Assignment 4

Rajdhani
DATE / /

Aim:-

Implement k-nearest neighbours algorithm on diabetes.csv dataset computes confusion matrix, accuracy, error, data precision and recall on the given datasets.

Requirements:-

Jupyter notebook, python installations, python libraries - pandas, sklearn, matplotlib.

Theory:-

K-Nearest Neighbours (KNN) algorithm:

K-Nearest Neighbour is one of the simplest machine learning algorithms based on supervised learning technique.
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 stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm.

K-NN algorithm can be used for regression as well as for classification but mostly it is used for the classification problems.

Rajdhani
DATE / /

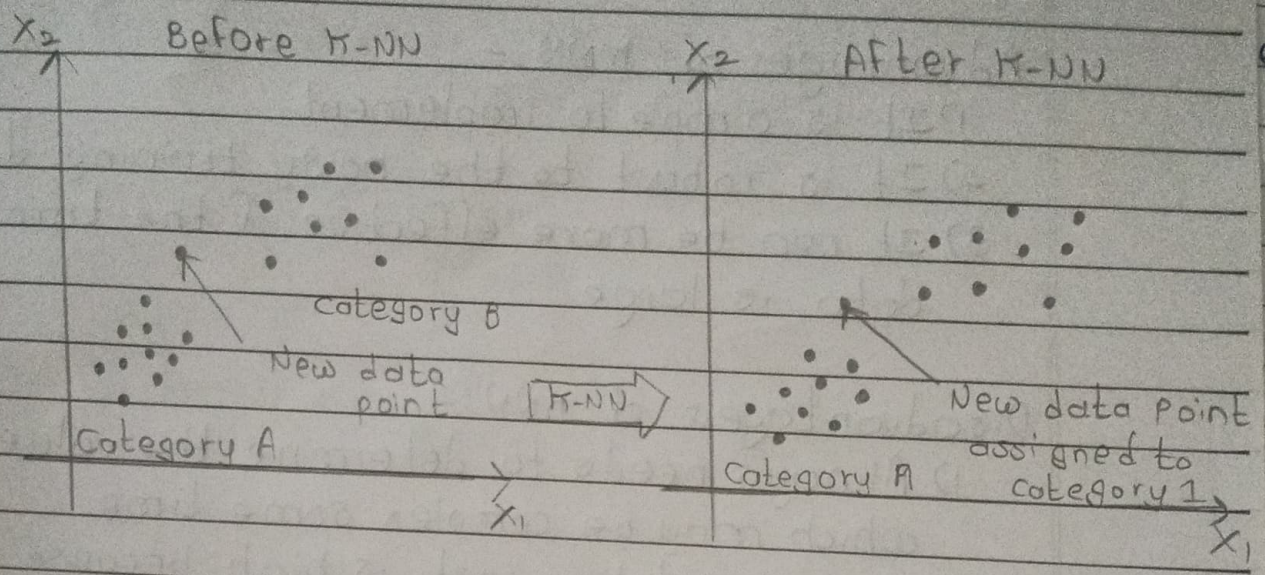
K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data

Example:-

Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog.
So for this identification, we can use the KNN algorithm, as it works on a similarity measure.
Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category

Rajdhani
DATE / /



How does KNN work?

Step 1: Select the number 'K' of the neighbours.

Step 2: Calculate the Euclidean distance of K number of neighbours.

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

Step 4: Among these K neighbours, 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 neighbour is maximum.

Step 6: Our model is ready.

Rajdhani
DATE / /

Advantages of KNN:-
1) It is simple to implement.
2) It is robust to the noisy training data.
3) It can be more effective if the training data is large.

Disadvantages of KNN:-
1) Always needs to determine the value of K which may be complex some time.
2) The computation cost is high because of calculating the distance between the data points for all the training samples.

Confusion Matrix:-
The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data.

n=total predictions	Actual : No	Actual : Yes
Predicted : No	True Negative	False Positive
Predicted : Yes	False Negative	True Positive

Calculation using the confusion matrix:-

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Rajdhani

DATE / /

$$\text{Error rate} = \frac{FP+FN}{TP+FP+FN+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F-measure} = \frac{2 * \text{recall} * \text{precision}}{\text{Recall} + \text{Precision}}$$

Correlations Matrix:

A correlation matrix is a table showing correlation coefficients between two variables. A correlation matrix is used to summarize data.

Distance metrices:

There are many different ways to compute distance as it is a fairly ambiguous notion Distance is as and the proper metrices to we is always by the dataset and task at hand

Conclusion:-

Hence, we have used KNN and computed confusion matrix, accuracy, error rate, precision and recall