NPTEL-DATA SCIENCE FOR ENGINEERS

ASSIGNMENT 8

Consider the information given below and answer questions 1 - 3.

A social networking company has a website where they publish advertisements about the various products. Their employees shop these products. They have collected the details of their employees and stored them in two different datasets, ***social_network_train.csv*** and ***social_network_test.csv*****. The objective is to** find whether the user buys a product by clicking the ad on the site or not. The data has two independent variables (***Age***, ***Estimated Salary***), and one response variable (***Purchased***). The response variable has two values **0** (not purchased) and **1** (purchased successfully).

Variables	Description
Age	Age of the user
EstimatedSalary	Estimated salary of the user
Purchased	0 (not purchased) or 1 (purchased successfully)

Read the datasets ***social_network_train.csv*** and ***social_network_test.csv*** as ***train_data*** and ***test_data,*** respectively (The variable ***Purchased*** should be considered as factor data, and not integer data)

Build a **kNN** model over ***train_data*** by considering 3 nearest neighbors.

1. What is the accuracy (in %) of the model over ***test_data***? **(Choose the appropriate range)**

Solution: c

```
> predictedknn <- knn(train = train_data[,-3],
+                     test = test_data[,-3],
+                     cl = train_data$Purchased,
+                     k  = 3)
> COnF_Matrix <-confusionMatrix(data = predictedknn,test_data$Purchased)
> COnF_Matrix
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 68 11
         1 14 27

               Accuracy : 0.7917
                 95% CI : (0.708, 0.8604)
    No Information Rate : 0.6833
    P-Value [Acc > NIR] : 0.005712
```

2. Total number of misclassified samples obtained from the prediction of the kNN model built in Q1 are: -

Solution: a

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 68 11
         1 14 27
```

The number of misclassified samples are 11 + 14 = 25

3. The sensitivity of the kNN model built in Q1 is: - **(rounded off to two decimal points)**

Solution: b

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 68 11
         1 14 27

               Accuracy : 0.7917
                 95% CI : (0.708, 0.8604)
    No Information Rate : 0.6833
    P-Value [Acc > NIR] : 0.005712

                  Kappa : 0.5286

 Mcnemar's Test P-Value : 0.689157

            Sensitivity : 0.8293
            Specificity : 0.7105
         Pos Pred Value : 0.8608
         Neg Pred Value : 0.6585
             Prevalence : 0.6833
         Detection Rate : 0.5667
   Detection Prevalence : 0.6583
      Balanced Accuracy : 0.7699

       'Positive' Class : 0
```

Consider the information given below and answer the questions 4 to 7.

Consider the dataset ***Wholesale_customers_data.csv***, which refers to clients of a wholesale distributor. The objective is to segment the clients of a wholesale distributor based on their annual spending in monetary units (m.u.) on diverse product categories.

Variables	Variable Description
Fresh	annual spending (m.u.) on fresh products
Milk	annual spending (m.u.) on milk products
Grocery	annual spending (m.u.) on grocery products
Frozen	annual spending (m.u.) on frozen products
Detergents_Paper	annual spending (m.u.) on detergents and paper products
Delicassen	annual spending (m.u.) on and delicatessen products
Channel	customers' Channel – 1-Horeca (Hotel/Restaurant/Café) or 2-Retail channel
Region	customers' Region – 1-Lisbon, 2-Oporto or 3-Other

Read the dataset ***Wholesale_customers_data.csv*** as ***data*** **and answer questions from 4 to 10**

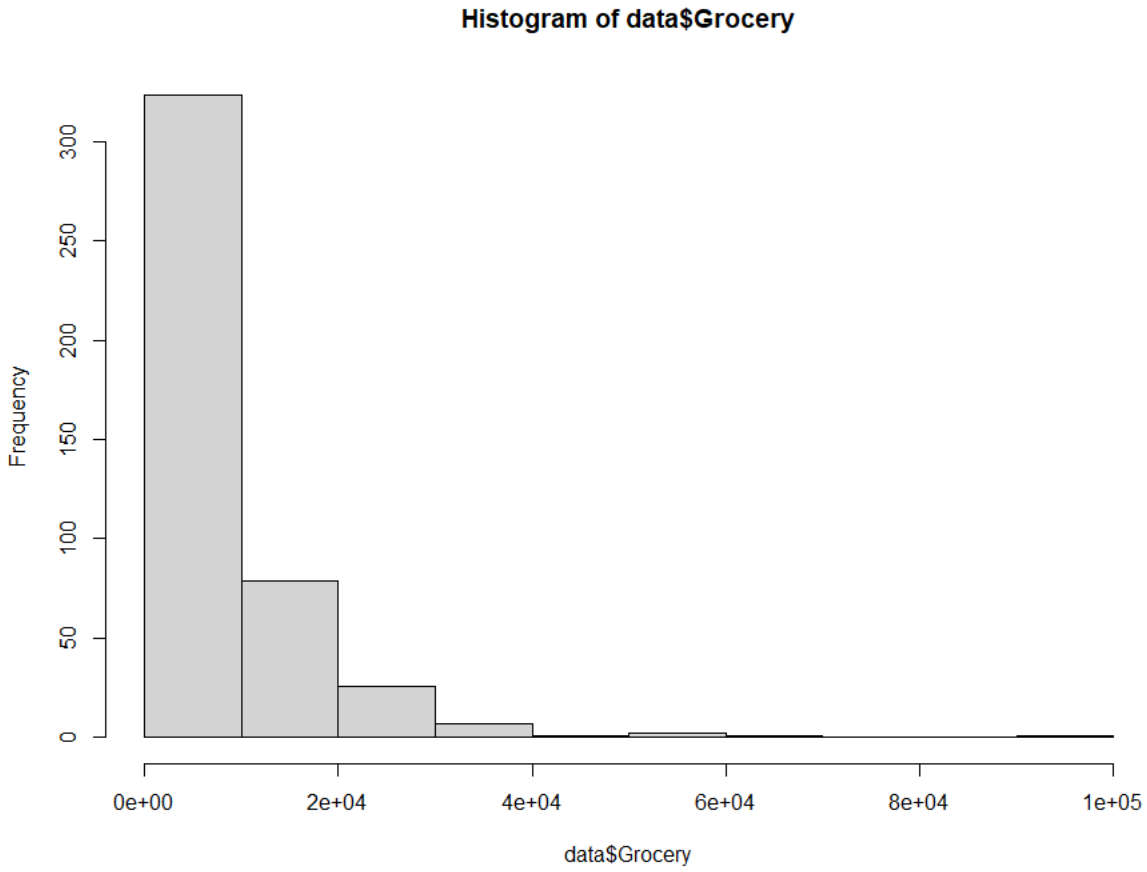
4. Which of the following products has the highest annual spending unit?

Solution: a

```
> summary(data)
    Channel          Region          Fresh             Milk          Grocery
 Min.   :1.000   Min.   :1.000   Min.   :     3   Min.   :   55   Min.   :    3
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:  3128   1st Qu.: 1533   1st Qu.: 2153
 Median :1.000   Median :3.000   Median :  8504   Median : 3627   Median : 4756
 Mean   :1.323   Mean   :2.543   Mean   : 12000   Mean   : 5796   Mean   : 7951
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 16934   3rd Qu.: 7190   3rd Qu.:10656
 Max.   :2.000   Max.   :3.000   Max.   :112151   Max.   :73498   Max.   :92780
     Frozen        Detergents_Paper    Delicassen
 Min.   :   25.0   Min.   :    3.0   Min.   :    3.0
 1st Qu.:  742.2   1st Qu.:  256.8   1st Qu.:  408.2
 Median : 1526.0   Median :  816.5   Median :  965.5
 Mean   : 3071.9   Mean   : 2881.5   Mean   : 1524.9
 3rd Qu.: 3554.2   3rd Qu.: 3922.0   3rd Qu.: 1820.2
 Max.   :60869.0   Max.   :40827.0   Max.   :47943.0
```

5. Which of the following interpretations are true with respect to the distribution of the variable ***Grocery***?

Solution: c



6. What is the percentage of customers buying the product from different region?

Solution: a

```
> region_split=round(table(data$Region)/sum(table(data$Region))*100)
> print(region_split)

 1  2  3
18 11 72
```

Segment the clients into four groups based on their annual spending on diverse product categories using K-Means clustering. Follow the conditions mentioned below before implementing the K_Means model to the data and answer questions from 7 to 10

- Drop the variables “Channel” and “Region”
- Normalize the data using scale() function
- Random number generator should be set to 123 using set.seed attribute (Should be executed before building the model)

7. The Within Cluster Sum-of-Squares (WCSS values for each cluster (in no specific order) are: -

Solution: a

```
> data=read.csv("Wholesale_customers_data.csv",header=T)
> data <- data[ -c(1,2)]
> data=scale(data)
> set.seed(123)
> dataCluster <- kmeans(data,4)
> dataCluster$withinss
[1] 440.1481 188.1085 490.5957 235.0199
```

8. The size of each cluster (in no specific order) is: -

Solution: b

```
> dataCluster$size
[1]  63  96  12 269
```

9. What is the Between Cluster Sum-of-Squares (BCSS) value of the K-means model? **(Choose the appropriate range)**

Solution: c

```
> dataCluster$betweenss
[1] 1280.128
```

10. What is the Total Sum-of-Squares value of the k-means model? **(Choose the appropriate range)**

Solution: a

```
> dataCluster$totss
[1] 2634
```

11. Elbow plot can be used to decide the optimal k value in both kNN and K-means clustering problems.

Solution: a

Yes, Elbow plot can be used to decide the optimal k value in both kNN and K-means clustering problems.

12. The most commonly used distance metric to calculate distance between centroid of each cluster and data points in K-means algorithm is

Solution: c

Euclidean distance is the most used distance metric to calculate distance between centroid of each cluster and data points in K-means algorithm.

13. A k-Means Clustering model becomes better as

Solution: c

A k-Means Clustering model is effective only when it clusters the data points such that the data points of similar class are closer to each other, and the data points of different classes are farther to the data points corresponding to other classes.