---
title: "Classifications"
output: html_notebook

Sagar Darnal
Manuel Romero

data taken from:
https://vincentarelbundock.github.io/Rdatasets/datasets.html

SmokeBan.csv

//getting data and saving it into df
```{r}
df <- read.csv("SmokeBan.csv", na.strings = "NA", header =TRUE)
str(df)
```

```{r}
df$smoker <- factor(df$smoker)
df$ban <- factor(df$ban)
df$education <- factor(df$education)
df$afam <- factor(df$afam)
df$hispanic <- factor(df$hispanic)
df$gender <- factor(df$gender)

head(df)
```

seeing if any rows have na
```{r}
sapply(df, function(x) sum(is.na(x)==TRUE))
```
no empty values
removing first column since it is unecessary
```{r}
df <- df <- df[,c(2,3,4,5,6,7,8)]
sapply(df, function(x) sum(is.na(x)==TRUE))

```

divide into train test data sets

```{r}
set.seed(1234)
i <- sample(1:nrow(df), 0.8*nrow(df), replace=FALSE)
train <- df[i,]
test <- df[-i,]
```
5 r functions
number of rows in data set
number of columns in data set
names of columns in data set
summary of data set
list structure of set
```{r}
print('number of rows in data set')
nrow(df)
print('number of columns in data set')
ncol(df)
print('names of columns in data set')
names(df)
print('summary of data set')
summary(df)
```

```
print('list structure of data set')
str(df)
```

```{r}
par(mfrow=c(1,2))
cdplot(train$smoker~train$age)
cdplot(train$smoker~train$ban)
```
Plots show that once people get much older they dont smoke, this might be due to health risks or possibly that smokers done typically make it past 80, this could be why there is such a steep drop off or there just not be too many data points of 80

creating a logistic regression model
```{r}
glm1 <- glm(smoker~., data=train, family="binomial")
summary(glm1)

```

Here we see that our Residual deviance is lower than our null variance, which is what we want since that means there is less lack of fit when using the entire model vs just using the intercept(null deviance).

NAIVE BAYES
using the same data, hence df is already cleaned and split
getting library e1071 and creating our model
if you havent downloaded package run in console :
install.packages('e1071', dependencies=TRUE)

```{r}
library(e1071)
nb1 <- naiveBayes(smoker~., data=train)
nb1
```

We see that the probability of not smoking is .758 and smoking is .242, which adds up to 1.
We then see that the if there is no ban on smoking the probability of smoking is higher than if there is a ban, from .6342348 down to .5253099 when there is a smoking ban. for age we see that there isn't much difference with only about a year of difference, with non smokers being one year older and there standard deviation being wider by also about a year.Education doesn't give us much information either but it does tell us that people with hs education make up a big part of smokers with some college being in second. However is it hard to say that education plays a part in smoking or not smoking, meaning that higher education doesn't mean more non smokers. Race and sex dont play into it much this is probabily due to the low number of data, there is a lot of non hispanics and non afam, if you are hispanic non smokers are at about .1152 and smokers goes to about .1043 and for afam it is very small difference about .002. And gender we see that females are less likely too smoke since it goes from .5707454 non smokers to .5268595 smokers for women and for men it goes up, non smokers are at .4292546 to smoking .4731405.

Running test data for glm1
```{r}
probs <- predict(glm1, newdata=test, type="response")
pred <- ifelse(probs>0.5, 1, 0)
acc <- mean(pred==test$smoker)
print(paste("accuracy = ", acc))
table(pred, test$smoker)
```

```{r}
library(caret)
confusionMatrix(as.factor(pred), reference=test$survived)
```

running test data for nb1
```{r}
p1 <- predict(nb1, newdata=test, type="class")
table(p1, test$smoker)
mean(p1==test$smoker)
```
we see that we got a higher non smoker for Naive Bayes than we did for Logistic regression and we actually got a mean value of 0,7565 versus the logistic regression of 0. Im not sure why this happened, possibly the data used is not an suitable for classifying as we saw in the bayes model that not all the factors are actually necessarily good factors. Even if we run the GLM with just age and ban we dont see any change other than AIC going higher, which is not what we want. So for this data set Naive Bayes is a better model even if just by a little.

```{r}
p2_raw <- predict(nb1, newdata=test, type="raw")
head(p2_raw, n=2)
```

```{r}
glm2 <- glm(smoker~age+ban, data=train, family="binomial")
summary(glm2)

probs2 <- predict(glm2, newdata=test, type="response")
pred2 <- ifelse(probs>0.5, 1, 0)
acc2 <- mean(pred2==test$smoker)
print(paste("accuracy = ", acc2))
table(pred2, test$smoker)
```

Naive Bayes does better with smaller data compared to Logistic Regression. Naive bayes is a generative classifier and logistic regression is a discriminitive classifier. Naive bayes has a higher bias but a lower variance than logistic regression.