**Assignment Problems: Day 8**

The following problems require access to the file named ‘Dataset\_Day8.csv’, provided with this assignment.

This is the same assignment data as Day 7.

The datasets include data from 768 women with several medical predictor variables and one target variable(‘Outcome’). The classification goal is to predict whether or not the patients in the dataset have diabetes or not.

A screenshot of a computer code

Description automatically generated

For columns:

*Glucose,   
BloodPressure,   
BMI,   
DiabetesPedigreeFunction*

If the column value is 0, then they should be considered as **missing data.**

Problems to solve –

1. Firstly, replace all Missing values with relevant figures.
2. Then remove all existing outliers and get the final data for classification.
3. Split the data into 70% training and 30% testing data. Then, use a k-Nearest Neighbor algorithm with target variable as ‘Outcome’.
   1. Print the default model performance metrics: Accuracy, Precision, Recall, F1Score
   2. Plot a Precision & Recall vs k(no. of neighbours) curve (both Prec and Rec on the same graph). Find the k for which F1-score is the highest. **Use any one Distance Metric for this problem.**
   3. Find the best distance metric, no. of neighbors combination for the kNN algorithm. (do not plot any curve)

**MUST ADD INSIGHTS TO EVERY OBSERVATION/OUTPUT YOU PROVIDE.**