ML CVDs Project Documentation

## **Data Preprocessing:**

Cleaning and preparation of data for training

* Data Cleaning:

Cleaning of data after viewing data description and statistics and data types of features:

-Checking for duplicated rows or rows with null values

-Checking for values of 0 in numerical features and replacing it with the mean of data due to it being normally distributed (Cholesterol, RestingBP)

-Detecting outliers using box plot for numeric features and dropping outliers to not affect the performance of the models using IQR method which defines a lower and upper bound for data to be accepted.

* Viewing distribution of data and creating correlation heatmap to indicate relationships between different features for e.g. the most features affecting heart disease is ExerciseAngina and Oldpeak
* Preprocessing:

-Separate categorical and numerical data to perform different types of preprocessing

StandardScaler for numerical data: Necessary to use for KNN model because it is distance-based, used to be able to take the variance and contribution of features in account equally independent of their difference in their scales and values, also it is useful for logistic regression and neural network models because they use gradient descent optimization which assumes data to be on a similar scale

OneHotEncoding for categorical data: Used to make models handle categorical data by converting it to numerical format and it is also better for distance-based algorithms than normal label encoding.

## **Model Selection:**

1. KNN

Classical and simple supervised machine learning algorithm can be used for classification problems by using voting to K nearest examples (K is odd)

1. Logistic Regression

Supervised machine learning algorithm used for binary classification problems by prediction probability of occurrence of a certain event, but it assumes and works best when there is a linear relationship between input features and the outcome

1. Deep Learning

Uses neural networks which are better when there is complex or non-linear relationship between features than classical models but can overfit in smaller datasets, uses sigmoid activation function for binary classification

## **Evaluation Criteria:**

* Accuracy:

Measures percentage of total correct predictions

Accuracy =

* Precision:

Measures percentage of correct positive predictions

Precision =

Used to measure degree of false positives and avoid it

* Recall:

Measures how many of the positive cases were correctly predicted

Recall =

Very important in medical applications to detect all people that have a disease

* F1-Score:

Balance between precision and recall

F1-score =

## **Explanation:**

The main goal of the project is to use machine learning models to solve binary classification problem of detecting if individuals have heart disease based on medical features , This starts by preparing the dataset by performing Exploratory Data Analysis to understand data distribution and identify relations between different features , and handling null rows and detecting outliers , then separating features to different types to perform suitable types of encoding and preprocessing , then selecting models and performing hyperparameters tuning before training on the dataset and testing then using different types of evaluation scores.

## **Comparison between approaches:**

* KNN:

Advantages: Does not assume specific distribution for input data

Disadvantages: Sensitive to choice of hyperparameters and needs data scaling because it is a distance-based algorithm

* Logistic Regression

Advantages: Simple model to implement and interpret

Disadvantages: Assumes linearity between features and the target variable

* Deep Learning

Advantages: Has ability to detect complex patterns and relations between features and better for large datasets

Disadvantages: Requires more computation and resources