Nomaan Khan

Projet 2 Dataset 2

Dataset description

```
Link for training data <"https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.d
ata">
Link for testing data <"https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.te
st">

This is data set of census data compiled by the Census Bureau of the United States.

In this data set I'm trying to predict whether a person earns more than 50K or less than or equa
l to 50K based on the
census data submitted.

In this project train dataset (ds2) is adult.data.txt and the testing dataset is adult.test.txt.

I have analysed this dataset using Naive Bayes, Decision Trees and Neural Networks.

Total Number of Rows = 48,801
Number of columns = 15

Number of rows for train = 32,560.
Number of rows for test = 16,281.

Attribute Description:

1.  age: continuous.
2.  workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Witho
ut-pay, Never-worked.
3.  fnlwgt: continuous.
4.  education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th,
7th-8th, 12th, Masters,                       1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5.  education-num: continuous.
6.  marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spo
use-absent,                                     Married-AF-spouse.
7.  occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialt
y, Handlers-cleaners,                          Machine-op-inspct, Adm-clerical, Farming-fishing,
Transport-moving, Priv-house-serv, Protective-serv,                          Armed-Forces.
8.  relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9.  race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US
(Guam-USVI-etc), India,                              Japan,Greece, South, China, Cuba, Iran, Hond
uras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico,                          Portugal, Ir
eland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary,
                  Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&T
obago, Peru, Hong,                                  Holand-Netherlands.
```

Reading in the data. The first line of the test file was the description, which is why I skipped it.

```
ds2 <- read.table(file = "C:/Users/ADMIN/Desktop/Academia 2.0/Junior/Spring 2018/Machine Learnin
g/HW/Project 2/adult.data.txt",
                  header = FALSE, sep = ",")
ds2_test <- read.table(file = "C:/Users/ADMIN/Desktop/Academia 2.0/Junior/Spring 2018/Machine Le
arning/HW/Project 2/adult.test.txt", skip=1, header = FALSE, sep = ",")
```

Data Cleaning

Since both data sets do not have column names, I have to manually enter them. Only one row has Holand-Netherland as country of origin in the train dataset, and keeping that row created an error(difference in levels between train and test) which is why I removed it. Each row in the test file ended with a '.' unlike the train file, this created a difference in levels between the test and train file, so I used the sub function to fix it.

```
dim(ds2)
```

```
## [1] 32561    15
```

```
colnames(ds2) <- c("age","workclass","fnlwgt","education","education_num","marital_status","occu
pation","relationship",
                   "race","sex","capital_gain","capital_loss","hours_per_week","native_country",
"wage")
colnames(ds2_test) <- c("age","workclass","fnlwgt","education","education_num","marital_status",
"occupation","relationship",
                   "race","sex","capital_gain","capital_loss","hours_per_week","native_country",
"wage")


ds2 <- ds2[ds2$native_country != " Holand-Netherlands",]
ds2 <- droplevels(ds2)

ds2_test$wage <- sub(" <=50K.",replacement = " <=50K", ds2_test$wage)
ds2_test$wage <- sub(" >50K.",replacement = " >50K", ds2_test$wage)



ds2_test$wage <- as.factor(ds2_test$wage)

str(ds2)
```

```
## 'data.frame':    32560 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449
...
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5
3 ...
##  $ occupation    : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ 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 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capital_gain  : int  2174 0 0 0 0 0 0 14084 5178 ...
##  $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native_country: Factor w/ 41 levels " ?"," Cambodia",..: 39 39 39 39 6 39 23 39 39 39 ...
##  $ wage          : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

```
dim(ds2)
```

```
## [1] 32560    15
```

Dataset Exploration

```
summary(ds2)
```

```
##       age                    workclass         fnlwgt
##  Min.   :17.00    Private         :22695   Min.   :  12285
##  1st Qu.:28.00    Self-emp-not-inc: 2541   1st Qu.: 117832
##  Median :37.00    Local-gov       : 2093   Median : 178363
##  Mean   :38.58    ?               : 1836   Mean   : 189783
##  3rd Qu.:48.00    State-gov       : 1298   3rd Qu.: 237055
##  Max.   :90.00    Self-emp-inc    : 1116   Max.   :1484705
##                  (Other)          :  981
##          education     education_num                  marital_status
##   HS-grad     :10501   Min.   : 1.00    Divorced              : 4443
##   Some-college: 7290   1st Qu.: 9.00    Married-AF-spouse     :   23
##   Bachelors   : 5355   Median :10.00    Married-civ-spouse    :14976
##   Masters     : 1723   Mean   :10.08    Married-spouse-absent:  418
##   Assoc-voc   : 1382   3rd Qu.:12.00    Never-married         :10682
##   11th        : 1175   Max.   :16.00    Separated             : 1025
##  (Other)      : 5134                    Widowed               :  993
##             occupation            relationship
##   Prof-specialty :4140    Husband       :13193
##   Craft-repair   :4099    Not-in-family : 8305
##   Exec-managerial:4066    Other-relative:  980
##   Adm-clerical   :3770    Own-child     : 5068
##   Sales          :3650    Unmarried     : 3446
##   Other-service  :3295    Wife          : 1568
##  (Other)         :9540
##                    race            sex         capital_gain
##   Amer-Indian-Eskimo:  311    Female:10770   Min.   :    0
##   Asian-Pac-Islander: 1039    Male  :21790   1st Qu.:    0
##   Black             : 3124                   Median :    0
##   Other             :  271                   Mean   : 1078
##   White             :27815                   3rd Qu.:    0
##                                              Max.   :99999
##
##   capital_loss     hours_per_week         native_country      wage
##  Min.   :   0.00   Min.   : 1.00    United-States:29170    <=50K:24719
##  1st Qu.:   0.00   1st Qu.:40.00    Mexico       :  643    >50K : 7841
##  Median :   0.00   Median :40.00    ?            :  583
##  Mean   :  87.24   Mean   :40.44    Philippines  :  198
##  3rd Qu.:   0.00   3rd Qu.:45.00    Germany      :  137
##  Max.   :4356.00   Max.   :99.00    Canada       :  121
##                                    (Other)      : 1708
```

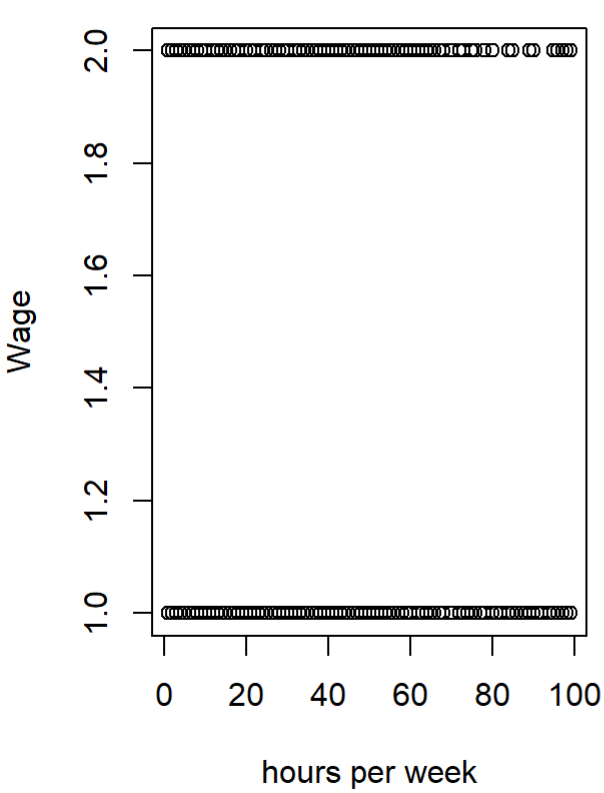
```
head(ds2)
```

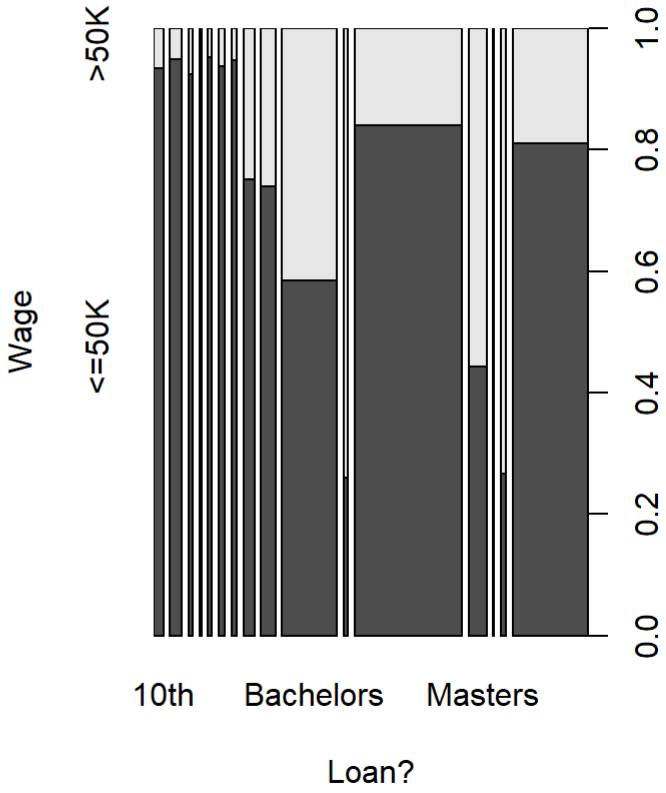
```
##   age         workclass fnlwgt  education education_num
## 1  39         State-gov  77516  Bachelors            13
## 2  50  Self-emp-not-inc  83311  Bachelors            13
## 3  38           Private 215646    HS-grad             9
## 4  53           Private 234721       11th             7
## 5  28           Private 338409  Bachelors            13
## 6  37           Private 284582    Masters            14
##        marital_status         occupation   relationship   race     sex
## 1       Never-married       Adm-clerical  Not-in-family  White    Male
## 2  Married-civ-spouse    Exec-managerial        Husband  White    Male
## 3            Divorced  Handlers-cleaners  Not-in-family  White    Male
## 4  Married-civ-spouse  Handlers-cleaners        Husband  Black    Male
## 5  Married-civ-spouse     Prof-specialty           Wife  Black  Female
## 6  Married-civ-spouse    Exec-managerial           Wife  White  Female
##   capital_gain capital_loss hours_per_week native_country   wage
## 1         2174            0             40  United-States  <=50K
## 2            0            0             13  United-States  <=50K
## 3            0            0             40  United-States  <=50K
## 4            0            0             40  United-States  <=50K
## 5            0            0             40           Cuba  <=50K
## 6            0            0             40  United-States  <=50K
```

```
dim(ds2)
```

```
## [1] 32560    15
```

```
par(mfrow=c(1,2))
plot(ds2$hours_per_week,ds2$wage, ylab = "Wage", xlab = "hours per week")
plot(ds2$education,ds2$wage, ylab = "Wage", xlab = "Loan?")
```





From the graphs we see hours per week is not a good predictor for wage but education generally is

Naive Bayes

Here I use the naive bayes algorithm on the data set.

```
library(e1071)
nb1 <- naiveBayes(wage~., data=ds2)
nb_pred <- predict(nb1, ds2_test[,-15])
mean(nb_pred==ds2_test[,15])
```

```
## [1] 0.8264234
```

Decision Trees

Here, I convert the factor native-country to numeric because tree() cannot handle a factor with over 32 levels.

```
library('tree')
```

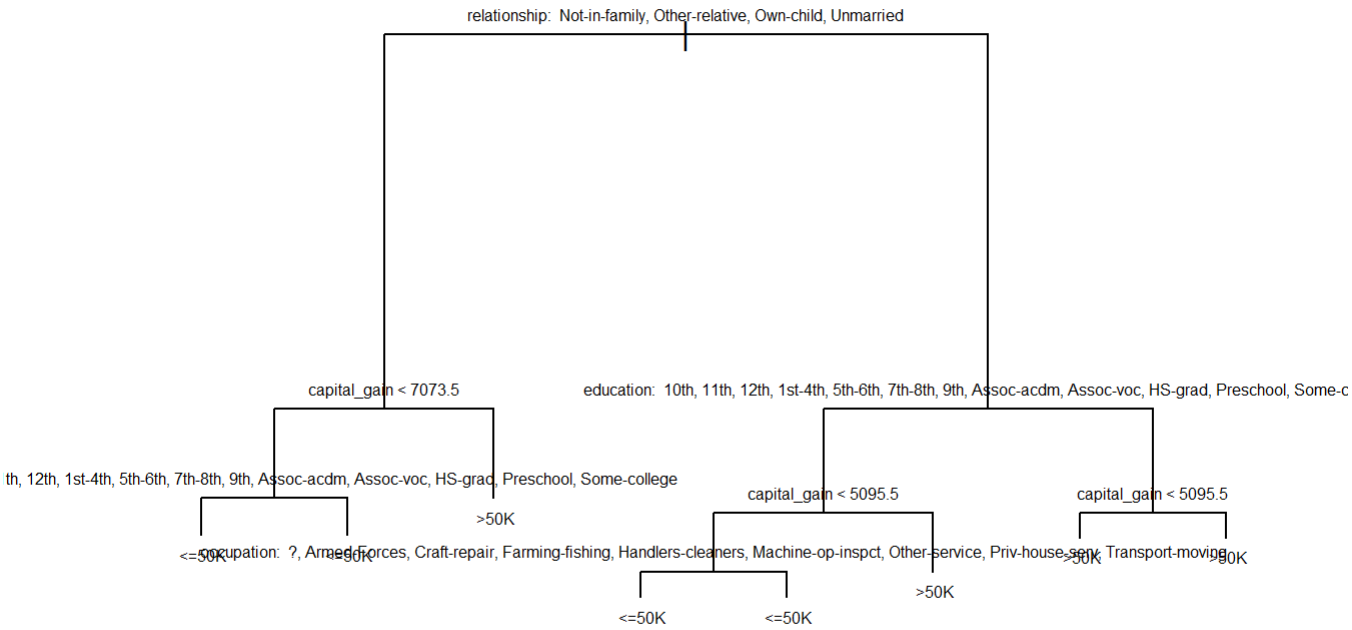
```
## Warning: package 'tree' was built under R version 3.4.4
```

```
ds2$native_country <- as.numeric(ds2$native_country)
ds2_test$native_country <- as.numeric(ds2_test$native_country)

tree.default = tree(wage~., ds2)
tree.pred <- predict(tree.default, ds2_test, type="class")
mean(tree.pred==ds2_test$wage, na.rm=TRUE)
```

```
## [1] 0.8445427
```

```
plot(tree.default)
text(tree.default, cex=0.5, pretty=0)
```

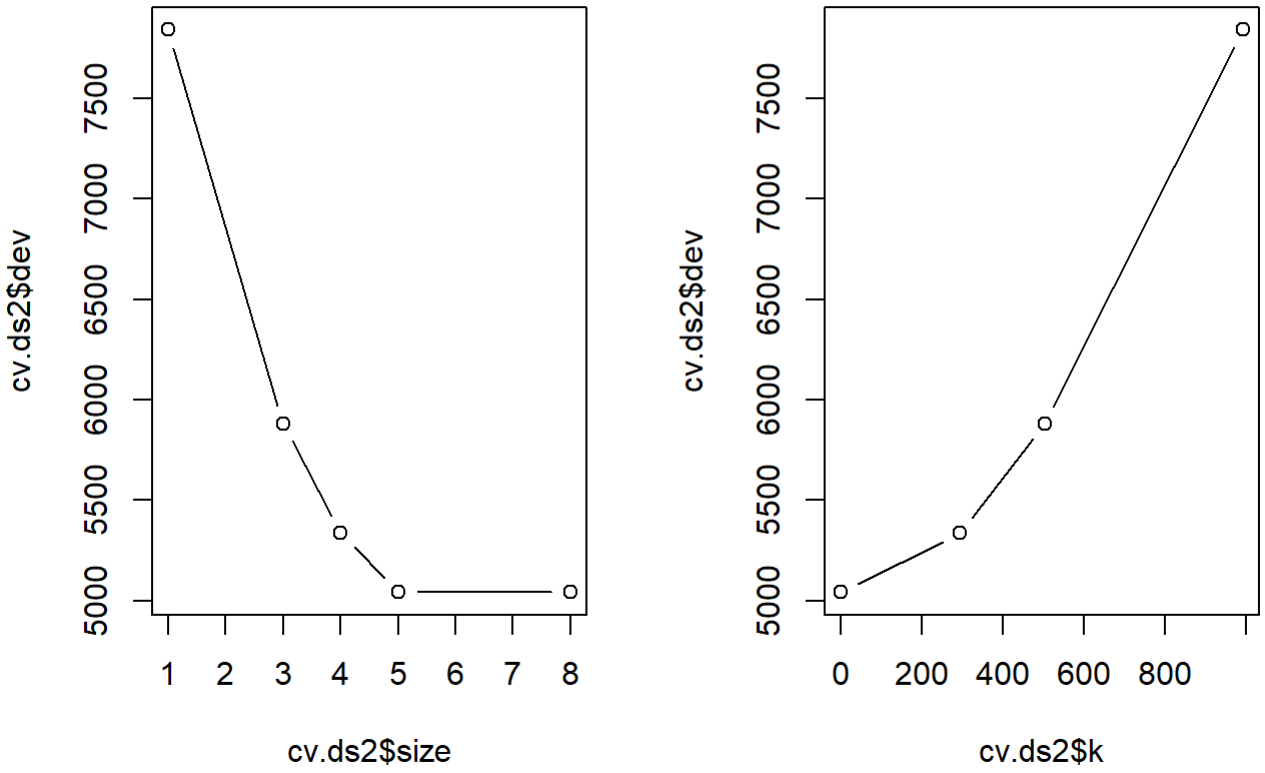


Cross Validating.

```
cv.ds2 = cv.tree(tree.default, FUN=prune.misclass)
cv.ds2
```

```
## $size
## [1] 8 5 4 3 1
##
## $dev
## [1] 5043 5043 5337 5880 7841
##
## $k
## [1] -Inf    0  294  502  991
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```
par(mfrow=c(1,2))
plot(cv.ds2$size, cv.ds2$dev, type="b")
plot(cv.ds2$k, cv.ds2$dev, type="b")
```



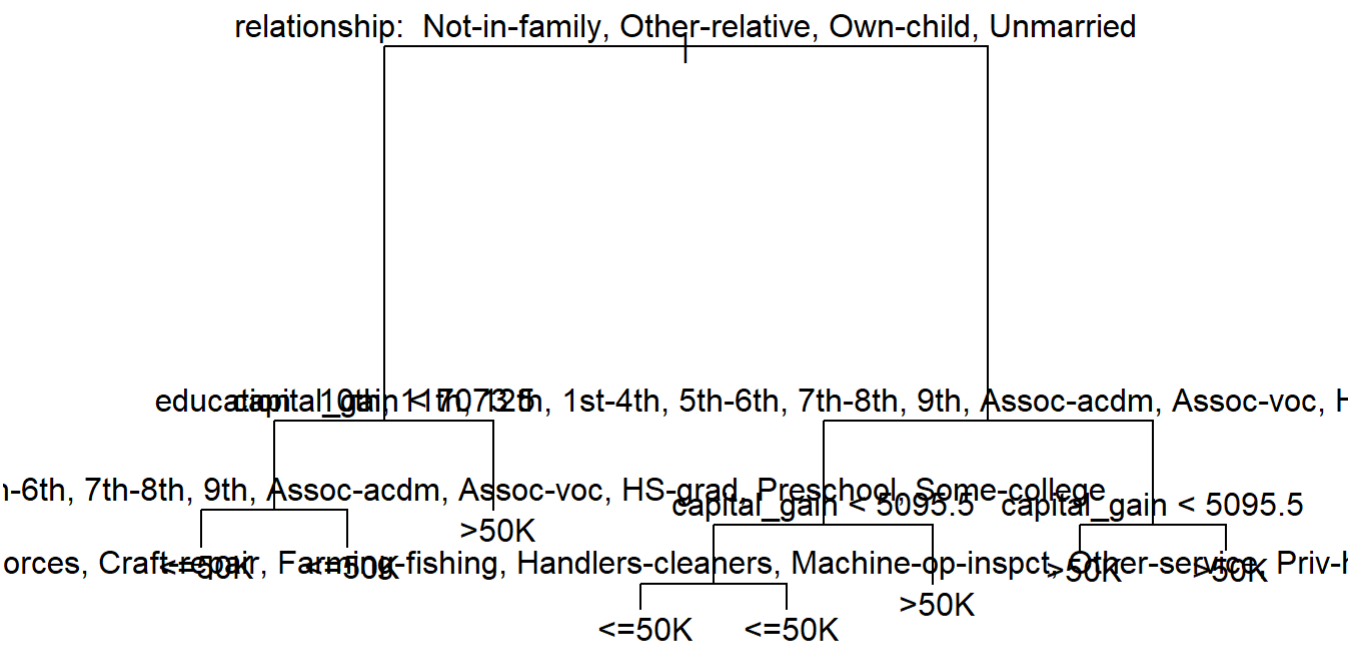
8 is the best choice.

Pruning.

```
prune.df = prune.misclass(tree.default, best=8)
prune.df.pred <- predict(prune.df, ds2_test, type="class")
mean(prune.df.pred==ds2_test$wage, na.rm=TRUE)
```

```
## [1] 0.8445427
```

```
plot(prune.df)
text(prune.df, pretty=0)
```



Neural Networks

Converting factors into numerics so that they can be used for neural network.

```
ds2_numeric <- ds2
ds2_numeric$workclass <- as.numeric(ds2_numeric$workclass)
ds2_numeric$education <- as.numeric(ds2_numeric$education)
ds2_numeric$marital_status <- as.numeric(ds2_numeric$marital_status)
ds2_numeric$occupation <- as.numeric(ds2_numeric$occupation)
ds2_numeric$relationship <- as.numeric(ds2_numeric$relationship)
ds2_numeric$race <- as.numeric(ds2_numeric$race)
ds2_numeric$sex <- as.numeric(ds2_numeric$sex)
ds2_numeric$native_country <- as.numeric(ds2_numeric$native_country)
ds2_numeric$wage <- as.numeric(ds2_numeric$wage)
str(ds2_numeric)
```

```
## 'data.frame':    32560 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : num  8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449
...
##  $ education     : num  10 10 12 2 10 13 7 12 13 10 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: num  5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : num  2 5 7 7 11 5 9 5 11 5 ...
##  $ relationship  : num  2 1 2 1 6 6 2 1 2 1 ...
##  $ race          : num  5 5 5 3 3 5 3 5 5 5 ...
##  $ sex           : num  2 2 2 2 1 1 1 2 1 2 ...
##  $ capital_gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native_country: num  39 39 39 39 6 39 23 39 39 39 ...
##  $ wage          : num  1 1 1 1 1 1 1 2 2 2 ...
```

```
ds2_test_numeric <- ds2_test
ds2_test_numeric$workclass <- as.numeric(ds2_test_numeric$workclass)
ds2_test_numeric$education <- as.numeric(ds2_test_numeric$education)
ds2_test_numeric$marital_status <- as.numeric(ds2_test_numeric$marital_status)
ds2_test_numeric$occupation <- as.numeric(ds2_test_numeric$occupation)
ds2_test_numeric$relationship <- as.numeric(ds2_test_numeric$relationship)
ds2_test_numeric$race <- as.numeric(ds2_test_numeric$race)
ds2_test_numeric$sex <- as.numeric(ds2_test_numeric$sex)
ds2_test_numeric$native_country <- as.numeric(ds2_test_numeric$native_country)
ds2_test_numeric$wage <- as.numeric(ds2_test_numeric$wage)
str(ds2_test_numeric)
```

```
## 'data.frame':    16281 obs. of  15 variables:
##  $ age           : int  25 38 28 44 18 34 29 63 24 55 ...
##  $ workclass     : num  5 5 3 5 1 5 1 7 5 5 ...
##  $ fnlwgt        : int  226802 89814 336951 160323 103497 198693 227026 104626 369667 104996
...
##  $ education     : num  2 12 8 16 16 1 12 15 16 6 ...
##  $ education_num : int  7 9 12 10 10 6 9 15 10 4 ...
##  $ marital_status: num  5 3 3 3 5 5 5 3 5 3 ...
##  $ occupation    : num  8 6 12 8 1 9 1 11 9 4 ...
##  $ relationship  : num  4 1 1 1 4 2 5 1 5 1 ...
##  $ race          : num  3 5 5 3 5 5 3 5 5 5 ...
##  $ sex           : num  2 2 2 2 1 2 2 2 1 2 ...
##  $ capital_gain  : int  0 0 0 7688 0 0 0 3103 0 0 ...
##  $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours_per_week: int  40 50 40 40 30 30 40 32 40 10 ...
##  $ native_country: num  39 39 39 39 39 39 39 39 39 39 ...
##  $ wage          : num  1 1 2 2 1 1 1 2 1 1 ...
```

```
library(neuralnet)
```

```
## Warning: package 'neuralnet' was built under R version 3.4.4
```

```
nn1 <- neuralnet(wage~age+workclass+fnlwgt+education+education_num+marital_status+occupation
                 +relationship+race+sex+capital_gain+capital_loss+hours_per_week+native_country,
ds2_numeric,
                  hidden=c(5,3), lifesign="minimal",
                  linear.output=FALSE, threshold=0.1)
```

```
## hidden: 5, 3    thresh: 0.1    rep: 1/1    steps:      24    error: 3920.57137   time: 2.85 s
ecs
```

```
temp_test <- subset(ds2_test_numeric, select=c("age","workclass","fnlwgt","education","education
_num","marital_status","occupation",
          "relationship","race","sex","capital_gain","capital_loss","hours_per_week","native_cou
ntry"))

nn1.results <- compute(nn1, temp_test)
results <- data.frame(actual=ds2_test_numeric$wage, prediction=nn1.results$net.result)
results$round <- round(results$prediction)
mean(results$round==results$actual)
```

```
## [1] 0.763773724
```

Model and Algorithm Analysis

Naive Bayes Accuracy: 0.8264234384.

Unpruned Tree accuracy: 0.8445427185 Pruned Tree accuracy: 0.8445427185

Neural Network accuracy: 0.763773724

The trees have the highest accuracy at 0.8445427185 followed by naive bayes and neural network produces the lowest accuracy.

This data set is well suited for trees and not very well suited for neural networks.