Pratik Chavan - 1001963580

Harshini Kandimalla - 1001960046

Pratik Patekar - 1001937948

Report

KNN CLASSIFER

Preprocessing:

The dataset contained to zero values. The columns pres,mass,plas,skin are the columns that have the zero values.

So replaced the 0 values with NaNs so that they would not create inconsistent data.

Here NaNs are treated as missing values: disregarded in fit, and maintained in transform

Used a for loop to iterate over the dataset.

KNN:

n__neighbor – This parameter selects the how many neighbors should be checked when an item is being classified

metric – This parameter is used to distribute weight based on distance.

We have 3 widely used distance type namely,. Euclidean , Manhattan, Minkowski.

Weight – This parameter is the weight function it can be uniform or distance

Algorithm – This parameter is to select the algorithm to compute the nearest neighbor.We can use auto, ball_tree, kd_tree, brute

Criteria for selecting the three attributes.

So, we have many columns that available to use in the dataset.

But we want the select the best 3 columns as our attributes.

These attributes have maximum correlation between them.

To define the model to select the attributes I used SelectKBest class.

For classification I used ‘chi2’ method as scoring function.

We want to select 3 features so used k = 3.

Found the test, Plas, age attributes with highest scores.

Observations:

From the first visualization where we have taken n =27 which is the optimal k values,

We can see that model has accuracy of 74% with 86 True positives and 28 True Negatives

For the second visualization we took n = 21 to test

This model has lower accuracy than when n was 27.

The accuracy here is 73.37% with 83 True Positives and 30 True negatives

In the third visualization we took n = 14

This model has accuracy 70.12% with 85 True Positives and 23 True Negatives

For the last visualization we took n = 7

Accuracy goes down to 69.48% with 78 True positives and 29 True negatives.

So as we go away from the optimal k value the accuracy goes down resulting in inaccurate prediction.