59435
   macro avg       0.60      0.76      0.59     59435
weighted avg       0.90      0.74      0.80     59435

Precision: 0.22
Recall: 0.77
Specificity: 0.74
F1 Score: 0.35
ROC AUC: 0.84
```

Figure 18. LogisticRegression

Accuracy 0.74: The overall percentage of correctly predicted cases in the dataset is 74%.

Precision 0.22: Out of all predicted “Has Heart Disease" cases only 22% were correct which indicates a high false positive rate.

Recall 0.77: The model correctly predicted 77% of actual “Has Heart Disease" cases.

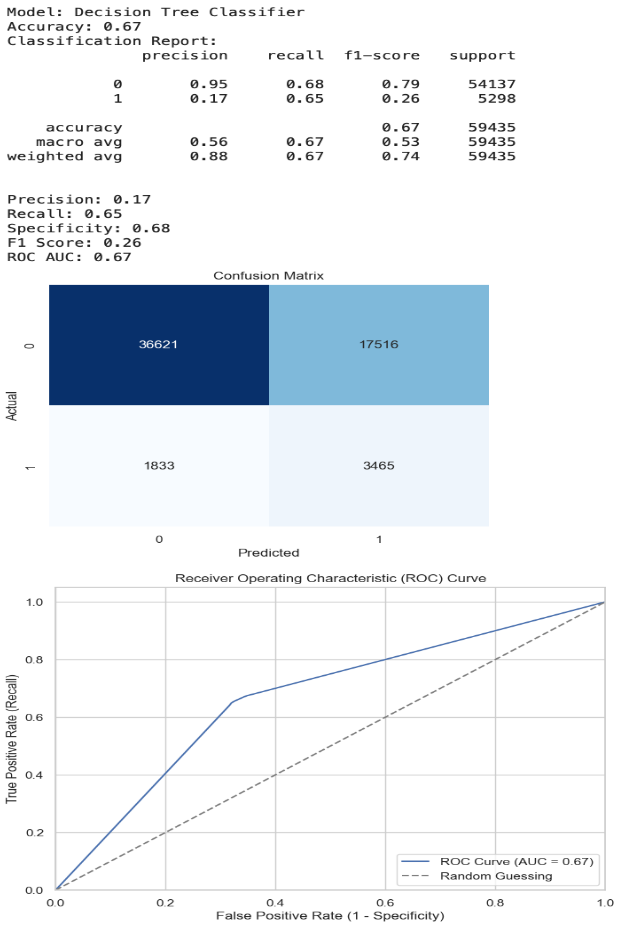
Specificity 0.74: The model correctly predicted 77% of actual “No Heart Disease" cases.

F1-Score 0.35: A low F1-score indicates poor balance between precision and recall for this model.

The AUC of 0.84 shows the model’s balance between True Positive Rate and False Positive Rate.

The model performs very well predicting negative cases 97% but it struggles predicting positive cases 22%. This highlights a struggle with class imbalance shown by a high number of false positives 14146 and false negatives 1198. The model can only detect 77% of actual positive cases which means that 23% of individual with heart disease will be misdiagnosed as "No Heart Disease”. The model shows strong ROC AUC of 0.84 indicating good overall balance between True Positive Rate and False Positive Rate.

E.2 Decision Tree Classifier:



```
Model: Decision Tree Classifier
Accuracy: 0.67
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.68      0.79     54137
           1       0.17      0.65      0.26      5298

    accuracy                           0.67     59435
   macro avg       0.56      0.67      0.53     59435
weighted avg       0.88      0.67      0.74     59435

Precision: 0.17
Recall: 0.65
Specificity: 0.68
F1 Score: 0.26
ROC AUC: 0.67
```

Figure 19. Decision Tree Classifier

Accuracy 0.67: The overall percentage of correctly predicted cases in the dataset is 67%.

Precision 0.17: Out of all predicted “Has Heart Disease" cases only 17% were correct which indicates a high false positive rate.

Recall 0.65: The model correctly predicted 0.65% of actual “Has Heart Disease" cases.

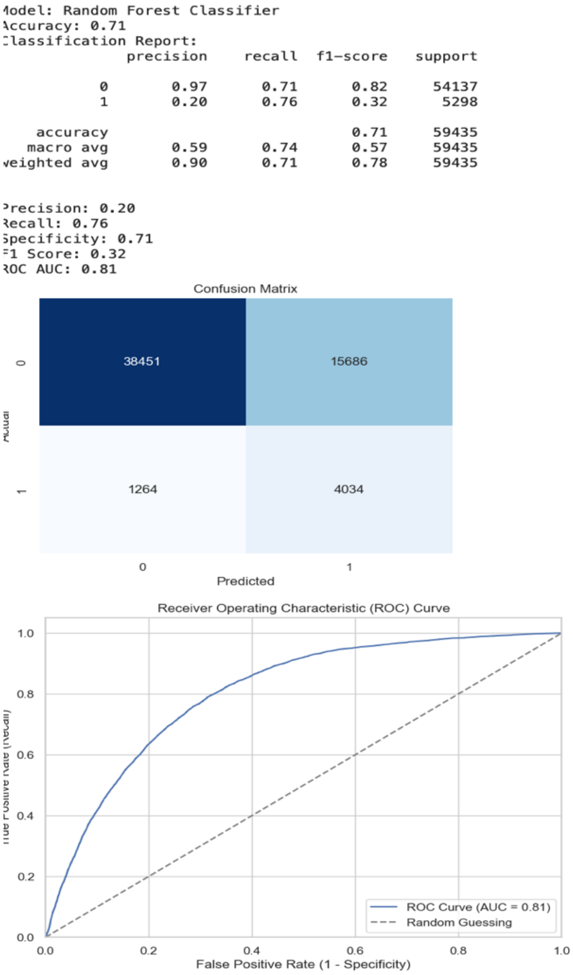
Specificity 0.68: The model correctly predicted 68% of actual “No Heart Disease" cases.

F1-Score 0.26: A low F1-score indicates poor balance between precision and recall for this model.

The AUC of 0.67 shows the model’s balance between True Positive Rate and False Positive Rate.

The model performs very well predicting negative cases 95% but it struggles predicting positive cases 17%. This highlights a struggle with class imbalance shown by a high number of false positives 17516 and false negatives 1833. The model can only detect 65% of actual positive cases which means that 35% of individual with heart disease will be misdiagnosed as "No Heart Disease”. The model shows a weak ROC AUC of 0.67 indicating a weak overall balance between True Positive Rate and False Positive Rate.

E.3 Random Forest Classifier:



```
1odel: Random Forest Classifier
\ccuracy: 0.71
:lassification Report:
              precision    recall  f1-score   support

           0       0.97      0.71      0.82     54137
           1       0.20      0.76      0.32      5298

    accuracy                           0.71     59435
   macro avg       0.59      0.74      0.57     59435
veighted avg       0.90      0.71      0.78     59435

?recision: 0.20
Recall: 0.76
Specificity: 0.71
=1 Score: 0.32
ROC AUC: 0.81
```

Figure 20. Random Forest Classifier

Accuracy 0.71: The overall percentage of correctly predicted cases in the dataset is 71%.

Precision 0.20: Out of all predicted “Has Heart Disease" cases only 20% were correct which indicates a high false positive rate.

Recall 0.76: The model correctly predicted 76 % of actual “Has Heart Disease" cases.

Specificity 0.71: The model correctly predicted 71% of actual “No Heart Disease" cases.

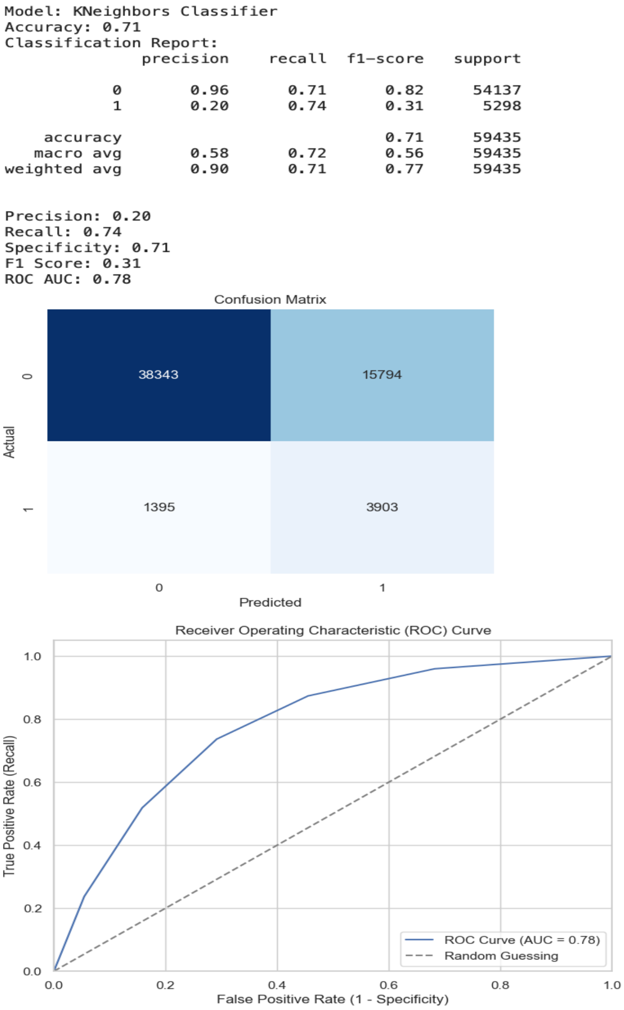
F1-Score 0.32: A low F1-score indicates poor balance between precision and recall for this model.

The AUC of 0.81 shows the model’s balance between True Positive Rate and False Positive Rate.

The model performs very well predicting negative cases 97% but it struggles predicting positive cases 20%. This highlights a struggle with class imbalance shown by a high number of false positives 15686 and false negatives

1264. The model can only detect 76% of actual positive cases which means that 24% of individual with heart disease will be misdiagnosed as "No Heart Disease”. The model shows good ROC AUC of 0.81 indicating good overall balance between True Positive Rate and False Positive Rate.

E.4 KNeighbors Classifier:



```
Model: KNeighbors Classifier
Accuracy: 0.71
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.71      0.82     54137
           1       0.20      0.74      0.31      5298

    accuracy                           0.71     59435
   macro avg       0.58      0.72      0.56     59435
weighted avg       0.90      0.71      0.77     59435

Precision: 0.20
Recall: 0.74
Specificity: 0.71
F1 Score: 0.31
ROC AUC: 0.78
```

Figure 21. KNeighbors Classifier

Accuracy 0.71: The overall percentage of correctly predicted cases in the dataset is 71%.

Precision 0.20: Out of all predicted “Has Heart Disease" cases only 20% were correct which indicates a high false positive rate.

Recall 0.74: The model correctly predicted 74% of actual “Has Heart Disease" cases.

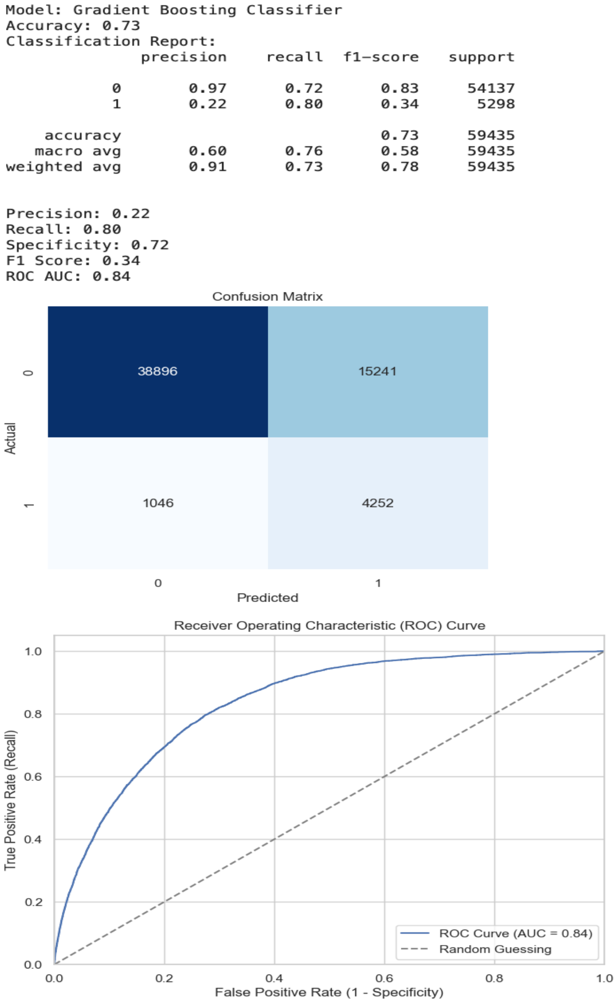
Specificity 0.71: The model correctly predicted 71% of actual “No Heart Disease" cases.

F1-Score 0.31: A low F1-score indicates poor balance between precision and recall for this model.

The AUC of 0.78 shows the model’s balance between True Positive Rate and False Positive Rate.

The model performs very well predicting negative cases 96% but it struggles predicting positive cases 20%. This highlights a struggle with class imbalance shown by a high number of false positives 15794 and false negatives 1395. The model can only detect 74% of actual positive cases which means that 26% of individual with heart disease will be misdiagnosed as "No Heart Disease”. The model shows weak ROC AUC of 0.78 indicating weak overall balance between True Positive Rate and False Positive Rate.

E.5 Gradient Boosting Classifier:



```
Model: Gradient Boosting Classifier
Accuracy: 0.73
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.72      0.83     54137
           1       0.22      0.80      0.34      5298

    accuracy                           0.73     59435
   macro avg       0.60      0.76      0.58     59435
weighted avg       0.91      0.73      0.78     59435

Precision: 0.22
Recall: 0.80
Specificity: 0.72
F1 Score: 0.34
ROC AUC: 0.84
```

Figure 22. Gradient Boosting Classifier

Accuracy 0.73: The overall percentage of correctly predicted cases in the dataset is 73%.

Precision 0.22: Out of all predicted “Has Heart Disease" cases only 22% were correct which indicates a high false positive rate.

Recall 0.80: The model correctly predicted 80% of actual “Has Heart Disease" cases.

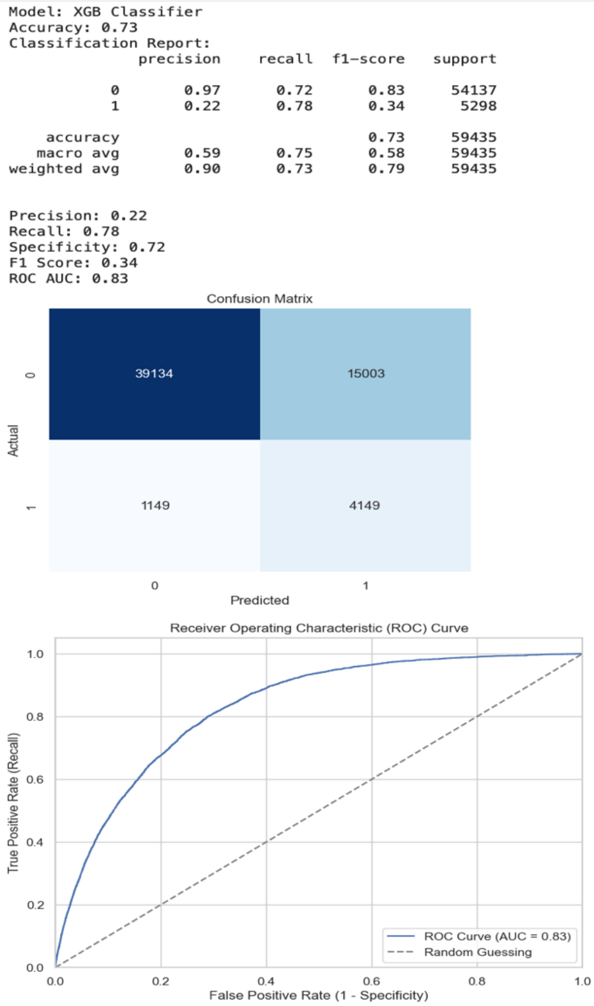
Specificity 0.72: The model correctly predicted 72% of actual “No Heart Disease" cases.

F1-Score 0.34: A low F1-score indicates poor balance between precision and recall for this model.

The AUC of 0.84 shows the model’s balance between True Positive Rate and False Positive Rate.

The model performs very well predicting negative cases 97% but it struggles predicting positive cases 22%.This highlights a struggle with class imbalance shown by a high number of false positives 15241 and false negatives 1046. The model is able to detect 80% of actual positive cases which means that 20% of individual with heart disease will be misdiagnosed as "No Heart Disease”. The model shows strong ROC AUC of 0.84 indicating good overall balance between True Positive Rate and False Positive Rate.

E.6 XGBoost Classifier:



```
Model: XGB Classifier
Accuracy: 0.73
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.72      0.83     54137
           1       0.22      0.78      0.34      5298

    accuracy                           0.73     59435
   macro avg       0.59      0.75      0.58     59435
weighted avg       0.90      0.73      0.79     59435

Precision: 0.22
Recall: 0.78
Specificity: 0.72
F1 Score: 0.34
ROC AUC: 0.83
```

Figure 23. XGBoost Classifier

Accuracy 0.73: The overall percentage of correctly predicted cases in the dataset is 73%.

Precision 0.22: Out of all predicted “Has Heart Disease" cases only 22% were correct which indicates a high false positive rate.

Recall 0.78: The model correctly predicted 78% of actual “Has Heart Disease" cases.

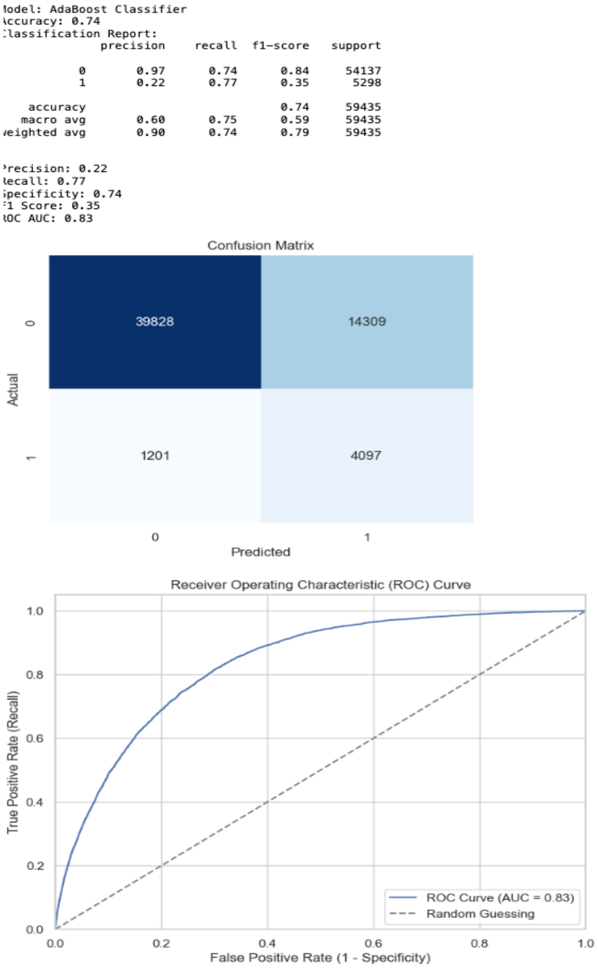
Specificity 0.72: The model correctly predicted 72% of actual “No Heart Disease" cases.

F1-Score 0.34: A low F1-score indicates poor balance between precision and recall for this model.

The AUC of 0.83 shows the model’s balance between True Positive Rate and False Positive Rate.

The model performs very well predicting negative cases 97% but it struggles predicting positive cases 22%. This highlights a struggle with class imbalance shown by a high number of false positives 4149 and false negatives 1149. The model can only detect 78% of actual positive cases which means that 22% of individual with heart disease will be misdiagnosed as "No Heart Disease”. The model shows strong ROC AUC of 0.83 indicating good overall balance between True Positive Rate and False Positive Rate.

E.7 AdaBoost Classifier:



```
Model: AdaBoost Classifier
Accuracy: 0.74
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.74      0.84     54137
           1       0.22      0.77      0.35      5298

    accuracy                           0.74     59435
   macro avg       0.60      0.75      0.59     59435
weighted avg       0.90      0.74      0.79     59435

Precision: 0.22
Recall: 0.77
Specificity: 0.74
F1 Score: 0.35
ROC AUC: 0.83
```

Figure 24. AdaBoost Classifier

Accuracy 0.74: The overall percentage of correctly predicted cases in the dataset is 74%.

Precision 0.22: Out of all predicted “Has Heart Disease" cases only 22% were correct which indicates a high false positive rate.

Recall 0.77: The model correctly predicted 77% of actual “Has Heart Disease" cases.

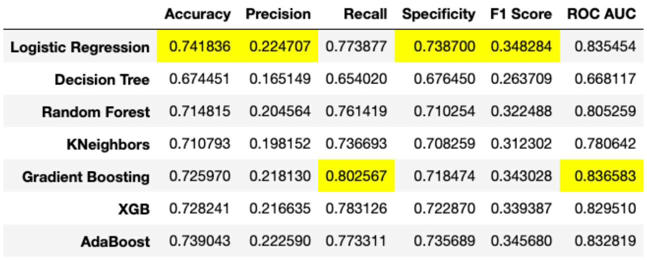
Specificity 0.74: The model correctly predicted 74% of actual “No Heart Disease" cases.

F1-Score 0.35: A low F1-score indicates poor balance between precision and recall for this model.

The AUC of 0.83 shows the model’s balance between True Positive Rate and False Positive Rate.

The model performs very well predicting negative cases 74% but it struggles predicting positive cases 22%. This highlights a struggle with class imbalance shown by a high number of false positives 14309 and false negatives 1201. The model can only detect 77% of actual positive cases which means that 33% of individual with heart disease will be misdiagnosed as "No Heart Disease”. The model shows strong ROC AUC of 0.83 indicating good overall balance between True Positive Rate and False Positive Rate.

IV.Conclusion



	Accuracy	Precision	Recall	Specificity	F1 Score	ROC AUC
Logistic Regression	0.741836	0.224707	0.773877	0.738700	0.348284	0.835454
Decision Tree	0.674451	0.165149	0.654020	0.676450	0.263709	0.668117
Random Forest	0.714815	0.204564	0.761419	0.710254	0.322488	0.805259
KNeighbors	0.710793	0.198152	0.736693	0.708259	0.312302	0.780642
Gradient Boosting	0.725970	0.218130	0.802567	0.718474	0.343028	0.836583
XGB	0.728241	0.216635	0.783126	0.722870	0.339387	0.829510
AdaBoost	0.739043	0.222590	0.773311	0.735689	0.345680	0.832819

Figure 25. Medal Evaluation Results

This study demonstrates the potential of machine learning to predict heart disease risk using demographic, health, and lifestyle data. While Logistic Regression Model renders highest accuracy, precession, specificity and F1 score it’s recommended to use Gradient Boosting as it renders the highest recall 0.80 and ROC AUC 0.84 which is crucial in Heath Care domain where missing a diagnosis can have severe consequences. Minimizing the number of false negatives is critical for predicting heart disease, while this might result in higher false positives, but these cases be investigated further by doing extra examinations and health checks. Gradient Boosting is rendering 20% false negative and its very high percentage for the model to be deployed in practical application but it’s the best performing model between the discussed models. The author explored different resampling methods, different feature selection and threshold adjustments for models but none of these steps rendered better performance, that’s why further research is required to optimize these models for real world applications. Future work could explore capturing additional features, advanced ML techniques or hybrid approaches to improve predictive accuracy.