Heart Disease Prediction Project Report

Introduction

The Heart Disease Prediction project aimed to develop an accurate classification model to predict the presence of heart disease based on various medical attributes. This report provides a comprehensive overview of the project, including data preprocessing, model development, and model evaluation.

Data Preprocessing

Data Collection

- The project utilized two main datasets: the training set (train.csv) and the test set (test_X.csv).
- The training set contained 11 feature columns, a ground-truth label column ("HeartDisease"), and a patient ID column ("PatientID").
- The test set contained 11 feature columns and a patient ID column.

Data Exploration

- Initial data exploration revealed the nature of the features, including numerical and categorical attributes.
- Categorical features included "Sex," "ChestPainType," "RestingECG," "ExerciseAngina," and "ST_Slope."

Data Preprocessing Steps

- Handling Missing Data: After tried a various of handling strategies including relacing with mode, mean, NKK, chose to delete the missing data.Appropriate strategies were applied to address missing data in the training set.
- Encoding Categorical Features: Categorical features were encoded into numerical labels using techniques: Label Encoding.
- Standardizing Numerical Features: Numerical features were standardized to ensure consistent scaling across the dataset.
- Feature Selection: Feature selection techniques include variance threshold were used to identify the most relevant attributes for model training.

Model Development

Model Selection

- Considering the task's requirements and the nature of the dataset, I evaluated several machine learning models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and Naive Bayes. The analysis revealed that discriminative models generally outperformed generative models. Among the discriminative models, Random Forest exhibited superior performance compared to Logistic Regression, leading to its selection for hyperparameter tuning.

Hyperparameter Tuning

- Hyperparameters of the Random Forest model were tuned using Grid Search to optimize model performance.
- Key hyperparameters tuned included the number of estimators, maximum depth, minimum samples split, minimum samples leaf, and maximum features.

Model Training

- The Random Forest classifier was trained on the preprocessed training data.
- The best hyperparameters were utilized for model training.

Model Evaluation

- Model evaluation was performed using various metrics such as accuracy, precision, recall, F1 score, and confusion matrix.
- Cross-validation was employed to assess the model's generalization performance.

Results

- The final Random Forest model achieved an F1 score of over 89.9%, indicating its strong performance in predicting heart disease.