Project 1: Classification

CSE 5334
Name: Hamza Reza Pavel
Student ID: 1001741797
Group: 2
Group Specific Information:
K Value for folding: 6
Train/Test Split: 70/30

Introduction

In this project, we will be doing analysis on 1990 Census dataset and using Decision trees and Naïve Bayes to train model for classification. We will further analyze the result of the individual models and compare them with each other.

The provided data set contains 32561 rows and 15 columns. In the following section we will perform some cursory analysis of the dataset, and preprocess the dataset based on our analysis to improve the accuracy of our classification models.

Cursory Analysis of the Dataset

The dataset contains 32561 rows and 15 columns. A short description of the columns is given in the table below:

Table 1: Short Description of dataset.

Column Name	Description
Age	The age of individuals. Continuous numeric data greater than 0.
Workclass	Factor type data that represents the work class an individual belongs to.
fnlwgt	Weight that represents number of people the given row of data represents.
Education	Factor type data representing the education level of people.
Education-num	The education level represented using numeric form.
Marital-status	Factor type data representing the marital status of the individual.
Occupation	Occupation of individuals. Factor type data.
Relationship	Relationship of individuals. Factor type data.
Race	Race of individuals. Factor type data.
Sex	Male/Female. Factor type data.
Capital-gain	Numeric data starting from 0 representing the capital gain of individual.
Capital-loss	Numeric data starting from 0 representing the capital loss of individuals.
Hours-per-week	Numeric data representing the hours worked per week by individuals.
Native-country	Native country of individuals. Factor type data.
Income-level	Discretized income level. Factor type data. Values are “<=50K” and “>50K”

From the description of the data, we can conclude that *fnlwgt* column is not relevant for our classification purpose. Similarly, *Education* and *Education-num* columns represent the same information in two different formats. We can keep one and drop the other when training our model.

Data Preprocessing

We have preprocessed the given dataset for ease of use with the R data-frames, and to improve the accuracy of our predication models. Following steps were taken to preprocess the dataset:

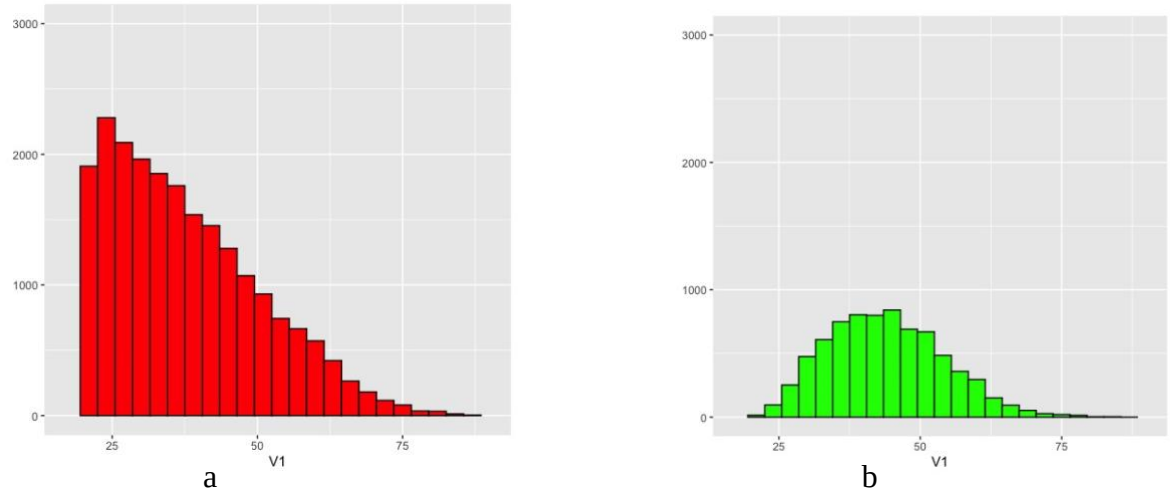
i. Using R, we removed the unknown values represented using "?".
ii. We replaced hyphens "-" with underscores "_".
iii. We remove any data-frame element that appears under a certain threshold in the dataset. For instance, the native-country name "*Holand-Netherlands*" appears once in the entire data-frame. We remove such tuples from dataset to reduce the complexity of our models.
iv. We remove any special characters such as "<", ">","&","=" etc. so that use of individual data-frame elements in R do not cause any issues.
v. We replace class elements "<=50K" with "LTE50K", and ">50k" with "GT50K"

After preprocessing the data, we split the dataset into *train* and *test* using random sampling and the ratio 70/30.

Table 2: Dataset dimensions.	
Cleaned Dataset(After Preprocessing)	30161
train	20107
test	10054

Analysis and Exploration of Features

In this section we are going to analyze interesting features in the dataset. First, we are going to analyze distribution of income based on age. For this, we plot distribution of income for ages in two graph. One graph represents people with income equal or below 50k, another graph represents people with income above 50k.



a b

Fig. 1: Distribution of income against age. a) Income below 50K. b)Income above 50k

An interesting point is, as age increase, the number of people having income below 50k decrease. The distribution of income above 50K seems like a normal distribution among different age group. From the distribution of the elements of the feature, we can say age is an important feature in classifying income.

Next, we look at the relationship between education level and income level of individuals.

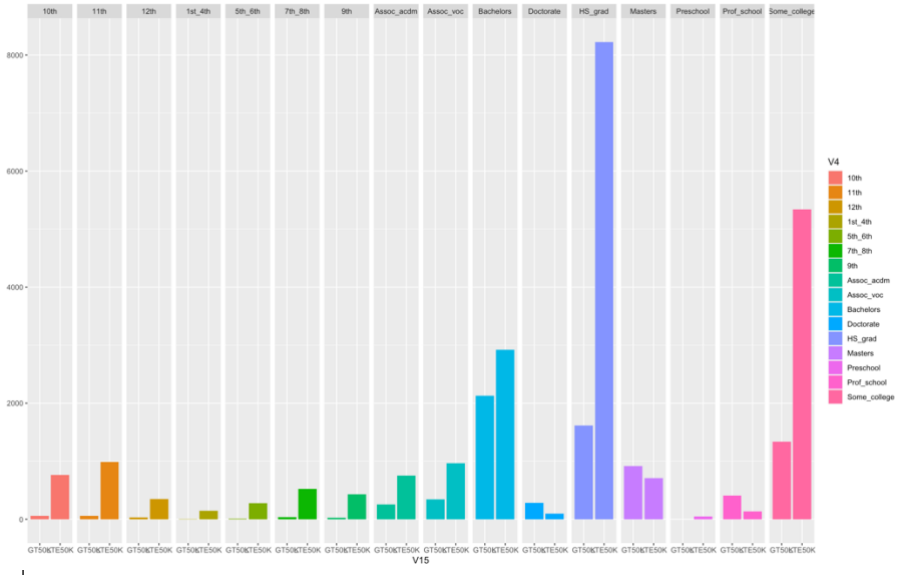


Fig. 2: Income level against education level.

One interesting thing to notice from this plot is, people with Doctorate, Masters, and Prof-School education tend to earn more than 50k in their respective groups. Also, most individual in the dataset has at least high school level education and least amount of people have Doctorate level degree.

Next we do analysis on relationship between gender and income level. This plot seems like a fair comparison. One interesting thing about this plot is, female population has significantly less population in the “>50K” income than their male counterpart.

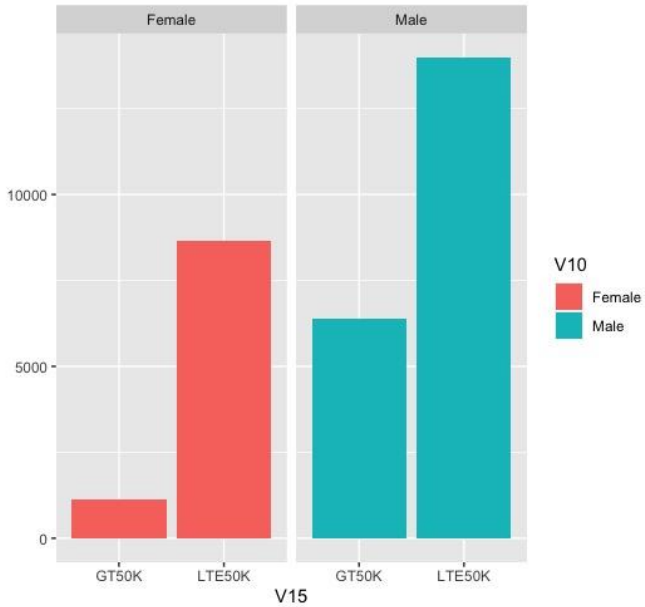


Fig. 3: Income distribution based on gender.

If we look at Figure 4, we are going to notice that most people in the sample dataset are from white demographic. Also, white demographic has the highest number of people with income above 50k. One interesting thing from this plot is, “Asian-Pac-Islander” has the highest ratio of people in income above 50k in their own demographic compared to other demographic.

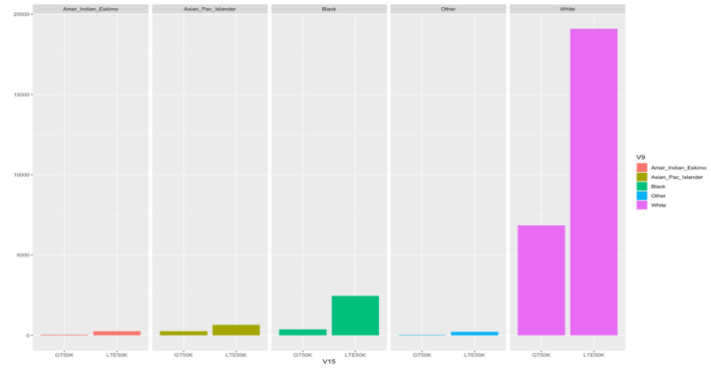


Fig. 4: Income distribution against race.

The Figure 5 shows income distribution among different occupation. Interesting thing about this graph is, people from Prof-specialty and Exec-manager has almost equal distribution among both income group. Most income disparity is the Adm-Clerical profession where a higher portion of people work in the <=50K income group.

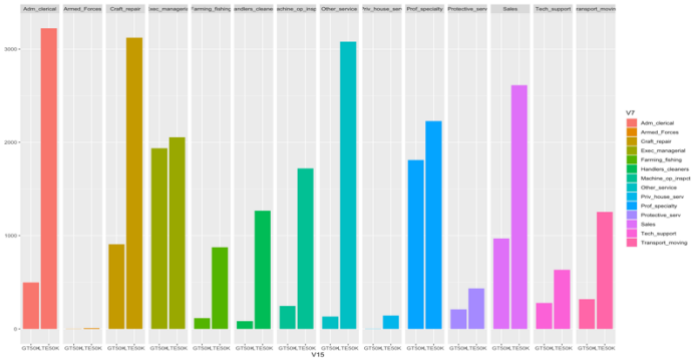


Fig. 5: Occupation vs Income.

The Figure 6 shows work-class vs income plot. From this plot, we can see that the Private workclass represent most of the tuples in the sample dataset. Self_emp are people who works for their own self where the limit for maximum income is without any bound. As a reason this is the group that has

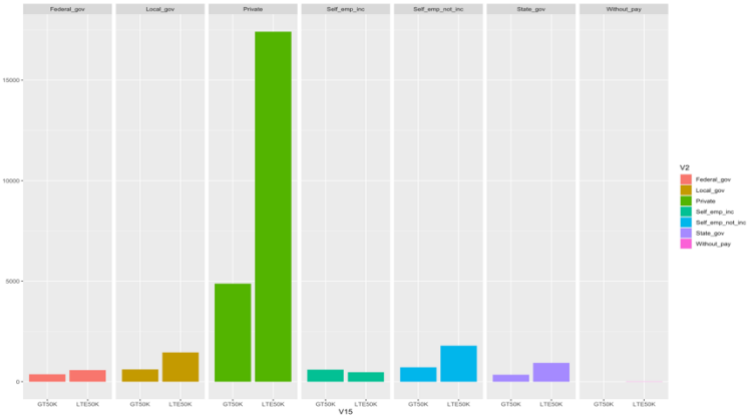


Fig. 6: Work-class vs Income.

more people in the “>50K” income group. As for other categories, federal_gov job has the highest ratio of people in “>50k” group among these categories.

Data Collection

We use *R* programming language and R studio IDE for our analysis. R has a wide variety of library for modelling decision tree and Naïve Bayes. We use the rpart package for decision tree and e1071 package for Naïve Bayes. We split out data set into training and testing set and use the same dataset for comparing results among the models.

Decision Tree (Information Gain): In our case, the value for K for cross validation is 6. We evaluate the model with K-fold cross validation and get the following accuracy for each fold of K:

Before withholding any column:

Iteration	Accuracy for Kth Fold(No column witheld)
1	0.8357292
2	0.827772
3	0.827772
4	0.808524
5	0.7964389
6	0.7672934

Accuracy using the testing data from split: 0.838

The confusion Matrix for Decision Tree using Information gain:

	Reference	
Prediction	>50K	<=50K
>50K	1180	368
<=50K	1261	7245

Precision = 0.8517517

Recall = 0.9516616

F1 Score = 0.8989391

With the *varImp* function we notice that column 11 is the root of the tree. We withheld this column and perform the analysis again to see the impact on the model.

After withholding column:

Iteration	Accuracy for Kth Fold(After withholding col 11)
1	0.8155364
2	0.8155364
3	0.8155364
4	0.8155364
5	0.8097184
6	0.7860983

Accuracy using the testing data from split: 0.8206

The confusion Matrix for Decision Tree using Information gain:

	Reference	
Prediction	>50K	<=50K
>50K	1207	570
<=50K	1234	7043

Precision = 0.8509122

Recall = 0.9251281

F1 Score = 0.8864695

Decision Tree(Gini): Same as before, we run K – fold cross validation on our decision tree using gini split, first without withholding any column then withholding the column in the root node. The results are given below:

Before withholding any column:

Iteration	Accuracy for Kth Fold(No column witheld)
1	0.8358287
2	0.827772
3	0.827772
4	0.808524
5	0.7964389
6	0.7672934

Accuracy using the testing data from split: 0.841

The confusion Matrix for Decision Tree using gini index:

	Reference	
Prediction	>50K	<=50K
>50K	1180	368
<=50K	1261	7245

Precision = 0.85

Recall = 0.952

F1 Score = 0.9

After withholding column: Same as before, we notice that col 11 is the root of the tree. We withhold this column and generate the result for the test dataframe.

Iteration	Accuracy for Kth Fold(No column witheld)
1	0.8173767
2	0.8173767
3	0.8165809
4	0.815437
5	0.8097184
6	0.795247

Accuracy using the testing data from split: 0.82

The confusion Matrix for Decision Tree using gini index:

	Reference	
Prediction	>50K	<=50K
>50K	1220	534
<=50K	1221	7079

Precision = 0.8528916

Recall = 0.9298568

F1 Score =0.8897128

Naïve Bayes:

Before withholding any column:

Mean for K-fold cross validation accuracy = 0.9682703

Accuracy for Split = 0.8254

The confusion Matrix for Naïve Bayes Model:

	Reference	
Prediction	>50K	<=50K
>50K	1220	534
<=50K	1221	7079

Precision = 0.9690682

Recall = 1

F1 Score = 0.9842912

After withholding column: Using varImp it seems that the most important feature is Col 8. We withhold this column and run our analysis on the model again.

Mean for K-fold cross validation accuracy = 0.9723483

Accuracy for Split = 0.8254

The confusion Matrix for Naïve Bayes Model:

	Reference	
Prediction	>50K	<=50K
>50K	1220	534
<=50K	1221	7079

Precision = 0.9675902

Recall = 1

F1 Score = 0.9835282

Roc Curve: We plot the following ROC curve for models using the evalMl library of R.

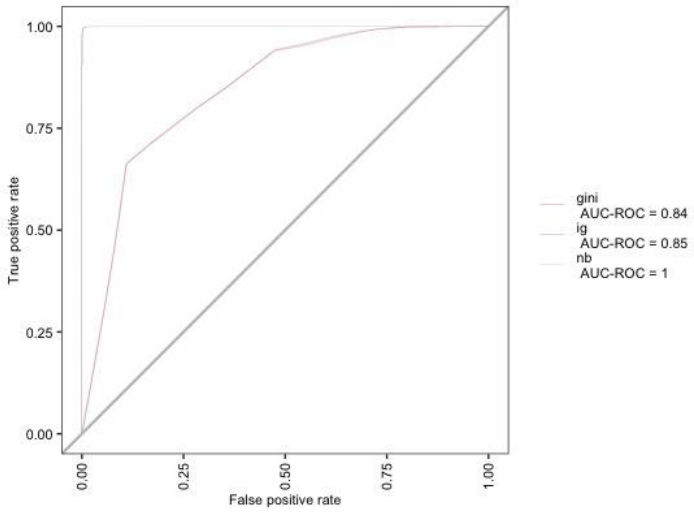
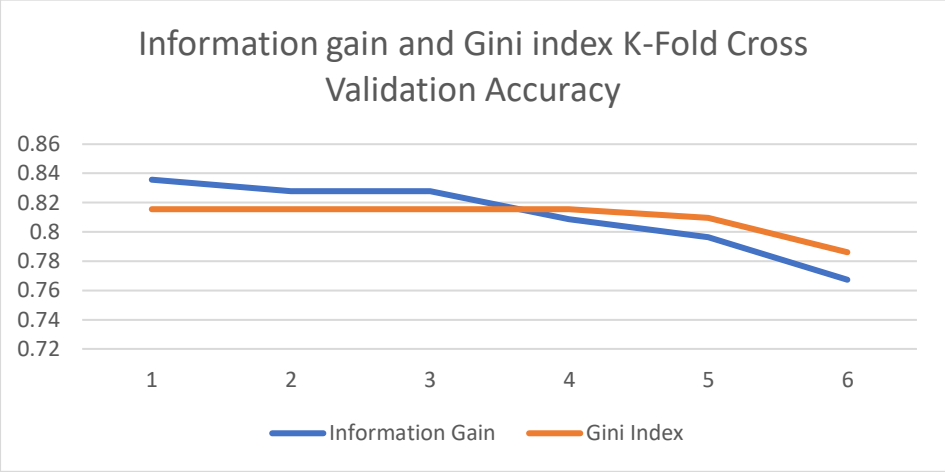


Fig. 7: ROC curve for the models.

Analysis Tasks

a. **Compare gini and information gain index:** The following plot shows the accuracy for each iteration of K-fold cross validation:

From observation we can see that the accuracy is within a very close threshold for both.
Mean for K-fold cross validation of IG index is 0.81058825 and gini index is 0.80966038.
Using the same splitted training and testing data, obtained accuracy for IG index is 0.8 and gini index is 0.84.



b. **Compare Naïve Bayes and Information Gain:** For information gain, the mean of K-Fold cross validation is 0.81, and mean for K-fold cross validation of Naïve Bayes is 0.96. The accuracy for testing data of split for Information gain is 0.83 and same for Naïve Bayes is 0.82. From observation it seems that the accuracy for split data for Information gain index and Naïve Bayes is very similar, but for K-fold cross validation mean, Naïve Bayes model performs better.

c. **Compare Gini Index with Naïve Bayes:** For gini index, the mean of K-Fold cross validation is 0.80, and the mean for K- fold CV for Naïve Bayes is 0.96. The accuracy for testing data for Gini index is 0.84, and Naïve Bayes is 0.82. Same as before, the cross validation mean accuracy for Naïve Bayes is better than Gini index.

d. **Compare the Split of each with Corresponding K-Fold:** The following table compares the result of each split with corresponding K-Fold:

Classification Technique	Accuracy of Split	Mean Accuracy of K-fold
Decision Tree(Information Gain)	0.838	0.81058825
Decision Tree(Gini index)	0.841	0.80966038
Naïve Bayes	0.8254	0.9682703

e. **Compare Result of dropped and non-dropped results:** The following table contains split and mean accuracy for dropped and non-dropped results along with dropped column:

Classification Technique	Column Dropped	Dropped Split Accuracy	Dropped Mean K-Fold Accuracy	Non Dropped Split Accuracy	Non Dropped Mean K-Fold accuracy
Decision Tree(Information Gain)	Column 11(capital-gain)	0.8206	0.80966038	0.838	0.81058825

Decision Tree(Gini Index)	Column 11 (capital-gain)	0.82	0.811956117	0.841	0.80966038
Naïve Bayes	Column 8 (relationship)	0.8254	0.9723483	0.8254	0.9682703

From the above table data, we can surmise that there is not much difference in split or mean accuracy for dropping the columns.

Overall Status

Finished the project in entirety by completing the mentioned task. Started the project by getting familiar with R, R studio, Decision tree, and Naïve Bayes for classification. Then performed the data analysis tasks. After that, played around with a small sample size of test data to better understand the model parameters. Finally, collected mentioned data in the project description for the entire dataset.

File Descriptions

Apart from the project report file (this file), two additional files are attached.
HW1_src.R: This R script file contains the source code of the project.
Data_Sheet.xlx: Contains the collected data and some plot used for analysis.

Division of Labor

The entire project was done alone.

Challenges Encountered

1. Obtaining a balanced and small sample of the dataset. As the entire dataset was too large, a small sample dataset was needed for analysis and studying the parameters. For this purpose, a small subset of dataset was extracted using random sampling.
2. Getting rid of outliers. There was a special case of outlier data(only a single tuple) which was present in the testing set, but no similar tuple was present in the training set. This particular column value was of factor type, hence causing error in evaluation of testing data. Removed the tuple from the data-frame.
3. Plotting the ROC curve. Tried to do this manually and failed. Finally used a built it library of R.