Assignment_3

Importing the Dataset

```
data<-read.csv("universalBank.csv")
```

```
summary(data)
```

```
##        ID            Age          Experience       Income          ZIP.Code
##  Min.   :   1   Min.   :23.00   Min.   :-3.0   Min.   :  8.00   Min.   : 9307
##  1st Qu.:1251   1st Qu.:35.00   1st Qu.:10.0   1st Qu.: 39.00   1st Qu.:91911
##  Median :2500   Median :45.00   Median :20.0   Median : 64.00   Median :93437
##  Mean   :2500   Mean   :45.34   Mean   :20.1   Mean   : 73.77   Mean   :93152
##  3rd Qu.:3750   3rd Qu.:55.00   3rd Qu.:30.0   3rd Qu.: 98.00   3rd Qu.:94608
##  Max.   :5000   Max.   :67.00   Max.   :43.0   Max.   :224.00   Max.   :96651
##      Family          CCAvg          Education        Mortgage
##  Min.   :1.000   Min.   : 0.000   Min.   :1.000   Min.   :  0.0
##  1st Qu.:1.000   1st Qu.: 0.700   1st Qu.:1.000   1st Qu.:  0.0
##  Median :2.000   Median : 1.500   Median :2.000   Median :  0.0
##  Mean   :2.396   Mean   : 1.938   Mean   :1.881   Mean   : 56.5
##  3rd Qu.:3.000   3rd Qu.: 2.500   3rd Qu.:3.000   3rd Qu.:101.0
##  Max.   :4.000   Max.   :10.000   Max.   :3.000   Max.   :635.0
##  Personal.Loan   Securities.Account   CD.Account         Online
##  Min.   :0.000   Min.   :0.0000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.000   Median :0.0000     Median :0.0000   Median :1.0000
##  Mean   :0.096   Mean   :0.1044     Mean   :0.0604   Mean   :0.5968
##  3rd Qu.:0.000   3rd Qu.:0.0000     3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.0000     Max.   :1.0000   Max.   :1.0000
##    CreditCard
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.294
##  3rd Qu.:1.000
##  Max.   :1.000
```

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```r
library(class)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ISLR)
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(FNN)
```

```
##
## Attaching package: 'FNN'

## The following objects are masked from 'package:class':
##
##     knn, knn.cv
```

```r
library(lattice)
```

#Removing ID and ZIP Code

```r
data$ID <- NULL
data$ZIP.Code <- NULL
data$Education = as.factor(data$Education)
```

creating a dummy dataset

```r
dummy_var <- as.data.frame(dummy.code(data$Education))
```

```
names(dummy_var) <- c("Education_1", "Education_2","Education_3")
```

Setting education to NULL

```
data$Education <- NULL
```

```
data_2<- cbind(data, dummy_var)
```

Dividing the dataset into train and test data

```
set.seed(1)
train.index <- createDataPartition(data_2$Personal.Loan, p= 0.6 , list=FALSE)
valid.index <- setdiff(row.names(data_2), train.index)
train.dataset <- data_2[train.index,]
valid.dataset <- data_2[valid.index,]
```

Generating the Test data

```
data_customer<- data.frame(Age = 40,
                        Experience = 10,
                        Income = 84,
                        Family = 2,
                        CCAvg = 2,
                        Mortgage = 0,
                        Securities.Account = 0,
                        CD.Account = 0,
                        Online = 1,
                        CreditCard = 1,
                        Education_1 = 0,
                        Education_2 = 1,
                        Education_3 = 0)
```

Data normalisation

```
train_norm <- train.dataset[,-7]
valid_norm <- valid.dataset[,-7]
data_norm <- data_customer

normalisation.values <- preProcess(train.dataset[, -7], method=c("center", "scale"))
train_norm <- predict(normalisation.values, train.dataset[, -7])
valid.normalisation.dataset <- predict(normalisation.values, valid.dataset[, -7])
data_norm <- predict(normalisation.values, data_norm)
```

```
summary(train_norm)
```

```
##       Age               Experience           Income             Family
##  Min.   :-1.97257   Min.   :-2.03718   Min.   :-1.4240   Min.   :-1.2058
##  1st Qu.:-0.82922   1st Qu.:-0.89531   1st Qu.:-0.7457   1st Qu.:-1.2058
##  Median :-0.03767   Median :-0.01695   Median :-0.2206   Median :-0.3368
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.84183   3rd Qu.: 0.86141   3rd Qu.: 0.5452   3rd Qu.: 0.5321
##  Max.   : 1.89723   Max.   : 2.00328   Max.   : 3.3022   Max.   : 1.4010
##      CCAvg             Mortgage       Securities.Account   CD.Account
##  Min.   :-1.1059   Min.   :-0.5679   Min.   :-0.3339    Min.   :-0.2381
##  1st Qu.:-0.7016   1st Qu.:-0.5679   1st Qu.:-0.3339    1st Qu.:-0.2381
##  Median :-0.2396   Median :-0.5679   Median :-0.3339    Median :-0.2381
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000    Mean   : 0.0000
##  3rd Qu.: 0.3380   3rd Qu.: 0.4423   3rd Qu.:-0.3339    3rd Qu.:-0.2381
##  Max.   : 4.6700   Max.   : 5.7216   Max.   : 2.9940    Max.   : 4.1985
##      Online          CreditCard       Education_1       Education_2
##  Min.   :-1.1863   Min.   :-0.6431   Min.   :-0.8462   Min.   :-0.6509
##  1st Qu.:-1.1863   1st Qu.:-0.6431   1st Qu.:-0.8462   1st Qu.:-0.6509
##  Median : 0.8427   Median :-0.6431   Median :-0.8462   Median :-0.6509
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.8427   3rd Qu.: 1.5544   3rd Qu.: 1.1814   3rd Qu.: 1.5358
##  Max.   : 0.8427   Max.   : 1.5544   Max.   : 1.1814   Max.   : 1.5358
##   Education_3
##  Min.   :-0.6312
##  1st Qu.:-0.6312
##  Median :-0.6312
##  Mean   : 0.0000
##  3rd Qu.: 1.5836
##  Max.   : 1.5836
```

Performing Knn classification, using K=1

```
knn_data <- class::knn(train = train_norm,test = data_norm,
                       cl = train.dataset$Personal.Loan, k = 1)

print(knn_data)
```

```
## [1] 0
## Levels: 0 1
```

Finding best K value

```
k_value <- data.frame(k = seq(1, 10, 1), accuracy = rep(0, 10))
```

```
for(i in 1:10) {
  knn_prediction <- class::knn(train = train_norm,
                         test = valid.normalisation.dataset,
                         cl = train.dataset$Personal.Loan, k = i)
  k_value[i, 2] <- confusionMatrix(knn_prediction,
                                   as.factor(valid.dataset$Personal.Loan))$overall[1]
}
which(k_value[,2] == max(k_value[,2]))
```

```
## [1] 3
```

```
k_value
```

```
##     k accuracy
## 1   1   0.9630
## 2   2   0.9565
## 3   3   0.9640
## 4   4   0.9595
## 5   5   0.9605
## 6   6   0.9575
## 7   7   0.9580
## 8   8   0.9575
## 9   9   0.9535
## 10 10   0.9550
```

##choosing k = 3

```
knn_prediction <- class::knn(train = train_norm,
                       test = valid.normalisation.dataset,
                       cl = train.dataset$Personal.Loan, k = 3)

confusionMatrix(knn_prediction, as.factor(valid.dataset$Personal.Loan), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1786   63
##          1    9  142
##
##                Accuracy : 0.964
##                  95% CI : (0.9549, 0.9717)
##     No Information Rate : 0.8975
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7785
##
##  Mcnemar's Test P-Value : 4.208e-10
##
##             Sensitivity : 0.6927
##             Specificity : 0.9950
##          Pos Pred Value : 0.9404
```

```
##          Neg Pred Value : 0.9659
##              Prevalence : 0.1025
##          Detection Rate : 0.0710
##    Detection Prevalence : 0.0755
##       Balanced Accuracy : 0.8438
##
##        'Positive' Class : 1
##
```

Confusion matrix for the best k value =3

```
newcustomer <- data.frame(Age = 40,
                          Experience = 10,
                          Income = 84,
                          Family = 2,
                          CCAvg = 2,
                          Mortgage = 0,
                          Securities.Account = 0,
                          CD.Account = 0,
                          Online = 1,
                          CreditCard = 1,
                          Education_1 = 0,
                          Education_2 = 1,
                          Education_3 = 0)

fitknn <-class::knn(train = train_norm,
                    test = newcustomer,
                    cl = train.dataset$Personal.Loan, k = 3)

fitknn
```

```
## [1] 1
## Levels: 0 1
```

Knn model tells that new customer will accept loan

```
data<- read.csv("universalBank.csv")
```

Loading packages

```
library(ISLR)
library(psych)
library(caret)
library(FNN)
library(class)
```

```
library(dplyr)
library(lattice)
```

Removing id and zipcode variables from the dataset

```
data$ID <- NULL
data$ZIP.Code <- NULL
data$Education = as.factor(data$Education)
```

Creating dummy dataframe

```
dummymod <- as.data.frame(dummy.code(data$Education))
```

##Renaming the data frame

```
names(dummymod) <- c("Education_1", "Education_2","Education_3")
```

Deleting education variable

```
data$Education <- NULL
```

##Main dataset

```
data_2 <- cbind(data, dummymod)
```

#Partitioning the dataset

```
set.seed(1)
train.index <- createDataPartition(data_2$Personal.Loan, p= 0.5 , list=FALSE)
valid.index <- createDataPartition(data_2$Personal.Loan, p= 0.3 , list=FALSE)
test.index <- setdiff(row.names(data_2), union(train.index, valid.index))

train.dataset <- data_2[train.index, ]
valid.dataset <- data_2[valid.index, ]
test.dataset <- data_2[test.index, ]
```

##Performing normalisation

```
train_norm <- train.dataset[,-7]
valid.normalisation.dataset <- valid.dataset[,-7]
test.normalisation.dataset <- test.dataset[,-7]

normalisation.values <- preProcess(train.dataset[, -7], method=c("center", "scale"))
train_norm<- predict(normalisation.values, train.dataset[, -7])
valid.normalisation.dataset <- predict(normalisation.values, valid.dataset[, -7])
test.normalisation.dataset <- predict(normalisation.values, test.dataset[, -7])
```

Performing Knn classification using K=3

```
knn.test.pred <- class::knn(train = train_norm,
                      test = test.normalisation.dataset,
                      cl = train.dataset$Personal.Loan, k = 3)

knn.train.pred <- class::knn(train = train_norm,
                             test = train_norm,
                             cl = train.dataset$Personal.Loan, k = 3)

knn.valid.pred <- class::knn(train = train_norm,
                             test = valid.normalisation.dataset,
                             cl = train.dataset$Personal.Loan, k = 3)
```

##Confusion matrix for K=3

```
confusionMatrix(knn.test.pred, as.factor(test.dataset$Personal.Loan), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1590   50
##          1    8  111
##
##                Accuracy : 0.967
##                  95% CI : (0.9576, 0.9749)
##     No Information Rate : 0.9085
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7754
##
##  Mcnemar's Test P-Value : 7.303e-08
##
##             Sensitivity : 0.68944
##             Specificity : 0.99499
##          Pos Pred Value : 0.93277
##          Neg Pred Value : 0.96951
##              Prevalence : 0.09153
##          Detection Rate : 0.06310
##    Detection Prevalence : 0.06765
##       Balanced Accuracy : 0.84222
##
##        'Positive' Class : 1
##
```

```
confusionMatrix(knn.train.pred, as.factor(train.dataset$Personal.Loan), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction    0    1
##          0 2263   54
##          1    5  178
##
##                Accuracy : 0.9764
##                  95% CI : (0.9697, 0.982)
##     No Information Rate : 0.9072
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8452
##
##  Mcnemar's Test P-Value : 4.129e-10
##
##             Sensitivity : 0.7672
##             Specificity : 0.9978
##          Pos Pred Value : 0.9727
##          Neg Pred Value : 0.9767
##              Prevalence : 0.0928
##          Detection Rate : 0.0712
##    Detection Prevalence : 0.0732
##       Balanced Accuracy : 0.8825
##
##        'Positive' Class : 1
##
```

```
confusionMatrix(knn.valid.pred, as.factor(valid.dataset$Personal.Loan), positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1347   43
##          1    3  107
##
##                Accuracy : 0.9693
##                  95% CI : (0.9593, 0.9775)
##     No Information Rate : 0.9
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8067
##
##  Mcnemar's Test P-Value : 8.912e-09
##
##             Sensitivity : 0.71333
##             Specificity : 0.99778
##          Pos Pred Value : 0.97273
##          Neg Pred Value : 0.96906
##              Prevalence : 0.10000
##          Detection Rate : 0.07133
##    Detection Prevalence : 0.07333
##       Balanced Accuracy : 0.85556
##
##        'Positive' Class : 1
##
```