Classification of Census Data Using Machine Learning Techniques

Shaurya Gaur (sg4040), Sunjana Ramana Chintala(sc4921), Manisha Rajkumar (mr4136), Marc Vazquez(mav21879)

EECS E6690: Statistical Learning for Biological and Information Systems

**Abstract**

***The objective of this report is to utilize an existing dataset and extend the scope of a previous study [1] by incorporating the machine learning techniques learnt in class. To begin with, we selected a dataset from the UCI Machine Learning Repository referred to as the Adult or the Census income dataset (***[***https://archive.ics.uci.edu/ml/datasets/adult***](https://archive.ics.uci.edu/ml/datasets/adult)***) and a paper based on Naive Bayes Trees [1].This dataset contains a number of categorical as well as numerical variables. We used these variables as a part of our exploratory as well as conclusive analysis to derive the final results. Using this dataset, we completed an initial pre-processing analysis and then applied techniques such as Principal Component Analysis (PCA), Naive Bayes Trees (NB Trees) and Support Vector Machines (SVM) to compute the accuracy of the model using each of these methods. As a next step, we compared the different accuracies that we obtained to determine the best approach for this model. We have compiled the results obtained from each of these methods in the Results section of this report***.

***Keywords: Naive Bayes Trees, Machine Learning, Principal Component Analysis***

**1. Introduction**

In this project, we will be analyzing the Adult data set extracted from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/adult>). This dataset contains the following variables:

| Variables | Type |
| --- | --- |
| 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. |

Table 1. Variables in the Adult/Census Income dataset

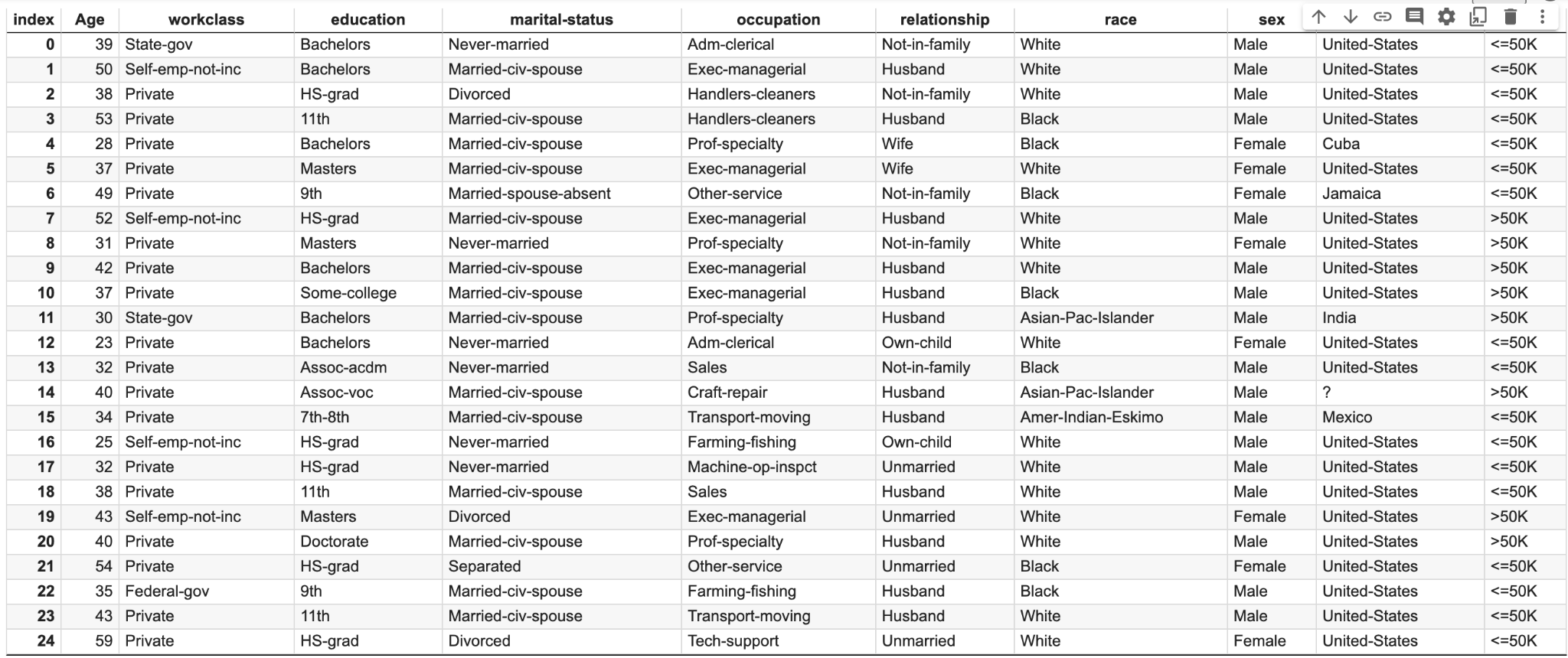


Fig. 1. This figure was generated using Python. It shows the different variables in the adult dataset. We have only selected certain variables which can be utilized to generate preliminary graphs and results to understand the basics of the dataset.

After performing some preliminary analysis on the dataset, we discovered the following results:

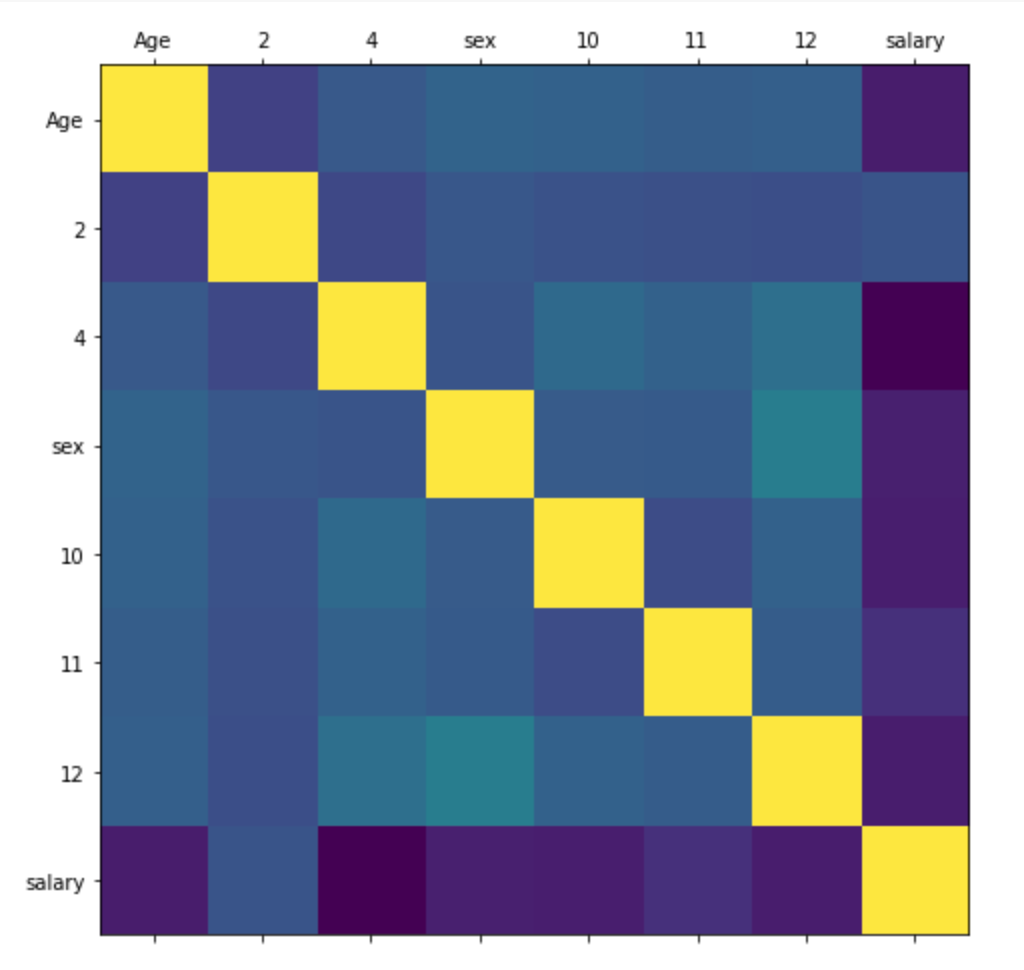


Fig. 2(a). The above figure shows the correlation between the different variables. In this above heatmap, we have explored the relationship between the following variables: Age, salary and, sex.

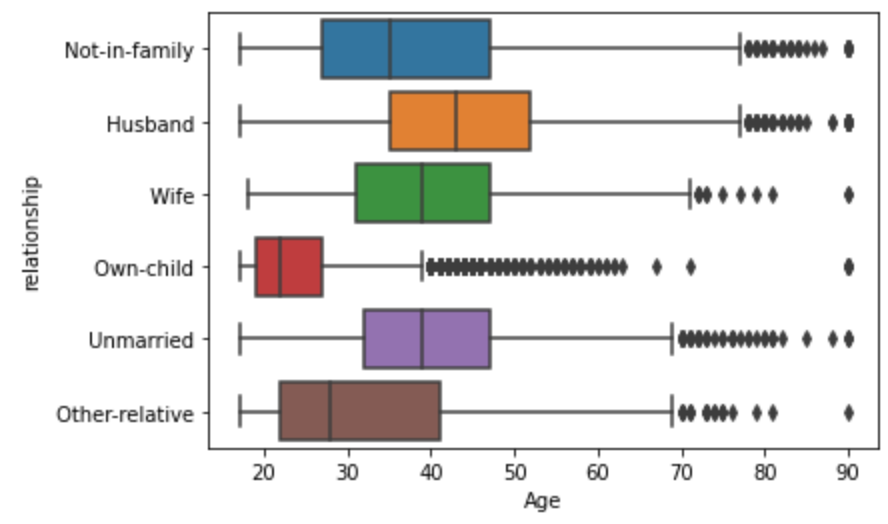


Fig. 2(b). This figure depicts the relationship between different age groups and their relationship status. We can see that the different relationship statuses have different median age groups.

As we can see from the above graphs, there exists a relationship between the variables in the dataset. However, this is not clear from any of the above analysis and we might not be able to conclude a definitive relationship between all the variables. To understand this better, we will conduct an in-depth analysis of the different machine learning approaches and methods learnt in class. This will be divided into two parts:

1. Reproduce the results from the paper that uses the Adult data set.

2. Specifically, we intend to apply techniques such as Support Vector Machines (SVM), Naive Bayes Trees (NB Trees) and Random Forest classifiers on our data and compare the final results with the accuracies obtained by implementing each of them.

**2. Related Work**

Prior to starting the project, we did extensive research and literature review to understand the work that was done previously in this direction:

1. **Scaling up the accuracy of Naive Bayes classifiers: A decision-tree hybrid. (R. Kohavi 2011)** ([chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Fwww.aaai.org%2FPapers%2FKDD%2F1996%2FKDD96-033.pdf&clen=677945&chunk=true](about:blank))

A Bayesian classifier is derived from pattern recognition research. For each class, the approach saves a probabilistic summary, which includes the conditional likelihood of each attribute value given the class, as well as the class's probability. This data structure approximates the perceptron's representational power by describing a single decision boundary in an instance space. The 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. The 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.

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

(chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2F[www.cc.gatech.edu%2F~isbell%2Freading%2Fpapers%2FRish.pdf&clen=257508&chunk=true](about:blank))

This paper outlines the major advantages and disadvantages of naive bayes classifier. Bayesian classifiers assign the most likely class to a given example described by its feature vector. Learning such classifiers can be greatly simplified by assuming that features are independent given class, that is, where is a feature vector andis a class. Despite this unrealistic assumption, the resulting classifier known as naive Bayes is remarkably successful in practice and is used in many different applications from medical diagnostics to predictions.

It is outlined in this paper that naive Bayes classifier work well in these conditions:

1. Concepts without noise: In this paper, we only use binary feature to explore this condition.
2. Concepts with noise: Generally, concepts can be noisy, i.e. can have nondeterministic and thus a non-zero Bayes risk. A natural extension of the conditions of Theorem 1 to noisy concepts yields low-entropy, or “extreme”, probability distributions, having almost all the probability mass concentrated in one state. The independence assumption becomes more accurate with decreasing entropy which yields an asymptotically optimal performance of naive Bayes.

# 3. Data

## 3.1 Dataset

The Dataset used for this study is from the UCI Machine Learning Repository. It is named as the “ADULT” Dataset or sometimes also referred to as the “CENSUS INCOME” Data.

The dimensions of the original dataset are given as below:

| Number of Rows | 32561 |
| --- | --- |
| Number of Columns | 15 |

## 3.2 Data Preprocessing

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. To conduct further analysis, an understanding of the kind of variables existing in the dataset is necessary. A quick look at the variables present in the dataset can be seen below.

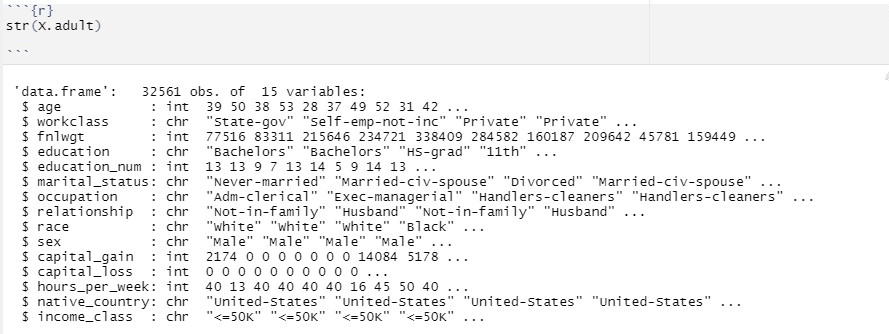


Fig 3. A glimpse into the data

The various characteristics of the dataset can be seen in Fig 3.1. From the figure it can be observed that the dataset has a mix of categorical and continuous variables.

## 3.3 Data Cleaning

The primary objective of data preprocessing is to prepare the dataset for further analysis. A good dataset is one that is free from errors and duplicates. To ensure that the dataset is ready for analysis, it is crucial to get rid of any missing or redundant values that exist within it.In the following sections, we will conduct analysis of the variable and eliminate any unnecessary datum that might cause trouble for pertinent data analysis.

### Categorical Variables:

From Fig 3.1, it can be observed that the dataset contains nine categorical variables. The following figures show the barplot for eight variables. The categorical variable “education” is not considered for this study. Instead the continuous variable “education\_num” is used for analysis.

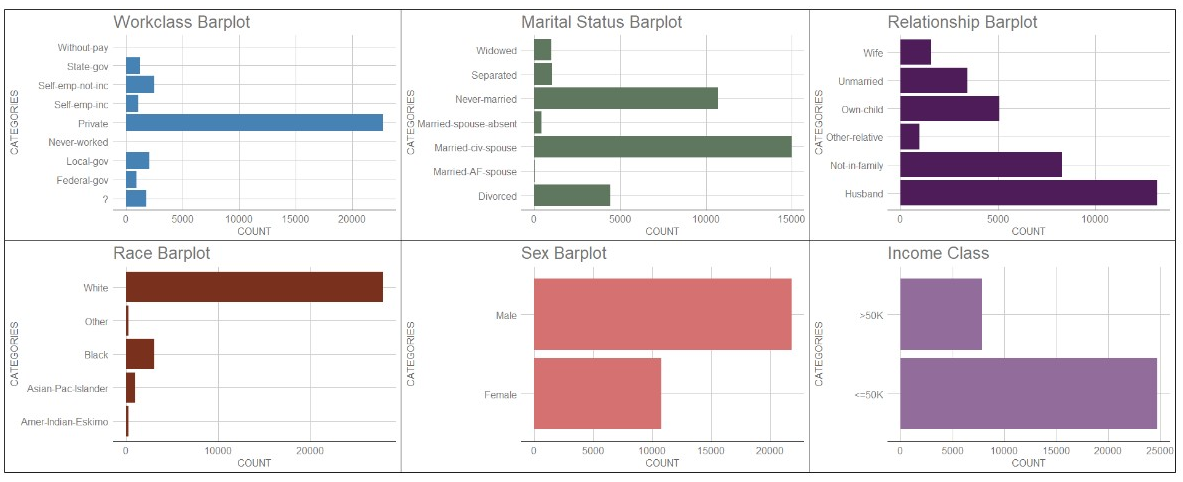


Fig 4. (a).

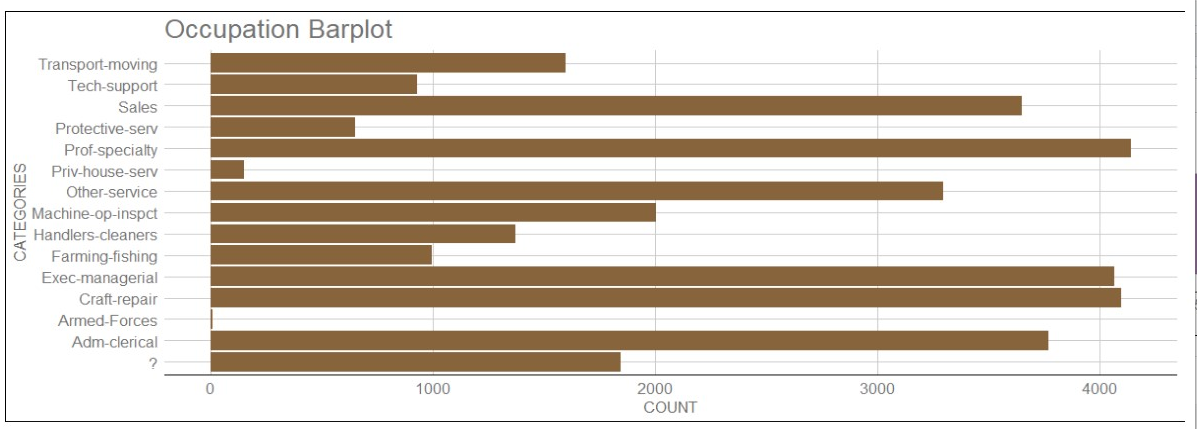


Fig. 4(b).

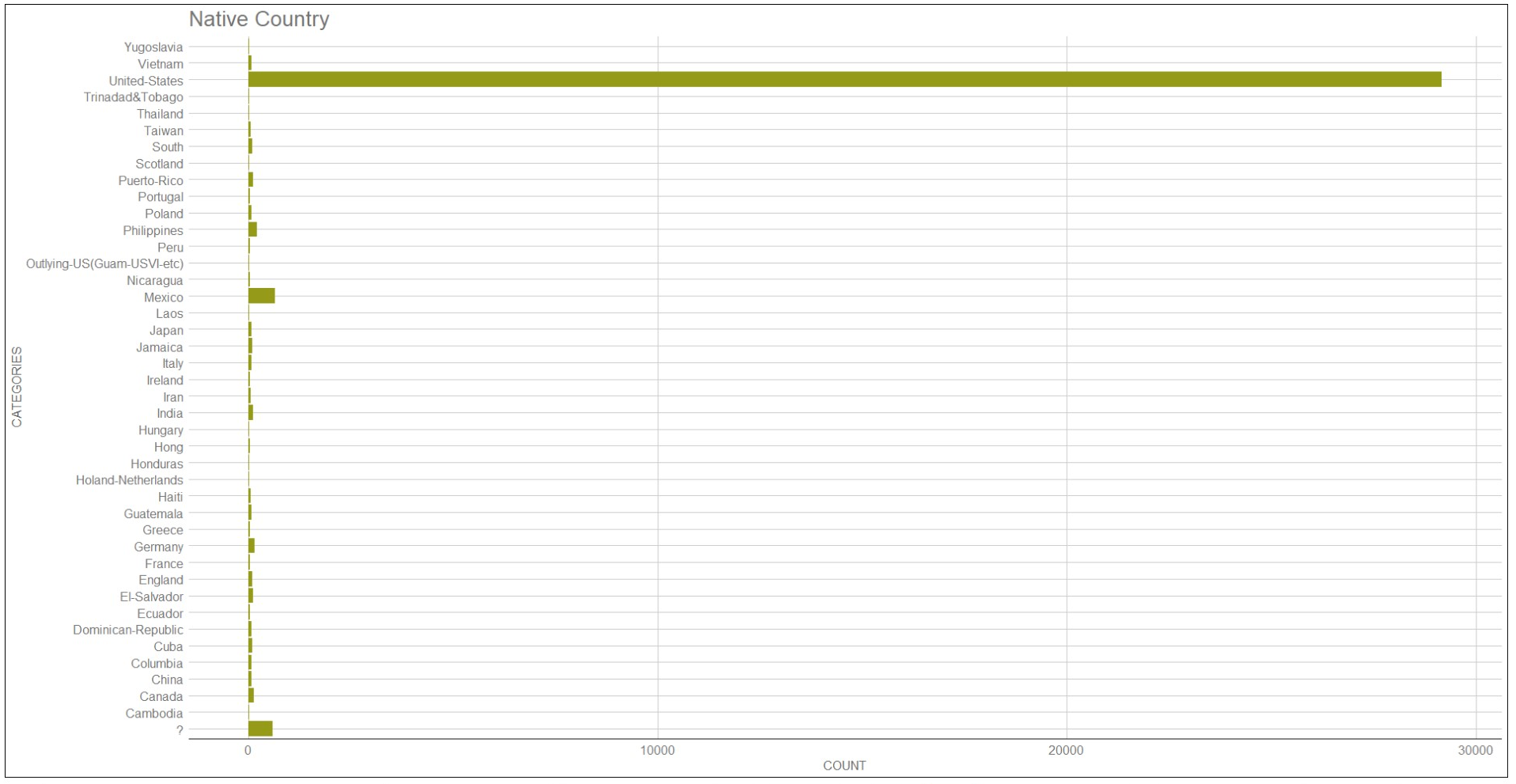


Fig. 4(c).

Some key observations from the bar plots for raw data are as follows:

* It can be observed that there are missing variable values in the dataset. Missing values dilute the efficacy of the dataset.
* It is also clear that 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.

These observations are essentially problems that need remedy. The first of these problems can be addressed by perusing the dataset for “?” or “NA” values. The second problem can be addressed by unifying similar subcategories and labelling these as a common subcategory.

The “?” values are mostly visible in workclass and occupation barplots. After removing the missing values along with “?” values, the comparison of the old and cleaned data can be observed in the following barplot.

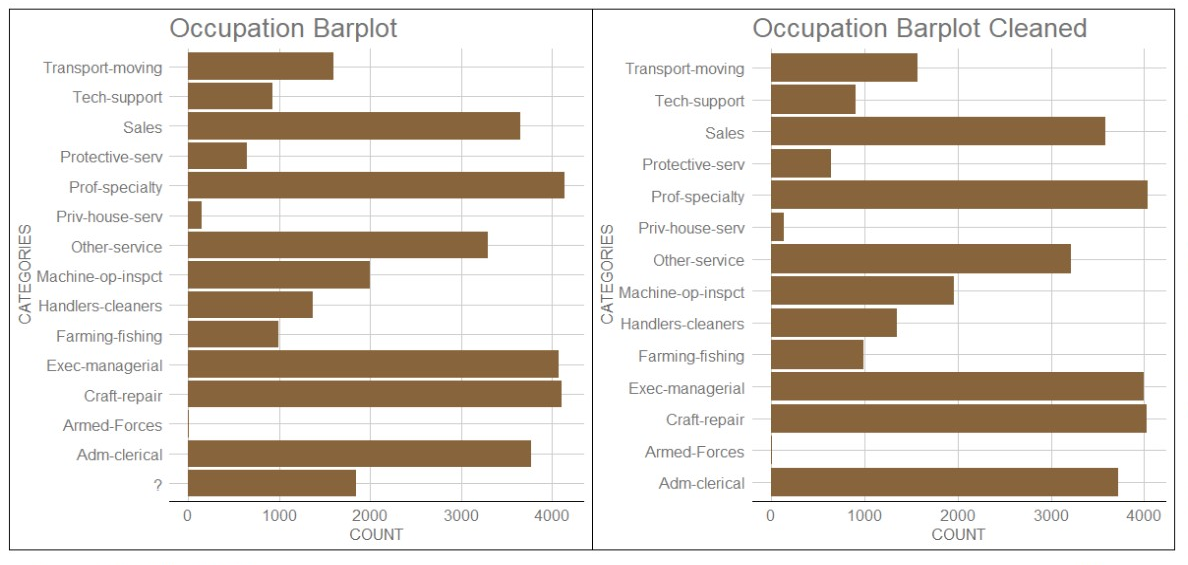


Fig. 4. (d). Side-by-side comparison of the raw versus the cleaned data set for the continuous variable - Occupation.

From the above barplot, it can be observed that the raw data contains “?” values. However, the cleaned data no longer contains that value. Hence, we can conclude that the data is now free from missing variables.

This brings us to the next challenge: Grouping Common Subcategories.

Consider the Occupation Variable. It has fourteen subcategories. These subcategories can be concisely grouped into eight categories namely White Collar, Service, Sales, Professional, Military, Admin( Administrative) and Other Occupations. The comparison of the Pre-Existing and Modified Data can be seen below.

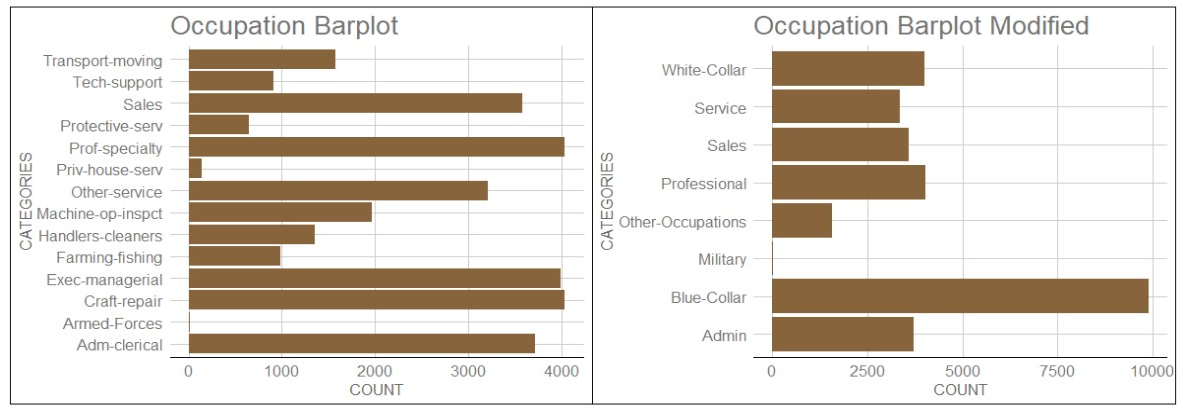


Fig. 4(e). Side-by-side comparison of grouped categories in the variable Occupation.

Similarly the comparison of Pre-Existing and Modified Values for variables like Workclass, Marital Status, Race are given below.

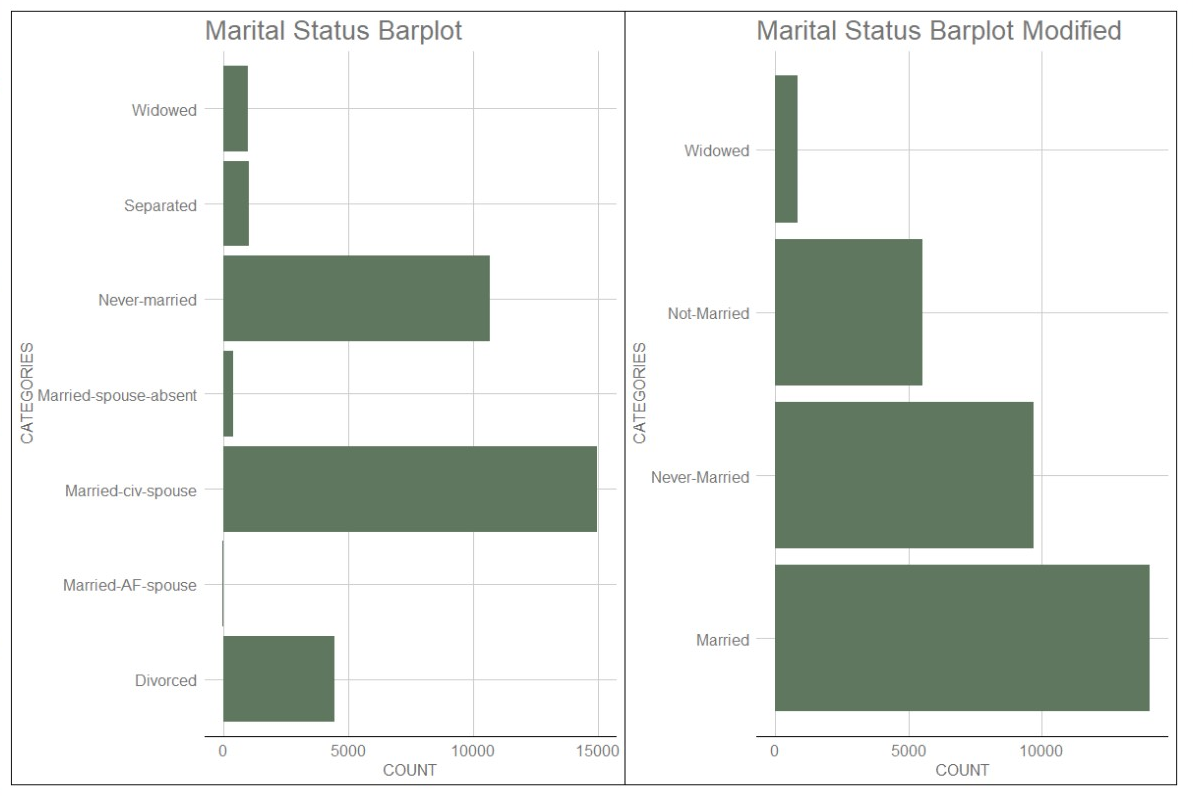
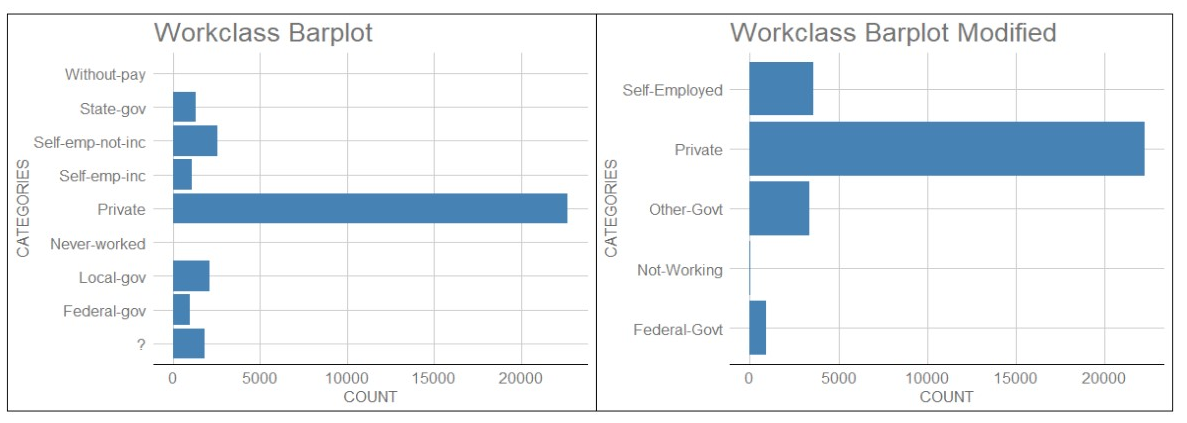


Fig. 4(g). Side-by-side comparison of Workclass and Marital Status barplots. On the left, we have the raw datasets for both the variables and on the right, we have the cleaned data for the same.

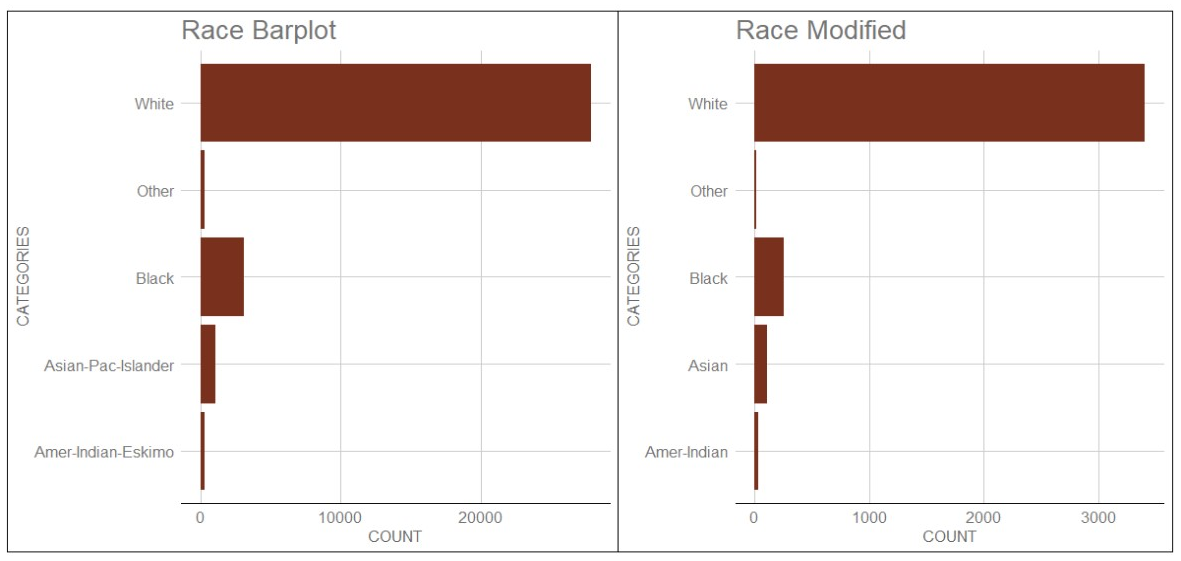


Fig. 4(h). Side-by-side comparison of the Race barplot. On the left, we have the raw data grouped by race and on the right, we have the cleaned data grouped into the same subcategories by Race.

### Continuous Variables :

There are five continuous variables in the raw dataset. They are age, education\_num, capital\_gain, capital\_loss and hours\_per\_week. The box plots for each of these variables with respect to income\_class are shown below.

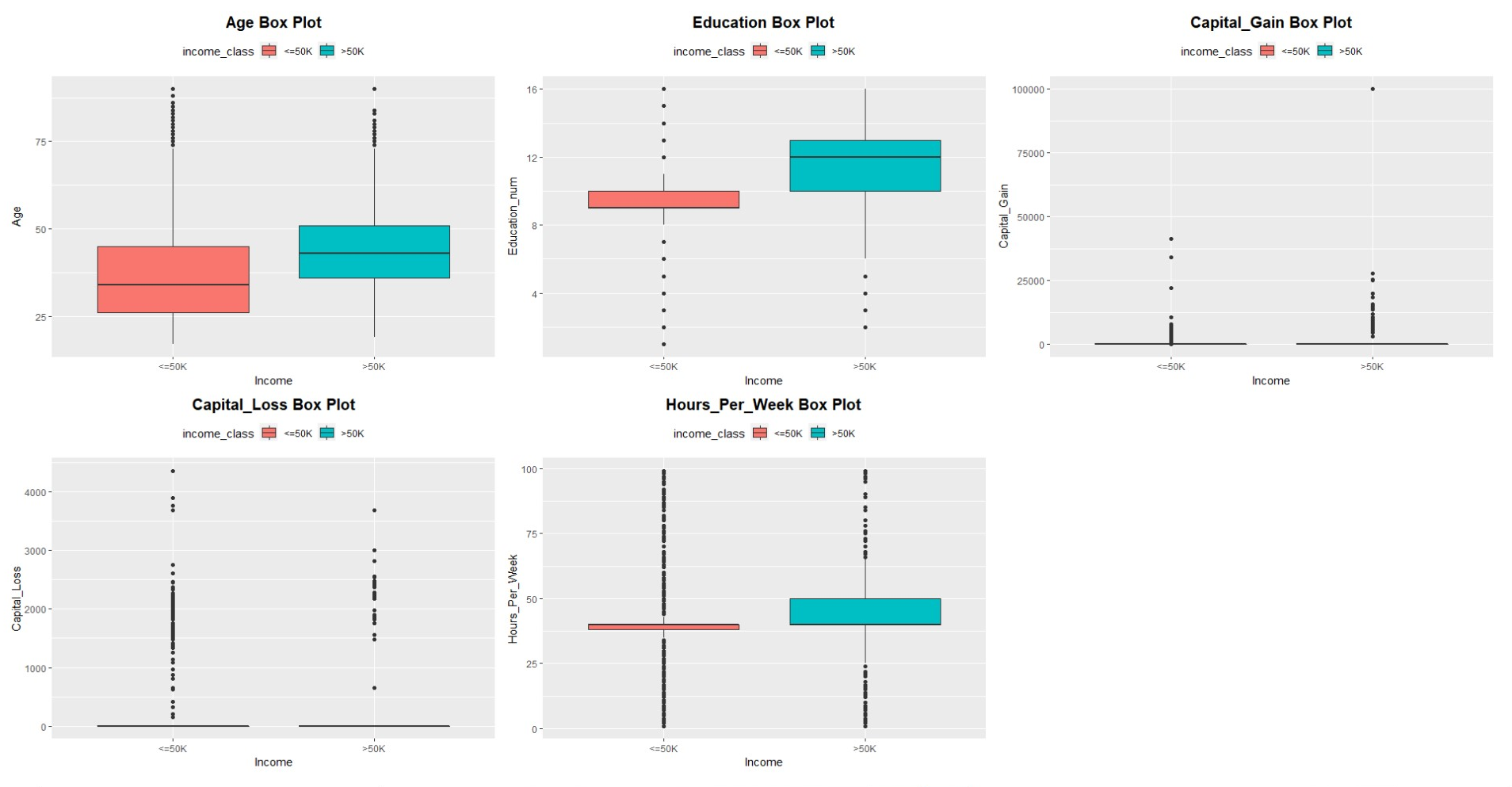


Fig. 5. From the boxplots, it can be observed that the distributions for Age, Education and Hours\_per\_week are well spread. However, that is not the case for the continuous variables “Capital\_gain” and “Capital\_loss”.

To begin with, the values that are close to zero are deleted from the dataset. The distribution obtained post that can be seen in the following box plot.

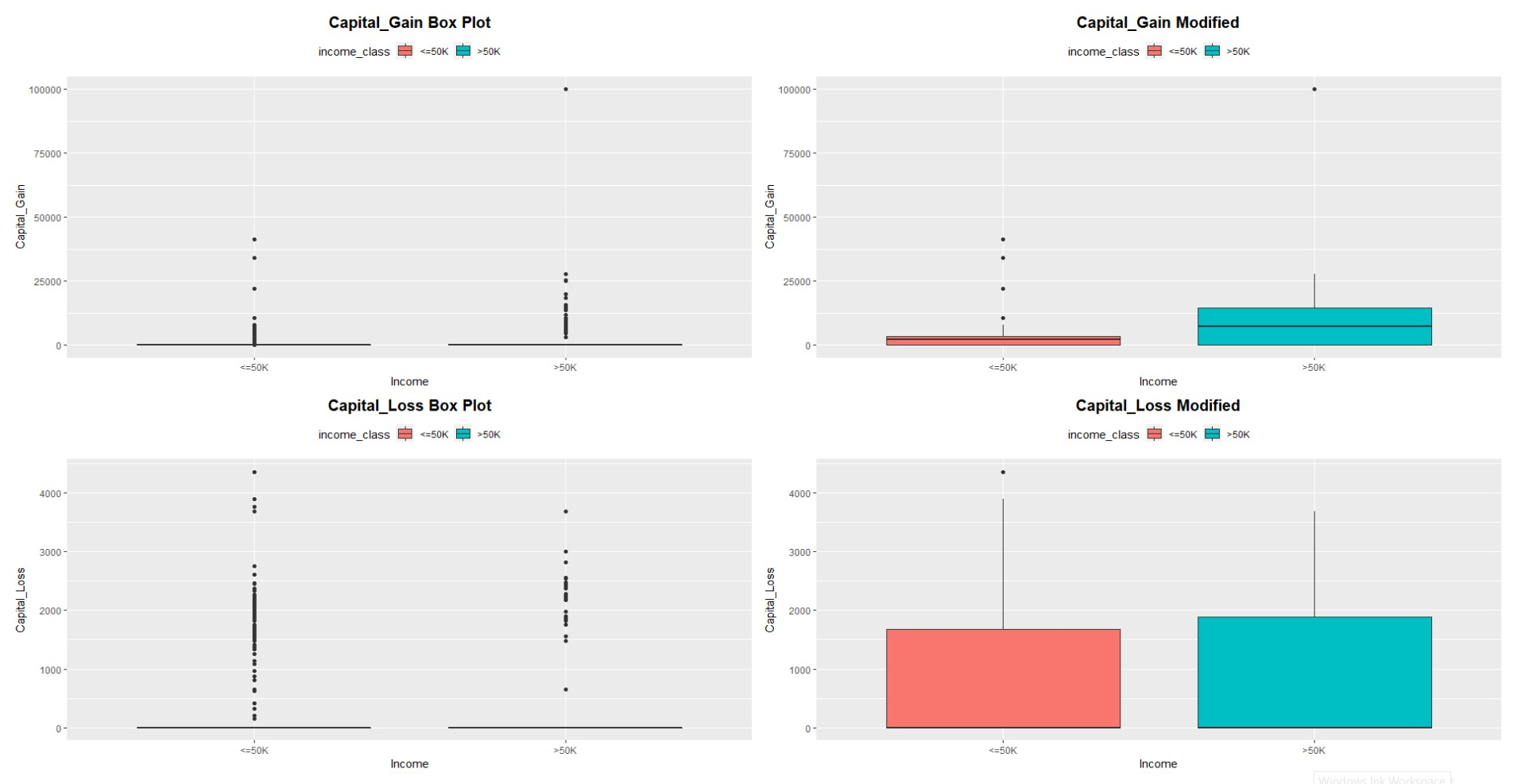


Fig. 6. Side-by-side comparison of the raw and cleaned data in the form of boxplots for the variable Capital\_Loss.

However, the mean value of capital\_gain is still very close to zero. To overcome this challenge, the interquartile distance is calculated and values that are greater than Q3 + 3IQR are labelled as outliers. The comparison of the new box plots obtained can be seen below.

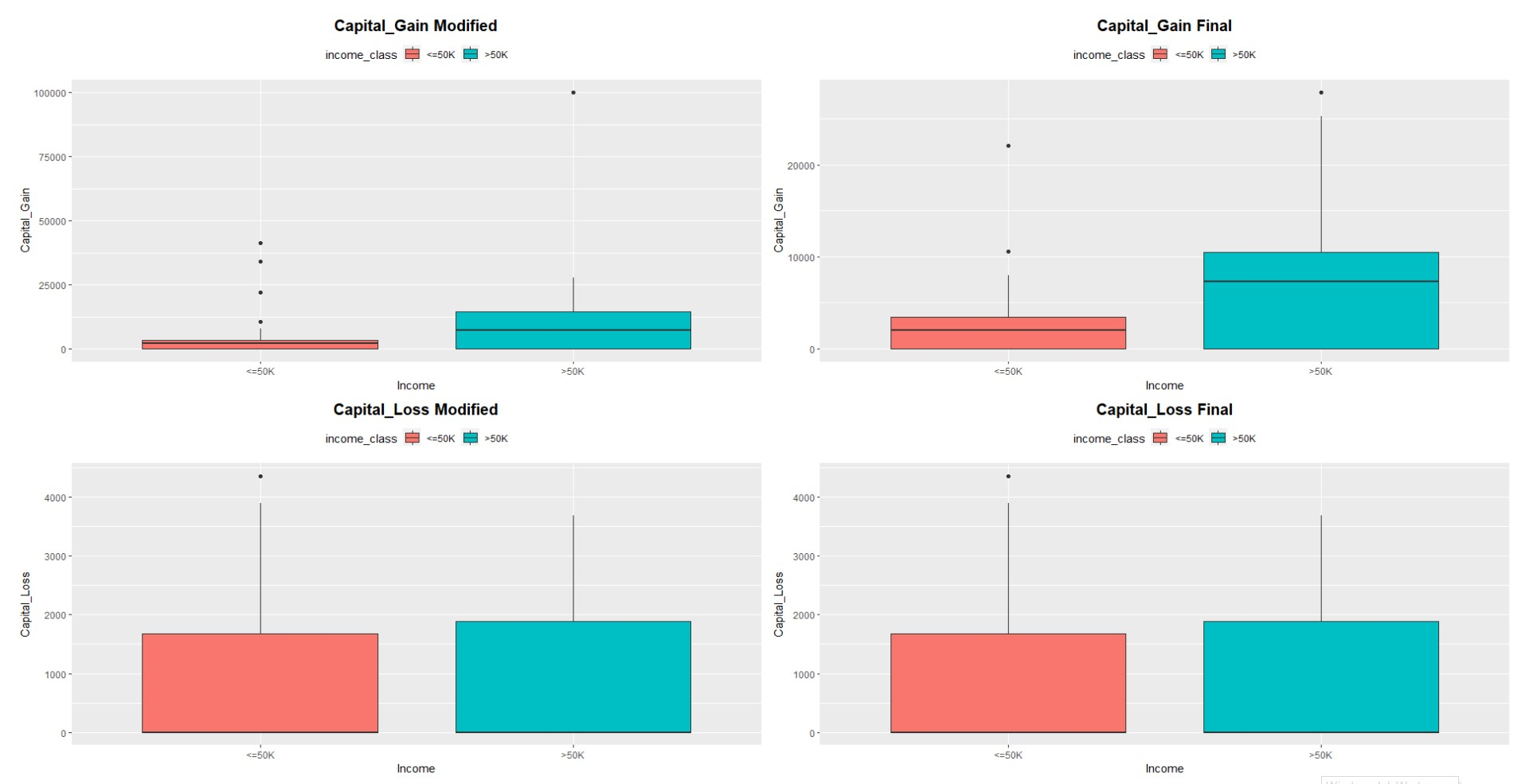


Fig. 7. Side-by-side comparison of the Interquartile ranges (IQR) for the variable Capital Loss in the form of boxplots.

With this implementation, the dataset is sufficiently cleaned and thus we move on to the next step which is Data Analysis.

## 3.4 Exploratory Data Analysis: PCA and MCA

### Principal Component Analysis:

Principal Component Analysis or PCA, is a dimensionality reduction technique which is mostly used to find the most important components or variables in large datasets. It is an effective technique that not only increases the interpretability of the dataset but also helps in maintaining the quality of the dataset.

In this study PCA is conducted to derive the most important continuous variables and to study the relationship between them.



Fig. 8. The ggfortify library in R was used to conduct PCA. The above figure shows variations of the PCA plot and the corresponding biplot of continuous variables.

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.

The biplot is plotted along with the income\_class to get a better understanding of which variables contribute mostly to high levels of income and which variables do not.

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.

### Multiple Correspondence Analysis :

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

###### 

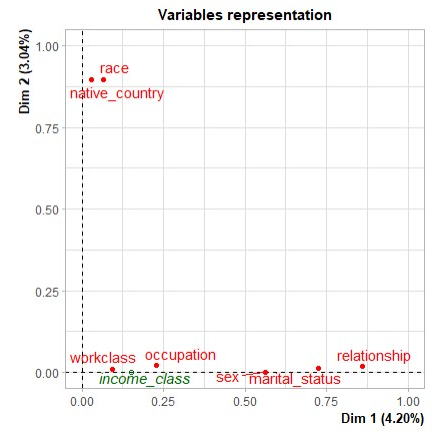


Fig. 9. MCA plot of the categorical variables.

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.

# 4. Methods

| 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% |
| --- |

Fig. 10 A summary of the accuracy scores of different Machine Learning algorithms[1]

The main paper detailed a method of combining naive bayes classification with decision trees to better classify on larger datasets. Harnessing the segmentation of decision trees and the ability of naive bayes to draw evidence from multiple predictors, the NBTree is shown to slightly improve results of classification on the Adult/Census dataset [1]. Despite being over 20 years old, the methods highlighted in this paper struck curiosity for us to attempt newer methods to improve the model’s performance on this classification task.

Figure 10 contains some info from the main paper, detailing some information about the data split, as well as the accuracy of some of the common techniques. We sought to confirm these results by building models using some of the same techniques, as well as trying new ones of our own.

Implementation of vanilla trees yielded similar results, while our naive-bayes classifier significantly outperformed the original paper. Expanding upon the tree based methods highlighted in the paper, we attempted to improve classification accuracy further using newer tree-based methods. The method of NBTree, described in the main paper, is a tree based method that takes the naive bayes classifier at each leaf node. By evaluating the success of applying naive-bayes classifiers at each decision point, the algorithm will choose to continue splitting or to implement the bayes classifier in an attempt to improve accuracy[1]. Another method we tried was tree augmented naive bayes, which similarly to NBTree augments naive bayes classifiers with trees. In this case, tree augmentation expands upon the naive independence assumption, allowing dependency between predictors only between the class and another predictor[8]. For this particular dataset, TANs performed worse than vanilla naive bayes, indicating that this dataset might not benefit from an alternate independence assumption.

**SVM**

Support Vector Machine is a supervised machine learning method that is responsible for finding the decision boundary to separate different classes and maximize the margin. In our implementation of SVM we used 3 different kernels to fit our model.

1. Linear Kernel

2. Quadratic Kernel

3. RBF Kernel

These kernels are different in the sense that they make different hyperplane decision boundaries to separate the classes.

The Linear kernel assumes that the data is linearly separable whereas the quadratic kernel assumes that the decision boundary is a mixture of quadratics. RBF kernels are the most generalized form of kernelization and is more widely used due to its similarity to the Gaussian distribution. The RBF kernel function for two points X₁ and X₂ computes the exponent of negative L2 norm by the variance, which is a hyperparameter. This kernel can be mathematically represented as follows:

When the points are the same, there is no distance between them and therefore they are extremely similar. When the points are separated by a large distance, then the kernel value is less than 1 and close to 0 which would mean that the points are dissimilar

**Neural Network Model**

We constructed a simple feedforward neural network with the nnet package in R. We first train the data and use the trained model to classify the data in the test set. Size describes the number of nodes that will be used in the hidden layer, in this case 10 nodes are used. Decay illustrates how quickly it decreases in gradient descent. Maxit is the maximum iteration to be carried out, in this case the maximum iteration to be carried out is 2000 iterations.

The model that we used has a single hidden layer to exploit the linear nature that was highlighted in the SVM Models.

**Random Forest Model**

Random forest in R is developed by an aggregating tree and this can be used for classification and regression. One of the major advantages is that it avoids overfitting. Random forest contains two parameters ntree and mtry.

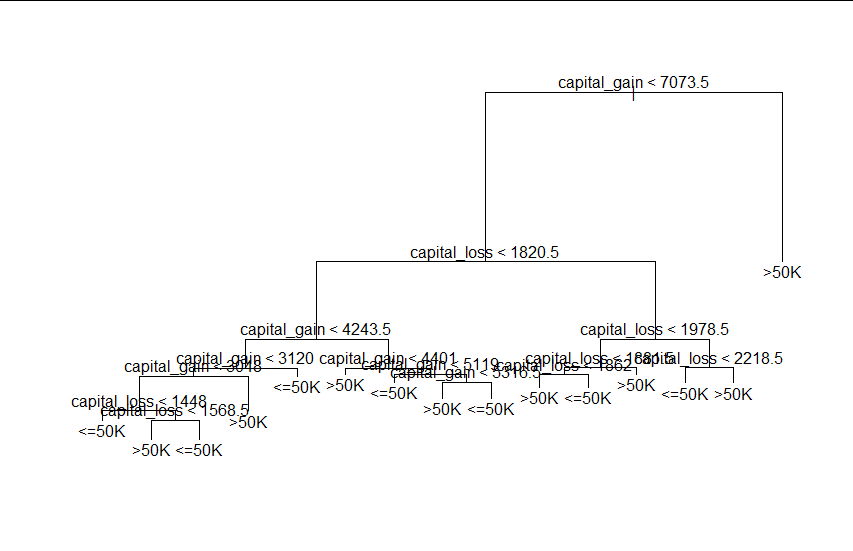
1. ntree- ntree by default it is 500 trees.

2. mtry- variables randomly sampled as candidates at each split. It is chosen as the square root of number of explanatory variables

The ntree value is however chosen by calculating the OOB (Out of Bag) Error for a range of values of ntree and picking the one which gives the least OOB error.

# 5. Results

Results obtained for tree based methods were conducted with two versions of the dataset, one using the original with all missing values and the other containing all complete observations.



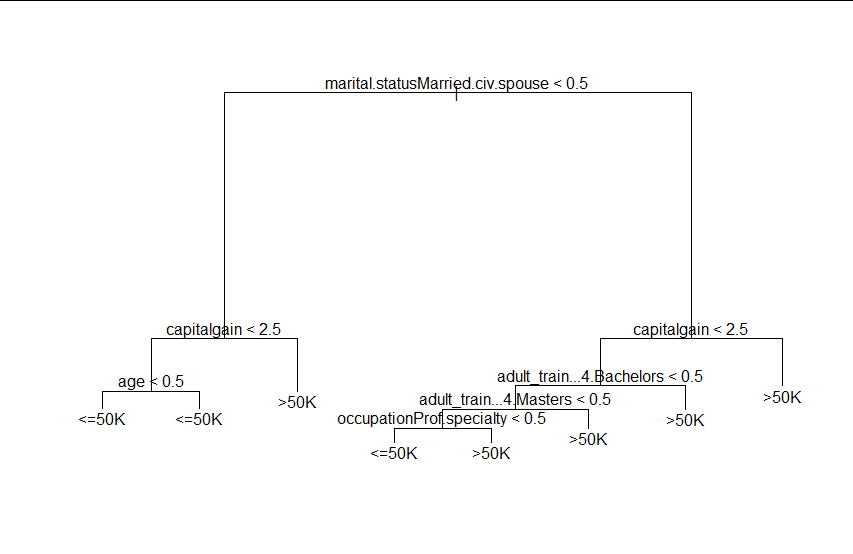
| Acc: 94.46% | <= 50K | > 50K |
| --- | --- | --- |
| <= 50K | 446 | 42 |
| > 50K | 19 | 636 |

Figure 11: Unpruned Decision Tree (Complete Observations Only)

# 

| Acc: 84.46% | <= 50K | > 50K |
| --- | --- | --- |
| <= 50K | 460 | 28 |
| > 50K | 148 | 507 |

Figure 12: Pruned Decision Tree (Complete Observations Only)



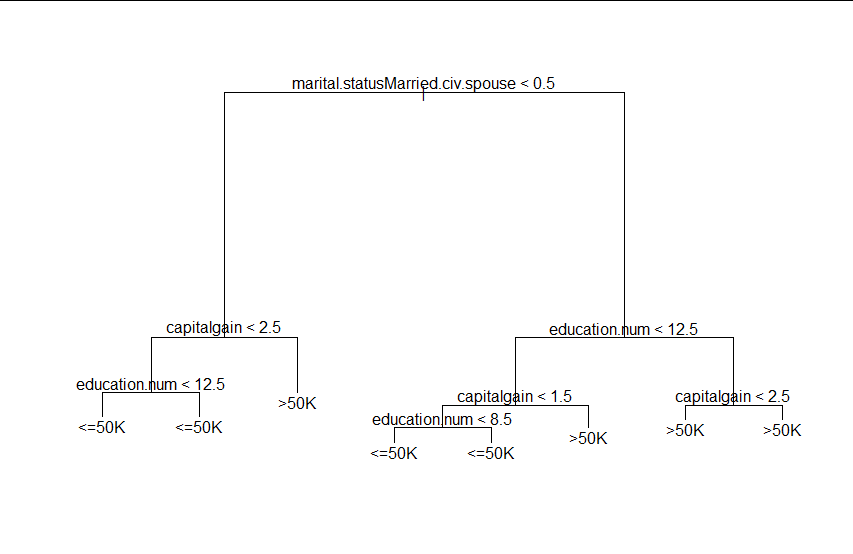


Figure 13: Weaker decision trees give insight into important variables and confirm EDA findings.

| Acc: 90.14% | <= 50K | > 50K |
| --- | --- | --- |
| <= 50K | 9824 | 122 |
| > 50K | 1322 | 3384 |

Figure 14: Naive-Bayes Classifier

| Acc: 85.07% | <= 50K | > 50K |
| --- | --- | --- |
| <= 50K | 10228 | 1269 |
| > 50K | 918 | 2237 |

Figure 15: Tree Augmented Naive-Bayes Classifier (TAN)

| Acc: 84.90% | <= 50K | > 50K |
| --- | --- | --- |
| <= 50K | 10133 | 1200 |
| > 50K | 1013 | 2306 |

Figure 16: NBTree

| Acc: 86.88% | <= 50K | > 50K |
| --- | --- | --- |
| <= 50K | 341 | 61 |
| > 50K | 64 | 487 |

Figure 17: Linear SVM

| Acc: 57.29% | <= 50K | > 50K |
| --- | --- | --- |
| <= 50K | 225 | 177 |
| > 50K | 230 | 321 |

Figure 18: Quadratic SVM

| Acc: 94.54% | <= 50K | > 50K |
| --- | --- | --- |
| <= 50K | 386 | 16 |
| > 50K | 36 | 515 |

Figure 19: Radial Basis SVM

| Neural Network: | 88.56% |
| --- | --- |
| Random Forest | 95.67% |

Figure 20: Other model accuracies

# 6. Conclusion

Motivated by the advances in Trees and Bayesian methods, we sought to expand upon the work described in the main paper and see if we can improve the prediction accuracy with new models and different techniques to pre-process the dataset. We did a thorough study of the dataset and cleaned it up to obtain the important features required for the prediction of the model. We tried to implement multiple models for the task and obtained the results discussed above.

Moving forward, we can explore more techniques to pre-process the data and we can try using more models or a combination of models to try and improve the accuracy further. Another suggestion would be to try and fine tune the Random Forest to obtain better results.

Finally, these models can be applied to a more recent dataset and study if the trends fall in line with the model’s predictions and if there are other features that are important to the accuracy of the model that can be included in the dataset.

# 7. References

[1] Fu, K. S., “Syntactic Pattern Recognition and Applications”, 1982.

[2] T.M. Cover and J.A. Thomas, “Elements of information theory”, 1991.

[3] R. Kohavi, “Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid”, 1996.

[4] P. Domingos and M. Pazzani, “On the optimality of the simple Bayesian classifier under zero-one loss", 1997.

[5] C. D. Manning, and H. Schutze, “Foundation of Statistical Natural Language Processing”, 1999.

[6] J. Hellerstein, Jayram Thathachar, and I. Rish, “Recognizing end-user transactions in performance management”, 2000.

[7] I. Rish, “An empirical study of the naive Bayes classifier”, 2000.

[8] H. Padmanabhan, “Comparing Analysis of Naive Bayes and Tree Augmented Naive Bayes Models”, 2014.