

Problem 1: Classification using k-NN algorithm

Link: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

- a. Find the total number of records present in the dataset.
- b. Find the attributes and types of attributes present in the dataset.
- c. Find the range of values for numerical variables present in the dataset.
- d. Retrieve Train and Test Dataset. Partition the original dataset into train and test dataset with the ratio of 8:2.
- e. Apply Minimum Distance classification. Use K-NN. Vary the values of K and obtain prediction.
- f. Vary the value of **K** from **1 to 20**, and perform classification for each value.
- g. For each value of **K**, **predict** the labels for the test data.
- h. Print the confusion matrix.
- i. Print TP, TN, FP, FN.
- j. Using TP, TN, FP, FN; calculate the accuracy, recall, precision, f1-score.
- k. Calculate MSE.
- l. Plot the Accuracy VS K-Value graph for the corresponding prediction. Which K value performs the best and why?

Problem 2: Classification using k-NN algorithm

Consider the following dataset

| ID | Height (cm) | Weight (kg) | Class |
|----|-------------|-------------|-------|
| A | 160 | 50 | 0 |
| B | 165 | 55 | 0 |
| C | 170 | 65 | 1 |
| D | 175 | 70 | 1 |
| E | 180 | 80 | 1 |

Test Point: Height = 172 cm, Weight = 66 kg

Goal: Predict the class using k-NN with different values of **K**. Do not use built-in library for implementing K-NN.