**Curneu MedTech Innovations Assessment -Task 2**

**K-nearest neighbors classifier – Diabetes dataset**

**Introduction:**

The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithms.  is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all of the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc.

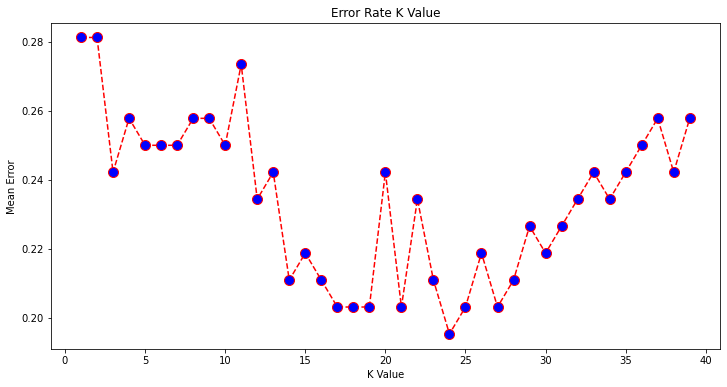
**Problem Statement:**

A labelled dataset is given with 9 variables such as BMI, age, insulin, glucose, skin thickness and etc in a csv. The task is to develop a best suitable machine learning model from scratch and find out the best value of k with highest r2 score. Also, predicting which patients could be affected by diabetes

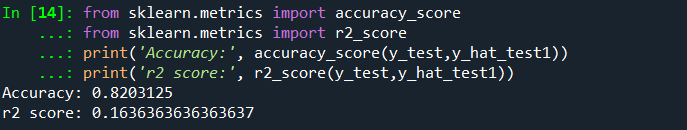
**Approach to the problem:**

Given dataset is read using pandas into a variable named diabetes\_dat. The description of the dataset and the first 5 rows of the dataset are printed. The correlation coefficient for every combination is also found to find the best possible combination of parameter. Since all the variable all of a very low correlation coefficient all the variables are taken as the features and the target variable is split from the dataset. The given dataset consists of numeric data and the dataset is labelled. Hence knn algorithm is used to predict the target variable. X and y variables are then split into training and test variables. X\_train and X\_Test variables are then standardized to make the variables on the same scale. Then the functions for euclidean distance and knn is built from scratch. The developed knn function is then made to run under a loop to find the best value of k by spotting the least error giving k value. The model is passed with the best k value and the X\_test variable is used to predict the y value. The accuracy and r2 value of the predicted y variable is found.

**Output:**

****

Interpreting the above graph, we find that minimum mean error occurs when the k values is 24. Therefor the best suitable value of k is to be chosen to be 24

****

The accuracy score for the test data set is 0.82 which means the model is 82% accurate to predict the target variable.