Project 2 - Classification

Initial Setup

```
df <- read.csv("adult.data.csv", header=TRUE, stringsAsFactors = FALSE)
sapply(df, function(x) sum(is.na(x)==TRUE))
```

```
##            age      workClass      education educationYears  maritalStatus
##              0              0              0              0              0
##     Occupation   relationship           race         gender           gain
##              0              0              0              0              0
##           loss          hours        country         income
##              0              0              0              0
```

```
str(df)
```

```
## 'data.frame':    32561 obs. of  14 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workClass     : chr  " State-gov" " Self-emp-not-inc" " Private" " Private" ...
##  $ education     : chr  " Bachelors" " Bachelors" " HS-grad" " 11th" ...
##  $ educationYears: int  13 13 9 7 13 14 5 9 14 13 ...
##  $ maritalStatus : chr  " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...
##  $ Occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners"
##  $ relationship  : chr  " Not-in-family" " Husband" " Not-in-family" " Husband" ...
##  $ race          : chr  " White" " White" " White" " Black" ...
##  $ gender        : chr  " Male" " Male" " Male" " Male" ...
##  $ gain          : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ loss          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours         : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ country       : chr  " United-States" " United-States" " United-States" " United-States" ...
##  $ income        : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
```

Data Cleaning

https://archive.ics.uci.edu/ml/datasets/Adult

The "date" column was removed from the CSV file prior to R code execution. This was done as the data was not useful in classifying the desired data.

This chunk of code will convert particular columns as factors, and will display the new structure of the data frame.

```
df$education <- as.factor(df$education)
df$maritalStatus <- as.factor(df$maritalStatus)
df$relationship <- as.factor(df$relationship)
```

```r
df$race <- as.factor(df$race)
df$gender <- as.factor(df$gender)
df$income <- as.factor(df$income)

df <- df[!is.na(df$workClass),]
df <- df[!is.na(df$education),]
df <- df[!is.na(df$Occupation),]
df <- df[!is.na(df$maritalStatus),]
df <- df[!is.na(df$relationship),]
df <- df[!is.na(df$race),]
df <- df[!is.na(df$gender),]
df <- df[!is.na(df$country),]
df <- df[!is.na(df$income),]

set.seed(1234)
i <- sample(1:nrow(df), nrow(df)*0.75, replace=FALSE)
train <- df[i,]
test <- df[-i,]

str(df)
```

```
## 'data.frame':    32561 obs. of  14 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workClass     : chr  " State-gov" " Self-emp-not-inc" " Private" " Private" ...
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ educationYears: int  13 13 9 7 13 14 5 9 14 13 ...
##  $ maritalStatus : Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ Occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners"
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ gender        : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ gain          : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ loss          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours         : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ country       : chr  " United-States" " United-States" " United-States" " United-States" ...
##  $ income        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

Data Exploration

```r
head(df)
```

```
##   age         workClass  education educationYears       maritalStatus
## 1  39         State-gov  Bachelors             13       Never-married
## 2  50  Self-emp-not-inc  Bachelors             13  Married-civ-spouse
## 3  38           Private    HS-grad              9            Divorced
## 4  53           Private       11th              7  Married-civ-spouse
## 5  28           Private  Bachelors             13  Married-civ-spouse
## 6  37           Private    Masters             14  Married-civ-spouse
##           Occupation   relationship   race  gender gain loss hours
## 1       Adm-clerical  Not-in-family  White    Male 2174    0    40
## 2    Exec-managerial        Husband  White    Male    0    0    13
```

```
## 3  Handlers-cleaners  Not-in-family  White    Male    0    0    40
## 4  Handlers-cleaners        Husband  Black    Male    0    0    40
## 5     Prof-specialty           Wife  Black  Female    0    0    40
## 6    Exec-managerial           Wife  White  Female    0    0    40
##          country income
## 1  United-States  <=50K
## 2  United-States  <=50K
## 3  United-States  <=50K
## 4  United-States  <=50K
## 5           Cuba  <=50K
## 6  United-States  <=50K
```

```
tail(df, n=10)
```

```
##       age     workClass     education educationYears       maritalStatus
## 32552  32       Private          10th              6  Married-civ-spouse
## 32553  43       Private     Assoc-voc             11  Married-civ-spouse
## 32554  32       Private       Masters             14       Never-married
## 32555  53       Private       Masters             14  Married-civ-spouse
## 32556  22       Private  Some-college             10       Never-married
## 32557  27       Private    Assoc-acdm             12  Married-civ-spouse
## 32558  40       Private       HS-grad              9  Married-civ-spouse
## 32559  58       Private       HS-grad              9             Widowed
## 32560  22       Private       HS-grad              9       Never-married
## 32561  52  Self-emp-inc       HS-grad              9  Married-civ-spouse
##              Occupation   relationship                race  gender  gain loss
## 32552  Handlers-cleaners        Husband  Amer-Indian-Eskimo    Male     0    0
## 32553              Sales        Husband               White    Male     0    0
## 32554       Tech-support  Not-in-family  Asian-Pac-Islander    Male     0    0
## 32555    Exec-managerial        Husband               White    Male     0    0
## 32556    Protective-serv  Not-in-family               White    Male     0    0
## 32557       Tech-support           Wife               White  Female     0    0
## 32558  Machine-op-inspct        Husband               White    Male     0    0
## 32559       Adm-clerical      Unmarried               White  Female     0    0
## 32560       Adm-clerical      Own-child               White    Male     0    0
## 32561    Exec-managerial           Wife               White  Female 15024    0
##       hours        country income
## 32552    40  United-States  <=50K
## 32553    45  United-States  <=50K
## 32554    11         Taiwan  <=50K
## 32555    40  United-States   >50K
## 32556    40  United-States  <=50K
## 32557    38  United-States  <=50K
## 32558    40  United-States   >50K
## 32559    40  United-States  <=50K
## 32560    20  United-States  <=50K
## 32561    40  United-States   >50K
```

```
summary(df)
```

```
##       age          workClass                 education     educationYears
##  Min.   :17.00   Length:32561        HS-grad     :10501   Min.   : 1.00
##  1st Qu.:28.00   Class :character    Some-college: 7291   1st Qu.: 9.00
```

```
##  Median :37.00   Mode  :character    Bachelors   : 5355   Median :10.00
##  Mean   :38.58                        Masters     : 1723   Mean   :10.08
##  3rd Qu.:48.00                        Assoc-voc   : 1382   3rd Qu.:12.00
##  Max.   :90.00                        11th        : 1175   Max.   :16.00
##                                      (Other)      : 5134
##                maritalStatus    Occupation              relationship
##   Divorced             : 4443   Length:32561       Husband       :13193
##   Married-AF-spouse    :   23   Class :character   Not-in-family : 8305
##   Married-civ-spouse   :14976   Mode  :character   Other-relative:  981
##   Married-spouse-absent:  418                      Own-child     : 5068
##   Never-married        :10683                      Unmarried     : 3446
##   Separated            : 1025                      Wife          : 1568
##   Widowed              :  993
##                   race          gender           gain            loss
##   Amer-Indian-Eskimo:  311    Female:10771   Min.   :    0   Min.   :   0.0
##   Asian-Pac-Islander: 1039    Male  :21790   1st Qu.:    0   1st Qu.:   0.0
##   Black             : 3124                   Median :    0   Median :   0.0
##   Other             :  271                   Mean   : 1078   Mean   :  87.3
##   White             :27816                   3rd Qu.:    0   3rd Qu.:   0.0
##                                              Max.   :99999   Max.   :4356.0
##
##      hours          country              income
##  Min.   : 1.00   Length:32561        <=50K:24720
##  1st Qu.:40.00   Class :character    >50K : 7841
##  Median :40.00   Mode  :character
##  Mean   :40.44
##  3rd Qu.:45.00
##  Max.   :99.00
##
```

```r
summary(df$relationship)
```

```
##        Husband  Not-in-family Other-relative      Own-child      Unmarried
##          13193           8305            981           5068           3446
##           Wife
##           1568
```

```r
contrasts(df$race)
```

```
##                     Asian-Pac-Islander Black Other White
##  Amer-Indian-Eskimo                  0     0     0     0
##  Asian-Pac-Islander                  1     0     0     0
##  Black                               0     1     0     0
##  Other                               0     0     1     0
##  White                               0     0     0     1
```

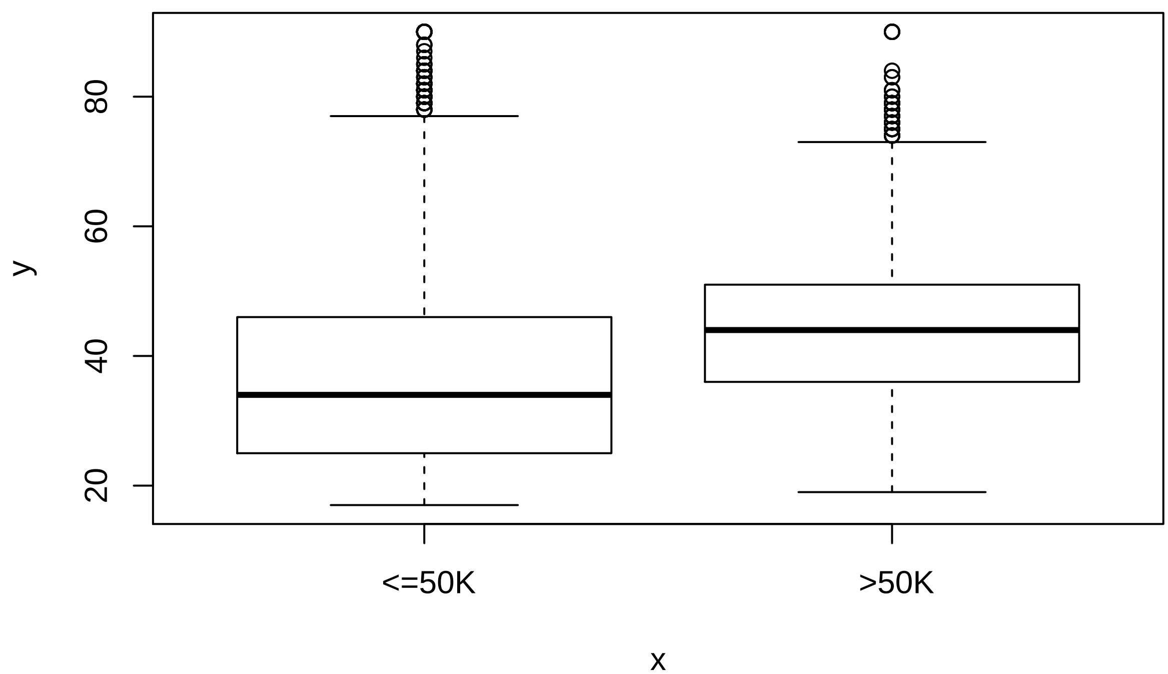
```r
contrasts(df$gender)
```

```
##         Male
##  Female    0
##  Male      1
```

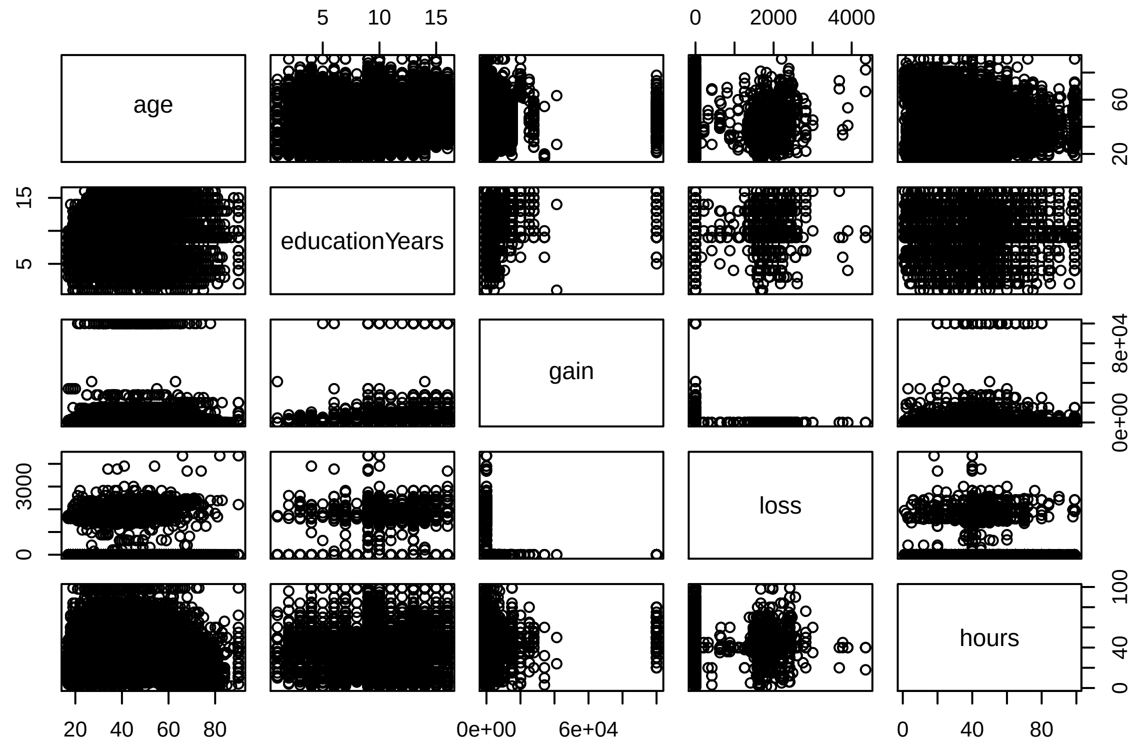
```r
cor(df$hours, df$gain)
```

```
## [1] 0.07840862
```

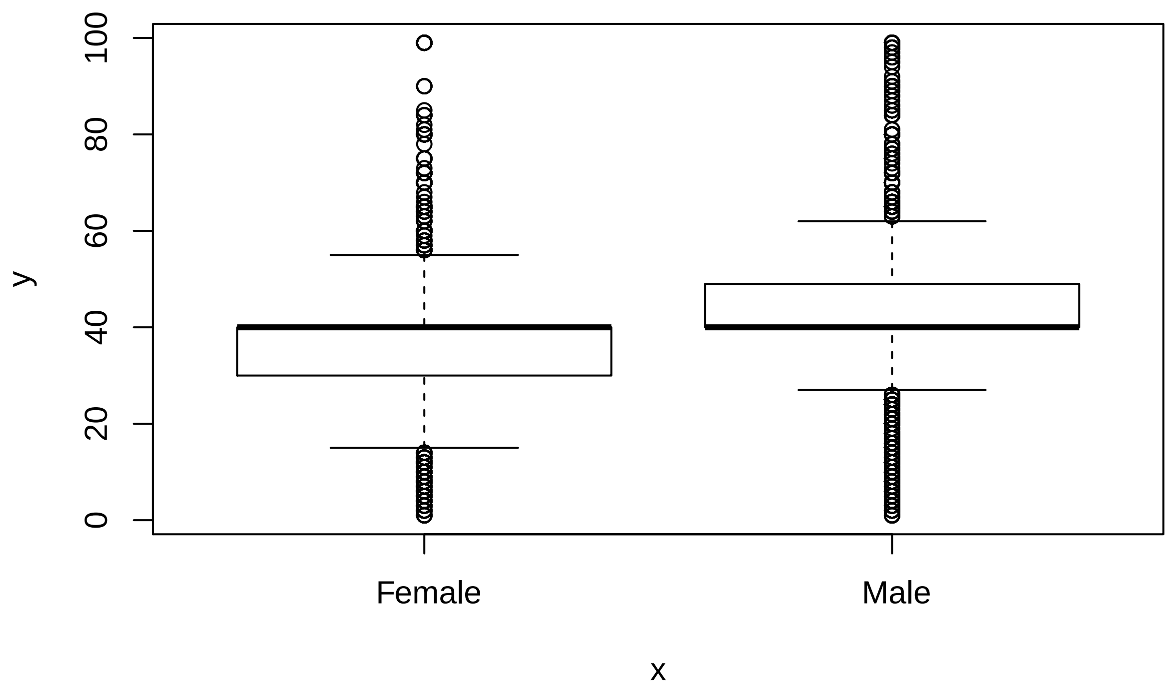
```r
plot(df$income, df$age)
```




```r
pairs(df[, c(1, 4, 10, 11, 12)])
```




```r
plot(df$gender, df$hours)
```



Naive Bayes Classification

Naive Bayes Classification was performed on this data frame because of it's popularity when it comes to classification algorithms. This algorithm in particular allows us to quantify how likely it is that an individual makes over $50k a year based on a number of variables.

Upon execution of the Naive Bayes algorithm, we are able to determine that it is much less likely for an individual to make >50k a year from within the dataset, rather than the latter. We are able to also view the correlation of data which remains at around 0.827. This is not the strongest reading, so we removed a variable to see if it may be responsible for the outcome. After the removal of educationYears from the data, we can see that the value decreases to 0.814. This shows that educationYears has a correlation in the odds of an individual making >50k a year.

```
library(e1071)
nb1 <- naiveBayes(income~., data=train)
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
```

```
## Y
##    <=50K     >50K
## 0.760647 0.239353
##
## Conditional probabilities:
##        age
## Y           [,1]     [,2]
##    <=50K 36.80861 14.06201
##    >50K  44.22618 10.43956
##
##        workClass
## Y                  ?  Federal-gov    Local-gov  Never-worked      Private
##    <=50K 0.0676177658 0.0231493943 0.0593808883  0.0003230148 0.7172005384
##    >50K  0.0244653550 0.0484174508 0.0802395210  0.0000000000 0.6337040205
##        workClass
## Y         Self-emp-inc  Self-emp-not-inc    State-gov  Without-pay
##    <=50K  0.0199730821      0.0743472409 0.0375235532 0.0004845222
##    >50K   0.0807527802      0.0881094953 0.0443113772 0.0000000000
##
##        education
## Y                10th         11th         12th      1st-4th      5th-6th
##    <=50K 0.0361238223 0.0448452221 0.0159353970 0.0062987887 0.0129744280
##    >50K  0.0073567151 0.0076988879 0.0046193328 0.0005132592 0.0017108640
##        education
## Y             7th-8th          9th   Assoc-acdm    Assoc-voc    Bachelors
##    <=50K 0.0260026918 0.0188425303 0.0331090175 0.0409152086 0.1288829071
##    >50K  0.0039349872 0.0035928144 0.0360992301 0.0444824636 0.2870829769
##        education
## Y           Doctorate      HS-grad      Masters    Preschool  Prof-school
##    <=50K 0.0043068641 0.3543472409 0.0316016151 0.0022072678 0.0066218035
##    >50K  0.0378100941 0.2112917023 0.1199315654 0.0000000000 0.0533789564
##        education
## Y         Some-college
##    <=50K  0.2369851952
##    >50K   0.1804961506
##
##        educationYears
## Y             [,1]     [,2]
##    <=50K  9.602584 2.454872
##    >50K  11.620701 2.350823
##
##        maritalStatus
## Y            Divorced  Married-AF-spouse  Married-civ-spouse
##    <=50K 0.1590847914       0.0005921938        0.3361507402
##    >50K  0.0612489307       0.0013686912        0.8544054748
##        maritalStatus
## Y         Married-spouse-absent  Never-married    Separated      Widowed
##    <=50K           0.0159892328   0.4140511440 0.0372543742 0.0368775236
##    >50K            0.0044482464   0.0602224123 0.0082121471 0.0100940975
##
##        Occupation
## Y                  ?  Adm-clerical  Armed-Forces  Craft-repair
##    <=50K 0.0679407806  0.1315208614  0.0003768506  0.1290444145
##    >50K  0.0244653550  0.0663815227  0.0001710864  0.1209580838
```

```
##        Occupation
## Y         Exec-managerial  Farming-fishing  Handlers-cleaners
##   <=50K     0.0841453567     0.0351547779       0.0511978466
##   >50K      0.2502994012     0.0124893071       0.0097519247
##        Occupation
## Y         Machine-op-inspct  Other-service  Priv-house-serv  Prof-specialty
##   <=50K        0.0708479139   0.1279676985     0.0057065949    0.0929205922
##   >50K         0.0335329341   0.0176218991     0.0001710864    0.2338751069
##        Occupation
## Y         Protective-serv        Sales  Tech-support  Transport-moving
##   <=50K     0.0176043069 0.1085868102  0.0254104980      0.0515746972
##   >50K      0.0277159966 0.1252352438  0.0355859709      0.0417450813
##
##        relationship
## Y           Husband  Not-in-family  Other-relative   Own-child   Unmarried
##   <=50K 0.295235532    0.303687752     0.037631225 0.200323015 0.129421265
##   >50K  0.754833191    0.109837468     0.003763901 0.007527802 0.027373824
##        relationship
## Y              Wife
##   <=50K 0.033701211
##   >50K  0.096663815
##
##        race
## Y         Amer-Indian-Eskimo  Asian-Pac-Islander       Black       Other
##   <=50K         0.011467026         0.030524899 0.110309556 0.009851952
##   >50K          0.004106074         0.035072712 0.052181352 0.003592814
##        race
## Y             White
##   <=50K 0.837846568
##   >50K  0.905047049
##
##        gender
## Y          Female      Male
##   <=50K 0.3849798 0.6150202
##   >50K  0.1527802 0.8472198
##
##        gain
## Y            [,1]       [,2]
##   <=50K  146.5799   918.0083
##   >50K  3946.2405 14323.2036
##
##        loss
## Y            [,1]     [,2]
##   <=50K  52.50627 310.0893
##   >50K  194.94645 594.3360
##
##        hours
## Y           [,1]     [,2]
##   <=50K 38.84011 12.29228
##   >50K  45.44756 10.95034
##
##        country
## Y                  ?     Cambodia       Canada        China     Columbia
##   <=50K 0.0181965007 0.0004845222 0.0029609690 0.0017765814 0.0019380888
```

```
##    >50K  0.0184773311 0.0008554320 0.0054747648 0.0027373824 0.0000000000
##         country
## Y                   Cuba  Dominican-Republic      Ecuador  El-Salvador
##    <=50K 0.0029609690         0.0027456258 0.0010228802 0.0034454913
##    >50K  0.0023952096         0.0003421728 0.0005132592 0.0013686912
##         country
## Y                England       France      Germany       Greece    Guatemala
##    <=50K 0.0024764468 0.0006460296 0.0039838493 0.0006998654 0.0027994616
##    >50K  0.0041060736 0.0011976048 0.0059880240 0.0010265184 0.0001710864
##         country
## Y                  Haiti  Holand-Netherlands     Honduras         Hong
##    <=50K 0.0016150740         0.0000538358 0.0004306864 0.0006460296
##    >50K  0.0006843456         0.0000000000 0.0001710864 0.0005132592
##         country
## Y                Hungary        India         Iran      Ireland        Italy
##    <=50K 0.0003768506 0.0023149394 0.0010228802 0.0009152086 0.0020457604
##    >50K  0.0003421728 0.0046193328 0.0022241232 0.0005132592 0.0034217280
##         country
## Y                Jamaica        Japan         Laos       Mexico    Nicaragua
##    <=50K 0.0026917900 0.0013997308 0.0006460296 0.0244952894 0.0011843876
##    >50K  0.0011976048 0.0023952096 0.0001710864 0.0039349872 0.0003421728
##         country
## Y          Outlying-US(Guam-USVI-etc)         Peru  Philippines       Poland
##    <=50K                 0.0006460296 0.0012382234 0.0058142665 0.0018842530
##    >50K                  0.0000000000 0.0003421728 0.0088964927 0.0010265184
##         country
## Y               Portugal  Puerto-Rico     Scotland        South       Taiwan
##    <=50K 0.0012920592 0.0036069987 0.0003768506 0.0026379542 0.0008613728
##    >50K  0.0006843456 0.0013686912 0.0005132592 0.0023952096 0.0020530368
##         country
## Y               Thailand  Trinadad&Tobago  United-States      Vietnam   Yugoslavia
##    <=50K 0.0007537012     0.0008613728   0.8910901750 0.0026379542 0.0003230148
##    >50K  0.0001710864     0.0003421728   0.9158254919 0.0005132592 0.0006843456
```

```r
p1 <- predict(nb1, newdata=test, type="class")
table(p1, test$income)
```

```
##
## p1        <=50K  >50K
##    <=50K   5755  1018
##    >50K     390   978
```

```r
print(paste("Mean =", mean(p1==test$income)))
```

```
## [1] "Mean = 0.827048274167793"
```

```r
p1_raw <- predict(nb1, newdata=test, type="raw")
head(p1_raw)
```

```
##             <=50K         >50K
## [1,] 7.136065e-01 0.2863935268
## [2,] 9.993981e-01 0.0006018519
```

```
## [3,] 9.429843e-01 0.0570156771
## [4,] 7.829707e-01 0.2170293057
## [5,] 1.808116e-48 1.0000000000
## [6,] 8.621895e-01 0.1378104643
```

```r
nb2 <- naiveBayes(income~.-educationYears, data=train)
p2 <- predict(nb2, newdata=test[,-4], type="class")
table(p2, test$income)
```

```
##
## p2        <=50K  >50K
##    <=50K   5788  1157
##    >50K     357   839
```

```r
print(paste("Mean =", mean(p2==test$income)))
```

```
## [1] "Mean = 0.814027760717357"
```

Decision Tree

Decision Tree Regression was performed on this data frame because it allows for the splitting of observations into proper partitions, allowing for the data to be placed into particular boundaries of which they belong. Although this algorithm is not the best when it comes to performance, it presents the data in a very strong and comprehensible manner, which is extremely beneficial and obtaining a visual understanding of the data frame.

Upon complete execution of the Decision Tree Regression algorithm, we can see that there is a moderate correlation between the predicted and actual data collected with a very low RMSE value. We can also see that the income of an individual has a strong relationship with their relationship status and education level. Overall, this algorithm proved to be the strongest.

```r
library(tree)

tree_df2 <- tree(income~., data=df)
```

```
## Warning in tree(income ~ ., data = df): NAs introduced by coercion
```

```r
tree_df2
```

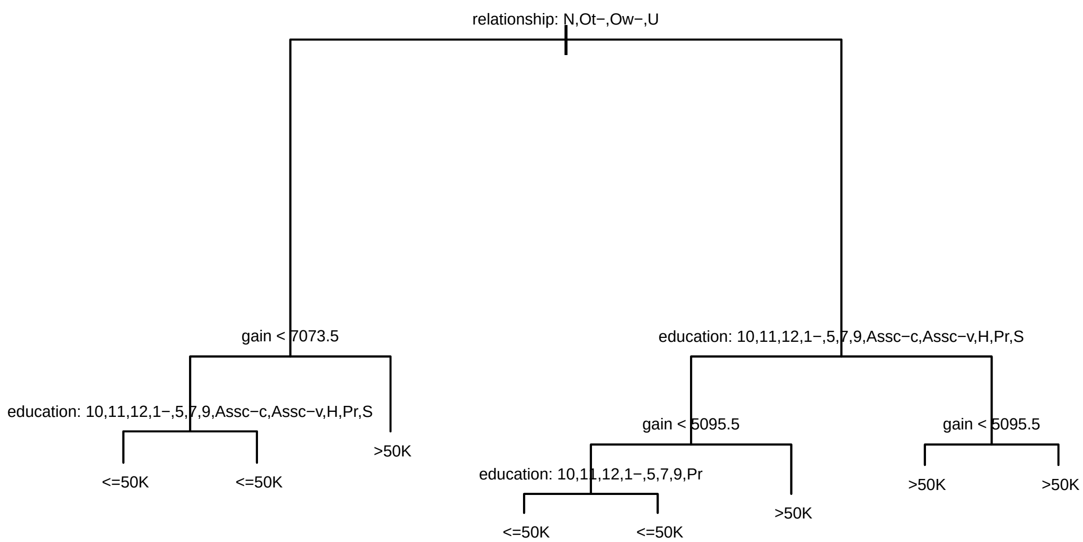
```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 32561 35950.00  <=50K ( 0.759190 0.240810 )
##    2) relationship:  Not-in-family, Other-relative, Own-child, Unmarried 17800  8674.00  <=50K ( 0.93
##      4) gain < 7073.5 17482  6929.00  <=50K ( 0.950120 0.049880 )
##        8) education:  10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, HS-gra
##        9) education:  Bachelors, Doctorate, Masters, Prof-school 3446  2900.00  <=50K ( 0.851132 0.14
##      5) gain > 7073.5 318   102.20  >50K ( 0.037736 0.962264 ) *
##    3) relationship:  Husband, Wife 14761 20320.00  <=50K ( 0.548608 0.451392 )
##      6) education:  10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, HS-grad
##       12) gain < 5095.5 9807 11980.00  <=50K ( 0.699806 0.300194 )
```

```
##          24) education:  10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Preschool 1597  1057.00  <=
##          25) education:  Assoc-acdm, Assoc-voc, HS-grad, Some-college 8210 10510.00  <=50K ( 0.661389
##        13) gain > 5095.5 522    98.91  >50K ( 0.019157 0.980843 ) *
##       7) education:  Bachelors, Doctorate, Masters, Prof-school 4432  5226.00  >50K ( 0.276399 0.72360
##        14) gain < 5095.5 3754  4737.00  >50K ( 0.325519 0.674481 ) *
##        15) gain > 5095.5 678    38.51  >50K ( 0.004425 0.995575 ) *
```

```r
summary(tree_df2)
```

```
##
## Classification tree:
## tree(formula = income ~ ., data = df)
## Variables actually used in tree construction:
## [1] "relationship" "gain"         "education"
## Number of terminal nodes:  8
## Residual mean deviance:  0.6999 = 22790 / 32550
## Misclassification error rate: 0.1555 = 5063 / 32561
```

```r
plot(tree_df2)
text(tree_df2, cex=0.5, pretty=1)
```




```r
set.seed(1234)
i <- sample(1:nrow(df), nrow(df)*0.75, replace=FALSE)
train_tree <- df[i,]
```

```r
test_tree <- df[-i,]

tree_df3 <- tree(income~., data=train_tree)
```

```
## Warning in tree(income ~ ., data = train_tree): NAs introduced by coercion
```

```r
pred_tree <- predict(tree_df3, newdata=test_tree, type="class")
```

```
## Warning in pred1.tree(object, tree.matrix(newdata)): NAs introduced by coercion
```

```r
table(pred_tree, test_tree$income)
```

```
##
## pred_tree  <=50K  >50K
##     <=50K   5880   972
##     >50K     265  1024
```

```r
mean(pred_tree == test_tree$income)
```

```
## [1] 0.8480531
```

SVM Classification

SVM classification was performed on this data frame because it can be used in a number of scenarios and is very versatile in implementation, and is especially useful in multi-class classification.

Upon completion of execution, we can see that the correlation value is very high and the RMSE is the lowest value when compared to the other classification algorithms used, making this the most useful, although it does have a slower execution time than the other algorithms.

```r
library(e1071)
svm1 <- svm(income ~ age + educationYears + race + gender + hours + gain + loss, data=train, kernel="li
pred_svm <- predict(svm1, newdata=test)

table(pred_svm, test$income)
```

```
##
## pred_svm  <=50K  >50K
##    <=50K   5993  1420
##    >50K     152   576
```

```r
print(paste("Mean =", mean(pred_svm==test$income)))
```

```
## [1] "Mean = 0.806903328829382"
```

Random Forest Ensemble Method

The random forest ensemble method was performed on this data frame because it works by aggregating the predictions made by multiple decision trees. This allows for us to develop a strong learning algorithm with fairly accurate results, although this algorithm has a slower runtime than most.

Upon the execution of Random Forest, we can see that there is a very low correlation between the predicted and actual data values. Although there is a low RMSE value, that does not make up for the lack of correlation.

```
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.
```

```
rf <- randomForest(income ~ age + education + educationYears + maritalStatus + relationship + race + gen

pred_rf <- predict(rf, newdata=test)

pred_rf <- as.numeric(pred_rf)
incomeNum <- as.numeric(test$income)

cor_rf <- cor(pred_rf, incomeNum)
print(paste("cor = ", cor_rf))
```

```
## [1] "cor =  0.59760775702497"
```

```
rmse_rf <- sqrt(mean(pred_rf - incomeNum)^2)
print(paste("rmse = ", rmse_rf))
```

```
## [1] "rmse =  0.0530647340621545"
```

Result Analysis

The algorithms ranked best to worst for this data frame are:

1. Decision Tree
2. Naive Bayes
3. SVM

The reason behind these rankings are based upon the fact that the RMSE and correlation values obtained from the Decision Tree algorithm are the strongest, followed by the naive bayes algorithm, then the SVM algorithm.

I believe the reason Decision Tree functioned best is based on it's ability to closely mirror the human-decision making process. By conditionally handling data to achieve the most logical fit, this showed to be the strongest algorithm between the others used. Although decision trees generally do not have the same level of accuracy when compared to many other algorithms, this case seems to be an exception.

Overall, this script was able to learn the relationship between the income of an individual and their various personal attributes. With the final results, we are able to determine that the likelihood of an individual making >50k a year has a strong correlation to their marital status, education, age, and gender. With this, we found that a married white male from the United States with a Bachelor's Degree has the greatest likelihood of making >50k a year compared to the other values within the data set.