DATA SCIENCE PROGRAMMING LAB (L3+L4)

ASSESSMENT 3

K-Nearest Neighbor IN R

Name: SWATHI D

Reg. no.: 22MID0035

Procedure for KNN classification

1. Load Libraries

```
library(dplyr)
library(stringr)
library(caTools)
library(class)
```

```
> library(dplyr)
> library(stringr)
> library(caTools)
> library(class)
>
```

2. Load Dataset

```
data <- read.csv("C:/Users/HP/Downloads/loan_data.csv", header = TRUE, check.names = FALSE)
head (data)
summary(data)
```

```
> data <- read.csv("C:/Users/HP/Downloads/loan_data.csv", header = TRUE, check.names = FALSE)
> head (data)
  credit.policy            purpose int.rate installment log.annual.inc   dti fico days.with.cr.line revol.bal
1             1 debt_consolidation   0.1189      829.10       11.35041 19.48  737          5639.958     28854
2             1        credit_card   0.1071      228.22       11.08214 14.29  707          2760.000     33623
3             1 debt_consolidation   0.1357      366.86       10.37349 11.63  682          4710.000      3511
4             1 debt_consolidation   0.1008      162.34       11.35041  8.10  712          2699.958     33667
5             1        credit_card   0.1426      102.92       11.29973 14.97  667          4066.000      4740
6             1        credit_card   0.0788      125.13       11.90497 16.98  727          6120.042     50807
  revol.util inq.last.6mths delinq.2yrs pub.rec not.fully.paid
1       52.1              0           0       0              0
2       76.7              0           0       0              0
3       25.6              1           0       0              0
4       73.2              1           0       0              0
5       39.5              0           1       0              0
6       51.0              0           0       0              0
```

```
> summary(data)
 credit.policy     purpose               int.rate        installment      log.annual.inc        dti
 Min.   :0.000   Length:9578         Min.   :0.0600   Min.   : 15.67   Min.   : 7.548   Min.   : 0.000
 1st Qu.:1.000   Class :character    1st Qu.:0.1039   1st Qu.:163.77   1st Qu.:10.558   1st Qu.: 7.213
 Median :1.000   Mode  :character    Median :0.1221   Median :268.95   Median :10.929   Median :12.665
 Mean   :0.805                       Mean   :0.1226   Mean   :319.09   Mean   :10.932   Mean   :12.607
 3rd Qu.:1.000                       3rd Qu.:0.1407   3rd Qu.:432.76   3rd Qu.:11.291   3rd Qu.:17.950
 Max.   :1.000                       Max.   :0.2164   Max.   :940.14   Max.   :14.528   Max.   :29.960
      fico         days.with.cr.line   revol.bal          revol.util      inq.last.6mths     delinq.2yrs
 Min.   :612.0   Min.   :  179      Min.   :       0   Min.   :  0.0    Min.   : 0.000   Min.   : 0.0000
 1st Qu.:682.0   1st Qu.: 2820      1st Qu.:    3187   1st Qu.: 22.6    1st Qu.: 0.000   1st Qu.: 0.0000
 Median :707.0   Median : 4140      Median :    8596   Median : 46.3    Median : 1.000   Median : 0.0000
 Mean   :710.8   Mean   : 4561      Mean   :   16914   Mean   : 46.8    Mean   : 1.577   Mean   : 0.1637
 3rd Qu.:737.0   3rd Qu.: 5730      3rd Qu.:   18250   3rd Qu.: 70.9    3rd Qu.: 2.000   3rd Qu.: 0.0000
 Max.   :827.0   Max.   :17640      Max.   : 1207359   Max.   :119.0    Max.   :33.000   Max.   :13.0000
    pub.rec          not.fully.paid
 Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:0.0000
 Median :0.00000   Median :0.0000
 Mean   :0.06212   Mean   :0.1601
 3rd Qu.:0.00000   3rd Qu.:0.0000
 Max.   :5.00000   Max.   :1.0000
```

3. Drop Unwanted Column

```
data <- subset(data, select = -c(purpose))
head(data,3)
```

```
> data <- subset(data, select = -c(purpose))
> head(data,3)
  credit.policy int.rate installment log.annual.inc   dti fico days.with.cr.line revol.bal revol.util inq.last.6mths
1             1   0.1189      829.10       11.35041 19.48  737          5639.958     28854       52.1              0
2             1   0.1071      228.22       11.08214 14.29  707          2760.000     33623       76.7              0
3             1   0.1357      366.86       10.37349 11.63  682          4710.000      3511       25.6              1
  delinq.2yrs pub.rec not.fully.paid
1           0       0              0
2           0       0              0
3           0       0              0
```

4. Train-Test Split

```
set.seed(255)

split = sample.split(data$not.fully.paid,
                 SplitRatio = 0.75)
train = subset(data,
             split == TRUE)
test = subset(data,
             split == FALSE)
```

```
> set.seed(255)
> split = sample.split(data$not.fully.paid,
+                 SplitRatio = 0.75)
> train = subset(data,
+             split == TRUE)
> test = subset(data,
+             split == FALSE)
```

5. Feature Scaling

```
train_scaled = scale(train[-13])
test_scaled = scale(test[-13])
```

```
> train_scaled = scale(train[-13])
> test_scaled = scale(test[-13])
```

6. Train KNN Classifier

```
test_pred <- knn(
   train = train_scaled,
   test = test_scaled,
   cl = train$not.fully.paid,
   k=10
)

actual <- test$not.fully.paid
```

```
> test_pred <- knn(
+    train = train_scaled,
+    test = test_scaled,
+    cl = train$not.fully.paid,
+    k=10
+ )
> actual <- test$not.fully.paid
```

7. Model Evaluation

```
cm <- table(actual,test_pred)
cm

accuracy <- sum(diag(cm))/length(actual)
sprintf("Accuracy: %.2f%%", accuracy*100)
```

```
> cm <- table(actual,test_pred)
> cm
      test_pred
actual    0    1
     0 1988   23
     1  373   10
> accuracy <- sum(diag(cm))/length(actual)
> sprintf("Accuracy: %.2f%%", accuracy*100)
[1] "Accuracy: 83.46%"
```