Assignment B-5

Problem Statement:

Implement K-Nearest Neighbours algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset.

K-Nearest Neighbours.

The k-nearest neighbours' algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. For classification problems, a class label is assigned on the basis of a majority vote—i.e. the label that is most frequently represented around a given data point is used.

- **Lazy learning algorithm** − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- **Non-parametric learning algorithm** − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.

Step 2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

Step 3 − For each point in the test data do the following −

- **3.1** − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
- **3.2** − Now, based on the distance value, sort them in ascending order.
- **3.3** − Next, it will choose the top K rows from the sorted array.
- **3.4** − Now, it will assign a class to the test point based on most frequent class of these rows.

Step 4 − End

Database Used:

https://www.kaggle.com/datasets/saurabh00007/diabetes.csv

Python: Colab, spider or similar platform _____

YT Ref: _____

Code (As attached) & Graphs (wherever applicable) _____

Metrics used for performance measurement: _____

Conclusion: In this experiment we classified patients into diabetic and non-diabetics based on age, Glucose, Insulin, BMI. K-nn algorithm finds the similarity distance based on number of features used for classification. The distance array collected is sorted by ascending order and top N samples decide the class of the patient.

K-NN algorithm for classification of diabetic patients.

This video demonstrates classification of patients as diabetic and non-diabetic. In this K-NN algorithm is used on daibetes.csv data from Kaggle. It is group B-5 assignment as per SPPU B.E. Computer -2019- LP-III. Write-up and code is available if requested through my email pp.halkarnikar@gmail.com