**Project Description**

1. **Problem Statement:**

Developing an advanced machine learning model for the early detection of heart disease (0 = no disease, 1 = disease) by leveraging patient demographics and clinical features. The project, part of the Google Solution Challenge, aims to utilize parameters such as age, chest pain type, and cholesterol levels to predict the likelihood of heart disease. Employing a diverse set of machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Deep Neural Network, Linear Discriminant Analysis (LDA), and AdaBoost, the objective is to explore various parameter configurations for each algorithm. The ultimate goal is to identify the optimal model that achieves the highest accuracy, facilitating proactive healthcare measures for individuals at risk of heart disease. This project aligns with the Google Solution Challenge's mission of addressing real-world problems through innovative and effective technological solutions.

1. **Data Gathering:**

Heart Disease Prediction dataset contains records about diagnosis of heart disease based on multiple factors like Age, Sex, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting Electrocardiographic Results, Maximum Heart Rate Achieved, Exercise-Induced Angina, Oldpeak (ST Depression), Slope of Peak Exercise ST Segment, Number of Major Vessels (Colored by Fluoroscopy) and Thalassemia. The data set dates from 1988 and consists of data from four countries: Cleveland, Hungary, Switzerland, and Long Beach V. There are 304 records based on 14 attributes (9 categorical and 5 continuous).

Name: Heart Disease Dataset

Number of instances: 304

Number of attributes: 14 ( 13 int64 and 1 float64)

1. **Explainatory Data Analysis**:

Delve into the exploratory data analysis section, where we unravel patterns, correlations, and crucial insights from the heart disease dataset. Visualizations and analyses provide a deep understanding of the data.

1. **Feature Engineering:**

* Outliers were replaced with operatiles and lower tails.
* All independent features were Normalized by MinMax Scaler.
* One Hot Encoding was performed on categorical variables.
* Highly Multicolinear features were removed from the model with the help of Variation Inflation Factor (VIF). The range of VIF in the model is between 1 and 5.

1. **Train Test Split:**

The model was split into training and testing sets with test\_size = 0.2

1. **Model Training:**

Model was trained on eight different algorithms with different parameter. They are as follows:

**Logistic Regression:**

Best params: {'C': 100, 'penalty': 'l2'}

Accuracy: 0.86

**Random Forest Algorithm:**

Best params: {'bootstrap': True, 'class\_weight': 'balanced', 'criterion': 'gini', 'max\_depth': None, 'max\_features': 'sqrt', 'min\_samples\_leaf': 2, 'min\_samples\_split': 2, 'n\_estimators': 140, 'verbose': 0}

Accuracy: 0.83

**Decision Tree:**

Best params: {'class\_weight': None, 'criterion': 'gini', 'max\_depth': None, 'max\_features': 'sqrt', 'min\_samples\_leaf': 2, 'min\_samples\_split': 2}

Accuracy: 0.78  
**K nearest neighbours:**

Best params: Best k= 7

Accuracy: 0.84

**Adaboost:**

Best params: {'learning\_rate': 0.01, 'n\_estimators': 100}

Accuracy: 0.81

**Naive Bayes:**

Accuracy: 0.54

**Linear Discriminant Analysis:**

Best params: {'n\_components': None}

Accuracy: 0.87

**Deep Neural Networks:**

Accuracy: 0.48

1. **Conclusion:**

Classification Model Evaluation:

Metric Class 0 Class 1

Precision 84% 90%

Recall 90% 84%

F1-Score 87% 87%

Support 29 32

Accuracy: 87%

Confusion Matrix:

Predicted 0 Predicted 1

Actual 0 26 3

Actual 1 5 27

In summary, the Linear Discriminant model demonstrates a good balance between precision and recall, with an overall accuracy of 87%. The confusion matrix provides a detailed breakdown of predicted and actual classifications for each class.