CS 513 B – KDD PROJECT PROPOSAL: Heart Disease Prediction

Project Group: 10

Problem Statement:

Heart disease and heart attack are significant health concerns worldwide. The proposed project aims to predict the risk of heart disease or heart attack using individuals' health and lifestyle habits, demographic and socioeconomic factors. The target variable is binary, with 1 representing the presence of a heart disease or heart attack event and 0 representing the absence of such an event.

Dataset:

The dataset comprises of 21 feature columns and a binary target variable: HeartDiseaseorAttack. The features include high blood pressure, high cholesterol, recent cholesterol check, body mass index (BMI), smoking habits, history of stroke, diabetes status, physical activity level, intake of fruits and vegetables, heavy alcohol consumption, access to healthcare, out-of-pocket healthcare costs, overall general health, mental health, physical health, difficulty with walking, sex, age, education level, and income. Feature reduction techniques like PCA and correlation may be employed to select essential features for implementation.

Source of Dataset:

https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset?select=heart_disease_health_indicators_BRFSS2015.csv

Implementation Strategy and algorithms used:

We have decided to implement and compare 7 different models among four different group members. We have chosen a few models from our course and a few from outside the course. The following are the models selected by us:

1. Decision Trees
2. Naive Bayes
3. K-nearest neighbour with Grid Search CV
4. Logistic Regression with Grid Search CV
5. Random Forest with Randomized Search CV
6. Support Vector Machine with Grid Search CV
7. AdaBoost Classifier

Model metrics and Evaluation:

Confusion Matrix: As the problem is a binary classification problem, a confusion matrix can help evaluate the model's performance in predicting true positives, true negatives, false positives, and false negatives.

Accuracy: A metric that measures the proportion of correct predictions out of the total number of predictions made.

Precision: A metric that measures the proportion of true positives out of the total number of positive predictions made by the model.

Recall: A metric that measures the proportion of true positives out of the total number of actual positives in the dataset.

F1 Score: A weighted harmonic mean of precision and recall, used to evaluate the balance between the two metrics.

Mean Squared Error (MSE): A metric that measures the average squared difference between the predicted and actual values.

Team Members:

Jeet Mehta - CWID 20015479

Mani Sai Prasad Masupalli - CWID 20015644

Moksha Sunilkumar Dave - CWID 20012110

Shloka Brijesh Singh - CWID 20015697