**DATA MINING ASSIGNMENT -2**

**KNN REPORT**

1. **Explain all the preprocessing done in detail**

Data preprocessing is a broad terminology that refers to several activities that data scientists employ to alter their data in a form that is more suitable for their purposes. When analyzing spatial data, you can scale it to be unit-independent, which means your algorithm won't care if the source measurements were in miles or centimeters. Preprocessing data, on the other hand, does not occur in a vacuum.

scaling and centering:

It is the simplest part of data transformation. It scales and centers a variable to get a mean of 0 and a standard deviation of 1. It ensures that the criterion for finding linear combinations of predictors is based on the amount of variance they explain, improving numerical stability. Data centering and scaling are required for models that involve finding linear combinations of predictors to describe response/predictors variance.

2. **Explain all the parameters of KNN in details**

The class label is classified using neighbors using the Nearest Neighbor method. One method is to identify all the training data that have attributes that are similar to those of the test data. We calculate the distance between a test example and the remainder of the data points in the training dataset when given a test example. This method involves first training a model with the training set then using that model to determine the testing set's class.

KNN algorithm using Minkowski distance:

It's a metric for vector spaces with real values. Minkowski distance can only be calculated in a normed vector space, which is a space where distances can be represented as a vector with a length that cannot be negative.

There are a few conditions that the distance metric must satisfy:

1. Non-negativity: d(x, y) >= 0
2. Identity: d(x, y) = 0 if and only if x == y
3. Symmetry: d(x, y) = d(y, x)
4. Triangle Inequality: d(x, y) + d(y, z) >= d(x, z)

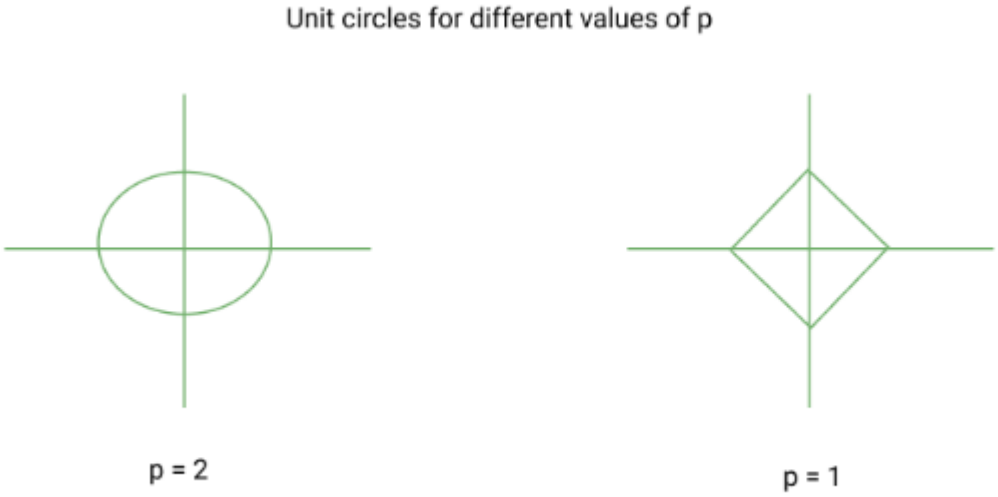
$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

The above formula for Minkowski distance is in a generalized form and we can manipulate it to get different distance metrices.

The p value in the formula can be manipulated to give us different distances like:

- p = 1, when p is set to 1 we get Manhattan distance
- p = 2, when p is set to 2 we get Euclidean distance

Let us consider a 2-dimensional space having three points P1 (X1, Y1), P2 (X2, Y2), and P3 (X3, Y3), the Minkowski distance is given by ( |X1 – Y1|p + |X2 – Y2|p + |X2 – Y2|p )1/p. In R, Minkowski distance is calculated with respect to vectors.




**3. Explain what was your criteria for selecting the three attributes**

If we see the features, we have skin, test, mass, class we need to select features that be long to skin or test but not one from each. The feature, which gives high accuracy mus t be selected. When we plot other features such as mass and class, we get less accurac y than skin and test. If we plot mass and class, “Over Fitting“ occurs. Skin and test ha s the least overfitting and high accuracy. Once it is done, skin and test give more accu racy than that of mass and class.

4. **Visualizations of the classifier in a 2D projection, and write your observations.**

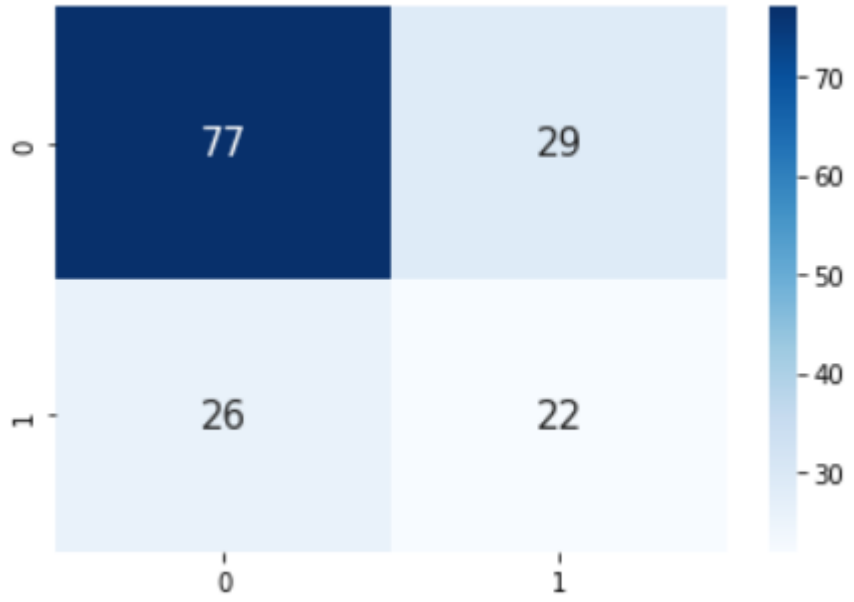
   Here there are 3 different background colors each representing 3 different class labels, represented as 0, 1, and 2.

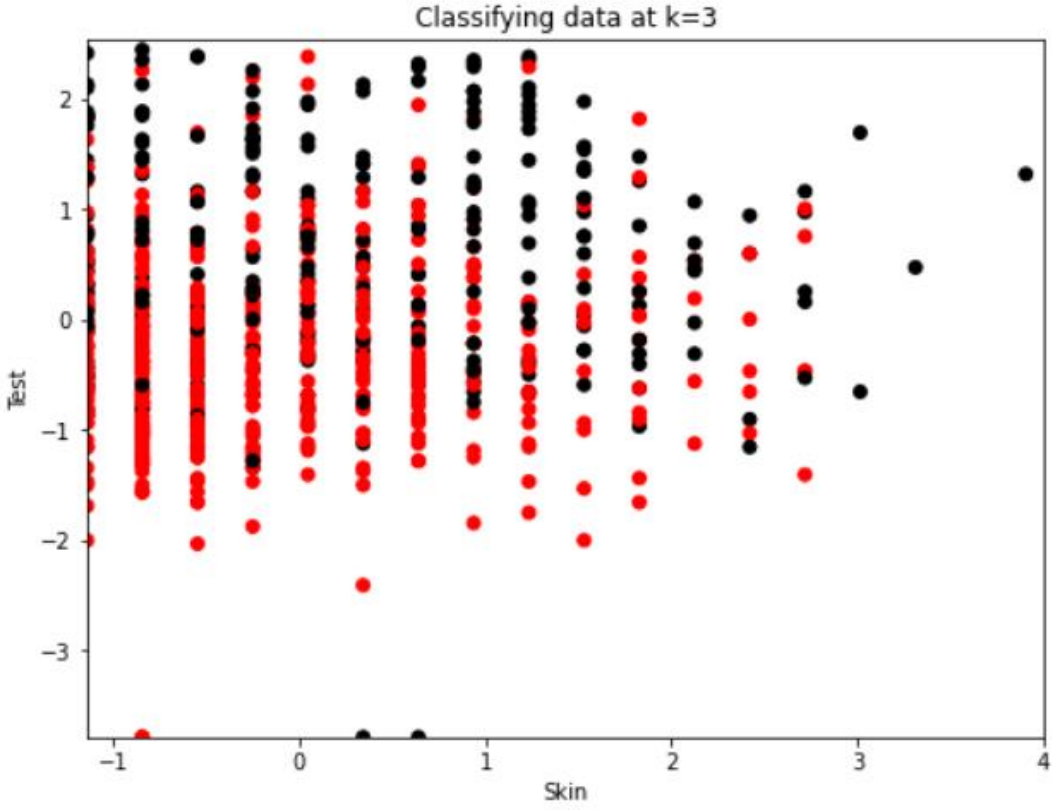
   Pink background color is for “0”

   Blue background color is for “1”

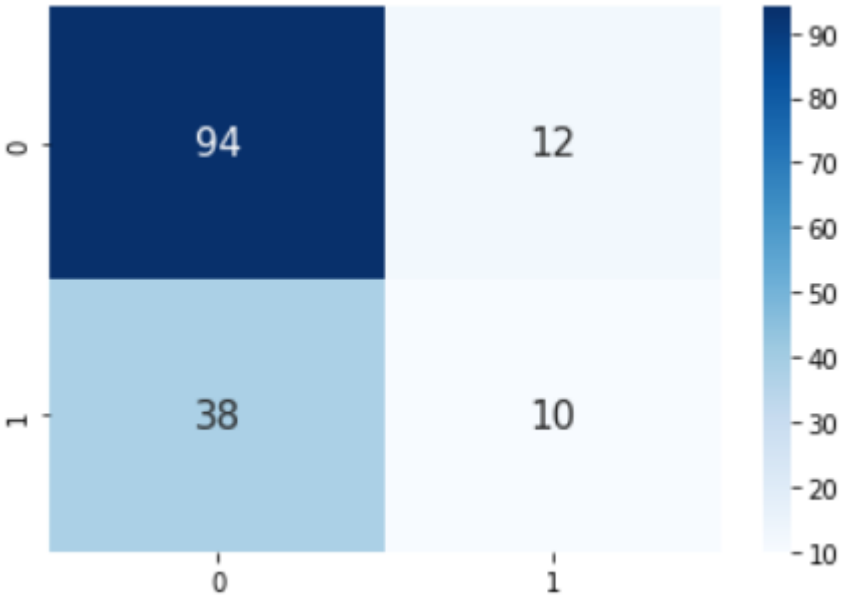
   Grey background color is for “2”

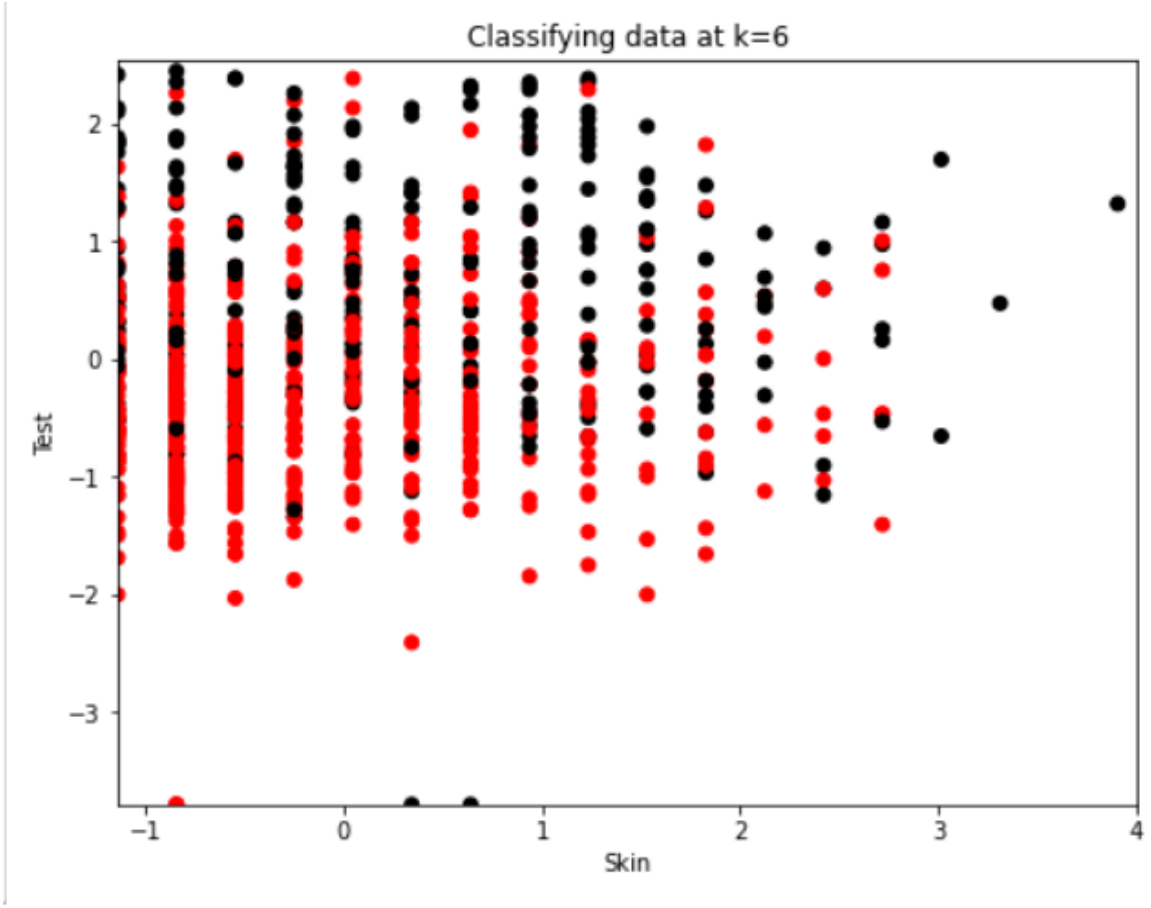
   Similarly, 3 different colored dots also represent the same class labels as mentioned above but these are of testing data class labels.



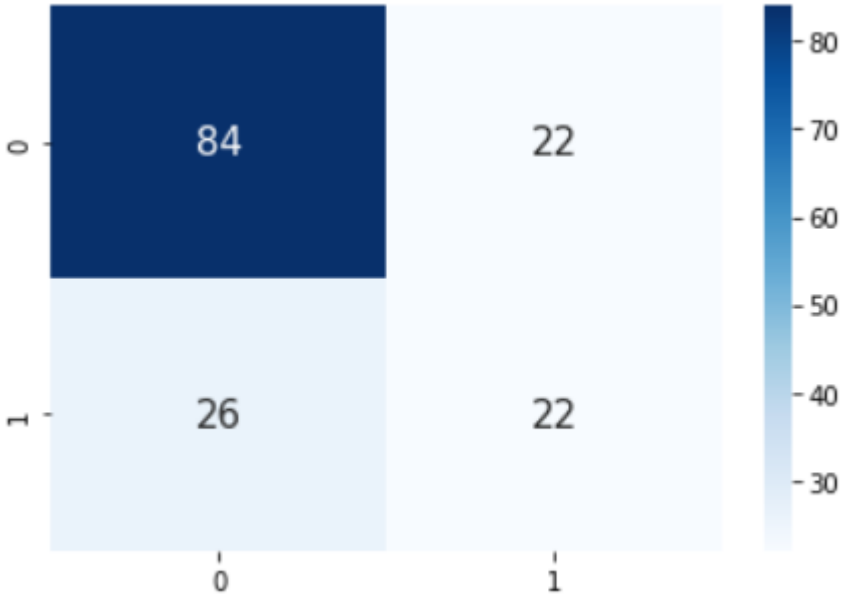


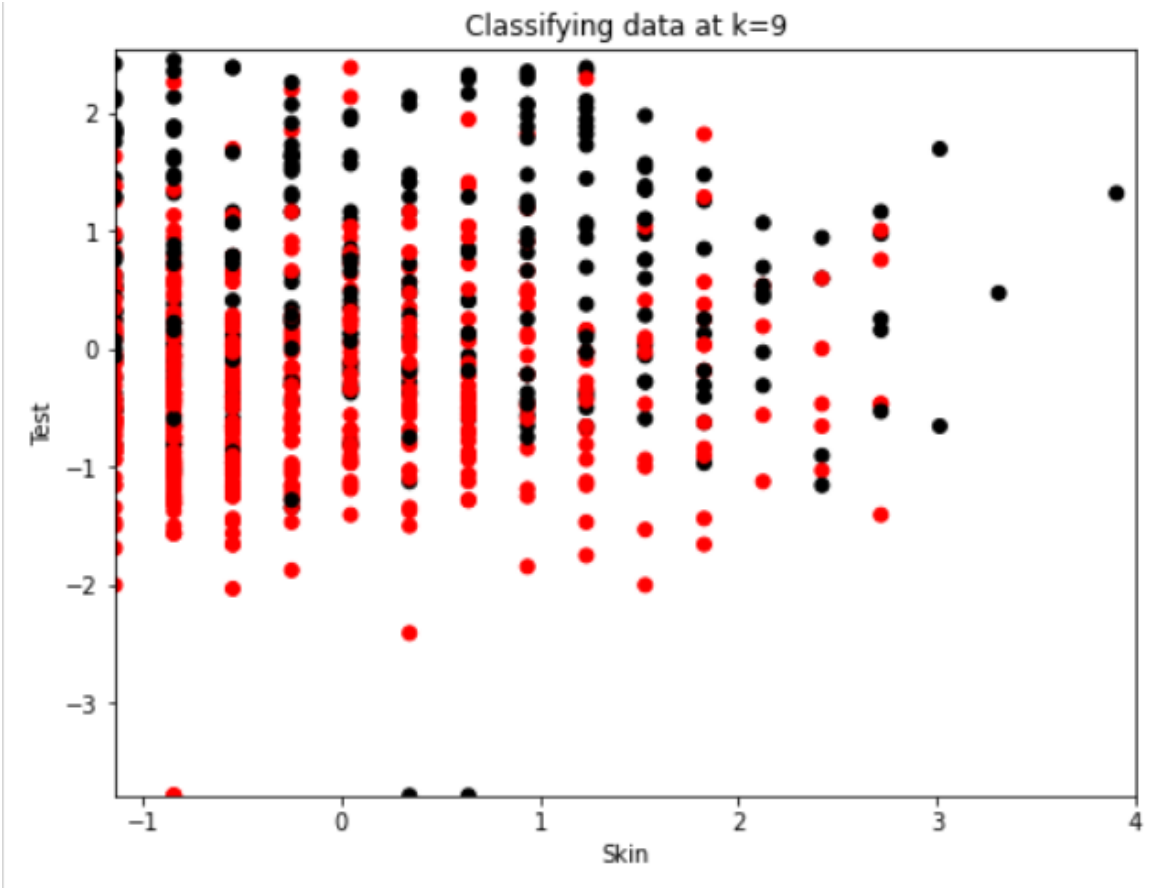

The above Figure has skin as x-axis and test as y-axis and k value is 3. The plot is a scatter plot as mentioned in the code.






The above Figure has skin as x-axis and test as y-axis and k value is 6. The plot is a scatter plot as mentioned in the code.




The above Figure has skin as x-axis and test as y-axis and k value is 9. The plot is a scatter plot as mentioned in the code.

## 5. Interpret and compare the results

- For k=3:

```
k-NN accuracy for testing set: 0.642857
k-NN score for training set: 0.817391
Accuracy: 0.6428571428571429

Confusion Matrix:
 [[77 29]
 [26 22]]

Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.73      0.74       106
           1       0.43      0.46      0.44        48

    accuracy                           0.64       154
   macro avg       0.59      0.59      0.59       154
weighted avg       0.65      0.64      0.65       154
```

Here, the column represents predicted values and rows represent actual values, which means in the 1st-row class 0 are of 106 values and there are 79 wrongly predicted classes. In the 2nd row class 1 are of a total of 48 values among which 26 are correctly classified and 22 are wrongly classified as class 2.

k-NN accuracy for testing set: 0.642857
k-NN score for training set: 0.817391

- For k=6:

```
k-NN accuracy for testing set: 0.675325
k-NN score for training set: 0.721739
Accuracy: 0.6753246753246753

Confusion Matrix:
 [[94 12]
 [38 10]]

Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.89      0.79       106
           1       0.45      0.21      0.29        48

    accuracy                           0.68       154
   macro avg       0.58      0.55      0.54       154
weighted avg       0.63      0.68      0.63       154
```

Here, the column represents predicted values and rows represent actual values, which means in the 1st-row class 0 are of 106 values and there are 94 wrongly predicted classes. In the 2nd row class, 1 is of a total of 48 values among which 38 are correctly classified as class 1 and 10 are wrongly classified as class 2.

k-NN accuracy for testing set: 0.675325
k-NN score for training set: 0.721739

- For k=9:

```
k-NN accuracy for testing set: 0.688312
k-NN score for training set: 0.717391
Accuracy: 0.6883116883116883

Confusion Matrix:
 [[84 22]
 [26 22]]

Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.79      0.78       106
           1       0.50      0.46      0.48        48

    accuracy                           0.69       154
   macro avg       0.63      0.63      0.63       154
weighted avg       0.68      0.69      0.68       154
```

Here, the column represents predicted values and rows represent actual values, which means in the 1st-row class 0 are of 106 values and there are 22 wrongly predicted classes. In the 2nd row class 1 are of a total of 48 values among which 26 are correctly classified and 22 are wrongly classified as class 2.

k-NN accuracy for testing set: 0.688312
k-NN score for training set: 0.717391

When we observe the accuracies of the 3 models k=9 has more accuracy than 0.68831 2 which means in this case more predicted values are the same as actual values than of models k=1 and k=2.