Practical - 05.

* Aim - Implement the K nearest neighbours algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error, rate, precesion and recall on the given dataset.

* Theory:

1] KNN -

- K-nearest neighbours is a non parametric, instance based learning method used for classification.
- It predicts the class of a test sample by looking at the majority class among its k - closest training samples in the feature space.

2] Algorithm steps -

1. Step 1: load the dataset and split it into feature X and target y.

Step 2: Split the data into training and testing sets in 80:20 ratio.

Step 3: Normalize/ standardize the features to improve distance calculation.

Step 4: Choose a suitable value of k (eg. 3, 5, 7

Step 5: -

For each test point, compute the euclidean distance from all training points.

Step 6: Identify the k nearest neighbours.

Step 7: Assign the test point the class that appears most frequently among its neighbours.

Step 8: After predictions, compute the confusion matrix, accuracy, error rate, precision, and recall.

3] Dataset description -

- Features -
  - Pregnancies.
  - Glucose
  - Blood pressure.
  - Skin Thickness.
  - Insulin.
  - BMI
  - Pedigree function.
  - Age
  - Outcome

4] Preprocessing steps -

- Load the dataset with Pandas.
- Split into feature matrix X and target vector y.

- Train test split (80%, 20% or 70% or 30%)
- Normalization / standardization - of features to improve distance computation (Standard scaler from scikit learn)

5) Model Building and evaluation -

- KNN model - Create KNeighbours Classifier (eg. k=5)
- Train on training set
- Predict on test set
- Metrics -
  - Confusion Matrix :

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

- Accuracy = (TP + TN) / (TP + FP + TN + FN)
- Error rate = 1 - accuracy.
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN).

CONCLUSION :

* By applying the K-nearest neighbours algorithm to the diabetes dataset -
we can classify patients as diabetic or non-diabetic based on medical measurements.

- Normalization of data improves the accuracy of distance based algorithms like KNN.
- Evaluation using accuracy, error rate, precision, recall and confusion matrix helps us understand the classifiers performance beyond just a single metric.
- With a properly chosen k, KNN provides a simple yet powerful baseline for medical diagnostic classification tasks.

* * * * * * * *