LP-III Machine Learning (2024-25) **Assignment 3:** Implement K-Nearest Neighbors algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset.	
Student Name:	**Roll No. :**
Batch:	**Division :**

Assignment 2:
Implement K-Nearest Neighbors algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset.

Title: Implement KNN classification algorithm to predict diabetes person using dataset available at https://www.kaggle.com/datasets/abdallamahgoub/diabetes.

Aim: Predict and Analyse Results of KNN algorithm for Classification.

Prerequisites: KNN.

Theory:

K-nearest Neighbours (KNN) Algorithm:

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

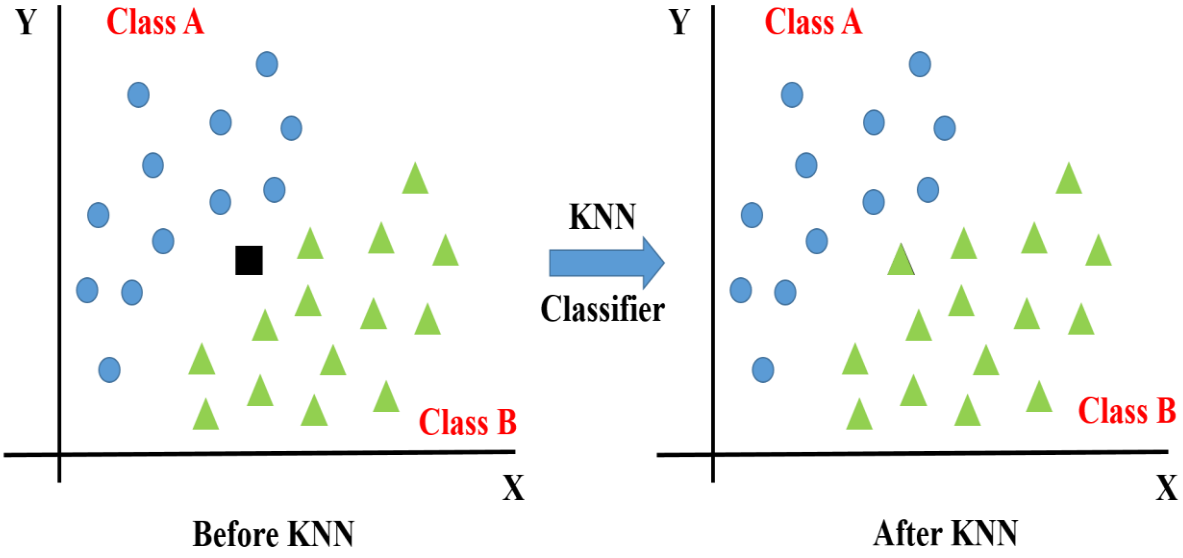

Figure 1: Visualization of KNN.

The impact of selecting a smaller or larger K value on the model

Larger K value: The case of underfitting occurs when the value of k is increased. In this case, the model would be unable to correctly learn on the training data.

Smaller k value: The condition of overfitting occurs when the value of k is smaller. The model will capture all of the training data, including noise. The model will perform poorly for the test data in this scenario.

When the problem statement is of 'classification' type, KNN tends to use the concept of "Majority Voting". Within the given range of K values, the class with the most votes is chosen. When the problem statement is of 'regression' type, KNN employs a mean/average method for predicting the value of new data. Based on the value of K, it would consider all of the nearest neighbours. The algorithm attempts to calculate the mean for all the nearest neighbours' values until it has identified all the nearest neighbours within a certain range of the K value.

Algorithm:

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Step-6: Our model is ready.

Conclusion: Using concept of KNN classification algorithms, we have classified diabetes person into two classes and evaluated KNN algorithm using evaluation metrics.

Questions:

1. What is confusion matrix?

2. Explain accuracy and error rate?

3. What is the significance of precision?

4. Explain Recall and F-1 Score?

5. Explain: 1. head() 2. shape 3. isnull() 4. drop()?