# MODULE 13

1. **Glass Dataset**

**Package**

**Package used in R**

Class - To run the KNN class

GGplot2 - Plot and Visualization

**Package used in python**

Pandas - Data Manipulation

numpy - Scientific Calculation

Matplotlib - Plot and Visualization

Sklearn - Used to import other Sklearn features

train\_test\_split - Split the data for training and test

accuracy\_score- Calculate the accuracy of the matrix

**Loading the data**

Loading the Glass dataset in R and Python

**Plot**

Ca: Most of the data lies in 8 to 9

RI: All the data lies within 1.51 to 1.53

Al: Majority of the data in 1.0 to 2.0

Na: Majority of the data in 12 to 14

Si: Majority of the data in 72.5 to 73.5

K : Majority of the data lies in 0 and 1

Ba: Majority of the data lies in 0

Fe: Majority of the data lies in 0

From the histogram, found that the data set has scaling problem on all the variable.

**EDA**

1. No Missing Value found
2. No Outliers Found
3. Min Max Method used to normalize all the variables in the dataset. Minimum value will be 0 and maximum value will be 1.

**Data Partitioning**

Normalized dataset is split into training and test data on non random sample. As the observation in the dataset already random.

Predicting variable is created as a training and test label.

**Modeling**

K = 1

glass\_test\_pred <- knn(train = glasstrain, test = glasstest,cl = glass\_train\_labels, k = 1)

Training data is tested in test data with the k value of 1.

Confusion matrix is done to check the Prediction and Accuracy of the model

Accuracy : 0.296

K =1

glass\_train\_pred <- knn(train = glasstrain, test = glasstrain, cl = glass\_train\_labels, k=1)

Training data is tested in Training data with the k value of 1.

Confusion matrix is done to check the Prediction and Accuracy of the model

Accuracy : 1

K = 2

glass\_train\_pred <- knn(train = glasstrain, test = glasstrain, cl = glass\_train\_labels, k=1)

Training data is tested in Training data with the k value of 1.

Confusion matrix is done to check the Prediction and Accuracy of the model

Accuracy : .84

Accuracy got reduced in K = 2

KNN Model on Loop

Loop Value is 1 to 15

Accuracy and error Plot is drawn from the Loop Model and to determine the best K value.

**K = 16 has high Accuracy on both test and test accuracy compared to other K values.**

1. **Zoo Dataset**

**Package**

**Package used in R**

Class - To run the KNN class

GGplot2 - Plot and Visualization

**Package used in python**

Pandas - Data Manipulation

numpy - Scientific Calculation

Matplotlib - Plot and Visualization

Sklearn - Used to import other Sklearn features

train\_test\_split - Split the data for training and test

accuracy\_score- Calculate the accuracy of the matrix

**Loading the data**

Loading the Zoo dataset in R and Python

**Plot**

Type :Type 1 has occupied the majority of the data.

Legs : Only 1 5 legged animal found in the dataset Starfish.

Breathes : 20 % of the dataset is water living animals.

Eggs : 60 % of the dataset is water living and reptile.

Feathers : Only 20 % dataset is related to Birds

Milk : 40% of the dataset is mammal

**EDA**

1. No Missing Value found
2. No Outliers Found
3. Dataset has no Scaling Issue as the majority of the variables are 0 and 1
4. Name Variable removed

**Data Partitioning**

Normalized dataset is split into training and test data on non random sample. As the observation in the dataset already random.

Training Data **1 to 80** and Testing data **81 to 101**

Predicting variable is created as a training and test label.

**Modeling**

K = 1

zoo\_test\_pred <- knn(train = zootrain, test = zootest,cl = zoo\_train\_labels, k = 1)

Training data is tested in test data with the k value of 1.

Confusion matrix is done to check the Prediction and Accuracy of the model

Accuracy : 0.904

K =1

zoo\_train\_pred1 <- knn(train = zootrain, test = zootrain, cl = zoo\_train\_labels, k=1)

Training data is tested in Training data with the k value of 1.

Confusion matrix is done to check the Prediction and Accuracy of the model

Accuracy : 1

K = 5

zoo\_train\_pred2 <- knn(train = zootrain, test = zootrain, cl = zoo\_train\_labels, k=5)

Training data is tested in Training data with the k value of 1.

Confusion matrix is done to check the Prediction and Accuracy of the model

Accuracy : .85

Accuracy got reduced in K =5

KNN Model on Loop

Loop Value is 1 to 5

Accuracy and error Plot is drawn from the Loop Model and to determine the best K value.

**K = 7 has high Accuracy on both test and test accuracy compared to other K values.**