Onur Kula

Final Report:

Heart Disease Prediction

Problem Statement

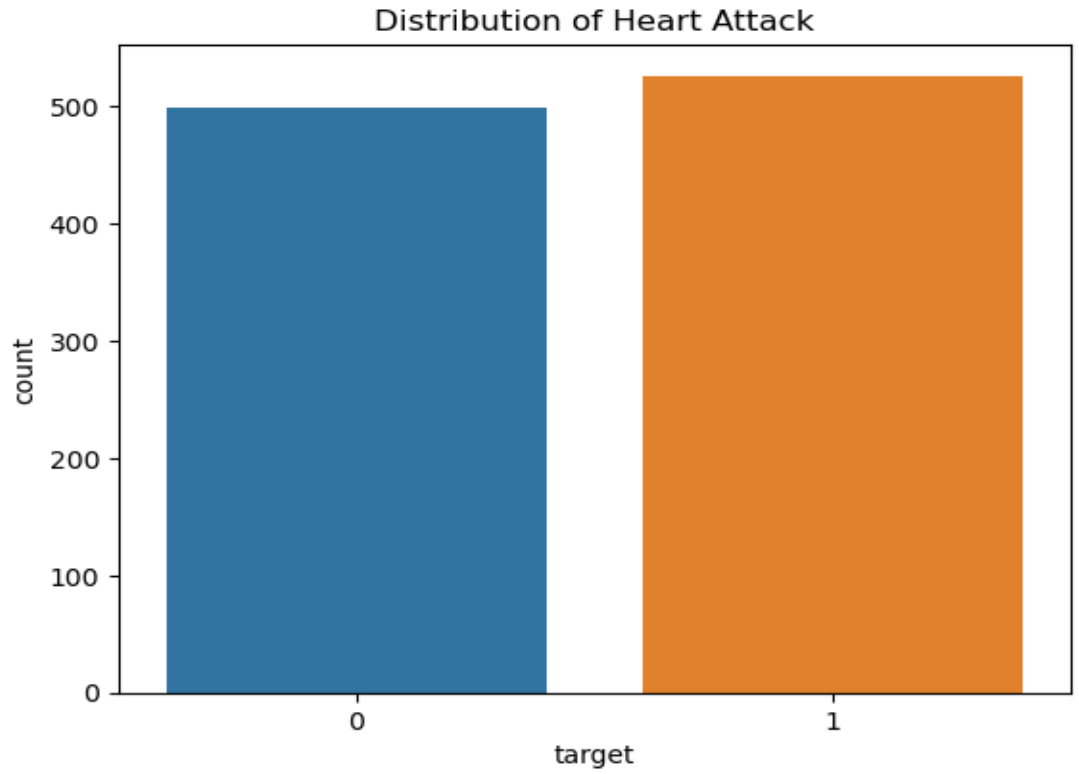
The primary objective of this project is to develop an effective predictive model for diagnosing heart disease within a 2-week timeline. Utilizing a comprehensive heart disease dataset curated from five popular sources (Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog), recorded over various periods, the project aims to create a robust predictive model. This model will be deemed successful if the predictions closely align with the real data, achieving an accuracy rate of 85% or more. Key variables, including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise induced angina, oldpeak (ST depression), the slope of the peak exercise ST segment, and the target class (presence or absence of heart disease), will be crucial in achieving this level of precision. The success criteria emphasize the importance of providing accurate and reliable predictions to contribute to a comprehensive understanding of factors influencing heart disease and enhancing early diagnosis and treatment.

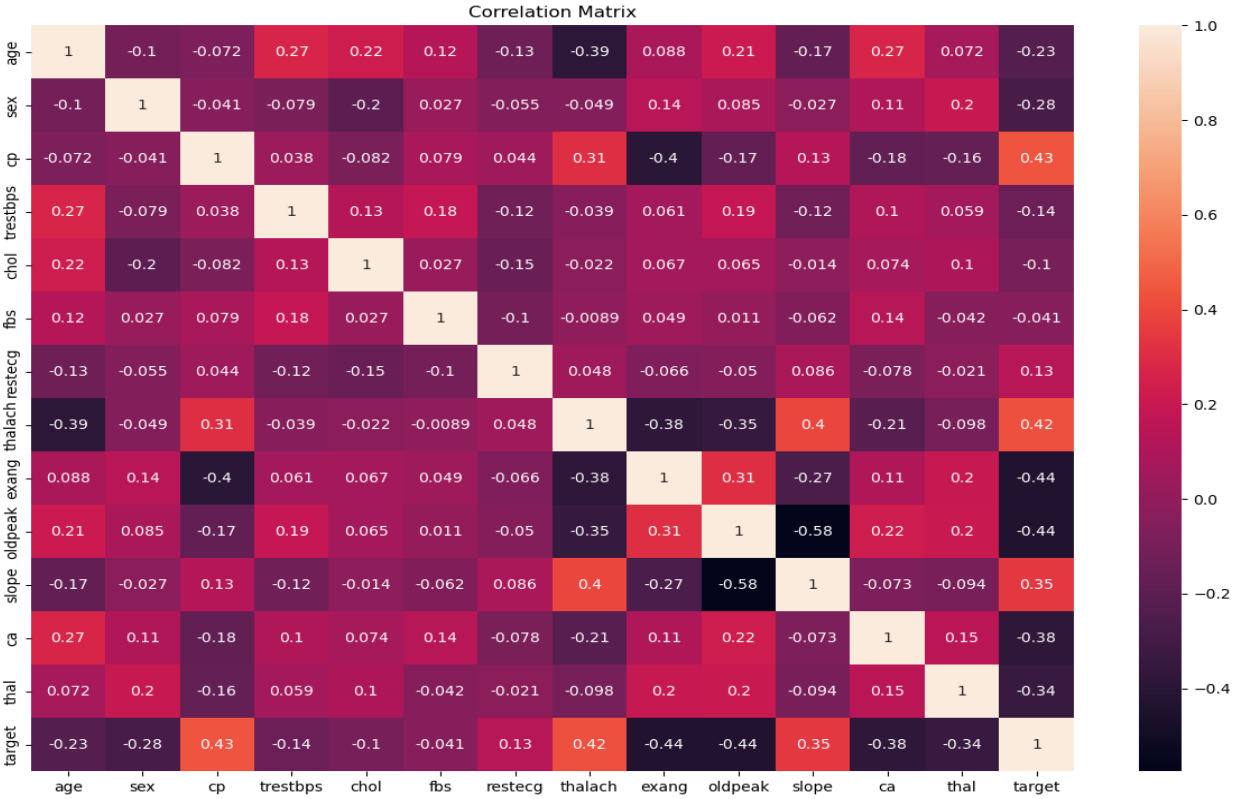
Data Wrangling

The heart disease dataset comprises various measurements from five different heart disease databases. The combined dataset contains 1025 rows and 13 features, including patient demographics, clinical measurements, and test results. Data wrangling steps included:

- Checking for and handling missing values
- Data type verification and conversion if necessary
- Removal of entries with impossible or erroneous values

After performing the necessary data wrangling steps, the refined dataset was prepared for analysis.


Distribution of Heart Attack
count
target
0
1
0
100
200
300
400
500


Correlation Matrix
age 1 -0.1 -0.072 0.27 0.22 0.12 -0.13 -0.39 0.088 0.21 -0.17 0.27 0.072 -0.23
sex -0.1 1 -0.041 -0.079 -0.2 0.027 -0.055 -0.049 0.14 0.085 -0.027 0.11 0.2 -0.28
cp -0.072 -0.041 1 0.038 -0.082 0.079 0.044 0.31 -0.4 -0.17 0.13 -0.18 -0.16 0.43
trestbps 0.27 -0.079 0.038 1 0.13 0.18 -0.12 -0.039 0.061 0.19 -0.12 0.1 0.059 -0.14
chol 0.22 -0.2 -0.082 0.13 1 0.027 -0.15 -0.022 0.067 0.065 -0.014 0.074 0.1 -0.1
fbs 0.12 0.027 0.079 0.18 0.027 1 -0.1 -0.0089 0.049 0.011 -0.062 0.14 -0.042 -0.041
restecg -0.13 -0.055 0.044 -0.12 -0.15 -0.1 1 0.048 -0.066 -0.05 0.086 -0.078 -0.021 0.13
thalach -0.39 -0.049 0.31 -0.039 -0.022 -0.0089 0.048 1 -0.38 -0.35 0.4 -0.21 -0.098 0.42
exang 0.088 0.14 -0.4 0.061 0.067 0.049 -0.066 -0.38 1 0.31 -0.27 0.11 0.2 -0.44
oldpeak 0.21 0.085 -0.17 0.19 0.065 0.011 -0.05 -0.35 0.31 1 -0.58 0.22 0.2 -0.44
slope -0.17 -0.027 0.13 -0.12 -0.014 -0.062 0.086 0.4 -0.27 -0.58 1 -0.073 -0.094 0.35
ca 0.27 0.11 -0.18 0.1 0.074 0.14 -0.078 -0.21 0.11 0.22 -0.073 1 0.15 -0.38
thal 0.072 0.2 -0.16 0.059 0.1 -0.042 -0.021 -0.098 0.2 0.2 -0.094 0.15 1 -0.34
target -0.23 -0.28 0.43 -0.14 -0.1 -0.041 0.13 0.42 -0.44 -0.44 0.35 -0.38 -0.34 1
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Pre-processing and Training Data Development

In this phase:

- Features were extracted from the dataset, excluding the target variable for analysis.
- StandardScaler was employed to standardize the dataset.
- The data were split into training and testing subsets, with the target variable designated as "target". This partitioning scheme was chosen to facilitate the prediction of heart disease presence.

Model Selection

Various classification models were trained and evaluated for predicting the presence of heart disease:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Classifier (SVM)
- K-Nearest Neighbors (KNN)
- Gradient Boosting Classifier
- XGBoost Classifier
- AdaBoost Classifier
- Naive Bayes Classifier
- MLP Neural Network

Each model was evaluated based on its accuracy in predicting the target variable on the test set.

Logistic Regression

The Logistic Regression model achieved an accuracy score of approximately 0.80 on the test data.

Decision Tree

The Decision Tree model achieved an accuracy score of approximately 0.99 on the test data.

Random Forest

The Random Forest model achieved an accuracy score of approximately 0.99 on the test data.

Support Vector Machine (SVM)

The SVM model achieved an accuracy score of approximately 0.89 on the test data.

K-Nearest Neighbors (KNN)

The KNN model achieved an accuracy score of approximately 0.83 on the test data.

Gradient Boosting

The Gradient Boosting model achieved an accuracy score of approximately 0.93 on the test data.

XGBoost

The XGBoost model achieved an accuracy score of approximately 0.99 on the test data.

AdaBoost

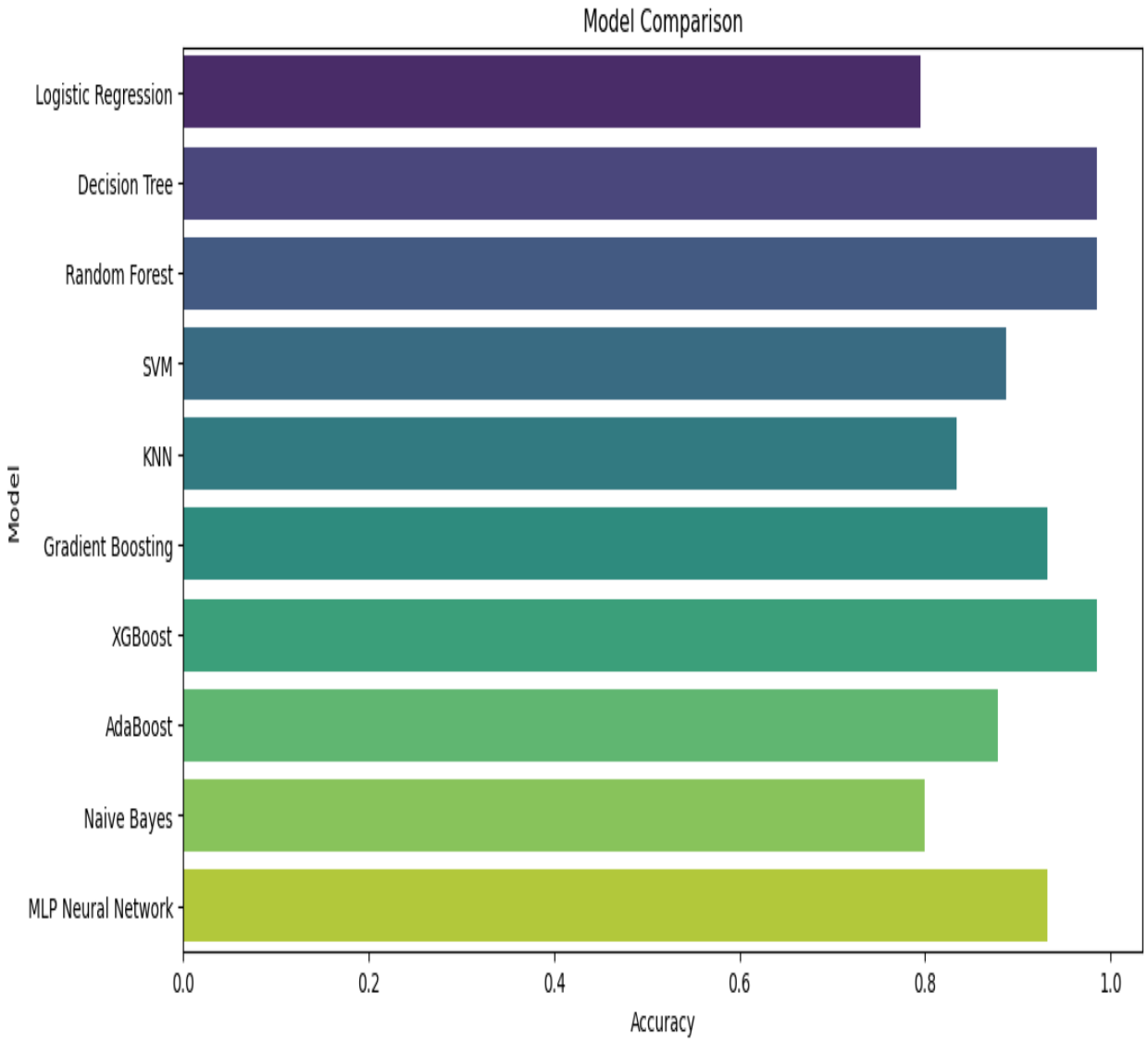
The AdaBoost model achieved an accuracy score of approximately 0.88 on the test data.

Naive Bayes

The Naive Bayes model achieved an accuracy score of approximately 0.80 on the test data.

MLP Neural Network

The MLP Neural Network model achieved an accuracy score of approximately 0.93 on the test data.



Conclusion

This project aimed to develop and compare the performance of various machine learning models in predicting the presence of heart disease based on a combined dataset from multiple sources. The results showed varying levels of performance across the models.

- The Random Forest Classifier, Decision Tree, and XGBoost Classifier outperformed other models with the highest accuracy.
- Logistic Regression and Naive Bayes had lower accuracy but provided valuable insights into the linearity of the data.

Future Directions

While the current models provided valuable insights and reasonably high accuracy, there are several areas for improvement and further exploration:

- **Feature Engineering:** Explore and engineer additional features that might contribute to the predictive power of the models. This could include interaction terms, polynomial features, or domain-specific features derived from the existing data.
- **Hyperparameter Tuning:** Further fine-t

-tuning of model hyperparameters using techniques such as GridSearchCV or RandomizedSearchCV to potentially improve model performance.

- **Cross-Validation:** Implement cross-validation techniques to ensure model robustness and generalizability.
- **Ensemble Methods:** Investigate combining multiple models to create ensemble methods that might yield better performance.
- **External Data Sources:** Incorporate additional data sources such as patient medical history, lifestyle factors, or genetic data to enhance model prediction capabilities.