STAT-627 Spring 2021 Project Report

**4/24/2021**

Yu-Jui Chang,Rongtao- Ding,Wei-Chen Lu,Jingyi Xu,Yiang Gu

**(1) Study of interest**

This study aims to predict whether the US Adult citizen’s income exceeds $50K/year based on the US Census income dataset from 1994. This dataset was extracted by Barry Becker from the 1994 Census database. What factors influence people’s income is an interesting topic for academic researchers who study about the US personal financial structure or even household financial income. This study may be helpful to other researchers to study further.

**(2) Our thoughts and questions**

After stepwise, we identified 12 factors in the data that could affect people's income, so we would like to analyze these factors to see which of them affected people's income. The predictors are age, workclass, fnlwgt,education,marital status, occupation, relationship, race, sex, capital gain, captial loss.

**(3) The analytics question**

We wanted to model what variables would affect the income status of adult U.S. citizens? By comparing the classification error rate of the model, the most appropriate model was selected to understand the impact of variables in the model on US Adult citizen’s income.

**(4) Method**

Because our response variable is binary (1: Income <= $50K/year, 0: Income > $50K/year), we used the logistic regression method, KNN, LDA, QDA, and Classification Tree method to classify the data. Before applying those data for analysis, we used stepwise variable selection to select the variables that we used for our best model. Then, we applied a cross-validation method to find the best model.

**(5) Dataset Description**

The adult dataset is made from the US Census income dataset by Barry Becker.

age: continuous. The age of an individual.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous. The continuous variable fnlwgt represents final weight, which is the number of units in the target population that the responding unit represents

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.

education-num: continuous. The highest level of education achieved in numerical form.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous. Capital gains for an individual.

capital-loss: continuous. Capital loss for an individual.

hours-per-week: continuous. Working hours for an individual per week.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad Tobago, Peru, Hong, Holland-Netherlands.

This dataset has 30162 observations after omitting missing values and 15 variables in total. There are 8 categorical variables and 7 variables are quantitative. Also, the income variable is our dependent variable that we would like to predict what makes people have more income.

# **(6) Load and tidy the data**

# 

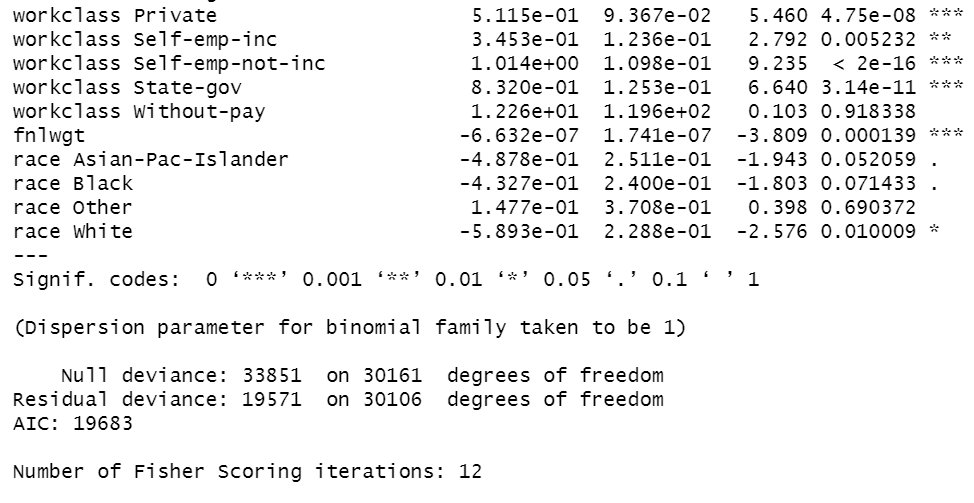
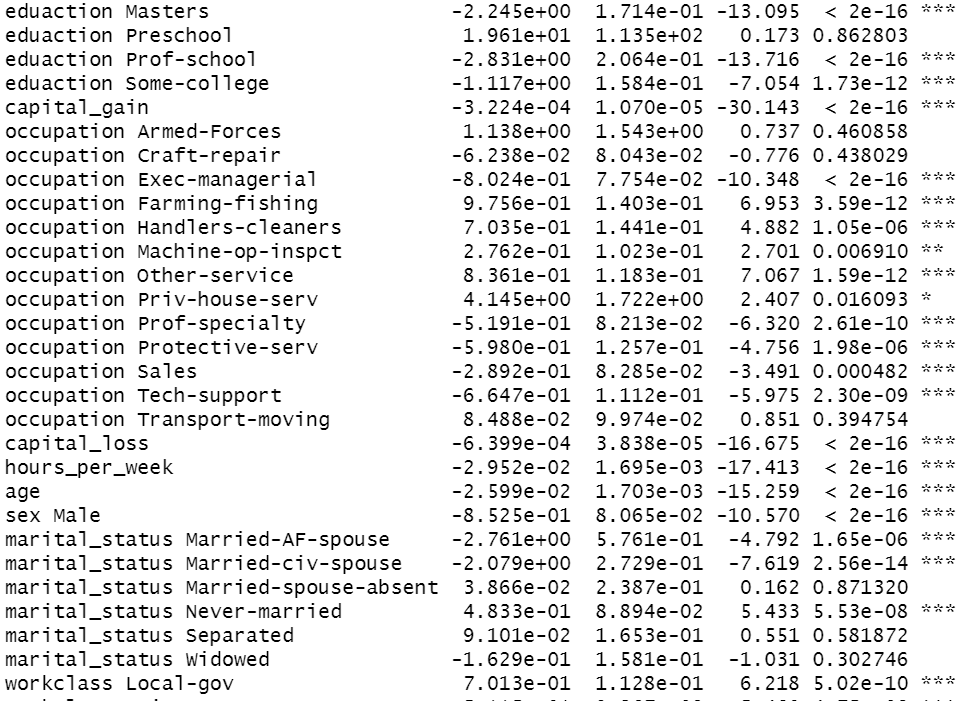
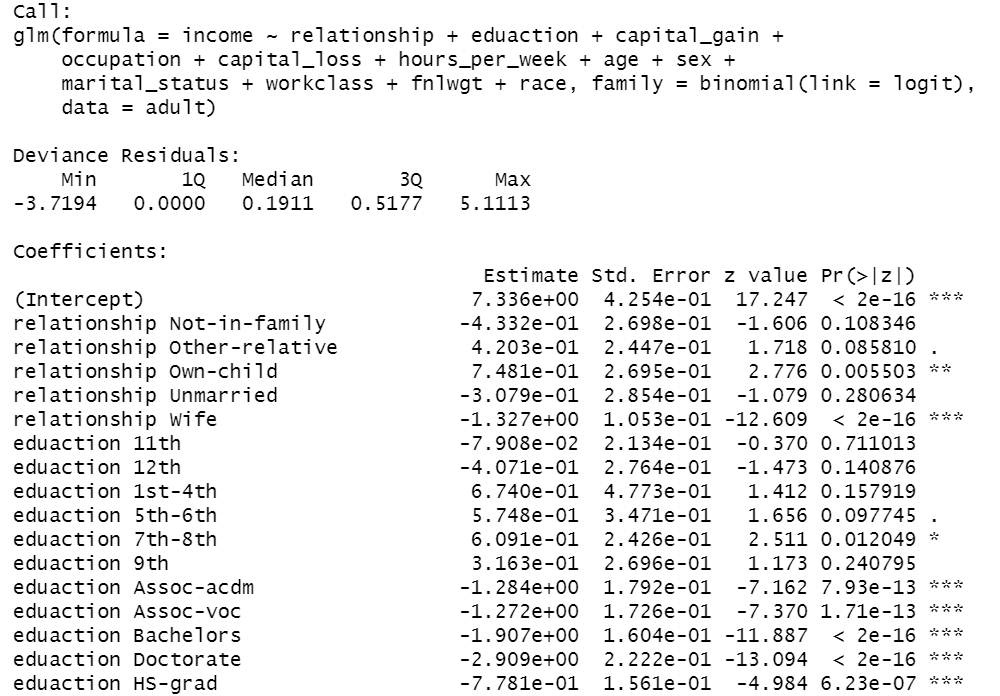
Since our dataset has some missing value, we clean those rows and maintain useful data to build our models.



**(7) Variable selection**

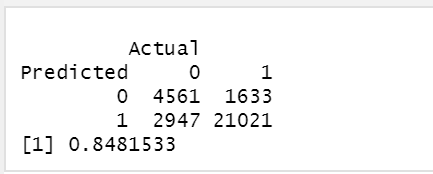
Our group used stepwise variable selection to avoid the multicollinearity from our full model. We applied the forward stepwise function from the full model, and have a reduced model with the lowest AIC as our selected predictors in the study. According to the result, we choose relationship, eduaction, capital\_gain, occupation, capital\_loss, hours\_per\_week, age, sex, marital\_status, workclass, fnlwgt, race for our predictors.

**(8) Logistic Regression**



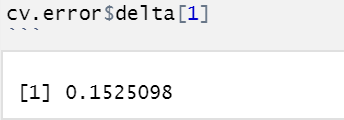
We fit the logic regression model, the residual deviance is 19571 and AIC is 19683.

# **Predict confusion matrix**



The predictive accuracy is 84.8%., thus the classification error rate is 15.2%.

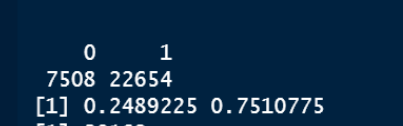
# **Cross-Validation**



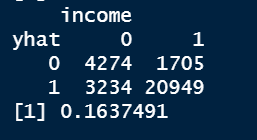
We use the K-fold(K=10) method to do cross-validation for the Logistic Regression model, the deviance of this model is 0.1525098.

# **(9) LDA classification**

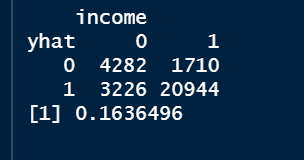
Prior probability: 0: 24.89% / 1: 75.11%



Confusion matrix & Classification error rate (16.3749%):



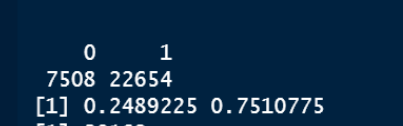
Find the best threshold. Get prior probability (0: 0.25/ 1: 0.75), confusion matrix and classification error rate (16.3649%).

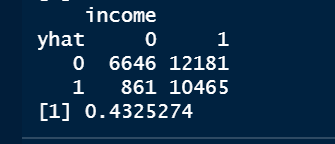


In LDA, we use relationship, eduaction, capital\_gain, occupation, capital\_loss, hours\_per\_week, age, sex, marital\_status, workclass, fnlwgt, and race 12 predictors. Prior probability are 0(income < 50K)= 0.25 and 1(income >= 50K)= 0.75 and classification error rate is 16.37%. After tuning to find the best threshold for prior probability, we get the classification error rate 16.36% which almost matches with origin.

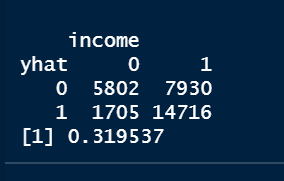
# **(10) QDA**

Prior probability: 0: 24.89% / 1: 75.11%



Confusion matrix & Classification error rate (43.25%):

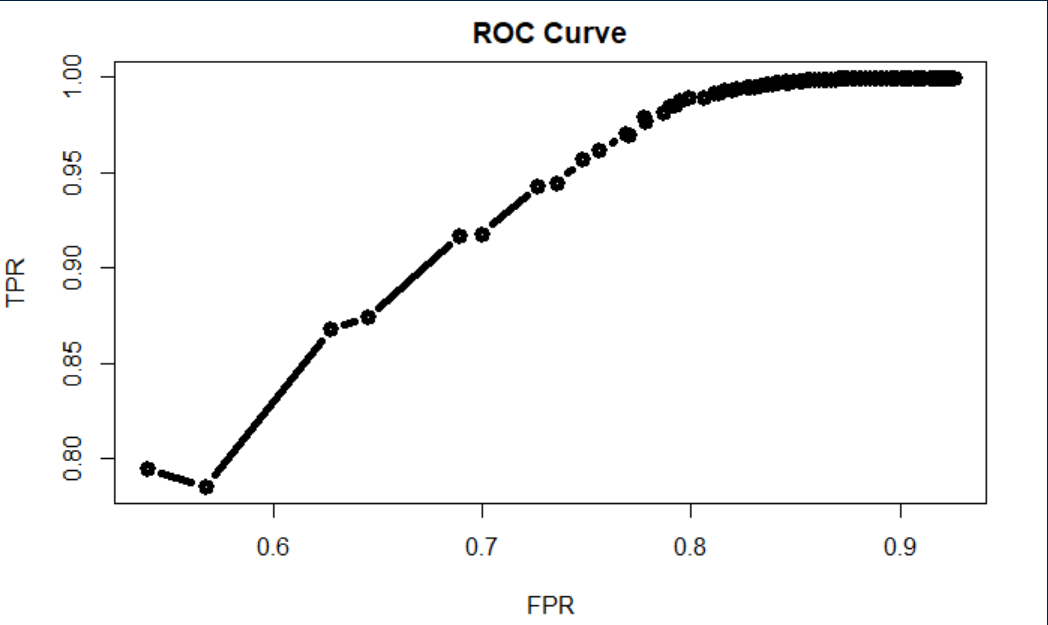
Find the best threshold. Get prior probability (0: 0.01/ 1: 0.99), confusion matrix and classification error rate (31.95%).

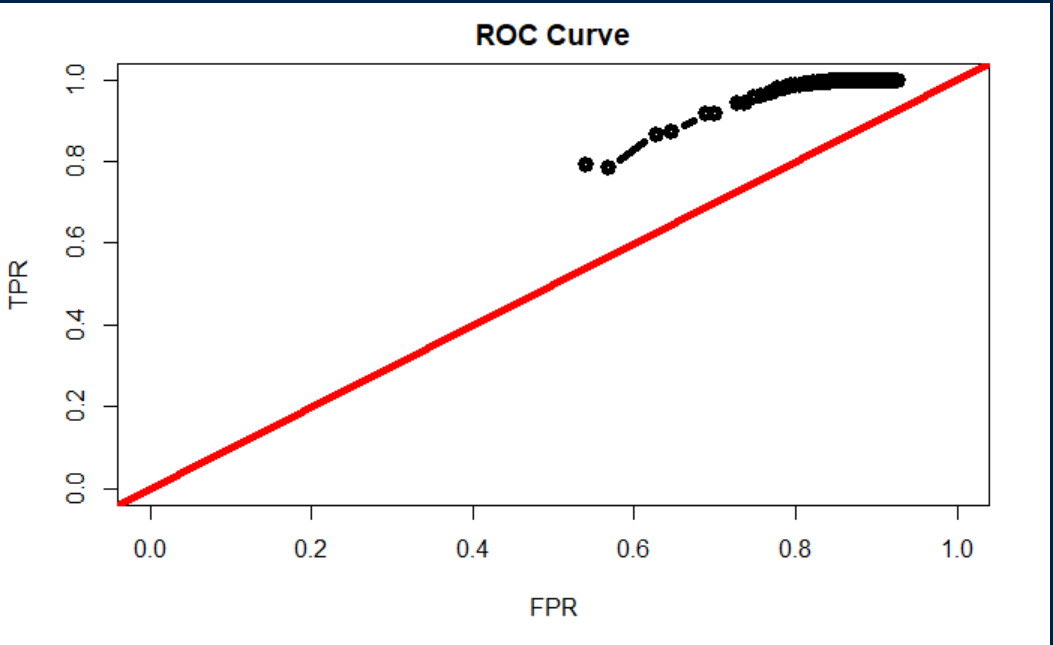


In QDA, at first, we use the same predictors like LDA but it has some errors. So, we choose age, education\_num, occupation, race, sex, and hours\_per\_week for our predictors. During QDA analysis it will cause some NA value so we would delete it, at preliminary modeling, our classification error rate is 43.25%. After tuning the threshold, we find prior probability with 0(income < 50K)= 0.01 and 1(income >= 50K)= 0.99 have lowest classification rate 31.95%.

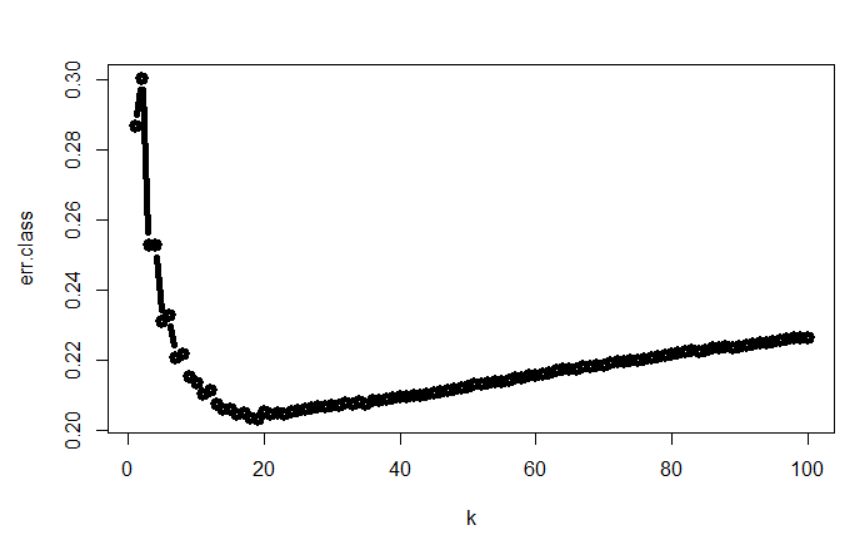
# **(11) KNN**

After we build the KNN model, we plot the ROC curve to verify if this is a good model or not and plot k verse to classification error rate to find the best K.

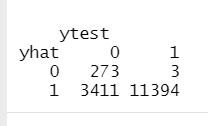




The ROC curve shows above the slope=1 line, so we could say this is a good model.



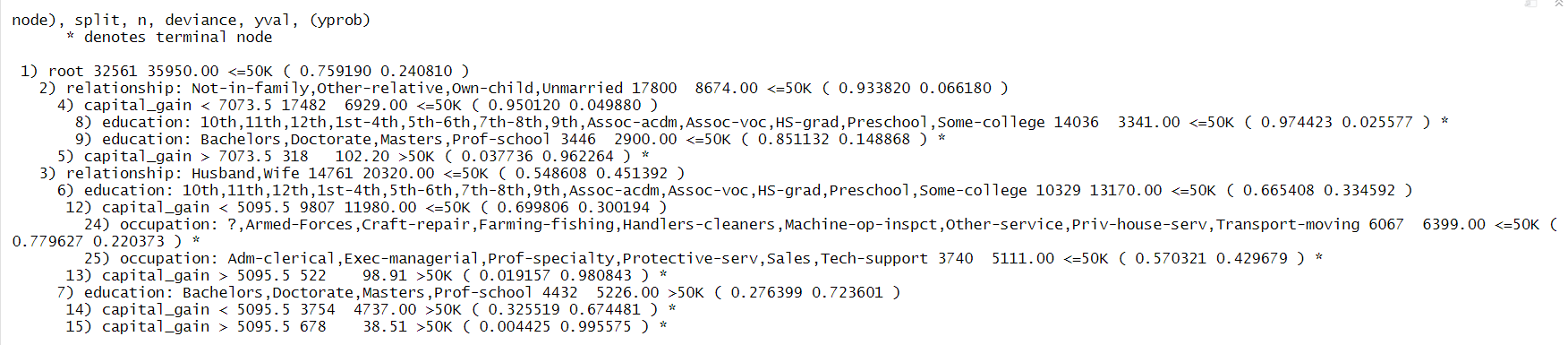
The best classification error rate performance when the K is 19.

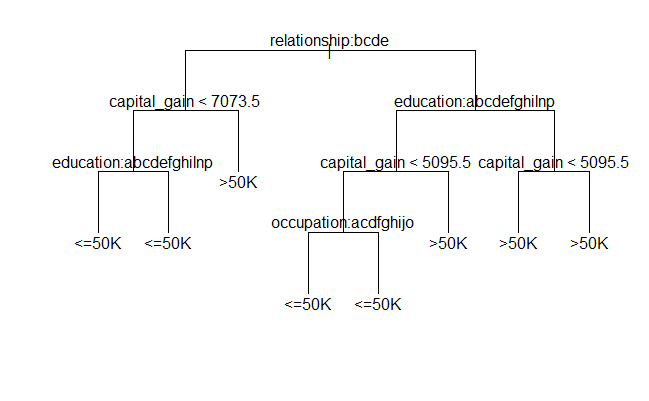


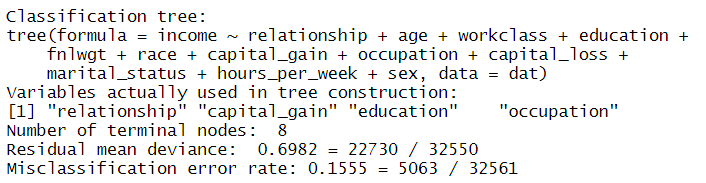
In KNN method, we use relationship, eduaction, capital\_gain, occupation, capital\_loss, hours\_per\_week, age, sex, marital\_status, workclass, fnlwgt, and race 12 predictors to build our KNN model. We find that K= 19 has the lowest classification error rate 20.33%. The KNN method’s performance is not as accurate as the logistic regression and LDA. Therefore, KNN is not our optimal method, at this point, for this data analysis.

# **(12) Classification trees**

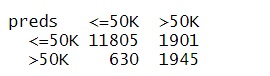
## 







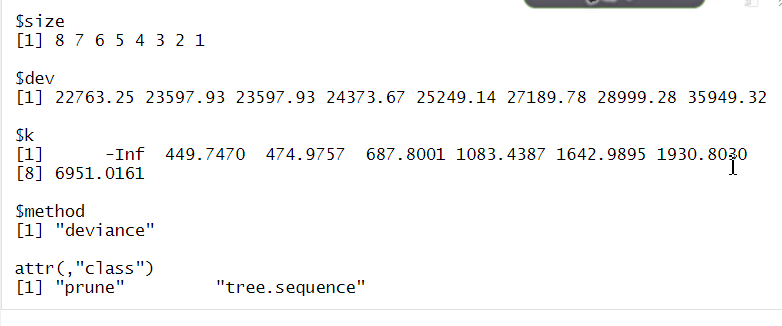
The misclassification rate of the training dataset is 0.1555.

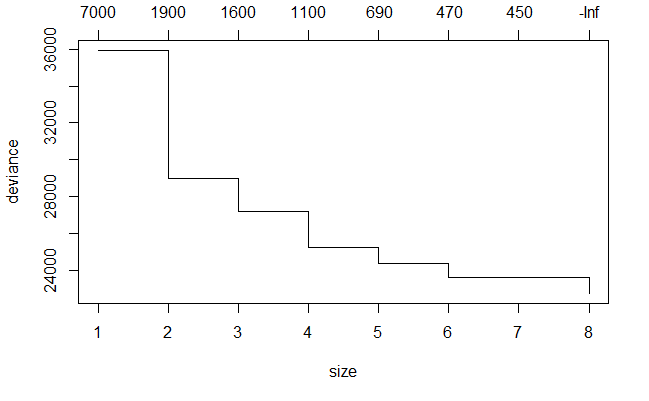


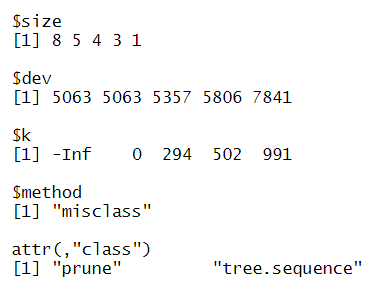


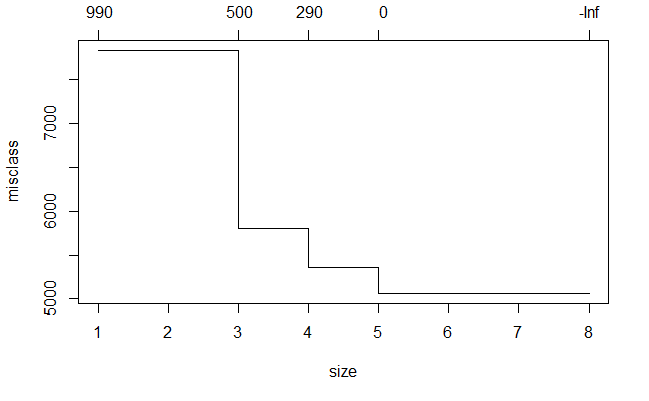
The misclassification rate of the test dataset is 0.1554.

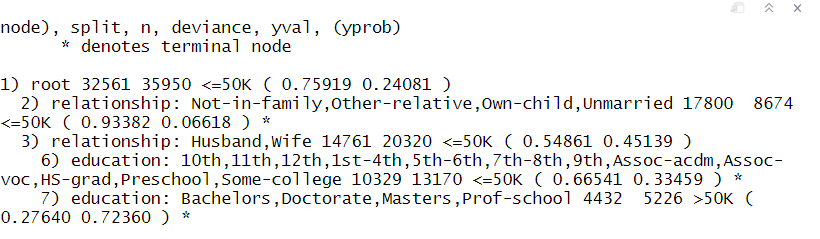
## **Pruning**











Therefore, it can be found that 24.08% of all the people’s income is greater than 50K. In the original model, capital\_gain, occupation, relationship and education influence people’s income to the greatest extent.By optimizing the classification of the tree, relationship and education influence people’s income status to the greatest extent in the final model.

**(13) Finding**

Based on above method results, the adult data analysis fit the logistic regression the best with the lowest prediction classification error rate. The logistic regression indicates that the education, occupation and relationship influences the response the most. We can know the group of people who have a wife, have a PhD background and work related to technology such as computer engineering that have the most possible probability to be classified as high income citizens in the US. The analysis result reveals that males have significantly more income than females on average. In other words, this data has a gender gap on income, which was a society issue back to 1996. Besides, older people are more likely to have higher income.

**(14) Conclusion**

Throughout this data analysis, the result is somehow reasonable to typical point of views about people with wealth. The education, occupation, race, sex and work class influence the income the most. This data collected the US citizen’s individual status in 1996. To our surprise, the 1996’s income structure seems to be familiar to modern US society. That is, the people who are older, have higher education, and have a family tend to be classified as a high income group in the US Adult data.

Our study also found that, on average, the more working hours per week the more income people have. The assumption of the anti- gender gap income was that males have more working hours per week than females. However, our study cannot justify or refute this argument and need further academic studies.

**(15) Reference**

Adult Data Set:<https://archive.ics.uci.edu/ml/datasets/Adult>