Similarity - Part 2

Exploring Data

Team: ASA190005, DLK190005,SCK190106, LDN190002

```
data <- read.csv(file="C:/Users/Abed/Desktop/employeeData.csv",header=TRUE,)
data <- na.omit(data)
data$Attrition <- as.factor(data$Attrition)



mean(data$Age)
```

```
## [1] 36.93336
```

```
median(data$Age)
```

```
## [1] 36
```

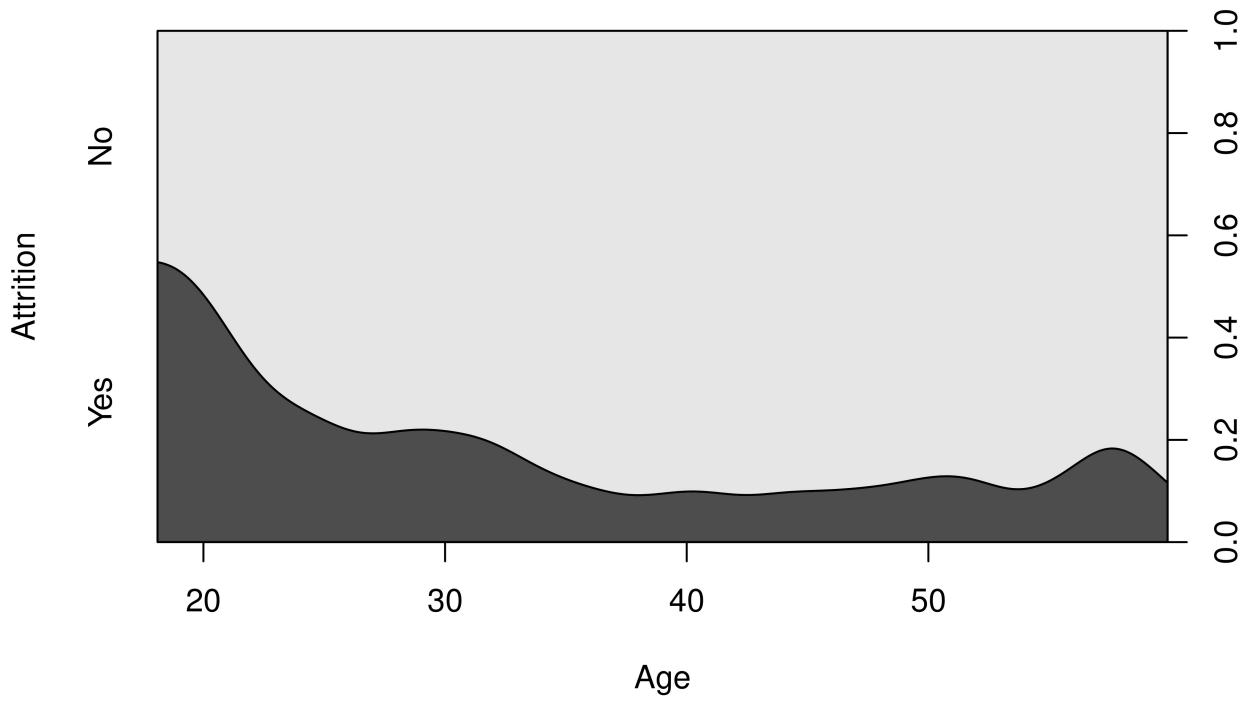
```
mean(data$DistanceFromHome)
```

```
## [1] 9.198996
```

```
median(data$DistanceFromHome)
```

```
## [1] 7
```

```
cdplot(Attrition ~ Age, data=data, xlab="Age",ylab="Attrition")
```



Logistic regression

```
data$Attrition <- as.numeric(as.factor(data$Attrition))
for(i in 1:length(data$Attrition)){
  if(data$Attrition[i] == 1) {
    data$Attrition[i] = 0
  }
  else {
    data$Attrition[i] = 1
  }
}
set.seed(10)
library(caTools)
split <- sample.split(data,SplitRatio=0.8)
train <- subset(data,split==TRUE)
test <- subset(data,split==FALSE)


lg <- glm(Attrition ~ Age,data = train, family = 'binomial')
summary(lg)
```

```
##
## Call:
## glm(formula = Attrition ~ Age, family = "binomial", data = train)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8944  -0.6382  -0.5337  -0.3994   2.4434
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.290988   0.200422   1.452    0.147
## Age         -0.055590   0.005736  -9.692   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3024.0  on 3469  degrees of freedom
## Residual deviance: 2920.1  on 3468  degrees of freedom
## AIC: 2924.1
##
## Number of Fisher Scoring iterations: 5
```

```
lrProb <- predict(lg,newdata=test,type="response")
pred1 <- ifelse(lrProb>0.5,1,0)
acc <- mean(pred1==test$Attrition)
print(paste("Accuracy of the logistic regression model is: ", acc))
```

```
## [1] "Accuracy of the logistic regression model is:  0.826754385964912"
```

KNN

```
library(caret)
```

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

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

```
data2 <- read.csv(file="C:/Users/Abed/Desktop/employeeData.csv",header=TRUE)
data2 <- data[c(1:2)]


data2$Attrition <- as.numeric(as.factor(data2$Attrition))
for(i in 1:length(data2$Attrition)){
  if(data2$Attrition[i] == 1) data2$Attrition[i] = 0
  else data2$Attrition[i] = 1
}


set.seed(2)
library(caTools)
library(class)
data2 <- na.omit(data2)
```

```
split2 <- sample.split(data2,SplitRatio=0.8)
train2 <- subset(data2,split2==TRUE)
test2 <- subset(data2,split2==FALSE)
trainX = train2[1:1] # Taking the Age paramater to predict Attrition
testX  = test2[1:1]



classifier_knn <- knn(train = trainX,
                      test = testX,
                      cl = train2$Attrition,
                      k = 1)


result <- classifier_knn == test2$Attrition
acc <- length(which(result == TRUE)) / length(result) * 100


print(paste("Accuracy of the knn model is: ", acc))
```

```
## [1] "Accuracy of the knn model is:  84.9840255591054"
```

Decision Trees

```
data <- read.csv(file="C:/Users/Abed/Desktop/employeeData.csv",header=TRUE)
data <- na.omit(data)
set.seed(2)


library(tree)
treeModel <- tree(Attrition ~ Age, data = train, method = "class")
pred <- predict(treeModel,newdata=test)

table <- table(test$Attrition,pred)
acc = sum(diag(table)) / sum(table)

print(paste("Accuracy of the Decision tree model is: ", acc))
```

```
## [1] "Accuracy of the Decision tree model is:  0.625"
```

Analysis

Logistic regression and KNN performed about equal. Decision Trees performed a bit worse.

For Logistic regression, it performed quite well, mainly because there is quite a good linear trend with age and attrition, as people get older, they tend to quit less and stick it out, so the model was able to glean that pretty well during training.

As for KNN, I believe it performed well here because of the lower dimensionality of the data. Also because, people of a certain age are more likely to do the same thing, so for the k nearest neighbors, the nearest neighbors are likelty to represent the likliehood of what that age demographic would do.

For Decision Tree, they did not perform very well. I would say this is because the size of the data resulted in an over fitting, since trees can easily get very complex if the size o data is too big.