**HEART DISEASE PREDICTION**

**INTERIM REPORT**

**Proposed Solution:**

The given problem can be interpreted as a binary classification based problem on whether a person has a heart disease or not. Hence, Classification based Machine Learning Algorithms such as Logistic Regression, Naive Bayes, K-Nearest Neighbours, Decision Tree, Support Vector Machines, etc can be used to solve this problem and the algorithm with the best accuracy can be chosen as the final solution based on different evaluation metrics.

**Evaluation metrics:**

The Evaluation metrics, used to find out how good a model is, are given below,

1. Classification report: The classification report function builds a text report showing the main classification metrics such as precision, recall, f1-score, support.
2. Confusion matrix: A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It shows the number of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN).
3. Accuracy score: Accuracy is one metric for evaluating classification models. It is defined as the fraction of predictions our model got right.

**Initial Exploratory Data Analysis:**

* The dataset has to be checked to see if there are any null values in any of the features and if any null values are found they have to be removed.
* The dataset has to be balanced. That is, the number of people that have a heart disease has to be close to the number of people that don’t have a heart disease. If it isn’t balance, then it should be balanced using various methods such as SMOOT, etc.
* Outliers in the features have to be removed from the dataset. This is done by standardizing the features with outliers.

**Summary of Initial Findings:**

* There are no null values in the dataset.
* The dataset is balanced.
* The feature ‘thalach’ is negatively co-related with all the other features.
* Without any feature engineering, KNN classifier based model showed the highest accuracy among all the models used.

**Challenges:**

* Finding the perfect value of number of neighbours for the KNN classifier model to get the highest accuracy possible
* Choosing between 2 different models that have the same level of accuracy.