Classification

2/18/2023

Read in Data * Subset the data set df with only column 4, 10, 15 which are qualitative value of education, sex and income

```
df <- read.csv("adult.csv", header = TRUE)
df$income<- factor(df$income)
df$education <- factor(df$education)
df$hours.per.week<- factor(df$hours.per.week)
df$sex<- factor(df$sex)
df<- df[,c(4,10,15) ]
str(df)
```

```
## 'data.frame':    32561 obs. of  3 variables:
##  $ education: Factor w/ 16 levels "10th","11th",..: 12 12 16 6 16 12 1 11 12 16 ...
##  $ sex      : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2 ...
##  $ income   : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 1 2 ...
```

- Divide data into train and test

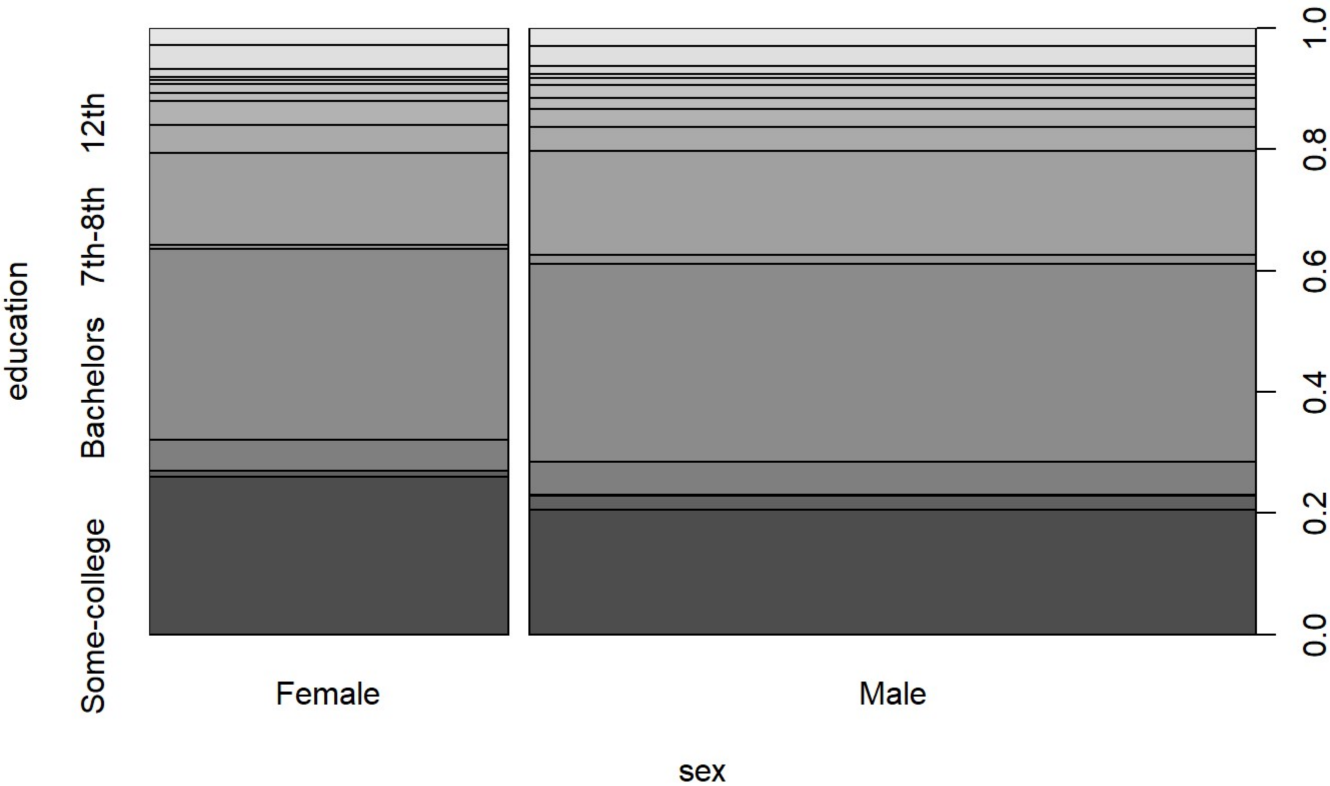
```
set.seed(1234)
i<- sample(1:nrow(df),0.8*nrow(df), replace = FALSE)
train <-df[i,]
test<-df[-i,]
df
```

education <fct>	sex <fct>	income <fct>
HS-grad	Female	<=50K
HS-grad	Female	<=50K
Some-college	Female	<=50K
7th-8th	Female	<=50K
Some-college	Female	<=50K
HS-grad	Female	<=50K
10th	Male	<=50K
Doctorate	Female	>50K
HS-grad	Female	<=50K
Some-college	Male	>50K

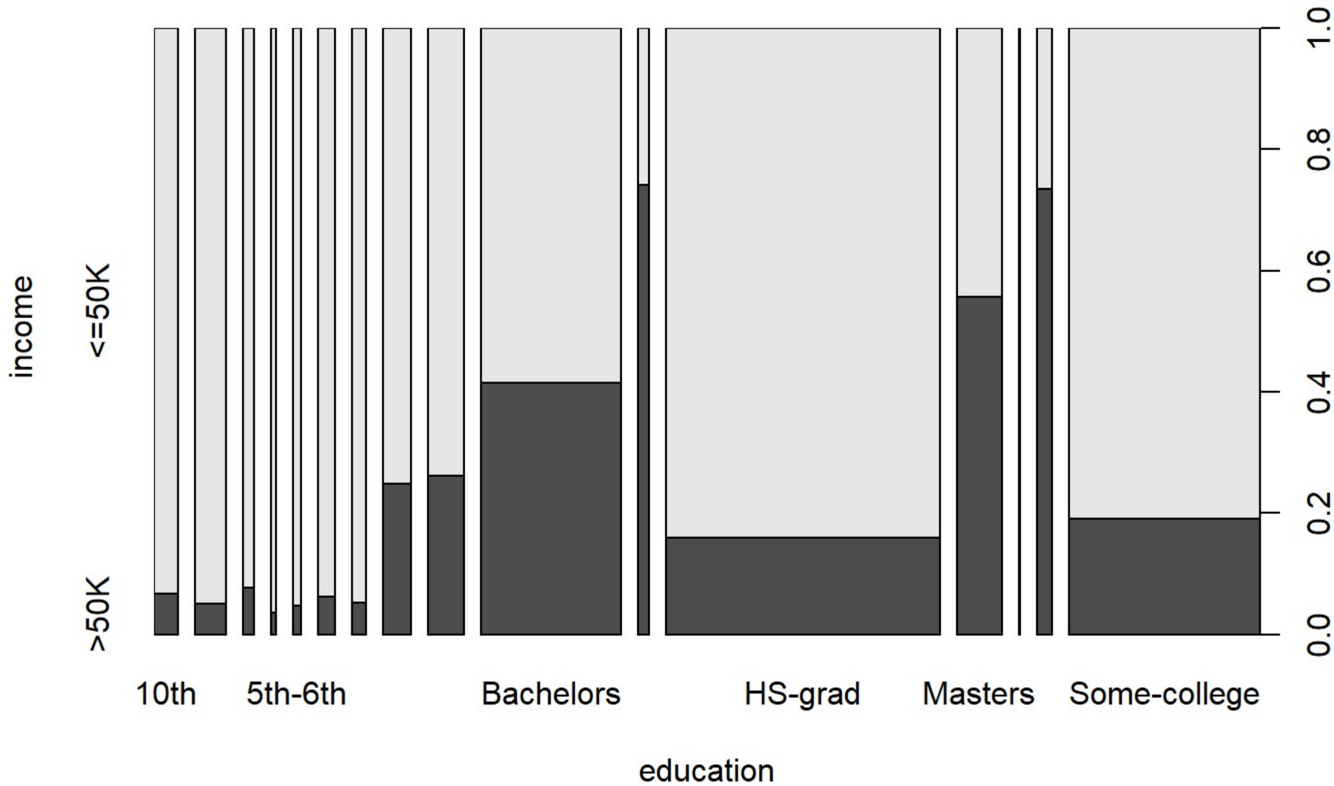
1-10 of 10,000 rows Previous **1** 2 3 4 5 6 ... 1000 Next

From the below graph. * female has a high education rate comparing to male * Education and income seems to have positive correlation

```
plot(df$education~df$sex, xlab = "sex",ylab = "education")
```




```
plot(df$income~df$education, xlab = "education", ylab = " income")
```



Contrast the factor classes. * Summarize the classes to see if there are N/A exists

```
contrasts(df$sex)
```

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

```
contrasts(df$education)
```

```
##              11th 12th 1st-4th 5th-6th 7th-8th 9th Assoc-acdm Assoc-voc
## 10th            0    0       0       0       0   0          0         0
## 11th            1    0       0       0       0   0          0         0
## 12th            0    1       0       0       0   0          0         0
## 1st-4th         0    0       1       0       0   0          0         0
## 5th-6th         0    0       0       1       0   0          0         0
## 7th-8th         0    0       0       0       1   0          0         0
## 9th             0    0       0       0       0   1          0         0
## Assoc-acdm      0    0       0       0       0   0          1         0
## Assoc-voc       0    0       0       0       0   0          0         1
## Bachelors       0    0       0       0       0   0          0         0
## Doctorate       0    0       0       0       0   0          0         0
## HS-grad         0    0       0       0       0   0          0         0
## Masters         0    0       0       0       0   0          0         0
## Preschool       0    0       0       0       0   0          0         0
## Prof-school     0    0       0       0       0   0          0         0
## Some-college    0    0       0       0       0   0          0         0
##              Bachelors Doctorate HS-grad Masters Preschool Prof-school
## 10th                 0         0       0       0         0           0
## 11th                 0         0       0       0         0           0
## 12th                 0         0       0       0         0           0
## 1st-4th              0         0       0       0         0           0
## 5th-6th              0         0       0       0         0           0
## 7th-8th              0         0       0       0         0           0
## 9th                  0         0       0       0         0           0
## Assoc-acdm           0         0       0       0         0           0
## Assoc-voc            0         0       0       0         0           0
## Bachelors            1         0       0       0         0           0
## Doctorate            0         1       0       0         0           0
## HS-grad              0         0       1       0         0           0
## Masters              0         0       0       1         0           0
## Preschool            0         0       0       0         1           0
## Prof-school          0         0       0       0         0           1
## Some-college         0         0       0       0         0           0
##              Some-college
## 10th                    0
## 11th                    0
## 12th                    0
## 1st-4th                 0
## 5th-6th                 0
## 7th-8th                 0
## 9th                     0
## Assoc-acdm              0
## Assoc-voc               0
## Bachelors               0
## Doctorate               0
## HS-grad                 0
## Masters                 0
## Preschool               0
## Prof-school             0
## Some-college            1
```

```r
sapply(df, function(x) sum(is.na(x) == TRUE))
```

```
## education       sex    income
##         0         0         0
```

- Build a generalized linear model glm1

```r
glm1<- glm(income~., data = train, family = binomial)
summary(glm1)
```

```
##
## Call:
## glm(formula = income ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8260  -0.6753  -0.4071  -0.1886   2.8802
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -3.75695    0.15300 -24.555  < 2e-16 ***
## education11th           -0.26349    0.21386  -1.232    0.218
## education12th            0.22926    0.25099   0.913    0.361
## education1st-4th        -0.42181    0.44505  -0.948    0.343
## education5th-6th        -0.37479    0.32182  -1.165    0.244
## education7th-8th        -0.07322    0.23209  -0.315    0.752
## education9th            -0.30688    0.27429  -1.119    0.263
## educationAssoc-acdm      1.73140    0.16943  10.219  < 2e-16 ***
## educationAssoc-voc       1.74349    0.16521  10.553  < 2e-16 ***
## educationBachelors       2.39465    0.15222  15.732  < 2e-16 ***
## educationDoctorate       3.84807    0.19903  19.334  < 2e-16 ***
## educationHS-grad         1.02207    0.15180   6.733 1.66e-11 ***
## educationMasters         3.05984    0.15943  19.192  < 2e-16 ***
## educationPreschool     -10.91608   82.40991  -0.132    0.895
## educationProf-school     3.64770    0.18401  19.823  < 2e-16 ***
## educationSome-college    1.30802    0.15265   8.569  < 2e-16 ***
## sexMale                  1.36688    0.04089  33.426  < 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: 28645  on 26047  degrees of freedom
## Residual deviance: 23891  on 26031  degrees of freedom
## AIC: 23925
##
## Number of Fisher Scoring iterations: 12
```

- Build the second model in Naiive Bayes nb1 predicting income with the variable of sex and education

- Calling nb1

```
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
## 0.7610565 0.2389435
##
## Conditional probabilities:
##        education
## Y               10th         11th         12th      1st-4th      5th-6th
##   <=50K 0.0358656174 0.0449455206 0.0160916061 0.0064063761 0.0129640839
##   >50K  0.0078727506 0.0072300771 0.0043380463 0.0009640103 0.0020886889
##        education
## Y            7th-8th          9th   Assoc-acdm    Assoc-voc    Bachelors
##   <=50K 0.0250706215 0.0190677966 0.0328389831 0.0396993543 0.1271186441
##   >50K  0.0054627249 0.0032133676 0.0348650386 0.0448264781 0.2839010283
##        education
## Y          Doctorate      HS-grad      Masters    Preschool  Prof-school
##   <=50K 0.0041364003 0.3564870864 0.0306698951 0.0020177563 0.0063054883
##   >50K  0.0401670951 0.2117609254 0.1238753213 0.0000000000 0.0539845758
##        education
## Y       Some-college
##   <=50K 0.2403147700
##   >50K  0.1754498715
##
##        sex
## Y          Female      Male
##   <=50K 0.3874596 0.6125404
##   >50K  0.1487789 0.8512211
```

- From the above data in the conditional probabilities suggested that among people that makes more than 50k, about 85% are male, where about 15% are female.
- In preschool education level, nearly 0 % probability rate to make 50k annually.
- Among people who had achieved their Bachelors, about 28% percent rate they can make more than 50k
- In the group of doctorate holders, the probability for them to make less than 50k is at 0.04 %.

Evaluate the logistic regression test set Accuracy prediction1 turns out to be around 0.78 in the generative data model.

```
probs <- predict(glm1, newdata=test, type = "response")

pred1<- ifelse(probs>0.5, 2,1)

accuracy1<- mean(pred1 == as.integer(test$income))

print(paste("glm accuracy = ",accuracy1))
```

```
## [1] "glm accuracy =  0.777521879318287"
```

```
table(pred1, as.integer(test$income))
```

```
##
## pred1    1    2
##     1 4406  959
##     2  490  658
```

- Test out the Naiive Bayes model
- Accuracy on naiive Bayes model turns out to be around 78%

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

```
##
## p1      <=50K >50K
##   <=50K  4761 1319
##   >50K    135  298
```

```
mean(p1 == test$income)
```

```
## [1] 0.7767542
```

Summary:

Comparing both model, logistic regression model glm1 and Naive Bayes model p1 both has similar accuracy rate reflecting how income behavior with variance of sex and different education levels in the sample. I think it’s becasue my data variables had great linear relationship. Tehrefore was effective under both models.