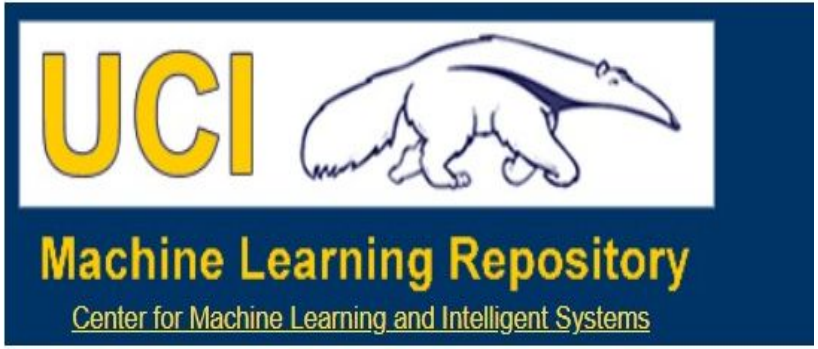
Classification of Census Data using Naive Bayes Trees

Shaurya (sg4040)
Sunjana (sc4921)
Manisha(mr4136)
Marc (mav2179)

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

DATA SET



From: UCI Machine Learning Repository

Description: Multivariate

Variety: Categorical, Integer

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	2341665

Paper/Study

- **Scaling up the accuracy of Naive Bayes classifiers: A decision-tree hybrid. (R. Kohavi 2011)**
- Bayesian classifier is derived from pattern recognition research. Easy to understand.
- Each class saves a probabilistic summary, including the conditional likelihood of each attribute value given the class, as well as the class's probability.
- The data structure approximates a perceptron’s representational power by describing a single decision boundary in an instance space.
- Probabilities recorded with the chosen class are updated when the algorithm encounters a new instance. This method is unaffected by the order of training cases or the existence of classification errors.
- Classifier utilizes an evaluation function when given a test instance to rank alternative classes based on their probabilistic summaries and assigns the instance to the highest scoring class.

Introduction

- Obtained the data set from UCI Machine Learning Repository
- Contains a variety of numerical as well as categorical variables as follows:

  Age: Continuous

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

  Fnlwgt: continuous

  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

  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-inspct, 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-loss: continuous.

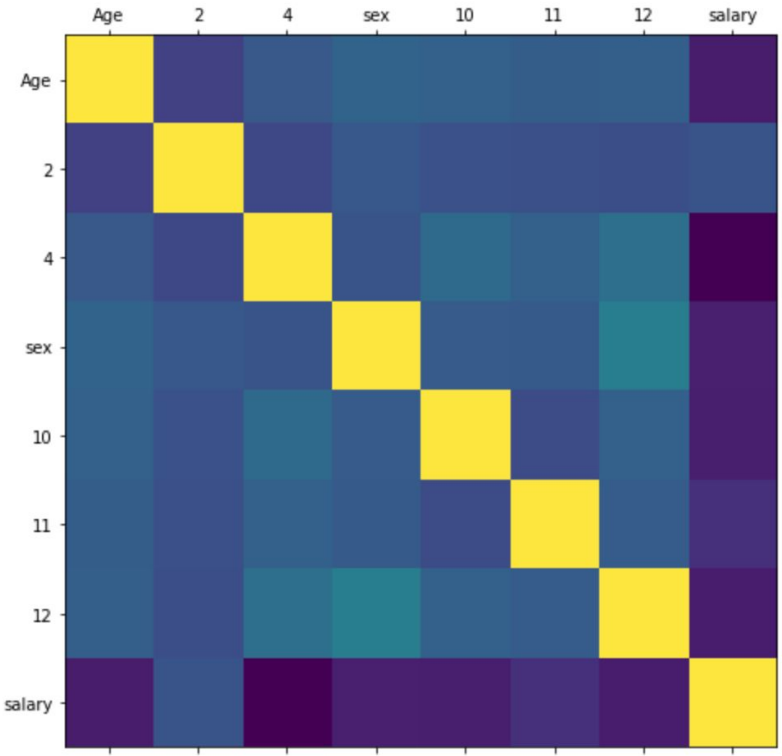
hours-per-week: continuous.

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, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

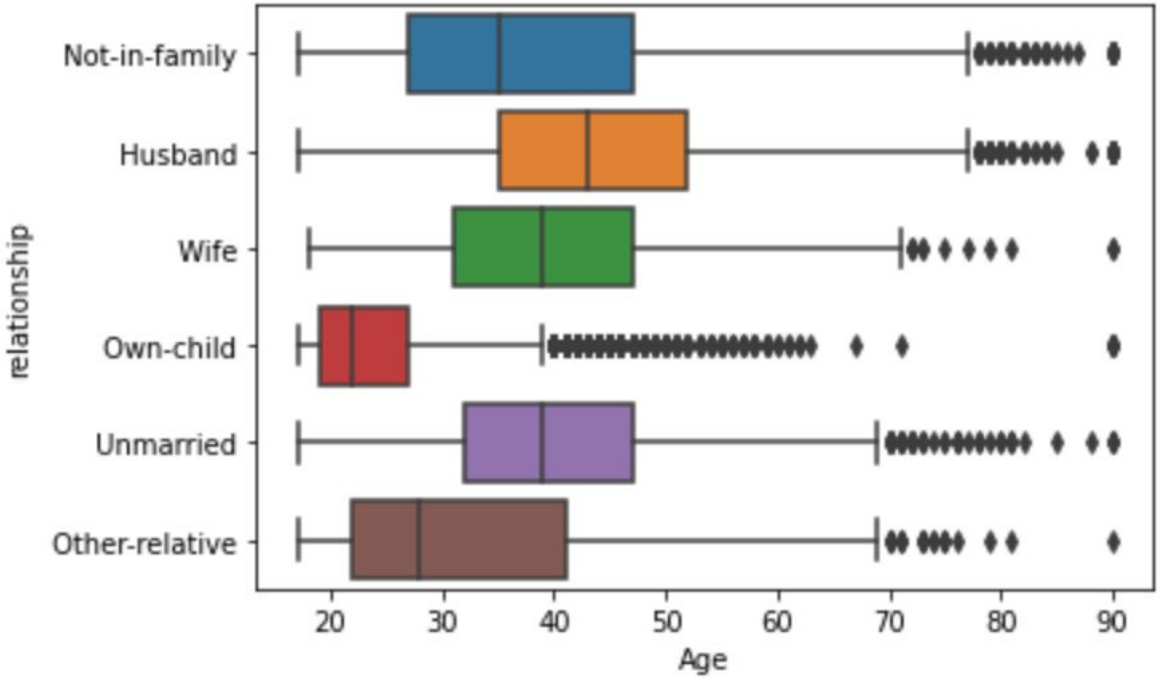
Sneak peek into the cleaned dataset using python

index	Age	workclass	education	marital-status	occupation	relationship	race	sex		
0	39	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	United-States	<=50K
1	50	Self-emp-not-inc	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	<=50K
2	38	Private	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	United-States	<=50K
3	53	Private	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	United-States	<=50K
4	28	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	Cuba	<=50K
5	37	Private	Masters	Married-civ-spouse	Exec-managerial	Wife	White	Female	United-States	<=50K
6	49	Private	9th	Married-spouse-absent	Other-service	Not-in-family	Black	Female	Jamaica	<=50K
7	52	Self-emp-not-inc	HS-grad	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	>50K
8	31	Private	Masters	Never-married	Prof-specialty	Not-in-family	White	Female	United-States	>50K
9	42	Private	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States	>50K
10	37	Private	Some-college	Married-civ-spouse	Exec-managerial	Husband	Black	Male	United-States	>50K
11	30	State-gov	Bachelors	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	India	>50K
12	23	Private	Bachelors	Never-married	Adm-clerical	Own-child	White	Female	United-States	<=50K
13	32	Private	Assoc-acdm	Never-married	Sales	Not-in-family	Black	Male	United-States	<=50K
14	40	Private	Assoc-voc	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	?	>50K
15	34	Private	7th-8th	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	Mexico	<=50K
16	25	Self-emp-not-inc	HS-grad	Never-married	Farming-fishing	Own-child	White	Male	United-States	<=50K
17	32	Private	HS-grad	Never-married	Machine-op-inspct	Unmarried	White	Male	United-States	<=50K
18	38	Private	11th	Married-civ-spouse	Sales	Husband	White	Male	United-States	<=50K
19	43	Self-emp-not-inc	Masters	Divorced	Exec-managerial	Unmarried	White	Female	United-States	>50K
20	40	Private	Doctorate	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States	>50K
21	54	Private	HS-grad	Separated	Other-service	Unmarried	Black	Female	United-States	<=50K
22	35	Federal-gov	9th	Married-civ-spouse	Farming-fishing	Husband	Black	Male	United-States	<=50K
23	43	Private	11th	Married-civ-spouse	Transport-moving	Husband	White	Male	United-States	<=50K
24	59	Private	HS-grad	Divorced	Tech-support	Unmarried	White	Female	United-States	<=50K

Heat Map Analysis depicting Correlation Matrix



Boxplot between Age and Relationship Status



Goal

- Main objective: Replicate the paper based on Naive Bayes Trees (NB Trees) (**Scaling up the accuracy of Naive Bayes classifiers: A decision-tree hybrid. (R. Kohavi 2011)**
- Apply Machine Learning techniques learnt in class to extend the scope of the aforementioned study
- Compare and contrast the accuracy scores obtained from applying the following ML algorithms:
    - Naive Bayes Trees (NB Trees)
    - Support Vector Machines (SVM)
    - Principal Component Analysis (PCA)

Related work done in this direction

- **Naive-Bayes classifiers (Langley, Iba, & Thompson 1992) are generally easy to understand and the induction of these classifiers is extremely fast.**

- **An empirical study of the naive Bayes classifier. (I. Rish 2000)**

- **Elements of information theory. (T.M. Cover and J.A. Thomas 1991)**

- **On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. (P. Domingos and M. Pazzani 1997)**

- **Recognizing end-user transactions in performance management. (J. Hellerstein, Jayram Thathachar, and I. Rish 2000)**

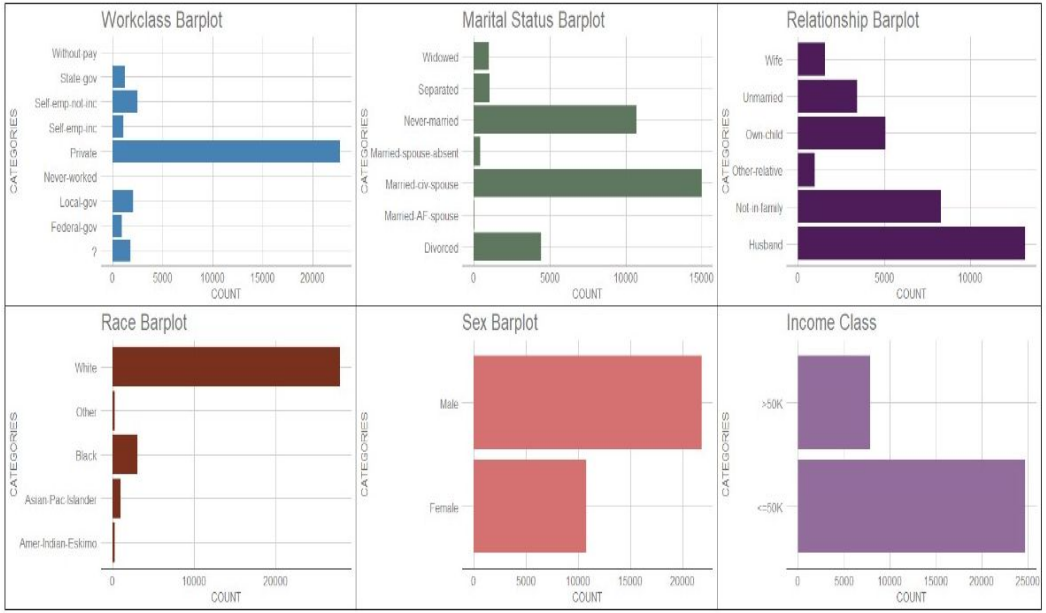
Data Preprocessing

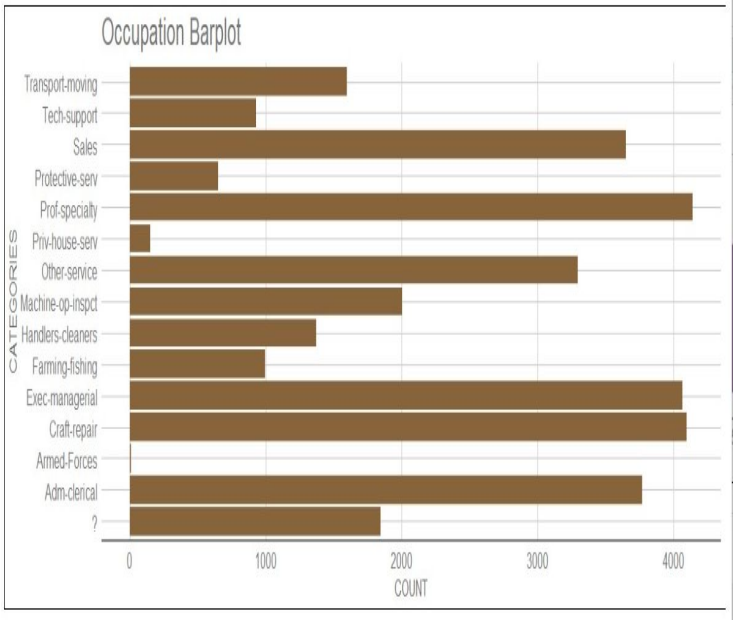
- The dimensions of the original dataset is obtained.
- The original dataset that is downloaded from the repository does not contain column names. To make the data more understandable, column names are embedded to the data. On doing so, the data appears in a more readable format.

Number of Rows	32561
Number of Columns	15

```
'data.frame':   32561 obs. of  15 variables:
 $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
 $ workclass     : chr  "State-gov" "Self-emp-not-inc" "Private" "Private" ...
 $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
 $ education     : chr  "Bachelors" "Bachelors" "HS-grad" "11th" ...
 $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
 $ marital_status: chr  "Never-married" "Married-civ-spouse" "Divorced" "Married-civ-spouse" ...
 $ occupation    : chr  "Adm-clerical" "Exec-managerial" "Handlers-cleaners" "Handlers-cleaners" ...
 $ relationship  : chr  "Not-in-family" "Husband" "Not-in-family" "Husband" ...
 $ race          : chr  "White" "White" "White" "Black" ...
 $ sex           : chr  "Male" "Male" "Male" "Male" ...
 $ 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: chr  "United-States" "United-States" "United-States" "United-States" ...
 $ income_class  : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
```

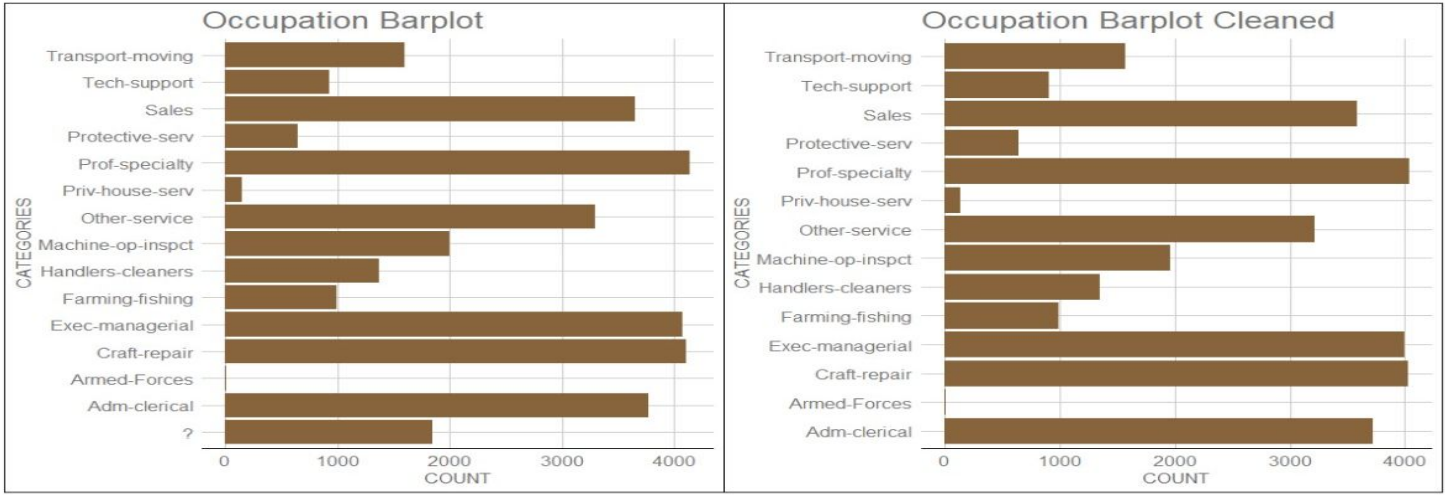
Data Cleaning : Categorical Variables





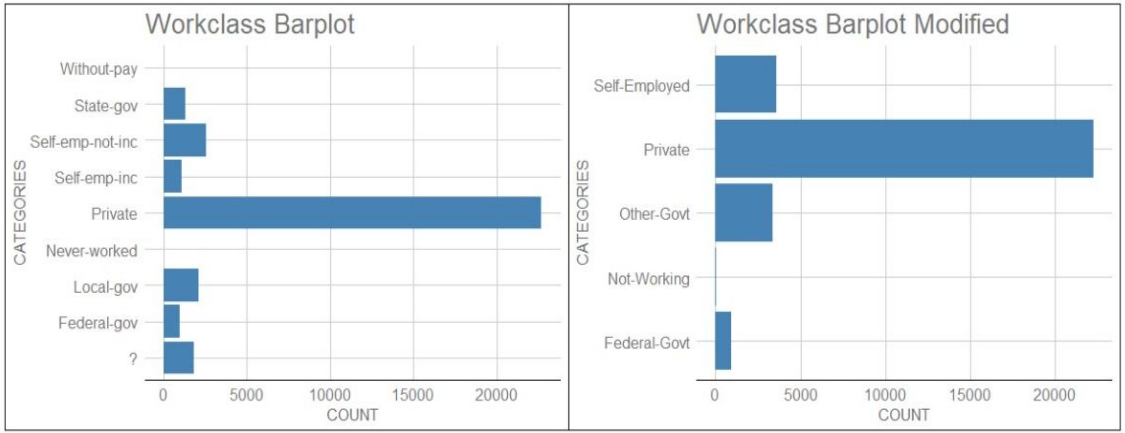
Data Cleaning : Categorical Variables

- A few Missing values and “?” values exist

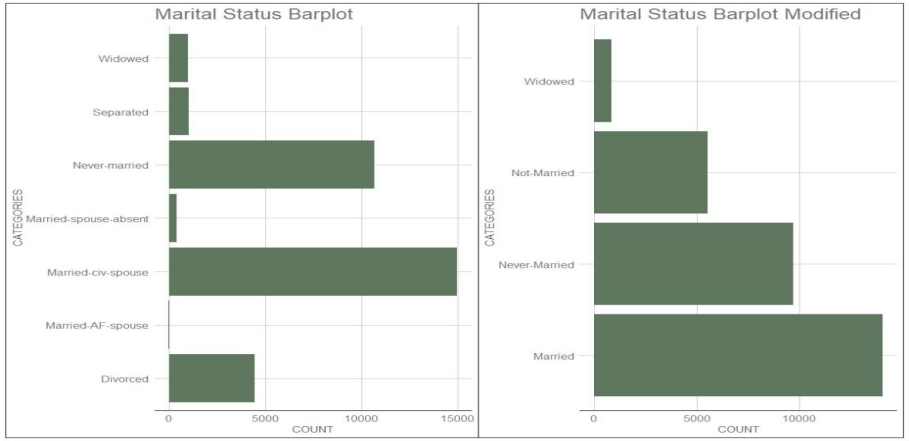


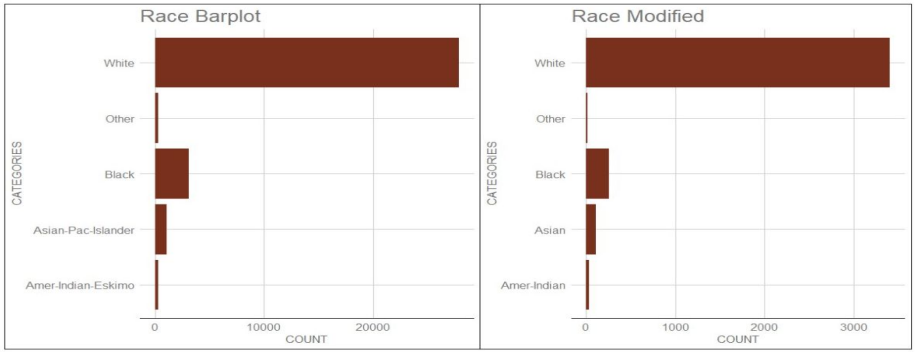
Data Cleaning : Categorical Variables

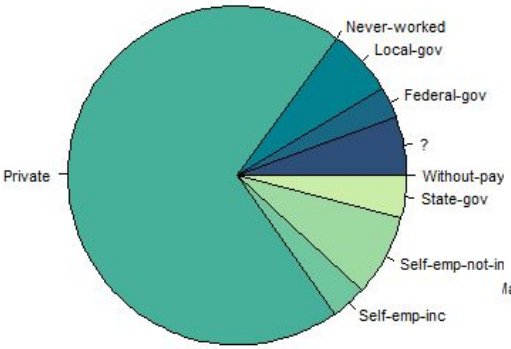
- There are multiple subcategories within categorical variables. This might lead to reduction inaccuracy of the models and might reduce the overall performance of the model that shall be used to conduct analysis.

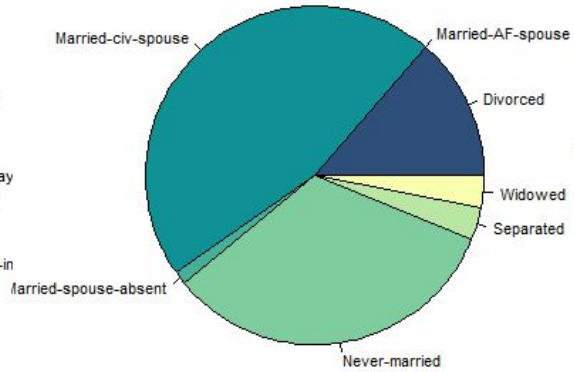


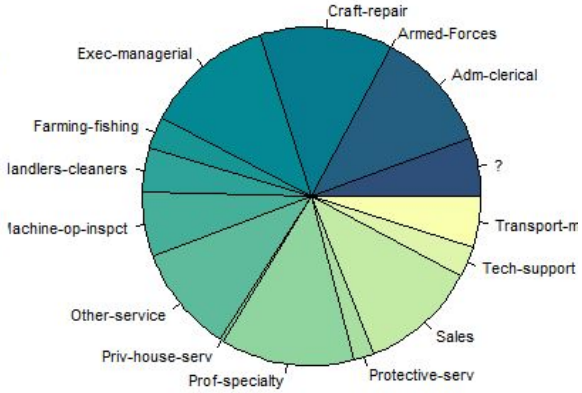
Data Cleaning : Categorical Variables

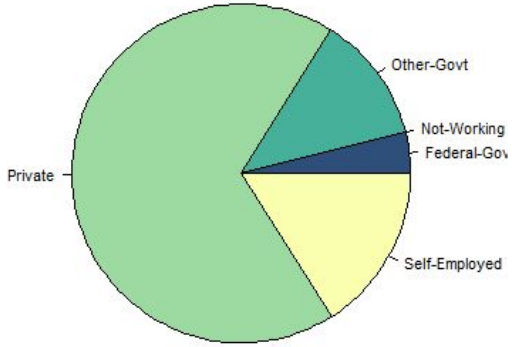


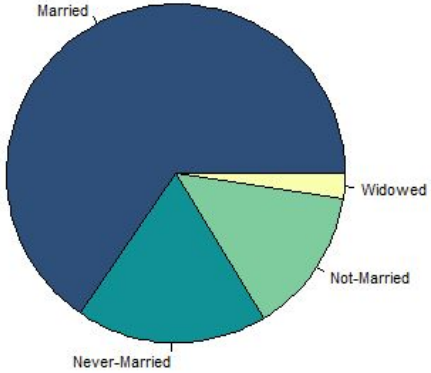


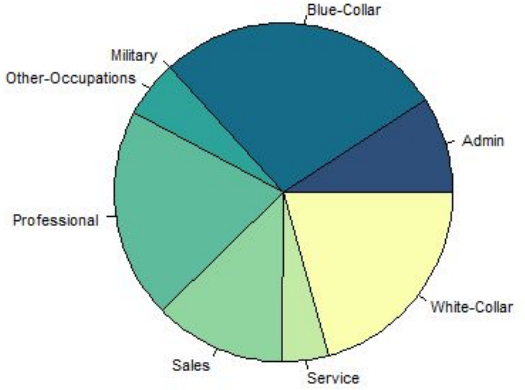

Never-worked
Local-gov
Federal-gov
?
Private
Without-pay
State-gov
Self-emp-not-in
Self-emp-inc


Married-civ-spouse
Married-AF-spouse
Divorced
Widowed
Separated
Married-spouse-absent
Never-married

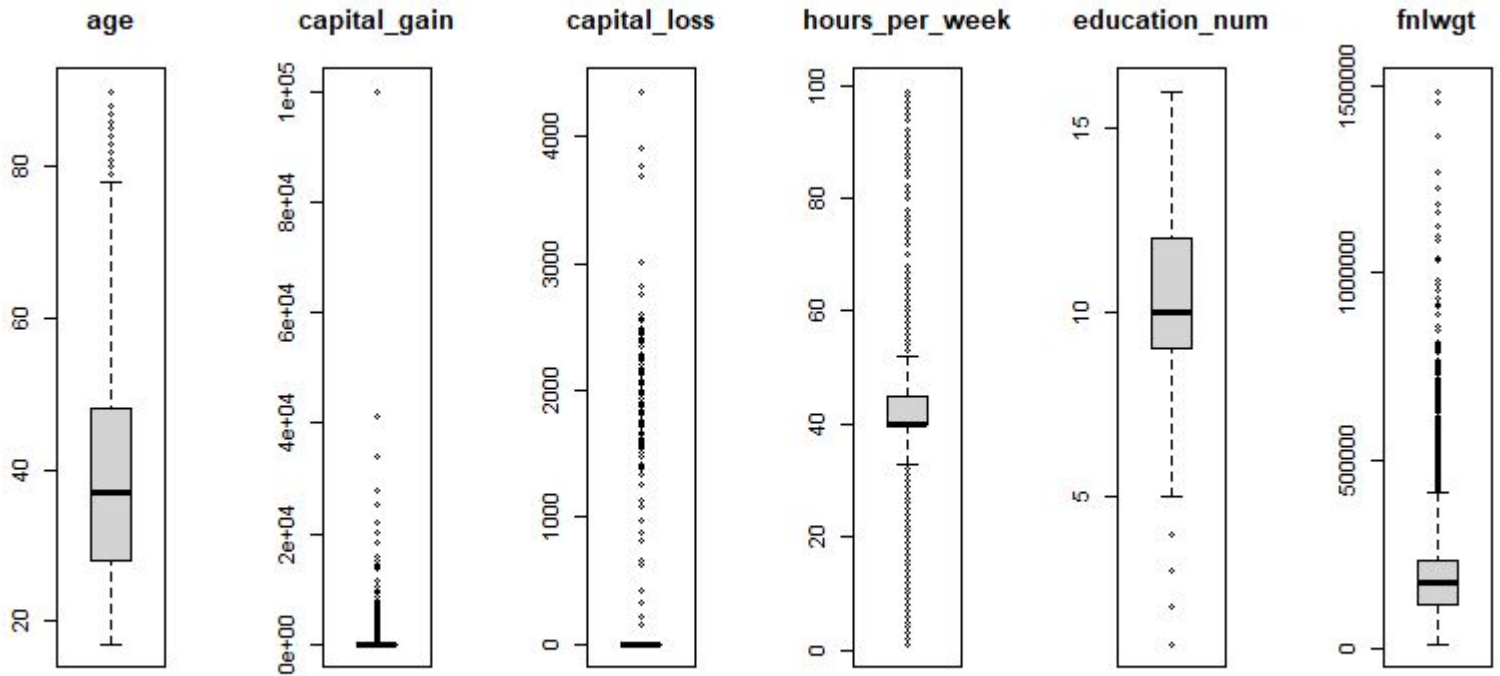

Craft-repair
Armed-Forces
Exec-managerial
Adm-clerical
Farming-fishing
?
landlers-cleaners
Transport-m
lachine-op-inspct
Tech-support
Other-service
Sales
Priv-house-serv
Protective-serv
Prof-specialty


Other-Govt
Not-Working
Private
Federal-Gov
Self-Employed

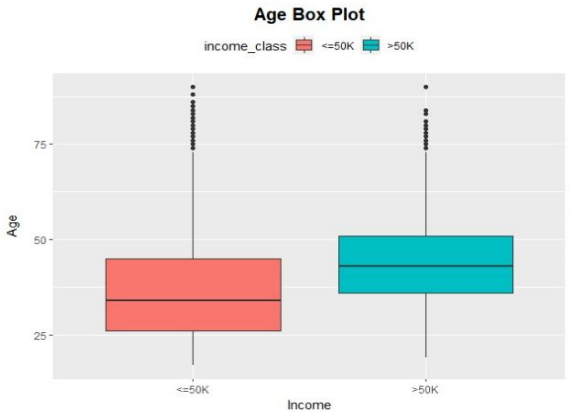

Married
Widowed
Not-Married
Never-Married

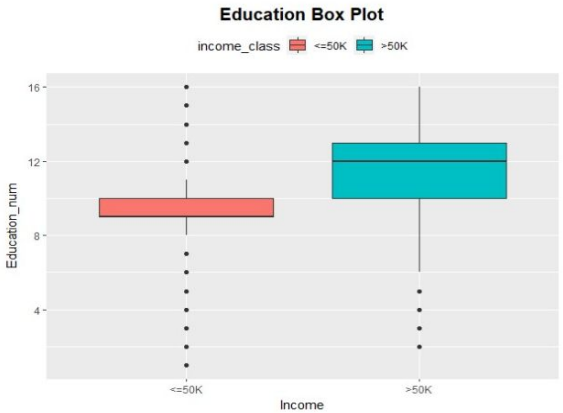

Blue-Collar
Military
Other-Occupations
Admin
Professional
White-Collar
Sales
Service

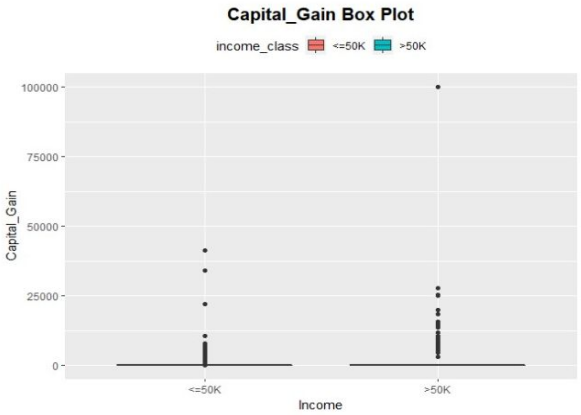
Data Cleaning : Continuous Variables

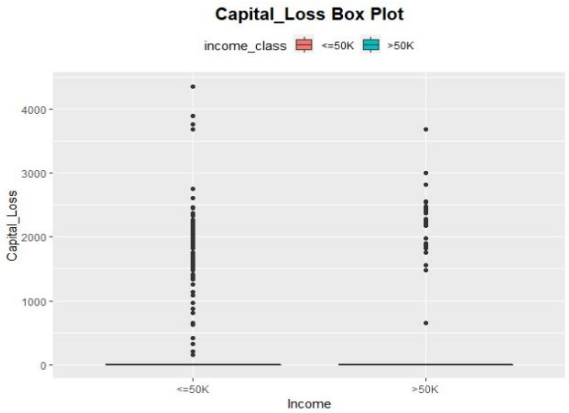


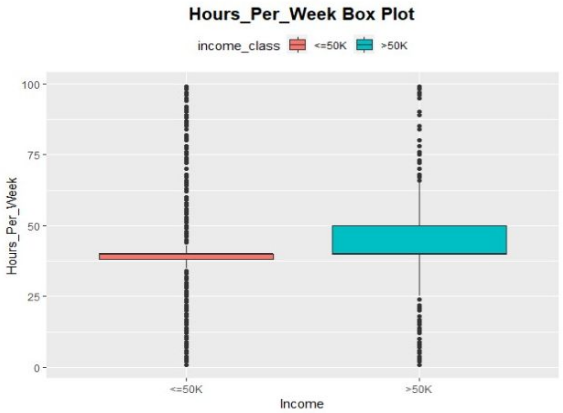
Data Cleaning : Continuous Variables



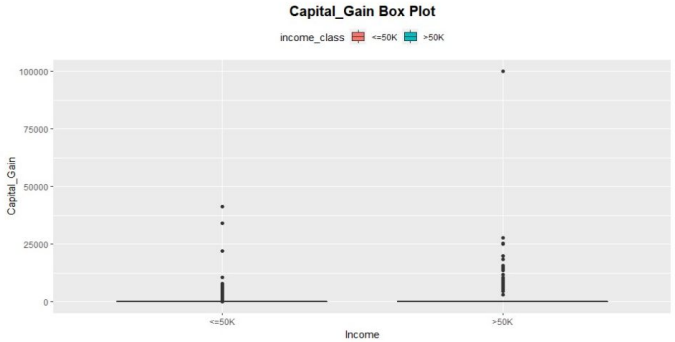


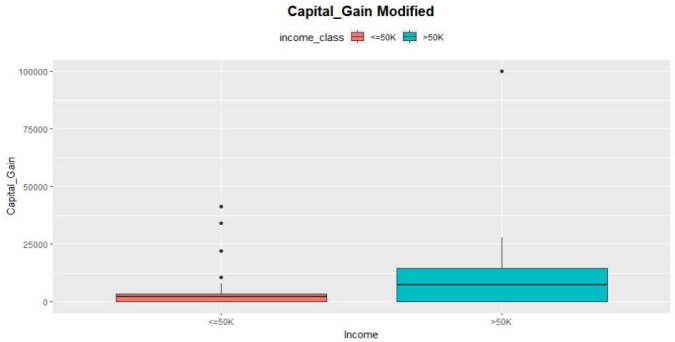


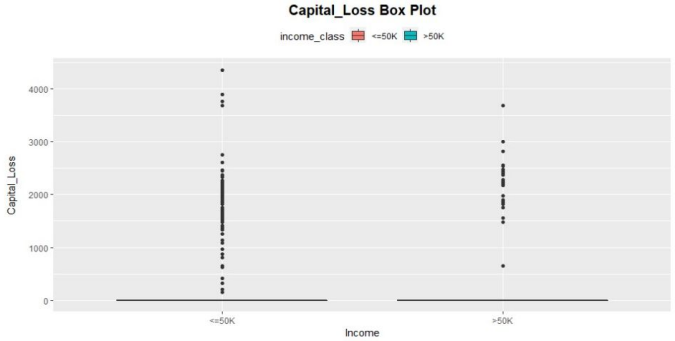




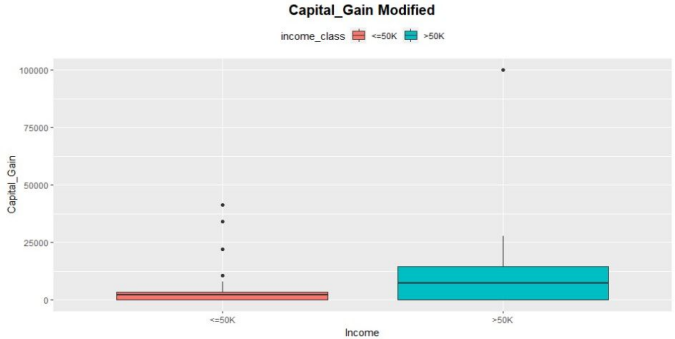
Data Cleaning : Continuous Variables

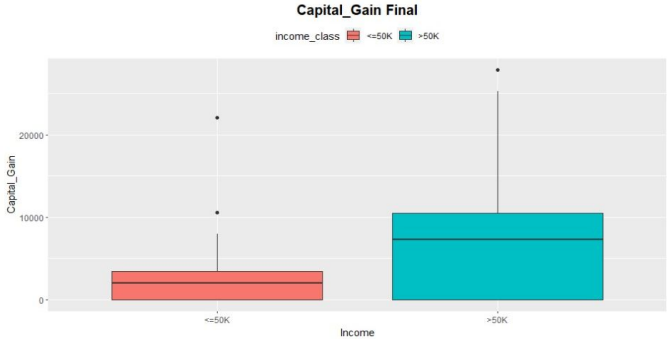


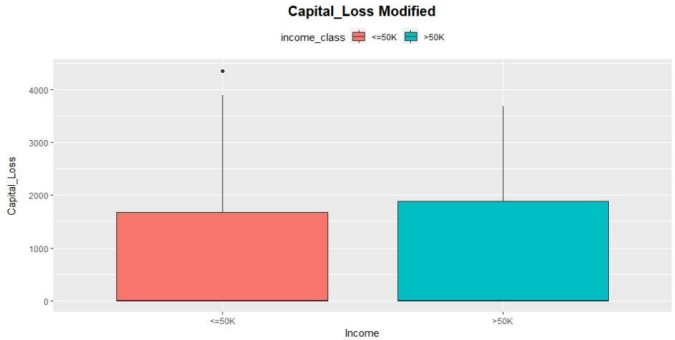


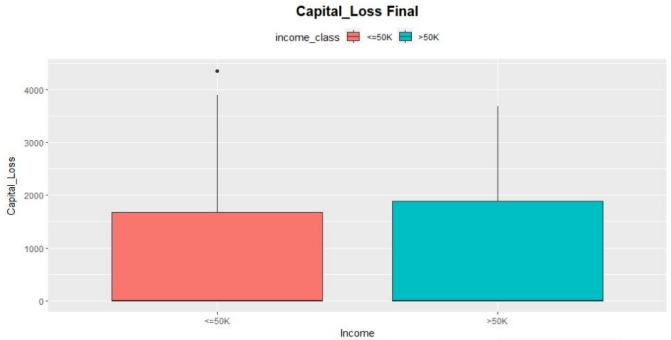


Data Cleaning : Continuous Variables

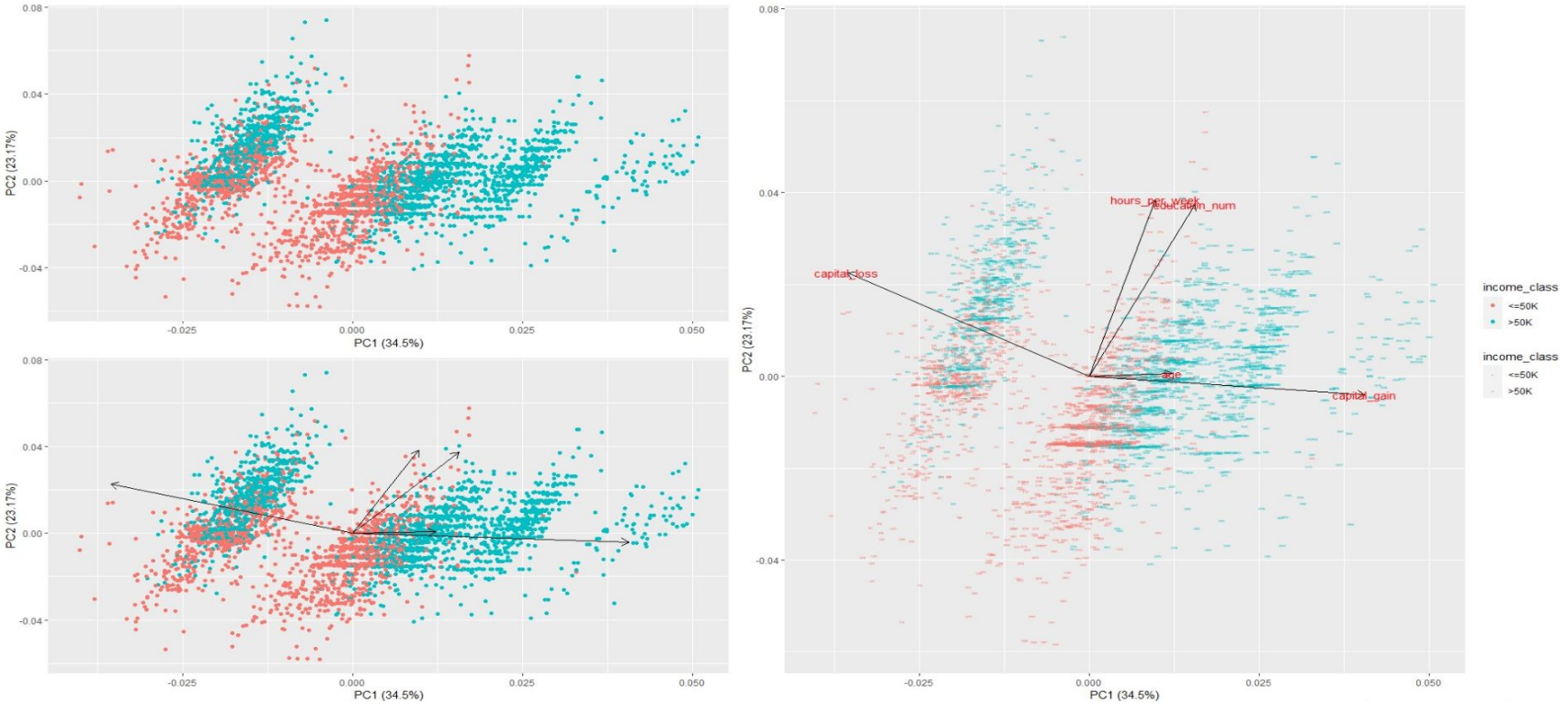




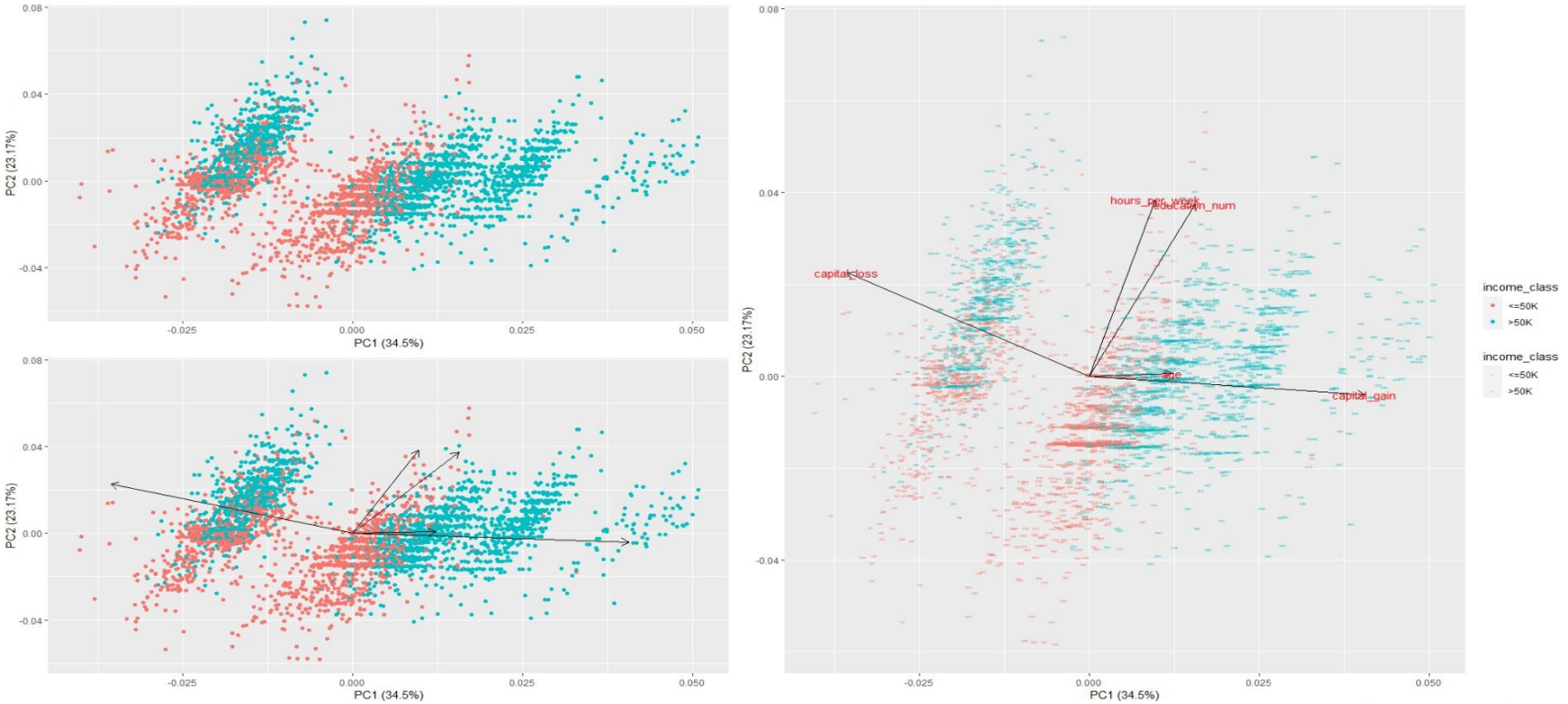




EDA : Principal Component Analysis



EDA : Principal Component Analysis



EDA : Principal Component Analysis

From the eigenvector distribution, the following can be derived:

- Age and Capital Gain have strong correlation.

It goes without saying that as an individual becomes older, it is more likely that the person's capital_gain increases.

- Education and Hours_Per_Week have strong correlation.

This is a significant observation. It means that when a person is well educated, he/she tends to spend more hours to work.

- Capital_loss has weak correlation with Age and Capital_Gain.

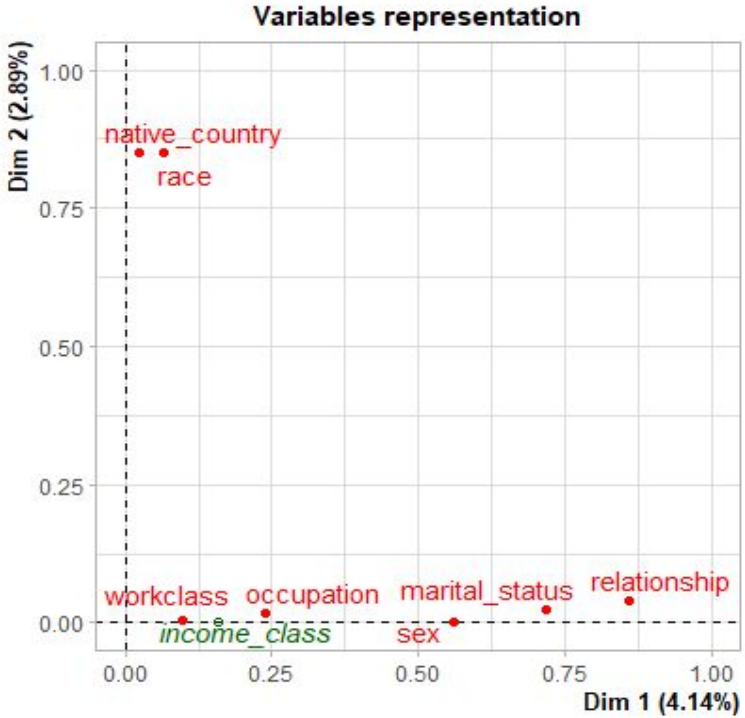
This means that as a person tends to grow older, Capital_loss is less likely to occur for him.

- Capital_loss has zero or no correlation with Hours_per_week and Education.

This means that when a person is well educated and puts in a greater number of hours to work, he/she is less likely to lose Capital.

From the biplot, it can be concluded that when Age, hours_per_week,education level and capital_gain are high, the income level of the individual is high.

EDA : Multiple Correspondence Analysis

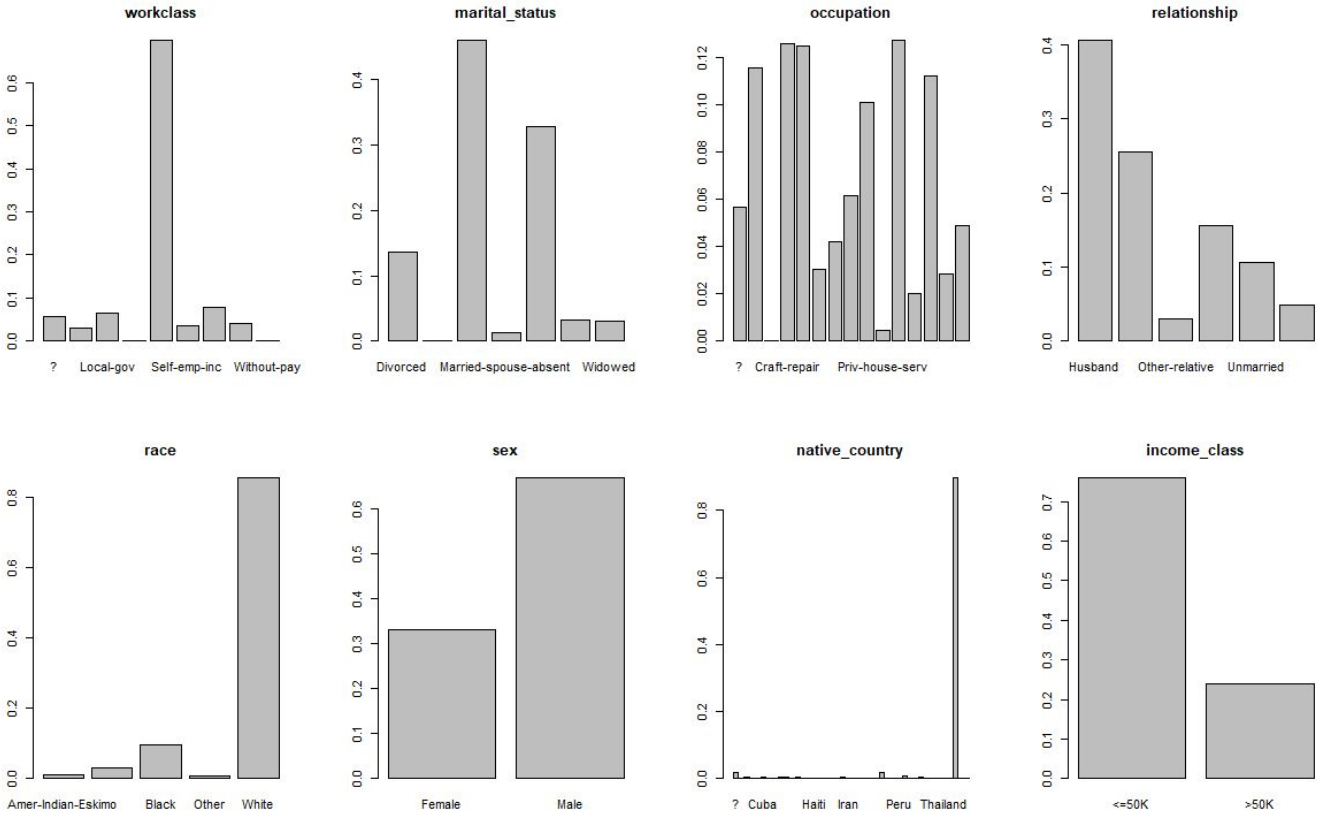


Multiple Correspondence Analysis or MCA is performed on the categorical variable to filter the components that mostly affect the income_class.

The MCA plot of the categorical variables is shown in Fig9. It can be seen that Income_class lies on the first dimension. This leads us to conclude that only the variables on Dimension1 have an effect on the income class.

Therefore, the most important variables are occupation, sex, marital_status and relationship.

Exploratory Data Analysis (EDA) - Box Plots of Categorical Variables



Previous results from NBTree Paper

Train/Test split: 66%/34%

Probability for the label '>50K' : 23.93% / 24.78% (without unknowns)

Probability for the label '<=50K' : 76.07% / 75.22% (without unknowns)

Decision trees: 84.46 - 85.54%

NBTree: 85.9%

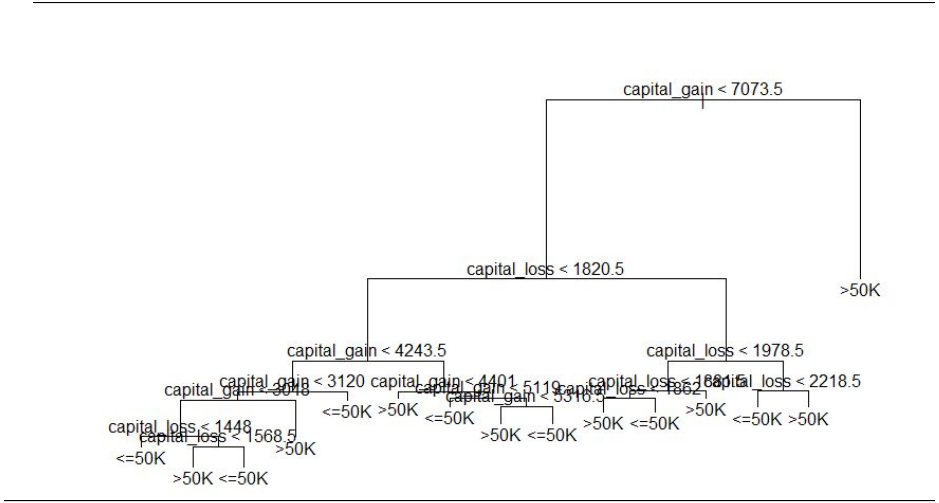
Naive-Bayes: 83.88%

Tree Based Methods

Decision Tree
Test Data

Predictions	<= 50K	> 50K
<= 50K	446	42
> 50K	19	636

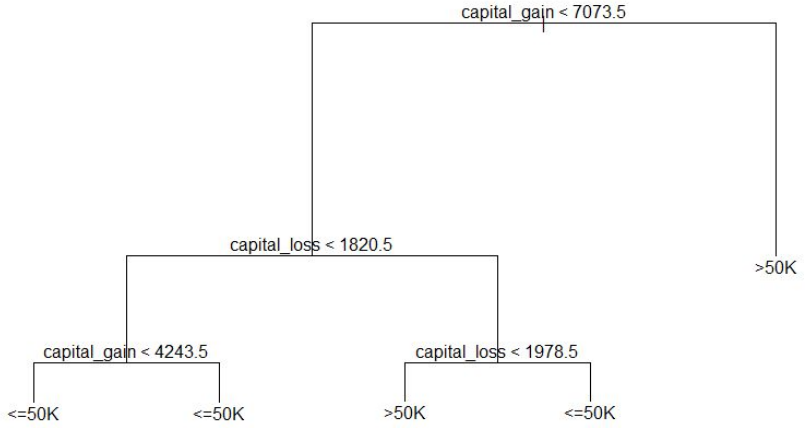
Accuracy: 94.46%



Pruned Decision Tree
Test Data

Predictions	<= 50K	> 50K
<= 50K	460	28
> 50K	148	507

Accuracy: 84.46%



Bayesian Methods/NBTree

Naive-Bayes
Test Data

Predictions	<= 50K	> 50K
<= 50K	9824	122
> 50K	1322	3384

Accuracy: 90.14%

Tree Augmented Naive-Bayes
Test Data

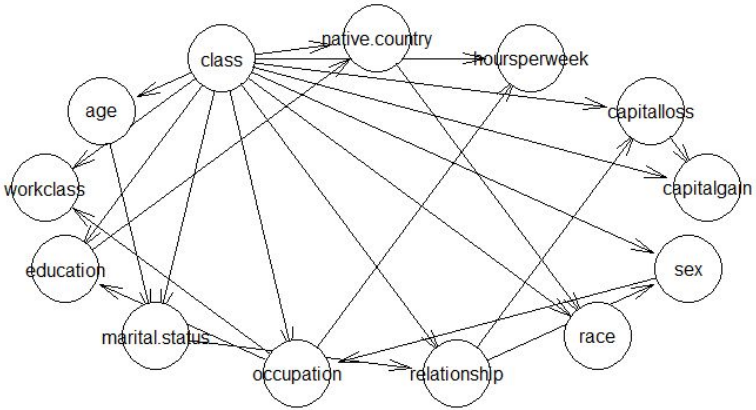
Predictions	<= 50K	> 50K
<= 50K	10228	1269
> 50K	918	2237

Accuracy: 85.07%

NBTree
Test Data

Predictions	<= 50K	> 50K
<= 50K	10133	1200
> 50K	1013	2306

Accuracy: 84.90%



SVM Model

Support Vector Machine is responsible for finding the decision boundary to separate different classes and maximize the margin.
Three different kernels were used to build SVMs

- Linear
- Quadratic
- RBF (Radial Basis Function)

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2})$$

Linear

Predictions	<= 50K	> 50K
<= 50K	341	61
> 50K	64	487

Accuracy: 86.88%

Quadratic

Predictions	<= 50K	> 50K
<= 50K	225	177
> 50K	230	321

Accuracy: 57.29%

RBF

Predictions	<= 50K	> 50K
<= 50K	386	16
> 50K	36	515

Accuracy: 94.54%

Neural Network Model

Used the nnet package to create a simple feed forward neural network.

- size - defines the number of neurons at the input (10 in our case)
- decay - how quickly the gradient decreases in Gradient Descent
- maxit - maximum number of iterations

We used a maxit of 2000 and decay of 0.01

We use a single hidden layer Neural Network model

Accuracy: 88.56%

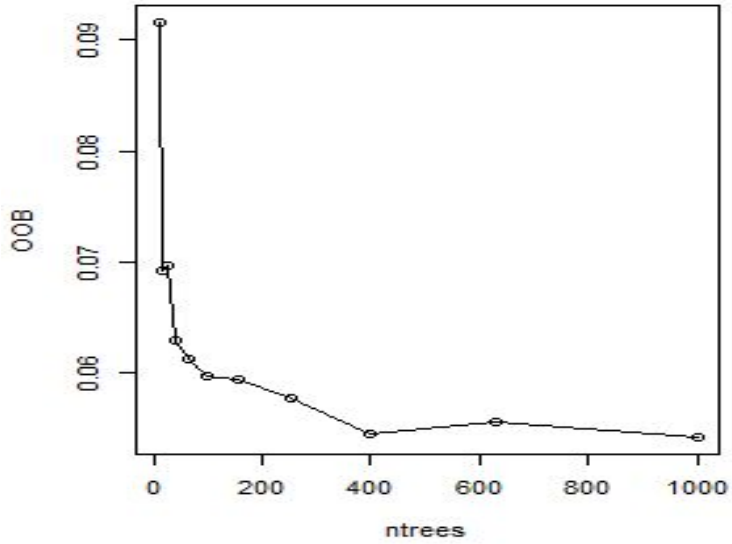
Random Forest Model

Random Forests are developed by an aggregating tree.

It has 2 parameters:

1. mtry - square root of exploratory variables
2. ntree - chosen based on OOB error

Accuracy: 95.67%



Conclusion

Models	Prediction Accuracy
SVM - Linear Kernel	86.88%
SVM - Quadratic Kernel	57.29%
SVM - RBF Kernel	94.54%
Neural Network	88.56%
Random Forest	**95.67%**
Decision Trees	94.46%
Pruned Decision Tree	84.46%
Naive-Bayes	90.14%
Tree Augmented Naive-Bayes	85.07%
Naive Bayes Tree (NBTree)	84.90%

Future Scope

- Explore more techniques to pre-process the data
- Try using more models or a combination of models to try and improve the classification accuracy further
- Fine tune the Random Forest to improve the accuracy
- Apply the models to a more recent dataset to see if the trends recorded in this dataset match with that of the current datasets.

THANK YOU!