Data mining course HW3

A)
1a)
(25-32)/10.36= -0.676
(40-32)/10.36= 0.772
(45-32)/10.36= 1.255
(30-32)/10.36= -0.193
(20-32)/10.36= -1.158

1b)
(5000-6600)/2408.31= -0.664
(4000-6600)/2408.31= -1.080
(6000-6600)/2408.31= -0.249
(8000-6600)/2408.31= 0.581
(10000-6600)/2408.31= 1.412

1c)
Standardized age: (35-32)/10.36=0.290
Standardized loan: (10000-6600)/2408.31=1.412

Euclidean distance to 1st row: (((0.290+0.676)^2)+((1.412+0.664)^2))^0.5=2.290
Euclidean distance to 2nd row: (((0.290-0.772)^2)+((1.412+1.080)^2))^0.5=2.538
Euclidean distance to 3rd row: (((0.290-1.255)^2)+((1.412+0.249)^2))^0.5=1.921
Euclidean distance to 4th row: (((0.290+0.193)^2)+((1.412-0.581)^2))^0.5=0.961
Euclidean distance to 5th row: (((0.290+1.158)^2)+((1.412-1.412)^2))^0.5=1.448

1d) Based on k=3, last three rows will be chosen. Two N and 1 Y. Thus, the predicted default status will be N

```
library(caret)
```

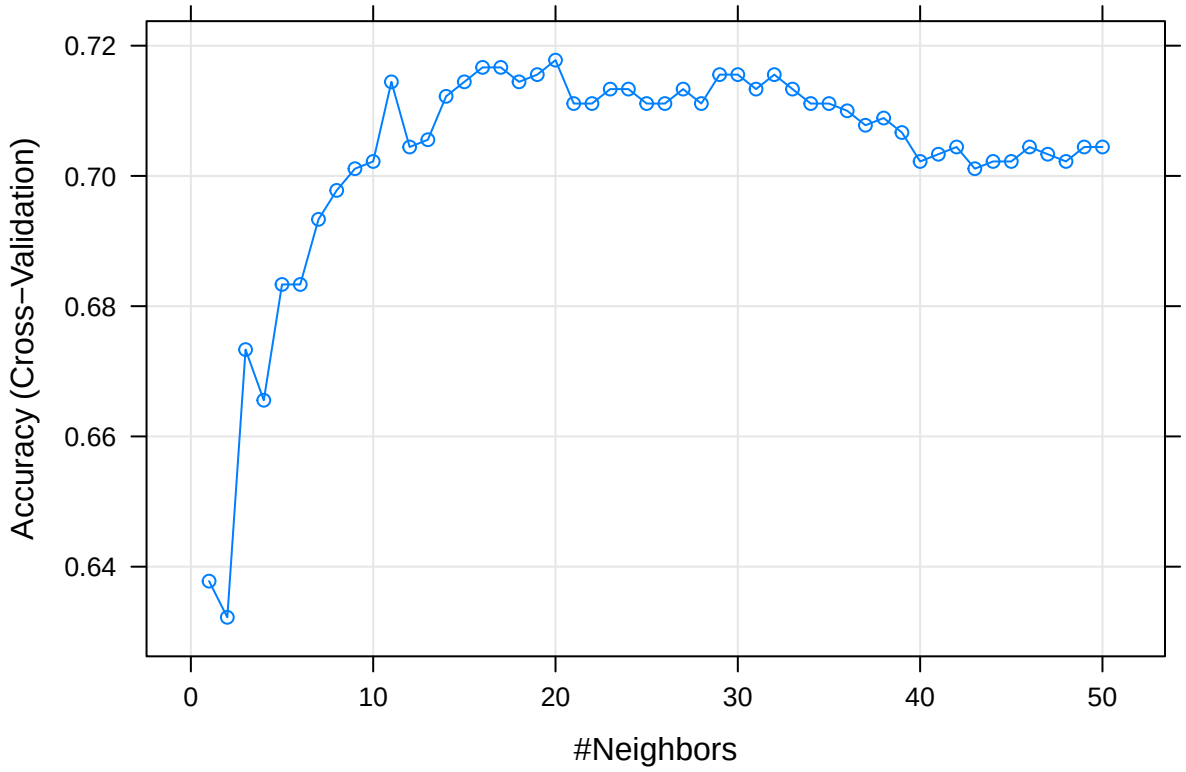
```
## Loading required package: lattice
```

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

```
credit_df <- read.csv("credit_data.csv")
credit_df$default = as.factor(credit_df$default)
```

```
set.seed(100)
knn_credit <- train(default ~ ., data=credit_df,
                    method = "knn",
                    preProcess = c("center", "scale"),
                    tuneGrid=expand.grid(k=1:50),
                    trControl = trainControl(method = "cv", number=10))
```

```
plot(knn_credit)
```




```
knn_credit
```

```
## k-Nearest Neighbors
##
## 900 samples
##  14 predictor
##   2 classes: 'no', 'yes'
##
## Pre-processing: centered (28), scaled (28)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 810, 810, 810, 810, 810, 810, ...
## Resampling results across tuning parameters:
##
```

```
##    k   Accuracy   Kappa
##    1  0.6377778  0.13300943
##    2  0.6322222  0.10956301
##    3  0.6733333  0.14703588
##    4  0.6655556  0.13273801
##    5  0.6833333  0.13849186
##    6  0.6833333  0.14368925
##    7  0.6933333  0.14206347
##    8  0.6977778  0.14108652
##    9  0.7011111  0.14357511
##   10  0.7022222  0.14553783
##   11  0.7144444  0.16210137
##   12  0.7044444  0.13477540
##   13  0.7055556  0.12695674
##   14  0.7122222  0.14043537
##   15  0.7144444  0.14190540
##   16  0.7166667  0.15505571
##   17  0.7166667  0.13683905
##   18  0.7144444  0.13046921
##   19  0.7155556  0.12818599
##   20  0.7177778  0.14526930
##   21  0.7111111  0.11263670
##   22  0.7111111  0.10771305
##   23  0.7133333  0.11204108
##   24  0.7133333  0.10973167
##   25  0.7111111  0.09497135
##   26  0.7111111  0.09965412
##   27  0.7133333  0.09714072
##   28  0.7111111  0.09002141
##   29  0.7155556  0.09881641
##   30  0.7155556  0.10125659
##   31  0.7133333  0.08196642
##   32  0.7155556  0.09439285
##   33  0.7133333  0.07916801
##   34  0.7111111  0.06949113
##   35  0.7111111  0.06949113
##   36  0.7100000  0.06201035
##   37  0.7077778  0.05226496
##   38  0.7088889  0.05704790
##   39  0.7066667  0.04730250
##   40  0.7022222  0.03324208
##   41  0.7033333  0.03814387
##   42  0.7044444  0.03474272
##   43  0.7011111  0.02521464
##   44  0.7022222  0.02454445
##   45  0.7022222  0.02450572
##   46  0.7044444  0.03180499
##   47  0.7033333  0.02688127
##   48  0.7022222  0.02173305
##   49  0.7044444  0.03173870
##   50  0.7044444  0.02900803
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 20.
```

A K value of 20 gives the highest accuracy

DMBA HW 3 Part C

1. The name of the dataset is hackathon_scouting.csv from the Big-Data-Cup-2021 on Github. The dataset consists of Stathletes-tracked junior hockey data from the Olympics and NCAA.
2. NCAA, any institutions relevant to competitive hockey, hockey players, any individual interested in hockey games would be able to gain insights from the dataset because it contains specific x, y coordinates of each shot that occurred in a game, which could offer insights on how to improve offensive hockey strategy.
3. An interesting supervised learning task would be to make a prediction of the best move the player could make based on the x, y coordinates of the player. The y variable of interest would be 'Event' because it is the variable that represents what type of action the player should make, such as shoot, goal, pass, incomplete play, takeaway, etc. Another interesting supervised learning task would be to make a prediction on which x, y coordinate the player should be located for optimal play. The y variable of interest could be a variable that combines the x, y coordinates because both variables are needed to determine the specific location.
4. The current dataset has information concentrated on the details of the game, such as the home team name, away team name, player name, date of the game. Although the information is very much relevant, additional data on the specifics of each shot could make the current dataset more powerful. We already have the type of shot, and the x, and y coordinates. It could be helpful to also know the speed, or the angle at which the shot occurred. The auxiliary data could be obtained by researching if such record exists, but I would assume it is impossible.