**Assignment 7**

Overdrawn.csv has data collected from "Sensation-Seeking, Risk-Taking, and Problematic Financial Behaviors of College Students," by Worthy S.L., Jonkman J.N., Blinn-Pike L. (2010).

The dataset contains following variables:

| | |
|---|---|
| Age | Age of the student (in years) |
| Sex | 0=male or 1=female |
| DaysDrink | Number of days drinking alcohol (in past 30 days) |
| Overdrawn | Has student overdrawn a checking account? 0=no or 1=yes |

Machine Learning algorithm

1. Logistic Regression
   →Applying the Logistic regression on the given data set since the outcome variable 'overdrawn' is 2 class variable, it has value 0 or 1.

   # Logistic regression for the OVerdrawn data set.

```
> Overdrawndt <- read.csv(file.choose(), header = TRUE, sep = ",")
> View(Overdrawndt)
>
> dim(Overdrawndt)
[1] 450   5
>
> #Removing the rows containing the N/A values
> Overdrawndt <- na.omit(Overdrawndt)
> dim(Overdrawndt)
[1] 437   5
>
> set.seed(1234)

>
> View(Overdrawndt)
> Overdrawndt$DyDrnkCat <- with(smplp, ifelse(Overdrawndt$DaysDrink < 7
, 0,
+                              ifelse(Overdrawndt$DaysDrink <= 14, 1,2))
)

> # Since the Overdrawn==1 class is rare class i.e. only 60 instances f
or Overdrawn==1 out of 450
> # Distributing the Overdrawn==1 class approximately equal to train an
d test data
>

> population <- sample(nrow(Overdrawndt), 0.75 * nrow(Overdrawndt))
> Overdrawn_train = Overdrawndt[population,]
> Overdrawn_test = Overdrawndt[-population,]
>
> summary(Overdrawndt$Overdrawn==1)
   Mode   FALSE    TRUE
logical     381      56
> summary(Overdrawn_train$Overdrawn==1)
   Mode   FALSE    TRUE
logical     287      40
> summary(Overdrawn_test$Overdrawn==1)
   Mode   FALSE    TRUE
```

```
logical      94      16
>
>
> mdl = glm(Overdrawn~Age+Sex+DyDrnkCat, family = binomial(link = "logi
t"),data = Overdrawn_train)
>
> summary(mdl)

Call:
glm(formula = Overdrawn ~ Age + Sex + DyDrnkCat, family = binomial(link
= "logit"),
    data = Overdrawn_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5172  -0.5385  -0.4291  -0.2814   2.6567

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.0114     2.4638  -3.658 0.000255 ***
Age           0.2901     0.1191   2.435 0.014894 *
Sex           1.1778     0.3974   2.964 0.003042 **
DyDrnkCat     0.8207     0.2145   3.827 0.000130 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 242.98  on 326  degrees of freedom
Residual deviance: 218.53  on 323  degrees of freedom
AIC: 226.53

Number of Fisher Scoring iterations: 5

>
> rs  = predict(mdl,newdata = Overdrawn_test,type="response")
>
> rs1 = ifelse(rs > 0.5,1,0)
>
> misClass = mean(rs1 != Overdrawn_test$Overdrawn)
>
> accry = 1 - misClass
> accry
[1] 0.8545455
```

2. kNN

→

We can kNN to classify the data for two class problems.

```
> names(Overdrawndt)
[1] "X"          "Age"        "Sex"        "DaysDrink" "Overdrawn" "DyDrnk
Cat"
>
> #Selecting only required columns
> Overdrawndt_new = Overdrawndt[c("Age","Sex","DaysDrink","Overdrawn")]
>
>
> mpldta = sample(2,nrow(Overdrawndt_new),replace = TRUE, prob = c(0.75
, 0.25))
```

```
>
>
> Overdrawndt_new.training = Overdrawndt_new[ind==1,1:3]
> Overdrawndt_new.test = Overdrawndt_new[ind==2,1:3]
>
> Overdrawndt_new.trainLabels = Overdrawndt_new[ind==1, 4]
> Overdrawndt_new.testLabels = Overdrawndt_new[ind ==2, 4]
>
>
> View(Overdrawndt_new.test )
> View(Overdrawndt_new.training )
>
> library(class)
>
> #kNN for  k=2
>
> Overdrawn_pred <- knn(train = Overdrawndt_new.training, test = Overdr
awndt_new.test, cl = Overdrawndt_new.trainLabels, k=2)
>
> library(gmodels)
> CrossTable(x=Overdrawn_pred, y=Overdrawndt_new.testLabels, prop.chisq
= FALSE)


   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  116


                | Overdrawndt_new.testLabels
Overdrawn_pred |         0 |         1 | Row Total |
---------------|-----------|-----------|-----------|
             0 |       100 |        10 |       110 |
               |     0.909 |     0.091 |     0.948 |
               |     0.943 |     1.000 |           |
               |     0.862 |     0.086 |           |
---------------|-----------|-----------|-----------|
             1 |         6 |         0 |         6 |
               |     1.000 |     0.000 |     0.052 |
               |     0.057 |     0.000 |           |
               |     0.052 |     0.000 |           |
---------------|-----------|-----------|-----------|
  Column Total |       106 |        10 |       116 |
               |     0.914 |     0.086 |           |
---------------|-----------|-----------|-----------|


>
> # Accuracy is [(101+0)/116 = 87%]
>
>
>
> #kNN for  k=3
>
> Overdrawn_pred <- knn(train = Overdrawndt_new.training, test = Overdr
awndt_new.test, cl = Overdrawndt_new.trainLabels, k=3)
>
```

```
> library(gmodels)
> CrossTable(x=Overdrawn_pred, y=Overdrawndt_new.testLabels, prop.chisq
= FALSE)


   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  116

                | Overdrawndt_new.testLabels
Overdrawn_pred |         0 |         1 | Row Total |
---------------|-----------|-----------|-----------|
             0 |       102 |        10 |       112 |
               |     0.911 |     0.089 |     0.966 |
               |     0.962 |     1.000 |           |
               |     0.879 |     0.086 |           |
---------------|-----------|-----------|-----------|
             1 |         4 |         0 |         4 |
               |     1.000 |     0.000 |     0.034 |
               |     0.038 |     0.000 |           |
               |     0.034 |     0.000 |           |
---------------|-----------|-----------|-----------|
  Column Total |       106 |        10 |       116 |
               |     0.914 |     0.086 |           |
---------------|-----------|-----------|-----------|


>
> # Accuracy is [(103+0)/116 = 88.79%]
>
>
> #kNN for  k=4
>
> Overdrawn_pred <- knn(train = Overdrawndt_new.training, test = Overdr
awndt_new.test, cl = Overdrawndt_new.trainLabels, k=4)
>
> library(gmodels)
```

```
> CrossTable(x=Overdrawn_pred, y=Overdrawndt_new.testLabels, prop.chisq
= FALSE)

   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  116

                | Overdrawndt_new.testLabels
Overdrawn_pred |         0 |         1 | Row Total |
---------------|-----------|-----------|-----------|
             0 |       105 |        10 |       115 |
               |     0.913 |     0.087 |     0.991 |
               |     0.991 |     1.000 |           |
               |     0.905 |     0.086 |           |
---------------|-----------|-----------|-----------|
             1 |         1 |         0 |         1 |
               |     1.000 |     0.000 |     0.009 |
               |     0.009 |     0.000 |           |
               |     0.009 |     0.000 |           |
---------------|-----------|-----------|-----------|
  Column Total |       106 |        10 |       116 |
               |     0.914 |     0.086 |           |
---------------|-----------|-----------|-----------|


>
> # Accuracy is [(104+0)/116 = 89.65%], this could be the case of model
overfitting
```