**Project Report: Heart Disease Classification using Random Forest**

[https://github.com/joshi39/AI-PROJECT--PREDICTING-HEART-DISEASES.git](https://github.com/joshi39/AI-FIRST-PROJECT.git)

Heart disease is one of the major health issues worldwide. Fatal outcomes can be prevented if it is detected and diagnosed early. This project aims to predict the presence of heart disease in patients using a Random Forest classifier-a machine learning classifier. The dataset used in this project contains various features of patients, such as age, sex, cholesterol levels, and blood pressure.

**Flow of the project:**

* Data loading
* Data splitting
* Hyperparameter tuning.
* Feature selection
* Model training
* Evaluation metrics

**Random Forest: Joshi Komaragiri**

A random forest classifier works on combining many decision trees to form a more accurate prediction. Each decision tree in random forest classifier can have a different subset of features when it comes to classifying several types of data. When building each tree, it applies bagging and feature randomness to create an uncorrelated forest of trees whose predictions by consensus are more accurate than any individual tree.

The random forest classifier uses each tree's strengths and increases overall accuracy by merging the predictions of all the decision trees.

Random forest classifiers are commonly used in applications such as image and speech recognition, text classification, and fraud detection.

**Logistic Regression: Priya Kaur**

Logistics regression is a regression analysis tool that helps predict the probability of a binary outcome based on input features. The input features, also known as 'predictors', are anything that influence the outcomes, such as gender, age, etc. This model usually has a threshold of 0.5. Probability score above 0.5 means the model predicts a positive outcome, and a score below 0.5 means the outcome predicted would be negative.

This tool is/can be widely used in the healthcare industry.

**Decision Tree: Pavan Muruturu**

Decision tree is based on analysing patterns in data. In simple terms, it is a flowchart with each decision to be made and its possible outcome. All of it is taken into consideration to reach the prediction. It’s clear and easy-to-follow representation makes it an ideal choice for a lot of people. Furthermore, since the decision tree has individual decision 'nodes', it is easy to add new data, making it flexible.

**Common steps followed:**

**Data Pre-processing:**

In this, we dropped the missing value, converted the categorical variable to a dummy variable, and normalized feature data using dummy data.

**Feature selection-Select from Model**

Sometimes we do not want to include all of a model's characteristics. We may want to reduce the model's variance or improve accuracy by including only the most important features.

A simple two-stage workflow in Scikit-Learn can be used to build a model with fewer features. First, all characteristics are used to train a random forest model. Then, we use this model to determine which characteristics are most crucial. Following that, we built a new feature matrix that only includes these features. The Select from Model method can be used to build a feature matrix that only includes features with significance greater than or equal to a predetermined threshold. Finally, we build a new model that incorporates just those characteristics.

**Hyperparameter tuning.**

Grid Search will also be used to fine-tune the Random Forest model's hyperparameters. To assess each hyperparameter combination's performance and choose the best one, we will utilize 5-fold cross-validation.

**Performance:**

The code calculates different evaluation metrics such as accuracy, sensitivity, precision, F1 score, and specificity using the testing sets predicted and actual values.

These metrics help to understand how well the model predicts the existence or absence of heart disease based on the features picked and the classifier used. The evaluation metrics will clarify the model's advantages and disadvantages in predicting heart disease.

**Results:**

The following evaluation metrics were obtained:

**Random Forest**

Accuracy: 0.8688524590163934

Sensitivity: 0.8947368421052632

Precision: 0.8461538461538461

F1 Score: 0.8695652173913043

Specificity: 0.8421052631578947

**Logistic regression**

Accuracy: 0.7658536585365854

Sensitivity: 0.8058252427184466

Precision: 0.7477477477477478

F1 Score: 0.7757009345794393

Specificity: 0.7254901960784313

AUC: 0.7656577193984391

**Decision tree**

Accuracy: 0.9853658536585366

Sensitivity: 0.970873786407767

Precision: 1.0

F1 Score: 0.9852216748768473

Specificity: 1.0

AUC: 0.9854368932038835

**Conclusion:**

Out of all the three models, Decision tree has a better performance metric and logistic regression is the least accurate of the three.