Heart Disease Prediction Using Machine Learning

Introduction

The goal of this project was to develop a machine learning model to predict the presence of heart disease. The dataset used in this project is the UCI Heart Disease dataset, which contains various patient information such as age, sex, cholesterol levels, and other medical indicators.

Data Preprocessing

The first step in the project was to preprocess the data. This involved several steps:

- Loading the dataset using pandas.
- Dropping unnecessary columns (like 'id').
- Handling missing values by replacing them with the mode for categorical variables and the mean for numerical variables.
- Encoding categorical variables using one-hot encoding.
- Scaling numerical variables using MinMaxScaler to bring them into a range between 0 and 1.
- Binarizing the 'num' column, which is the target variable indicating the presence of heart disease.

Model Training and Evaluation

After preprocessing the data, it was split into training and testing sets. A Random Forest model was trained on the training data. The model's hyperparameters were tuned using Grid Search with cross-validation to find the best combination that maximized accuracy.

The performance of the model was evaluated based on its accuracy on the training and testing data.

Visualization and Interpretation

The model was visualized using a decision tree plot, which provided insights into the decision-making process of the model. A feature importance plot was also generated, which showed the relative importance of each feature in making predictions.

Conclusion

This project demonstrated the effectiveness of machine learning in predicting heart disease. The model achieved good performance and provided valuable insights into the factors that contribute to heart disease. Future work could involve trying different machine learning models, further tuning the model's hyperparameters, or using more advanced techniques for handling missing data.