Executive Summary

Introduction

As one of the leading causes of global mortality, detecting heart disease in its early stages is pivotal for enhancing patient outcomes and halting its progression. The conventional diagnostic methods often come with substantial costs and time requirements. Thus, there exists a pressing need for a cutting-edge predictive model that can evaluate the risk of heart disease utilizing easily accessible patient information.

Objective

The objective of this project is to design and build a predictive model capable of accurately determining the probability of an individual having heart disease. The focus is on leveraging machine learning techniques to create a model that can analyze relevant features and provide reliable predictions. The model should demonstrate high accuracy and generalizability, ensuring its effectiveness on new, unseen data.

Significance

Solving this challenge holds great significance. Early prediction aids in timely intervention and prevention, optimizing healthcare resources. The cost-effectiveness of predictive models reduces unnecessary procedures, benefiting patients and healthcare systems. Targeted public health efforts can stem from aggregated data, and research also advances through model insights. Ultimately, this challenge addresses a crucial healthcare issue using advanced machine learning, promising improved interventions and better health outcomes.

Dataset

The dataset used for this project is a folder containing both training and testing dataset. The training data contains 7303 rows in total while testing contains 2697 rows. The features include the following;

- age
- sex
- chest pain type (4 values)
- resting blood pressure

- serum cholestoral in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

Data Preprocessing

The data was cleaned and preprocessed. For example, columns like resting blood pressure, age and serum cholestoral in mg/dl were converted to categorical data. Also, columns like maximum heart rate achieved and oldpeak = ST depression induced by exercise relative to rest were scaled.

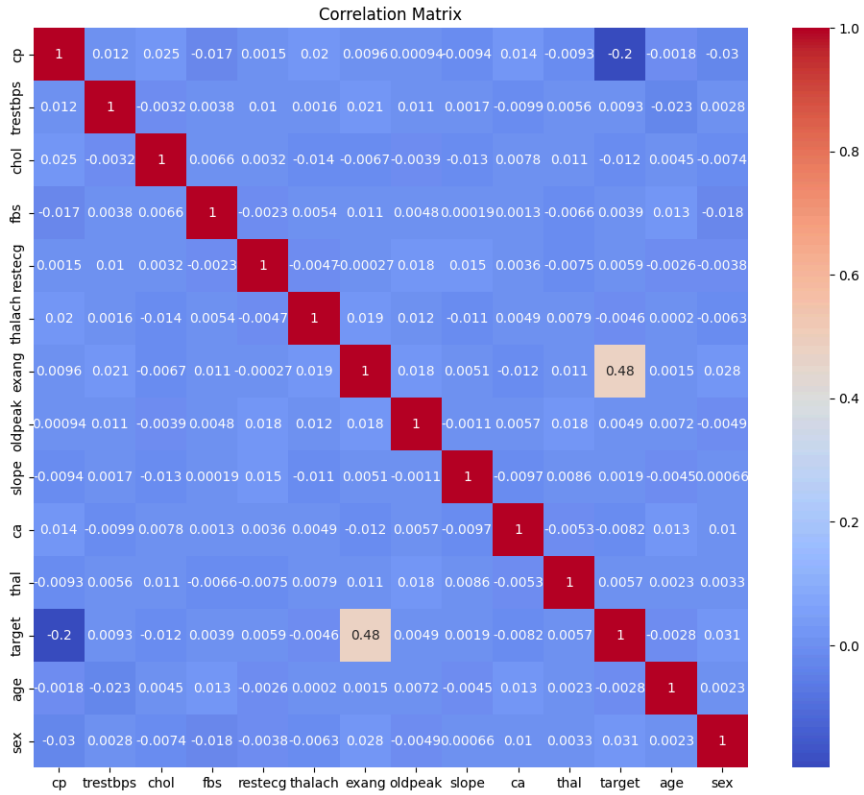

Fig 1 Correlation Matrix

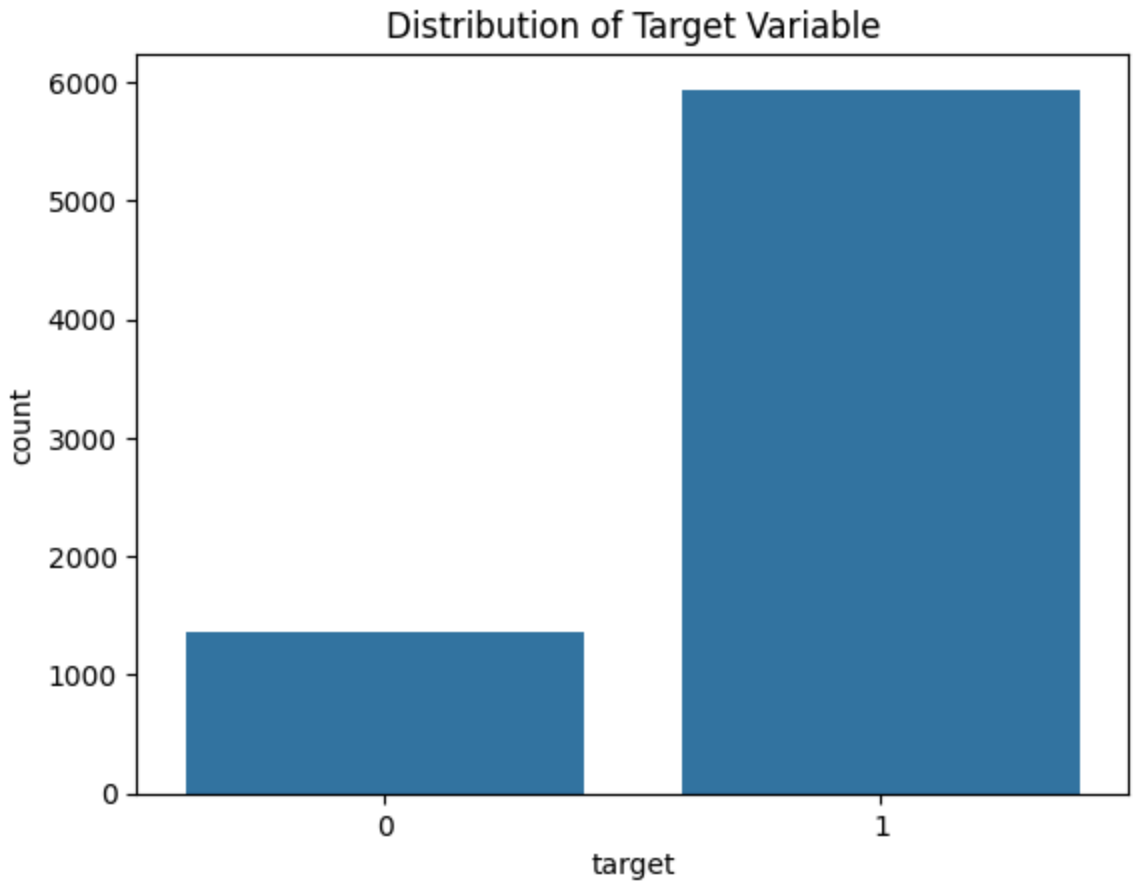

Fig 2 Target distribution

EDA

To understand the dataset, correlation and distributions were observed. It was noticed that 5941 of the patients were likely to have heart disease while 1362 didn't. It was also noted that exercise induced angina had the highest positive correlation with having heart diseases or not.

Feature Selection

The top 8 features were selected by applying SelectKBest with ANOVA F-value as the scoring function. These features were;

- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholestoral in mg/dl

- resting electrocardiographic results (values 0,1,2)
- exercise induced angina
- number of major vessels (0-3) colored by flourosopy

Model Building

Stacking ensemble was used on three models -lightgbm, catboost and logistic regression- in which logistic regression was the final estimator. This ensemble was trained on the training data and tested on the testing data.

Evaluation

Accuracy was used to evaluate the model. Our ensemble model gave an accuracy of 0.82.

Conclusion

Ensemble technique on machine learning models suggests an improved performance for this project. Stacking the models performed better than individual models. With its performance, our model is capable of detecting heart disease in its early stages while improving the heart status of a patient.