Preparing to Model the Data

By Saba Akram

1/2/2021

# Read in the data, partition Training and Testing Data.  
Clem3Training <- read.csv("D:/Data Mining and Predictive Analysis/Data sets/adult/Clem3Training")  
adult<-Clem3Training  
choose<-runif(length(adult$income),0,1)  
training<-adult[choose<=0.75,]  
testing<-adult[choose>0.75,]  
adult[1:5,1:3]

## age workclass demogweight  
## 1 39 State-gov 77516  
## 2 50 Self-emp-not-inc 83311  
## 3 38 Private 215646  
## 4 53 Private 234721  
## 5 28 Private 338409

training[1:5,1:3]

## age workclass demogweight  
## 1 39 State-gov 77516  
## 2 50 Self-emp-not-inc 83311  
## 3 38 Private 215646  
## 4 53 Private 234721  
## 5 28 Private 338409

testing[1:5,1:3]

## age workclass demogweight  
## 11 37 Private 280464  
## 14 32 Private 205019  
## 15 40 Private 121772  
## 17 25 Self-emp-not-inc 176756  
## 20 43 Self-emp-not-inc 292175

# Remove the target variable, Income, from the testing data.  
names(testing)

## [1] "age" "workclass" "demogweight" "education"   
## [5] "education.num" "marital.status" "occupation" "relationship"   
## [9] "race" "sex" "capital.gain" "capital.loss"   
## [13] "hours.per.week" "native.country" "income"

## Target variable is in column 15.  
testing<-testing[,-15]  
names(testing)

## [1] "age" "workclass" "demogweight" "education"   
## [5] "education.num" "marital.status" "occupation" "relationship"   
## [9] "race" "sex" "capital.gain" "capital.loss"   
## [13] "hours.per.week" "native.country"

## Target variable is no more in the testing data.  
  
# Remove the partitioning Variable, Part from both the datasets.  
testing<-testing[,-15]  
names(testing)

## [1] "age" "workclass" "demogweight" "education"   
## [5] "education.num" "marital.status" "occupation" "relationship"   
## [9] "race" "sex" "capital.gain" "capital.loss"   
## [13] "hours.per.week" "native.country"

names(training)

## [1] "age" "workclass" "demogweight" "education"   
## [5] "education.num" "marital.status" "occupation" "relationship"   
## [9] "race" "sex" "capital.gain" "capital.loss"   
## [13] "hours.per.week" "native.country" "income"

training<-training[,-16]  
names(training)

## [1] "age" "workclass" "demogweight" "education"   
## [5] "education.num" "marital.status" "occupation" "relationship"   
## [9] "race" "sex" "capital.gain" "capital.loss"   
## [13] "hours.per.week" "native.country" "income"