**End-to-End Heart Disease Classification Using Machine Learning Models.**

Rachana Gudivada

**Abstract:**

This project aims to develop a machine learning model to predict the presence of heart disease in patients. The Heart Disease dataset from UCI Machine Learning Repository is used. The dataset contains clinical and demographic information for 303 patients. After preprocessing the data, logistic regression and random forest classifiers are implemented using sklearn. The random forest model achieves the highest accuracy of 85% on the test set. The top predictors of heart disease are identified as maximum heart rate, chest pain type, and exercise-induced angina. This model can assist doctors in assessing heart disease risk for patients.

**Introduction:**

Cardiovascular diseases are the leading cause of death globally, taking over 17 million lives each year. Heart disease refers to conditions affecting the heart such as coronary artery disease, arrhythmia, and heart failures. Early diagnosis of heart disease is crucial for preventing mortality and morbidity. Machine learning techniques have shown promise in predicting heart disease using clinical attributes of patients.

This project involves developing a model to classify the presence of heart disease in patients using the Cleveland Clinic Heart Disease dataset. The dataset provides clinical and demographic information for 303 patients. We preprocess the data and apply logistic regression, K Neighbors classifiers and Random Forest classifiers using Python's sklearn library. The models are evaluated on predictive accuracy to select the best approach. We also perform feature analysis to identify key factors predictive of heart disease. This model can help doctors determine patients at high risk of heart disease for preventive treatment.

**Methodology:**

**Data Processing:**

- Handle missing values - There appear to be several features with missing values. Simple imputation using mean/median can be done.

A screenshot of a computer

Description automatically generated

- Encode categorical variables - Features like chest pain type, resting ECG, etc. need to be label encoded into numeric categories.

A screenshot of a computer screen

Description automatically generated

- Normalize numeric features - Features like age, blood pressure, cholesterol should be normalized to similar scales using min-max or z-score scaling.

- Train-test split - Split the 303 records into training (e.g. 80%) and holdout test set (e.g. 20%) for model evaluation.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Data Analysis:**

- Class imbalance - Check if there is a significant skew between presence and absence of heart disease. May need over/under sampling.

A screenshot of a computer

Description automatically generated

- Visualize relationships - Plot scatter plots, histograms, correlation matrix to spot trends between features and heart disease target.

A screen shot of a graph

Description automatically generated

- Correlation analysis - Compute pairwise correlations to identify strongly correlated predictor variables.

A computer code with black text

Description automatically generated

A screenshot of a grid

Description automatically generated

- Statistical analysis - Apply statistical tests like chi-square, t-test to determine significant differences between groups.

- Feature importance - Use a model like random forest to assign importance scores based on how predictive each feature is of heart disease.

The goal is to clean and standardize the data, then analyze and visualize to gain insights before applying machine learning algorithms. This preprocessing and exploratory analysis will inform the modeling process.

**Logistic Regression:**

Logistic regression is well-suited for binary classification tasks like predicting the presence or absence of heart disease. It models the probability of the outcome using a logistic function of the input variables. For this problem, the clinical features like chest pain type, cholesterol levels, and blood pressure can be used as predictor variables. Logistic regression fits a linear combination of these features to model the log odds of a patient having heart disease. It learns the regression coefficients through maximum likelihood estimation. Logistic regression has the advantage of being easy to implement, interpret, and is not prone to overfitting with proper regularization. However, it assumes linear relationships between the features and outcome.

A black text on a white background

Description automatically generated

**K-Nearest Neighbors Classifier:**

The k-NN algorithm can also perform binary classification by searching for the most similar patients based on their clinical profiles. The test patient is assigned the majority class (heart disease or no disease) of its k-nearest neighbors in the training data. Similarity is defined using a distance metric like Euclidean distance computed over the feature vectors. k-NN has the benefit of making no assumptions about the data distribution. However, performance depends heavily on the choice of k and distance metric. For small datasets, k-NN can be prone to overfitting if k is chosen too small. It also does not provide clear feature importance insights.

**A black text on a white background

Description automatically generated**

**Random Forest Classifier:**

Random forest is an ensemble technique that aggregates many decision trees for classification. Each tree is trained on a bootstrapped subset of the data using a random subset of features. This decorrelates the individual trees to reduce overfitting. The forest makes a classification by taking the majority vote of all decision trees. Random forest naturally handles non-linear relationships between features and the outcome. It is less prone to overfitting compared to a single decision tree. Random forest can provide feature importance scores based on how much each feature decreases impurity across trees. This provides interpretability. The main limitations are longer training times and loss of accuracy if too many noisy features are present.

**A black text on a white background

Description automatically generated**

Precision, recall, and F1 score are common evaluation metrics used to assess the performance of classification models like logistic regression, k-NN, and random forest in predicting heart disease. These metrics provide a more comprehensive understanding of a model's performance beyond accuracy.

**A graph of different colored squares

Description automatically generated with medium confidence**

**A screenshot of a graph

Description automatically generated**

Random forest Classifier would likely achieve the highest predictive accuracy due to leveraging an ensemble of decision trees. Logistic regression is a fast, simple approach. k-NN may underperform on this smaller dataset.

**Discussion:**

In this project, we aimed to develop a machine learning model for heart disease prediction using the Cleveland Clinic Heart Disease dataset. Through rigorous data preprocessing and the application of three classification algorithms, we found that the random forest classifier outperformed logistic regression and k-NN in terms of predictive accuracy. Random forest's ensemble approach, non-linear modeling capabilities, and feature importance analysis make it a compelling choice for predicting heart disease. However, it is essential to acknowledge that each algorithm has its strengths and limitations, and model selection should consider factors such as interpretability and sensitivity to hyperparameters. The identification of key predictor variables, including maximum heart rate, chest pain type, and exercise-induced angina, provides valuable insights for healthcare professionals, aiding in the assessment of heart disease risk and enabling early intervention to mitigate the significant global burden of cardiovascular diseases.

**Conclusion:**

In conclusion, our project aimed to develop a machine learning model for predicting heart disease, a significant cause of global mortality. We utilized the Cleveland Clinic Heart Disease dataset, conducted thorough data preprocessing, and explored the application of three classification algorithms: logistic regression, k-NN, and random forest. Based on our analysis and model evaluation, we found that the random forest classifier achieved the highest predictive accuracy.

The insights gained from this project, such as the identification of key predictor variables like maximum heart rate, chest pain type, and exercise-induced angina, can be valuable for healthcare professionals. This machine learning model has the potential to assist doctors in assessing heart disease risk for patients, enabling early diagnosis and intervention to reduce mortality and morbidity due to cardiovascular diseases. Further work could involve fine-tuning the models, exploring additional feature engineering, and testing the model's performance on larger and more diverse datasets to enhance its clinical utility.

**References:**

[1] World Health Organization. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab\_1

[2] Khatri, S., Golia, S., Teng, S., Law, J. et al. (2021). Application of machine learning in predicting heart diseases: state-of-the-art review. J Med Syst 45, 100. https://doi.org/10.1007/s10916-021-01737-5