ML Final Project Report – Heart Disease Prediction

Author: Doganer Kartum

Index: 97285

Filename: ML25\_Doganer\_Kartum\_97285\_report.py

# 1. Project Topic Selection

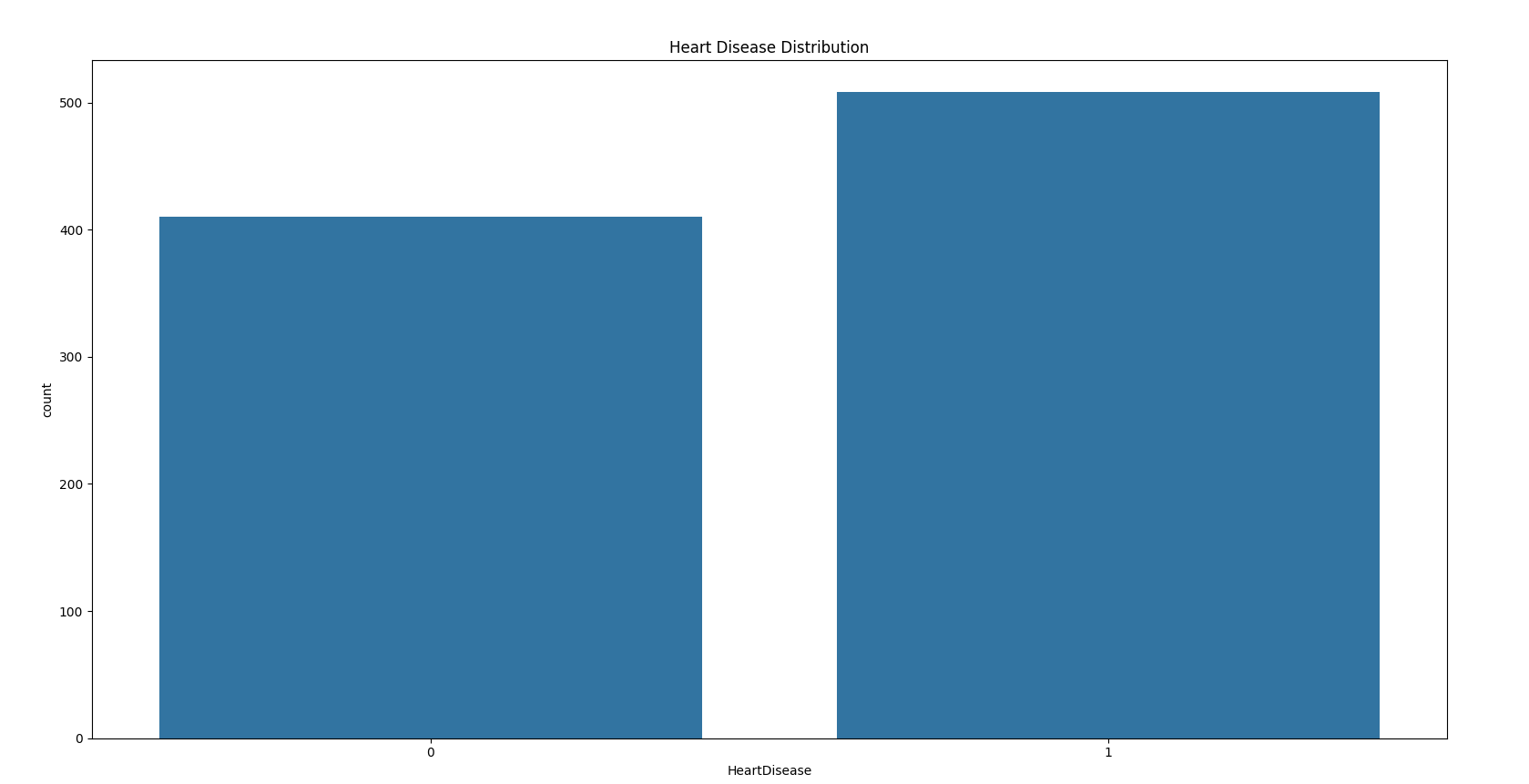
Dataset: Heart Disease UCI Dataset on Kaggle  
Problem Type: Classification  
Objective: To develop a machine learning model that can predict the presence of heart disease based on medical data such as age, blood pressure, cholesterol level, etc.

# 2. Data Analysis & Preprocessing

The dataset contains 918 records with 12 features and one target column 'HeartDisease'. The features include both numerical and categorical attributes representing clinical indicators.

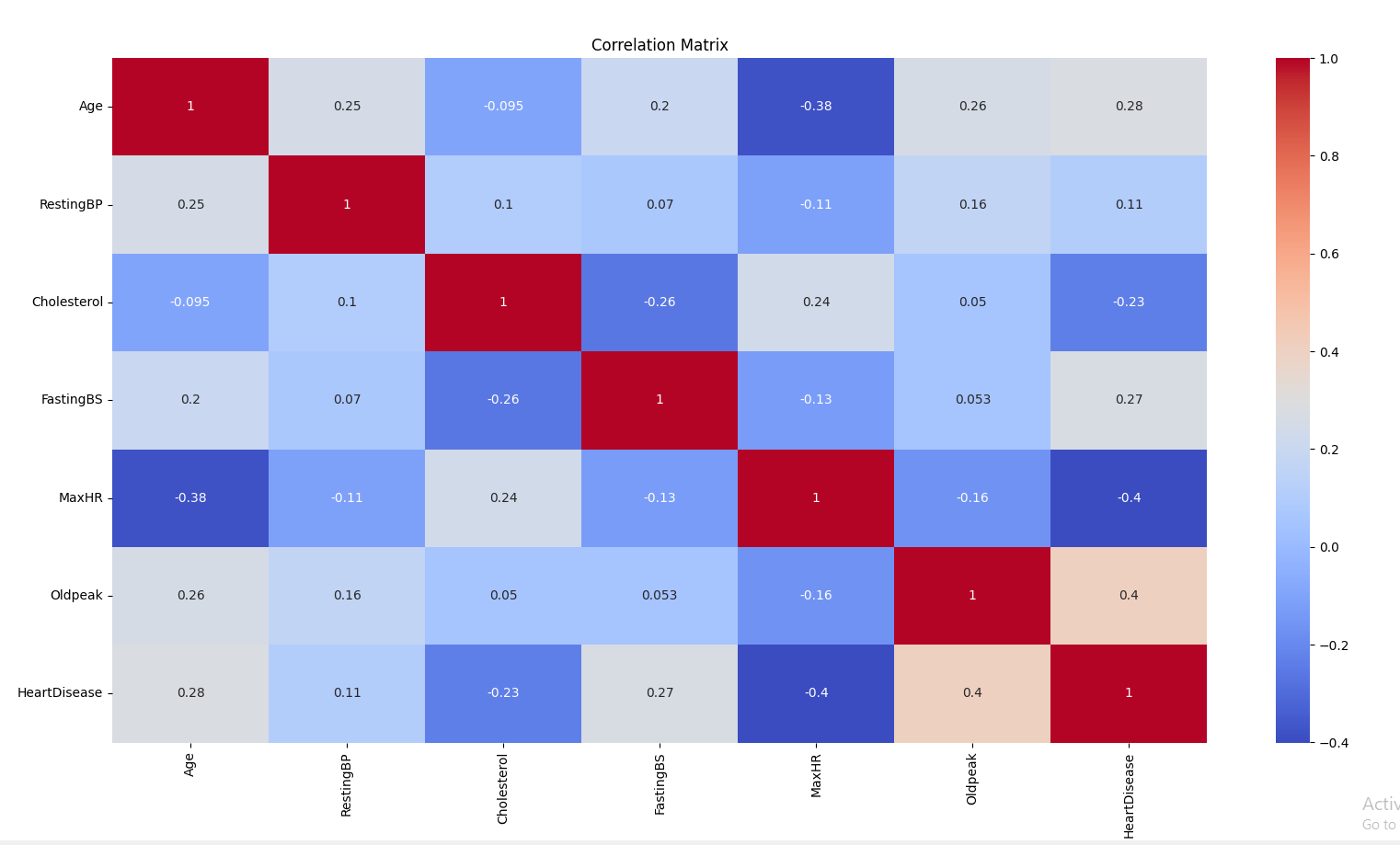
## 2.1 Target Distribution

The target variable ('HeartDisease') is fairly balanced: approximately 510 patients with heart disease and 408 without. This is shown in the countplot below.



## 2.2 Correlation Matrix

The correlation heatmap shows how numerical features relate to each other and to the target. Key observations:  
- Oldpeak and HeartDisease: 0.40 (moderate positive correlation)  
- MaxHR and HeartDisease: -0.40 (moderate negative correlation)  
- FastingBS and HeartDisease: 0.27  
- Cholesterol and HeartDisease: -0.23



These correlations suggest that lower MaxHR and higher Oldpeak may be linked to heart disease.

## 2.3 Cleaning & Encoding

All categorical features (Sex, ChestPainType, ExerciseAngina, ST\_Slope) were encoded using LabelEncoder. No missing values were found.

## 2.4 Standardization/Normalization

Standardization was not applied because Random Forest does not require it.

# 3. Model Selection & Training

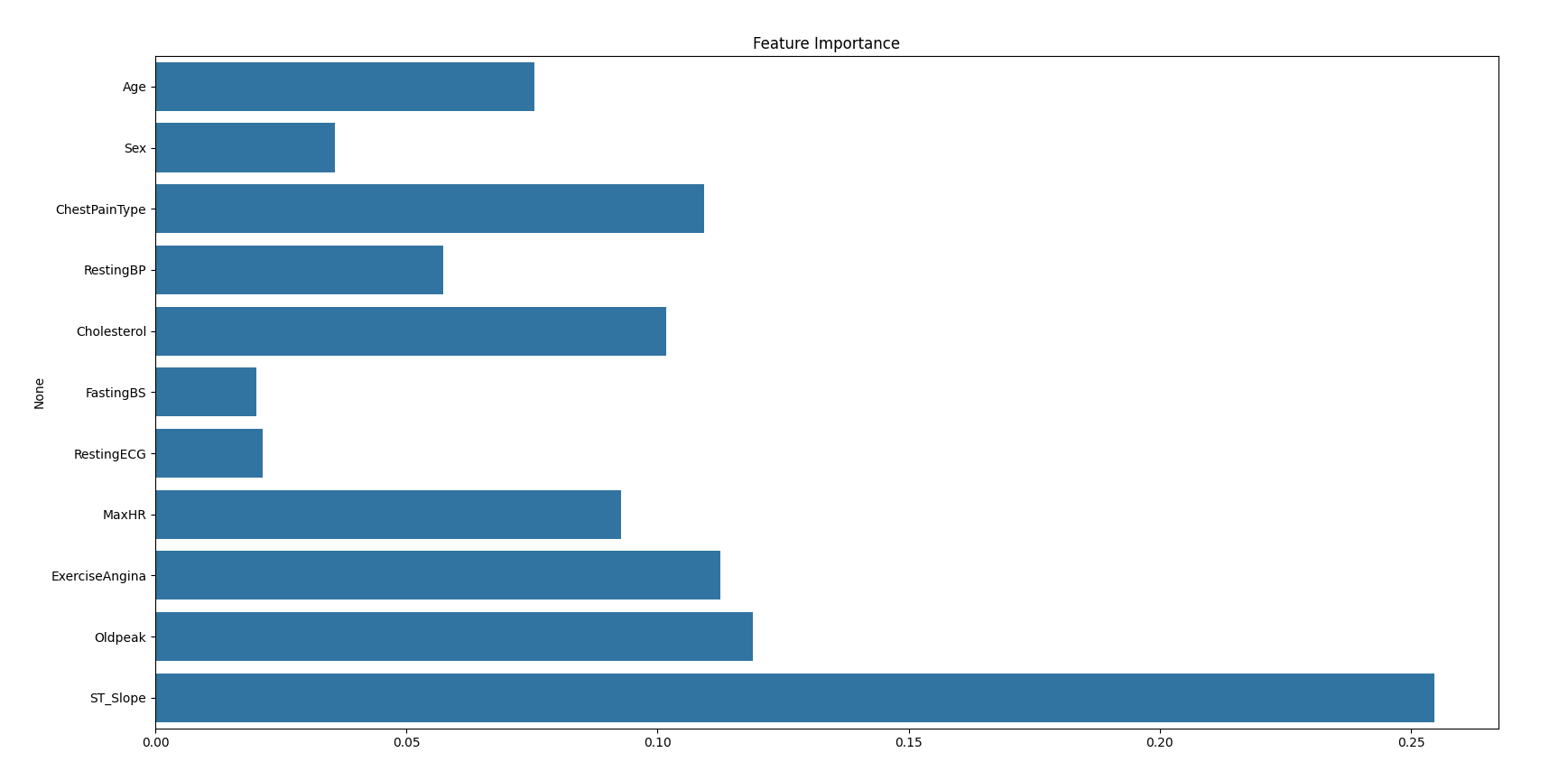
RandomForestClassifier was chosen due to its robustness. GridSearchCV was used for hyperparameter tuning, and 5-fold cross-validation achieved ~0.88 average accuracy. Data was split using an 80/20 ratio.

# 4. Results & Evaluation

Model Accuracy: ~88%  
Confusion Matrix: Low false positive/negative rates.  
AUC Score: ~0.91

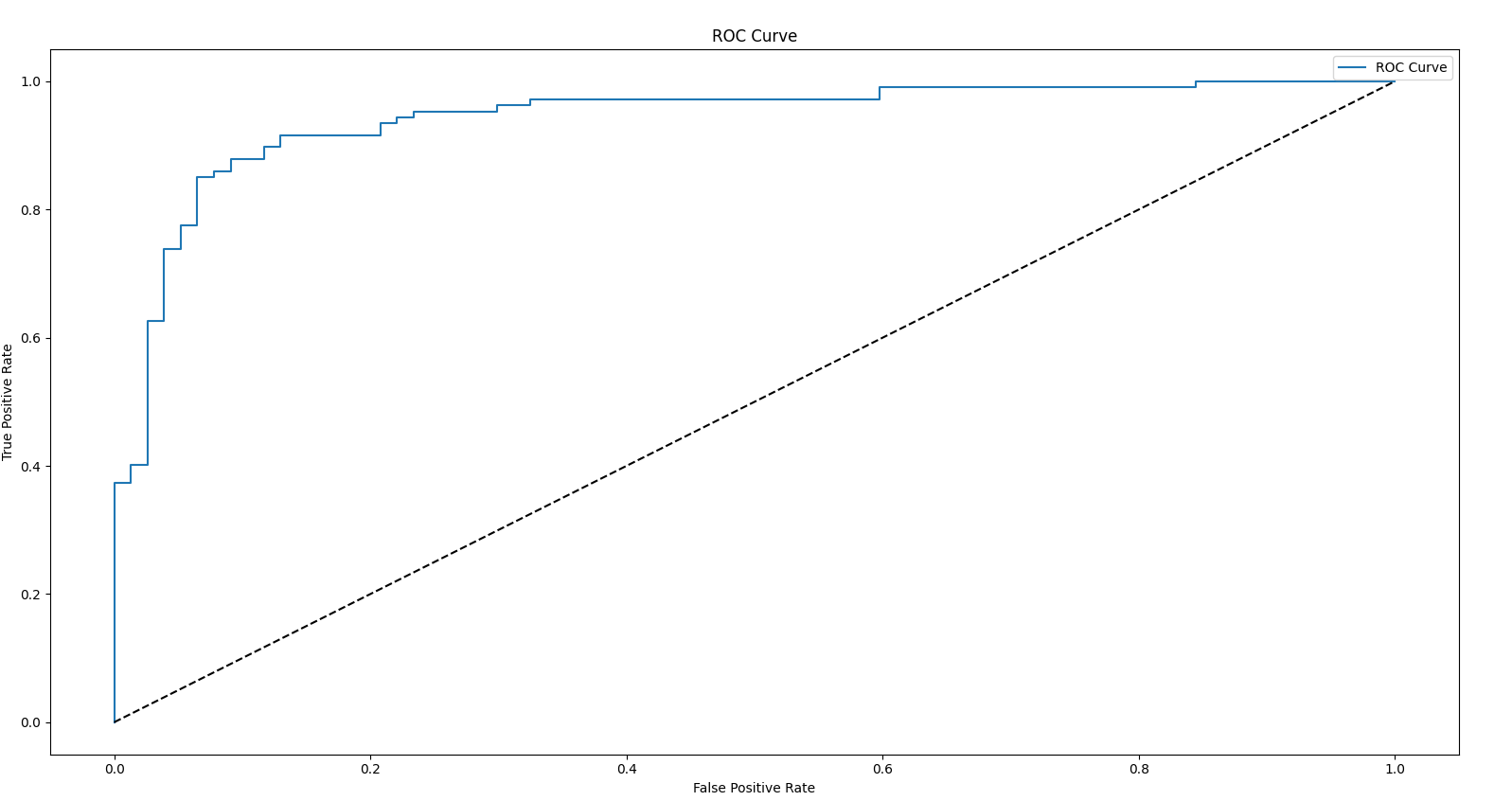
## 4.1 ROC Curve

The ROC curve shows strong separation. Higher TPR even at low FPR.



## 4.2 Feature Importance

Top predictive features include ST\_Slope, Oldpeak, ChestPainType, ExerciseAngina, and Cholesterol.



# 5. Final Conclusions

The model performed very well with high AUC and accuracy. Future improvements could involve trying new algorithms like XGBoost or building an interactive app for real-time use.

# 6. File Submission

ML25\_Doganer\_Kartum\_97285\_code.py  
ML25\_Doganer\_Kartum\_97285\_report.py  
Both will be submitted on Moodle after Professor’s approval.