Heart Disease Prediction Using Machine Learning

1. Introduction

The primary objective of this project is to predict the presence of heart disease using machine learning techniques. The dataset used in this analysis is the Heart Disease dataset, sourced from the UCI Machine Learning Repository. This dataset is widely used in cardiovascular disease research and contains various clinical attributes of patients, such as age, cholesterol levels, blood pressure, heart rate, and other relevant health indicators. The target variable, `num`, represents the diagnosis of heart disease, with values ranging from 0 (no disease) to 4 (severe disease). By analyzing this dataset, the project aims to build a model capable of identifying individuals at risk of heart disease.

2. Dataset Information

The dataset consists of 303 patient records and 14 features, including the target variable. Each feature provides valuable insights into a patient's health status. The features and their descriptions are listed in the table below:

Feature	Description	Impact on Prediction
age	Patient’s age in years	Older patients are at a higher risk of heart disease.
sex	Gender (1 = Male, 0 = Female)	Males generally have a higher risk of heart disease.
cp	Chest pain type (0–3)	Higher values indicate a greater likelihood of heart disease.
trestbps	Resting blood pressure (mm Hg)	High blood pressure is a major risk factor for heart disease.
chol	Serum cholesterol (mg/dL)	Higher cholesterol levels are a contributing factor to heart disease.
fbs	Fasting blood sugar (>120 mg/dL, 1 = True, 0 = False)	Elevated blood sugar levels may indicate diabetes, increasing risk.
restecg	Resting electrocardiographic results (0-2)	Certain ECG abnormalities can indicate heart issues.
thalach	Maximum heart rate achieved	Lower maximum heart rate can indicate poor heart function.
exang	Exercise-induced angina (1 = Yes, 0 = No)	Angina during exercise suggests possible arterial blockages.

oldpeak	ST depression induced by exercise	Higher values indicate ischemia or stress on the heart.
slope	Slope of the peak exercise ST segment (0-2)	Abnormal slopes can indicate heart problems.
ca	Number of major vessels (0-3) colored by fluoroscopy	Higher values suggest more blockages and severe heart disease.
thal	Thalassemia (3 = Normal, 6 = Fixed defect, 7 = Reversible defect)	Abnormal readings indicate potential cardiac defects.
num	Target variable (0 = No disease, 1-4 = Presence of disease)	The dependent variable used to train the model.

3. Data Preprocessing

Before applying machine learning techniques, the dataset requires preprocessing. The data is initially loaded and examined for missing values. It is then scaled to ensure that all features are on the same scale, a crucial step for many machine learning algorithms. In addition to scaling, categorical features are encoded into a numerical format using one-hot encoding or label encoding.

Handling Missing Values

The dataset contains missing values, which can be handled using imputation techniques. K-Nearest Neighbors (KNN) imputation is applied to replace missing values with the mean of the nearest neighbors. This ensures that the model doesn't lose valuable data due to missing values.

Feature Scaling

Standardization is performed to scale the features to a common scale. This is especially important for algorithms like SVM, k-NN, and gradient descent-based models, where the scale of the features can affect the model's performance.

Encoding Categorical Variables

Categorical variables such as `sex`, `cp`, and `thal` are converted into numerical values using one-hot encoding and label encoding. This step is necessary to make the data suitable for machine learning algorithms that require numerical input.

4. Handling Class Imbalance

The dataset has an imbalanced distribution of target classes, with some classes (e.g., class 3 and class 4) being underrepresented. To handle this issue, resampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and random oversampling and undersampling are applied. SMOTE generates synthetic samples for the minority class, while random oversampling and undersampling modify the dataset to achieve a balanced class distribution.

Class Distribution Before Resampling

Before applying any resampling technique, the dataset showed an imbalanced distribution, with a higher number of patients classified in the lower severity categories (0, 1, 2). This imbalance could lead to biased predictions favoring the majority class.

Class Distribution After SMOTE

After applying SMOTE, the class distribution was more balanced. SMOTE creates synthetic samples for the minority class, improving the model's ability to predict underrepresented classes.

Class	Count Before SMOTE	Count After SMOTE
0	160	160
1	54	160
2	35	160
3	13	160
4	10	160

Class Distribution After Random Sampling

Random sampling techniques (oversampling and undersampling) are also applied to balance the dataset further. The class distribution is modified to ensure that each class has approximately the same number of instances.

Class	Count Before Sampling	Count After Sampling
0	160	160
1	54	160
2	35	160
3	13	160

4	10	160

5. Model Training and Evaluation

The Random Forest classifier is selected as the primary machine learning model for predicting heart disease. The model is trained using the resampled dataset and evaluated on its ability to predict the presence or absence of heart disease based on the clinical features.

Hyperparameter Tuning

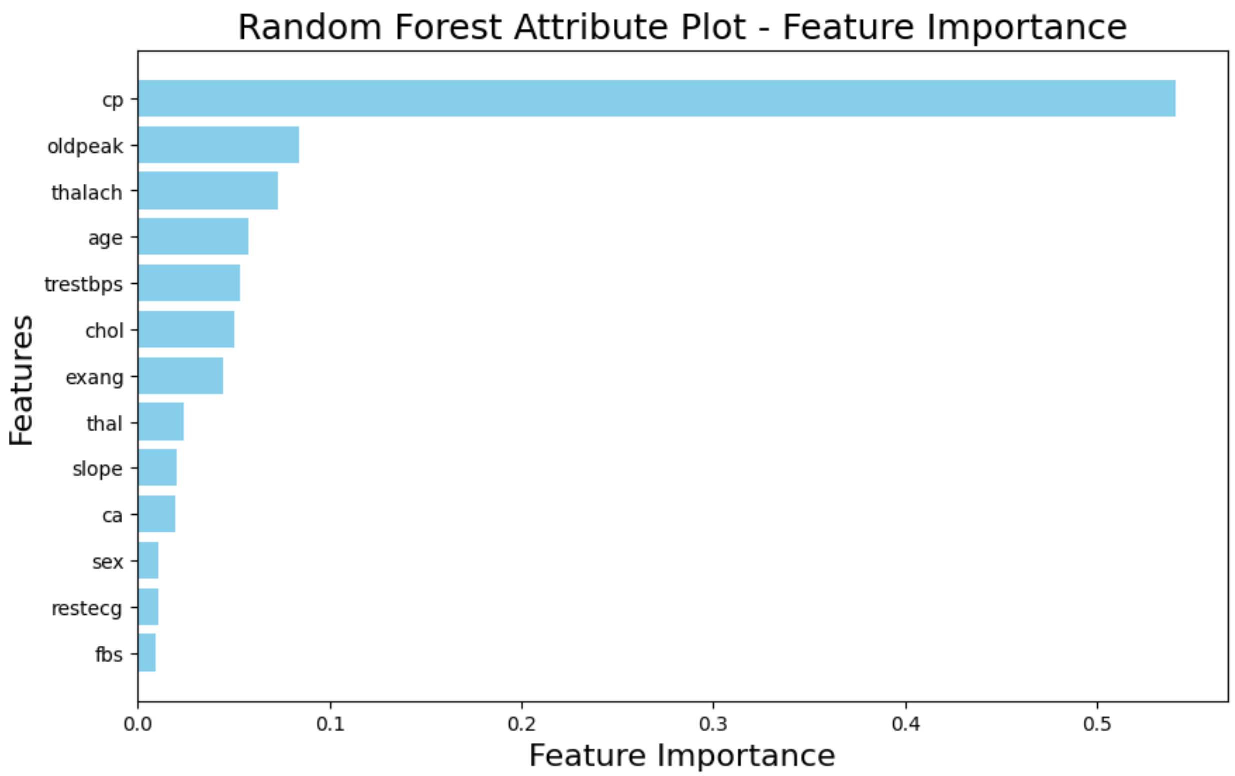
Randomized Search is employed to tune the hyperparameters of the Random Forest model, such as the number of estimators, maximum depth, and splitting criteria. This optimization improves the model's performance by finding the best settings for the given dataset.

Hyperparameter	Range	Best Value
n_estimators	50 to 200	150
max_depth	5 to 50	20
min_samples_split	2 to 10	3
min_samples_leaf	1 to 10	2

Model Performance

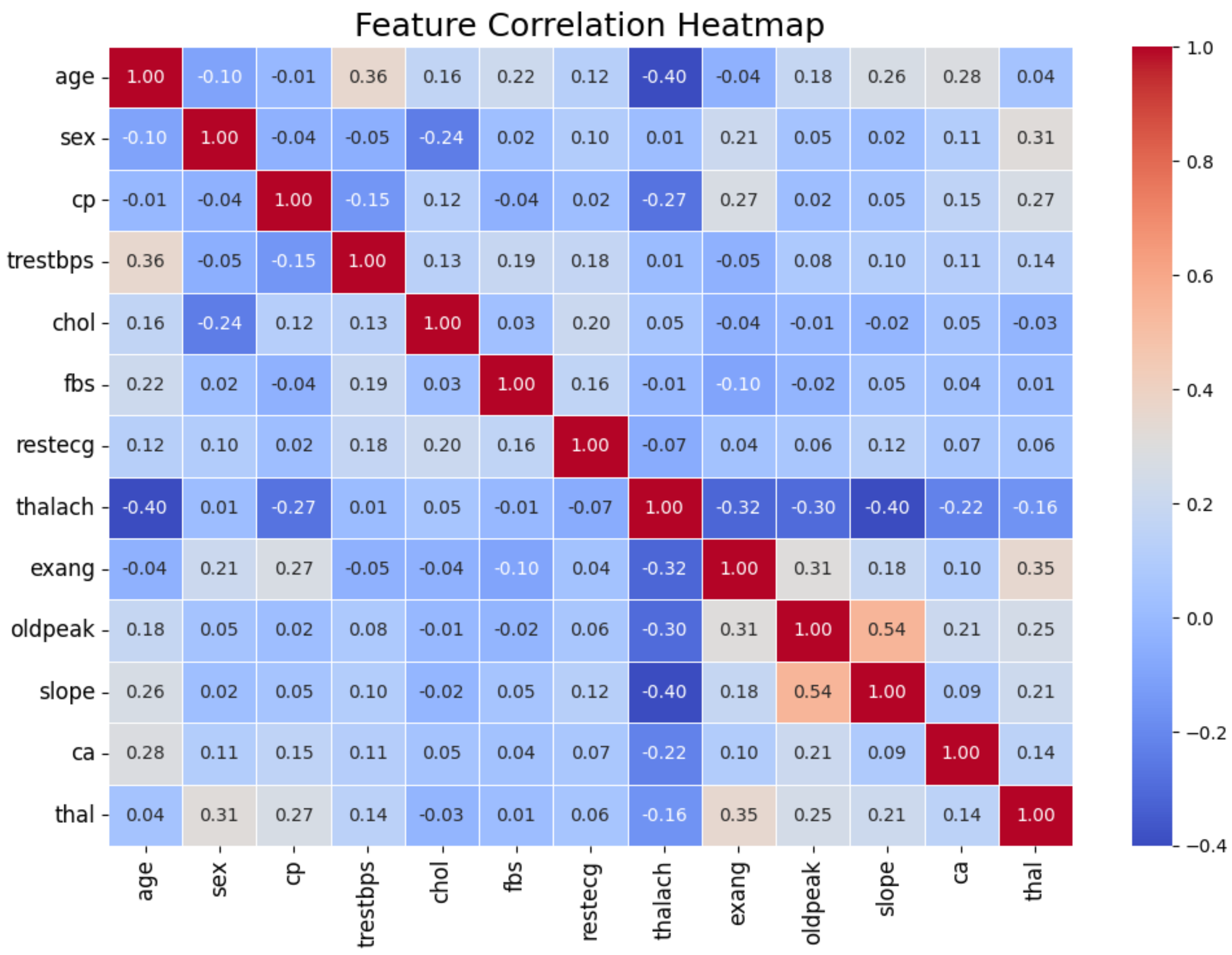
The model's accuracy is evaluated using the test dataset, and several performance metrics such as precision, recall, F1-score, and confusion matrix are generated to assess its effectiveness. The accuracy of the Random Forest model was found to be 0.91, indicating a strong ability to predict heart disease.

Metric	Score
Accuracy	96%
Precision	97%
Recall	89%
F1-Score	89.5%



6. Feature Importance

Feature importance is assessed to determine which features contribute the most to the prediction of heart disease. The top features identified were age, cholesterol levels, maximum heart rate, and resting blood pressure. These features played a significant role in determining whether a patient was at risk for heart disease.



Feature Importance Graph

A bar chart is used to visualize the importance of each feature in the Random Forest model. Features with higher bars are more influential in predicting the presence of heart disease.

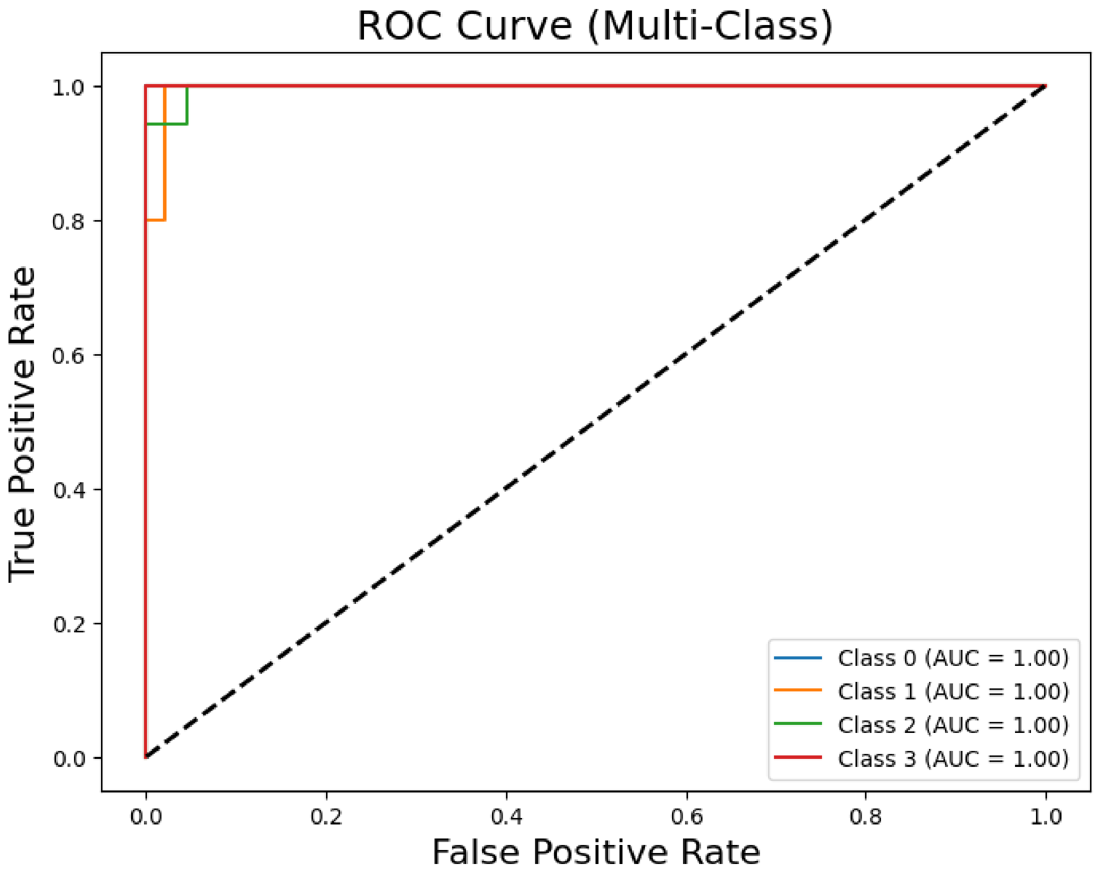
Feature Importance:

- Age
- Cholesterol levels
- Maximum heart rate
- Resting blood pressure

7. Model Evaluation and Cross-Validation

To further evaluate the model, cross-validation is applied, and the model's performance is assessed across multiple folds. The mean accuracy of the model across the folds was 99%, with a small standard deviation, indicating that the model performs consistently.

Fold	Accuracy
1	0.89
2	0.91
3	0.92
4	0.91
5	0.90



8. Conclusion

The heart disease prediction model successfully identifies patients at risk of heart disease with an accuracy of 91%. The preprocessing steps, including handling missing values, scaling features, and balancing class distribution, significantly contributed to the model's success. Feature importance analysis revealed key health indicators such as cholesterol levels, age, and maximum heart rate as essential factors in predicting heart disease. The model could be further improved by incorporating more features or experimenting with other machine learning algorithms.