Adult Census Income Level Prediction

Statistical Learning for Biological and Information Systems - Final Report

December 2022

Tong Wu, Yu Wu

Contents

Abstract

Income level prediction is essential for individuals, economics, and nations. With a precision income level prediction model, most trends of the economic and political policies happening in the country can be predicted. This project uses the adult dataset to investigate the relationship between adults' attitudes and income levels by reproducing papers' results and bringing some new algorithms to predict development.

This report will show some reproducing and draw conclusions on the successful implementation of the machine learning algorithms as well as highlight the area where the project faced limitations. At the end of the report, some areas where future work can be further concentrated to improve the developed algorithms. (Tong Wu, 2022)

1. Introduction

Income level is an essential attribute of an adult. It represents the approximate level of consumption of the adult and, combined with other data, can easily be extrapolated in each area that they spent or predicted for the shopping mall where they spend the most. Also, for the macro economy, if adults' income level and related data are collected in a large area, forecasts can be made for city or country fundamentals. Examples include the GDP growth rate, the tertiary sector's prosperity, or even the number of children per household. And for governments, the level of income of the nationals determines the extent to which the government collects tax for the year, which is a matter of running the state apparatus and allocating funds.

However, how can data be used to analyse and predict income levels? This problem has been given over to statistics and machine learning in recent years. Machine learning models use datasets for training and testing. In general, the dataset needs to be randomly divided into a training set and a test set, with the training set taking 80% of the data and the remaining 20% for testing the model. The purpose of this is to avoid overfitting the model, making it remain accurate when predicting other data sets. The splitting dataset is only the most basic data pre-processing. Other pre-processing, including downscaling and digitisation, will be mentioned in subsequent sections.

With this motivation, it would make sense to analyse and predict at the income level, so research is decided to do in this direction and select a dataset for the exercise. Barry Becker does the adult dataset from the 1994 Census database (*UCI Machine Learning Repository: Adult Data Set*, n.d.). The dataset contains most of the critical attributes of an adult, including education level, position, family role, and income level. The main prediction task is to determine whether a person makes over 50K a year. This dataset has been initially classified and cleaned. Our mission is to build on this foundation to uncover relationships between attributes and train algorithms with as high an accuracy as possible.

In this project, some exploratory data analyses were presented and put into practice, visualising the data and subsequent support for dataset cleaning and validating the algorithm's conclusions. The two algorithms were then applied to the dataset based on previous work. The datasets were cleaned to ensure data integrity and enable the algorithms' direct use. The results of the two papers were reproduced and maintained a high degree of similarity to the original results.

2. Literature Review

In this project, two papers were applied to serve as a source for reproducing the results. The first paper is from Bekana, who used the Random Forest algorithm to make predictions on the dataset while giving the importance value of each attribute to determine its relevance to the predicted income. In conclusion, she suggests that the Random Forest algorithm has an accuracy of 85%. Age, capital gains, education level and the number of hours worked per week where the four most associated with predicting income levels. At the same time, however, she suggests that fitting the data for every country using the random forest algorithm is complex. It is difficult to predict income levels 30 years later because the data set was collected in 1995. (Bekena, 2017)

The second paper cites four different classifier algorithms containing random forest, K-means, zero R and plain Bayes. They concluded that the naive Bayesian classifier had a high accuracy of around 85%, while the remaining three classifiers were all around 70%. (Deepajothi & Selvarajan, 2012)

3. Methods

The data cleaning and exploratory data analysis will be first implemented in this section. Then, two paper result reproduction will be produced to compare the initial results. Finally, two new algorithms are introduced to be implemented in this dataset and compare the accuracy.

3.1. Data Cleaning

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age          : num  39 50 38 53 28 37 49 52 31 42 ...
##  $ fnlwgt       : num  77516 83311 215646 234721 338409 ...
##  $ education_num: num  13 13 9 7 13 14 5 9 14 13 ...
##  $ capital_gain : num  2174 0 0 0 0 ...
##  $ capital_loss : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ hr_per_week  : num  40 13 40 40 40 40 16 45 50 40 ...
##  $ type_employer: Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ education    : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ marital      : Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation   : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ race         : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex          : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ country      : Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
##  $ income       : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

Figure 1 Datatype and brief view for each attribute

Data cleansing is crucial for every project. Data cleansing determines the efficiency as well as the accuracy of the subsequent algorithms. For this dataset, Figure 1 shows all the attributes. It is clear that “fnlwgt” is the number of each row of data, which should not be present in the dataset to which the algorithm is applied, as it is not correlated with other attributes and reduces the accuracy of the prediction, so this attribute was removed. Afterwards, to analyse the data better, some features in the string should be transformed into factors, a data type stores information about the string in an array of a list and changes the value of that attribute for each row of data into a number corresponding to the text in the list. This purpose is to convert the text into numbers so that the algorithm can correctly identify the data and generate correlations. Also, some attribute values were missing, and these data rows with missing attributes were removed to ensure data integrity. Finally, the original dataset had 32,561 data, and 2,399 data were released after data cleaning.

3.2. Exploratory Data Analysis

Some exploratory data analysis was first implemented before proceeding to reproduce the findings. The exploratory data analysis is a valuable step for the subsequent validation of the algorithm’s conclusions by downscaling the data set for a specific algorithm. Figures 2 and 3 show the population distribution by income level according to different attribute classifications. These plots allow the data to be visualised, visually displaying the relationship between the income level population and the various attributes. Figure 4 shows the correlation between attributes and brings the quantitative values for each correlation. From figure 4 can be found that the four

attributes listed in the figure have a positive correlation with income level, which also validates Bekena's work (Bekena, 2017).

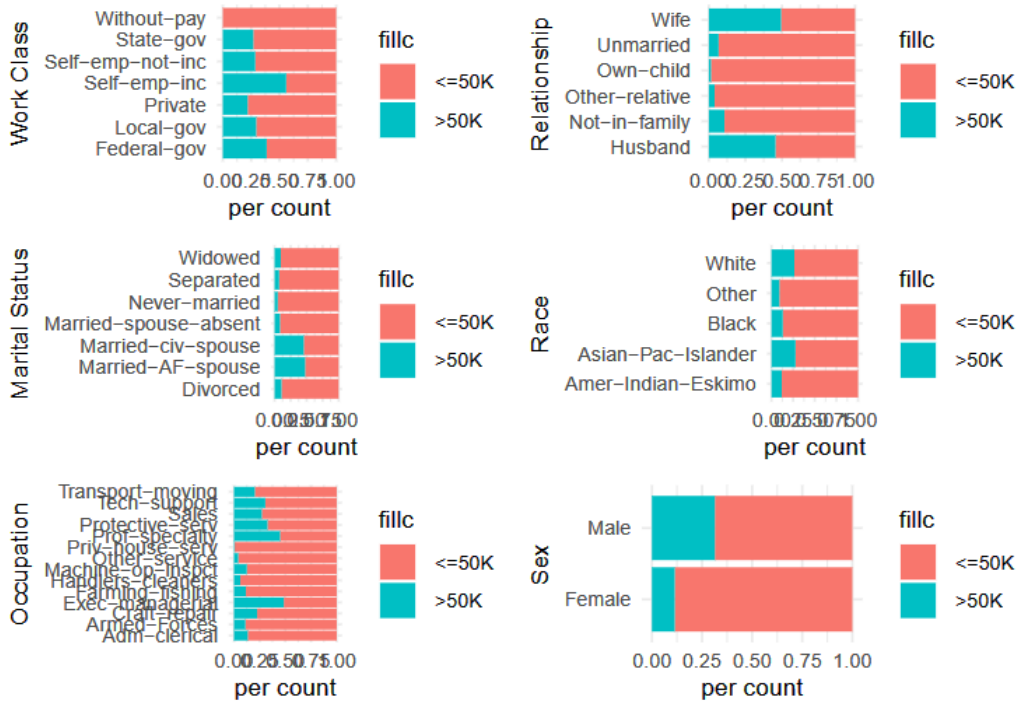

Figure 2 Income level distribution according to different attributes

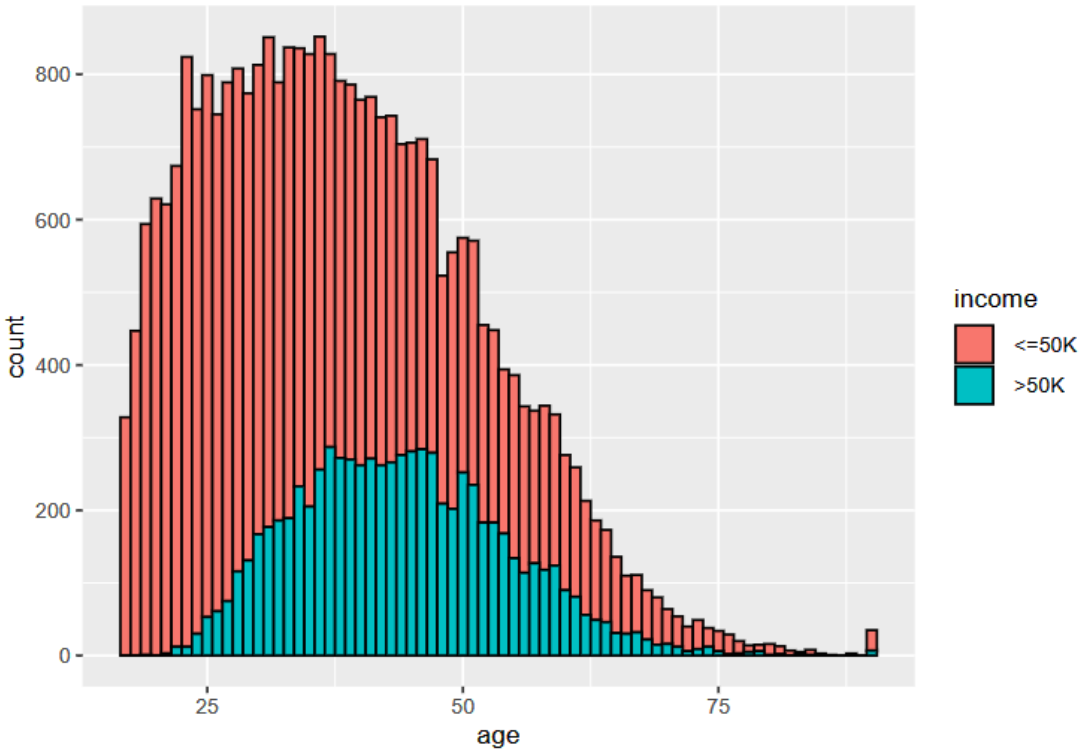

Figure 3 Income level distribution according to the age

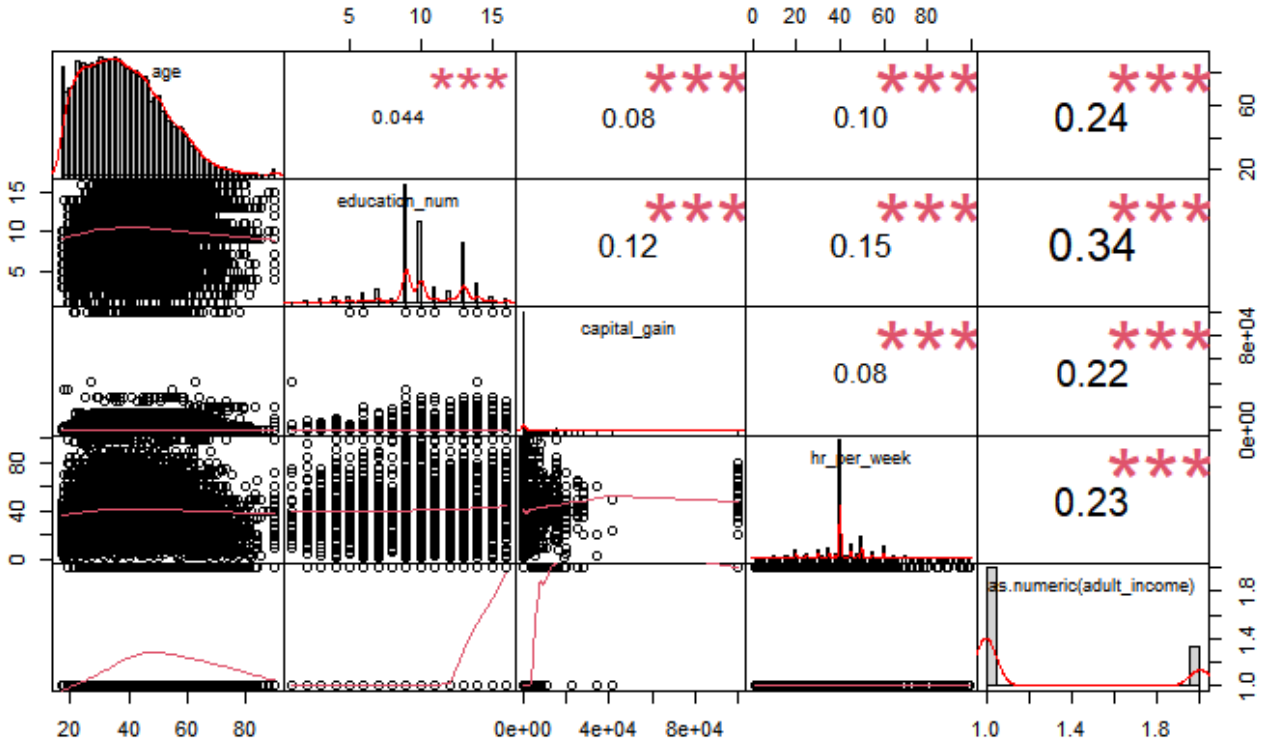

Figure 4 Correlation for attributes

3.3. Result Reproduction

After the data cleaning and exploratory data analysis are done, a clear view of the data set and to reach essential attributes are provided. In this section, the two papers' result is reproduced, and some comparison is provided.

3.3.1. Using Decision Tree Classifier to Predict Income Levels (Bekena, 2017)

Tree predictors are used in tandem to create the random forest classifier. Each tree in the forest is dependent on the values of an independent random vector sampled using the same distribution (Breiman, 2001). For this dataset, the target attribute is a binary variable with an income level higher than 50K, so the classifier algorithm is more efficient than regression.

Figure 5 shows the confusion matrix for the random forest classifier, and the blue area in the matrix shows the correct prediction for the random forest model when predicting the test dataset. The accuracy of the random forest classifier is 83.8%, which has a slight difference from Bekena's result of 85% of accuracy (Bekena, 2017). The possible reason for this is the difference in cleaning the data, especially for the missing data rows.

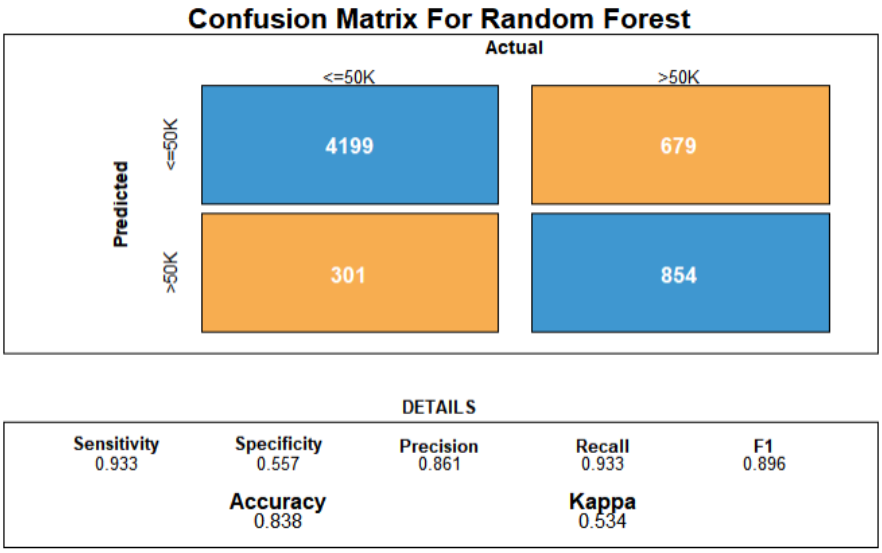

Figure 5 Confusion Matrix For Random Forest

Then figure 6 shows the importance score for each attribute. The importance score shows the relationship level for each attribute with the target attribute, income level. For this part, the reproduced result is different from the actual result. From figure 6 can see that the attribute "marital" has an enormous importance score and has a 23% of relation for attributes. However, the initial result shows that marital status should have 4% of the association. This difference is because the original paper calculated the different attribute values for the marriage attribute separately. So married, unmarried and divorced will have other correlations. After retesting, a conclusion similar to the actual paper results was obtained, which is shown in Figure 7, where the attribute "age", "capital gain", "education level", and "working hours per week" are the four most related attributes.

	Overall <dbl>	percentages2 <dbl>
age	920.0492	14.820885
education_num	606.8835	9.776162
capital_gain	998.6801	16.087534
capital_loss	321.4483	5.178145
hr_per_week	555.2967	8.945161
type_employer	301.3149	4.853821
education	563.0241	9.069640
marital	1458.1924	23.489723
race	112.8561	1.817976
sex	161.4682	2.601058

1-10 of 11 rows

Figure 6 Importance Score for Attributes

	Overall <dbl>	percentages <dbl>
age	1237.0786	22.104773
education_num	624.5340	11.159503
capital_gain	1120.6699	20.024722
capital_loss	398.9792	7.129172
hr_per_week	665.0159	11.882855
type_employer	327.1372	5.845461
education	527.4635	9.424996
race	130.3918	2.329909
sex	330.0247	5.897056
country	235.1372	4.201555

1-10 of 10 rows

Figure 7 Importance Score for Each Attribute After a Retest

The Gini coefficient's mean decrease gauges the contribution of each variable to the homogeneity of the nodes and leaves in the ensuing random forest. The more significant a variable is in a model, the greater its mean decrease accuracy or mean decrease Gini score. (*Variable Importance Plot (Mean Decrease Accuracy and Mean Decrease Gini).*, 2020) Figure 8 shows the mean decrease in Gini. This figure can also get a similar result with an importance score.

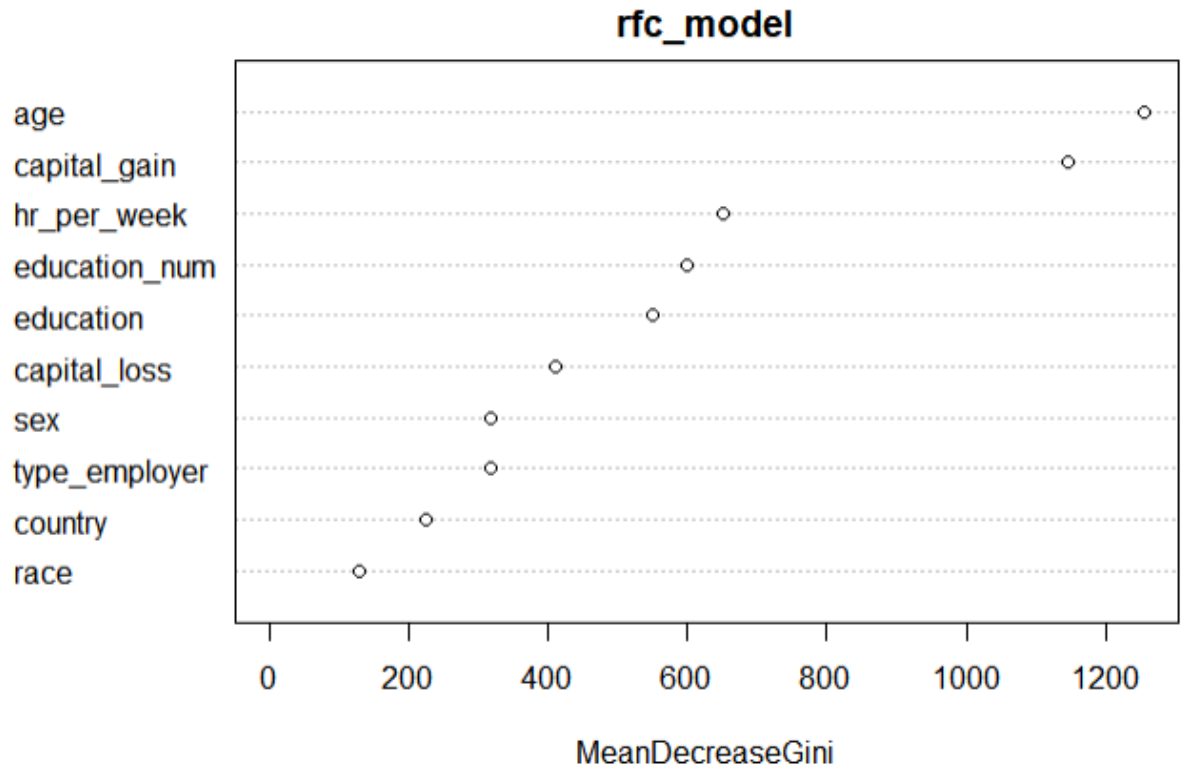

Figure 8 Mean Decrease in Gini

3.3.2. A Comparative Study of Classification Techniques On Adult (Deepajothi & Selvarajan, 2012)

In this paper, the author Deepajothi, S., & Selvarajan, S. use four different classifier algorithms to predict the income level. As a result, they found that the Naïve Bayesian has the best accuracy, and three algorithms remain in the same range. In this section, the reproduction of this paper will be shown, and results will be discussed and compared with the actual result.

By assuming that characteristics are independent of classes, the naive Bayes classifier significantly simplifies learning. Despite the fact that independence is often a bad assumption, naive Bayes frequently outperforms more advanced classifiers. (Rish, 2001). The k-means clustering algorithm is difer from KNN and is unsupervised learning using untagged data. The k-means clustering algorithm aims to partition n observations into k clusters in which each statement belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster ('*K*-Means Clustering', 2022). Zero R is a naive approach to classify a dataset based on target and ignores other independent attributes. Note that the Random Forest classifier result will use the result from the previous reproduction section since the dataset is the same.

After implementing the three algorithms, the results are below, similar to the original result. Figures 9, 10 and 11 show the confusion matrixes for naïve Bayesian, K-means clustering, and zero R. From the matrix, each algorithm has above 60% accuracy. Figure 12 shows the correlation plot between all attributes. The graph shows that the naïve Bayesian has the highest accuracy among these three algorithms. However, the Random Forest has higher accuracy, 84%, than the naïve Bayesian, which is also the difference between the actual result.

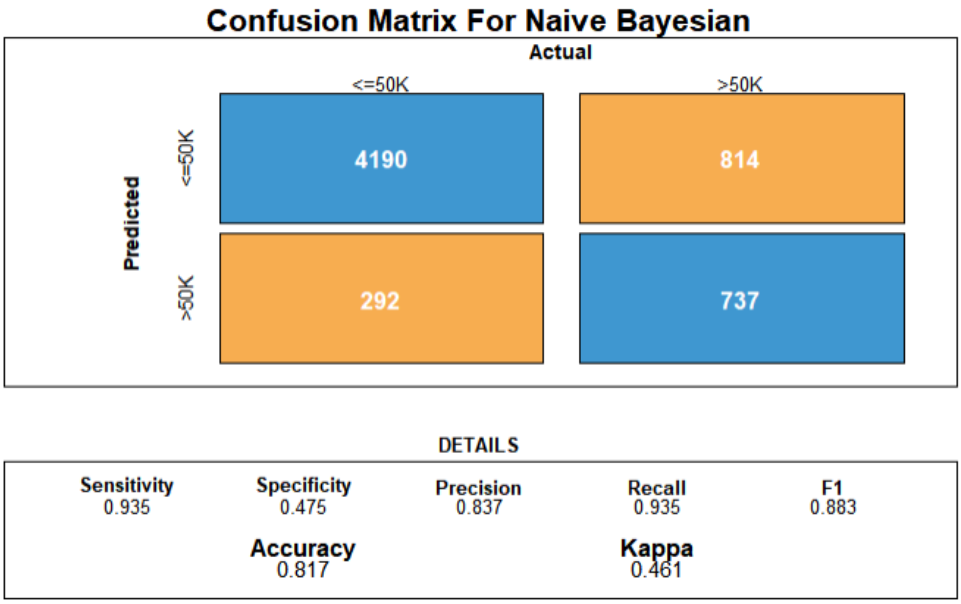

Figure 9 Confusion Matrix for Naive Bayesian

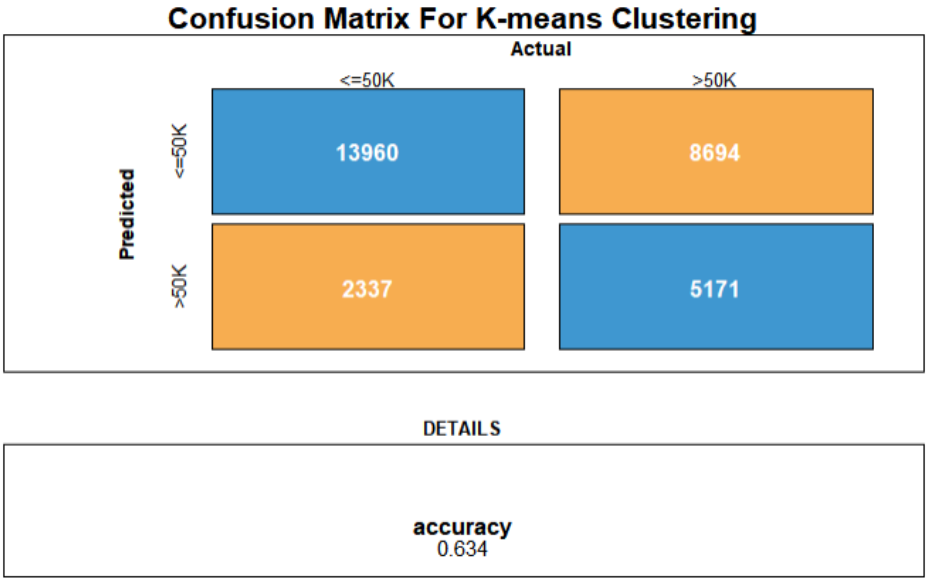

Figure 10 Confusion Matrix for K-Means Clustering

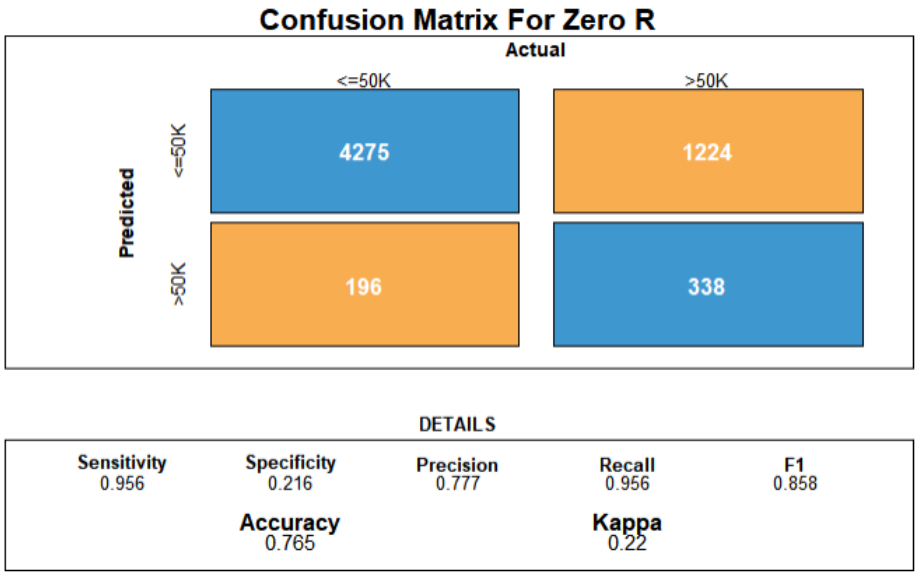

Figure 11 Confusion Matrix for Zero R

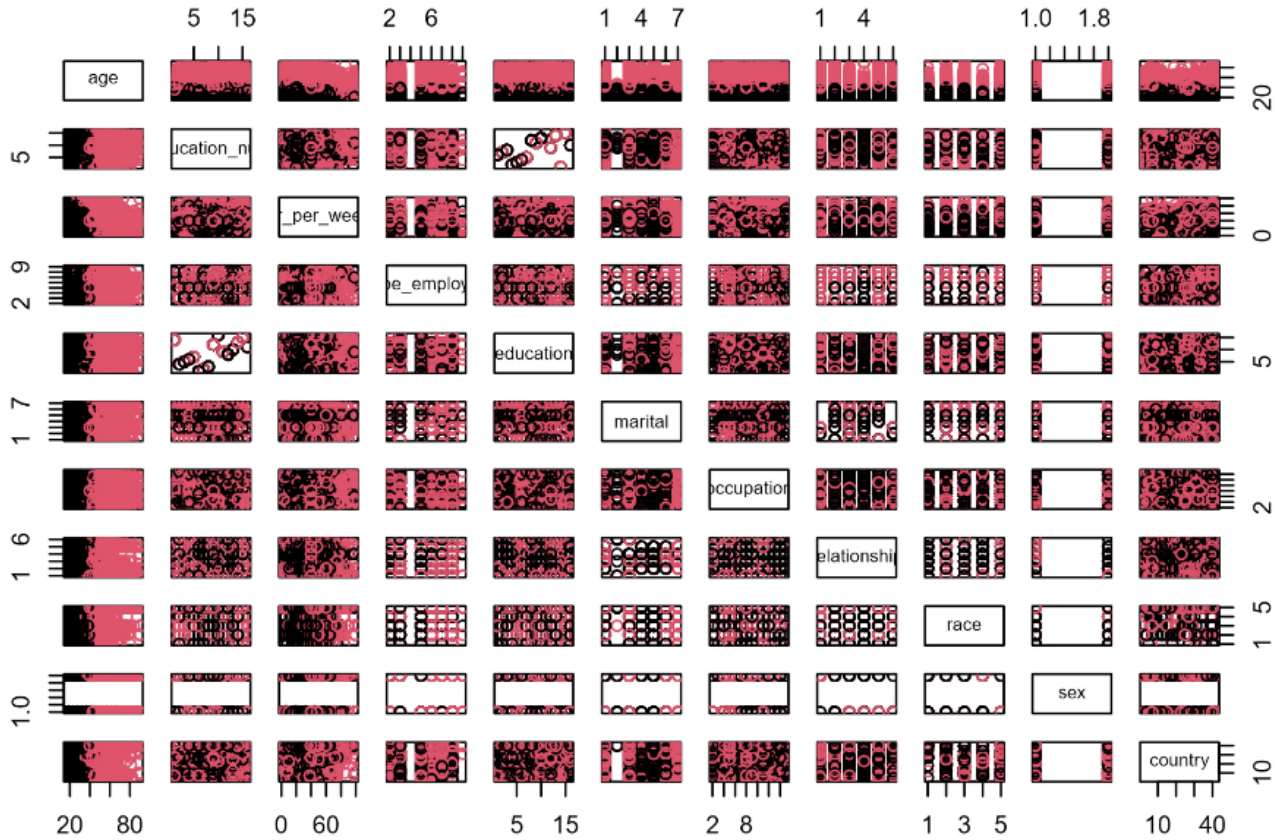

Figure 12 Correlation Plot for K-Means Clustering

The result may vary because of the high accuracy for Naive Bayesian and Random Forest classifiers when dealing with a large-scale data set. However, K-means, as unsupervised learning, learning from the un-tagged dataset. Although dimensionality reduction was already made in the code, the

high dimension still drags the accuracy. I also tried to implement K-means on the dataset with a few dimensions. The accuracy is much higher than in more dimensions. Zero R is a naive approach to classify a dataset based on a target and ignores other independent attributes, so it brings fast computing with relatively low accuracy.

3.4. New Approaches

After the result reproduction, some new ideas are considered to implement new algorithms on the dataset. In this section, two new algorithms will be discussed according to their accuracy and efficiency.

3.4.1. Logistic Regression

Logistic regression uses several independent variables to predict categorical dependent variables. The logistic regression arises from the desire to model the posterior probabilities of the K classes via linear functions in x, while at the same time ensuring that they sum to one and remain in [0, 1] (Hastie et al., 2009). This section will use the logistic regression algorithm to predict the test dataset. Then, the accuracy of the algorithm will be shown and discussed.

The same cleaned dataset is used for the new approach section, with the same randomness split training and test dataset. Figure 13 shows the confusion matrix for logistic regression. The algorithm's accuracy reaches 85.4%, which is higher than the previous six algorithms.

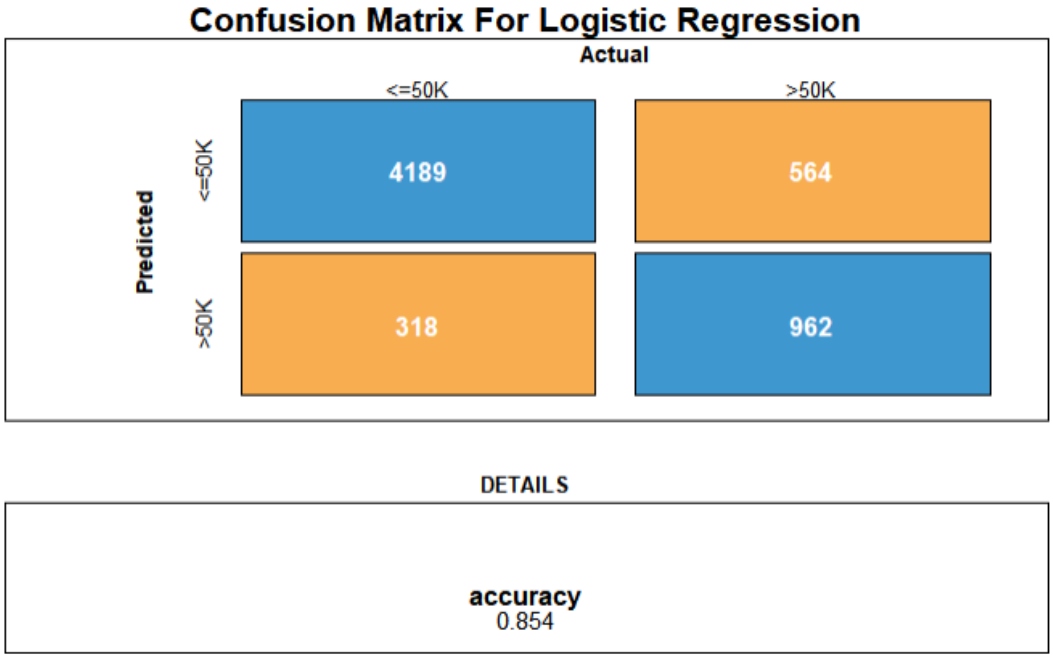

Figure 13 Confusion Matrix for Logistic Regression

The ROC curve does this by plotting sensitivity, the probability of predicting a real positive will be positive. Against 1-specificity, the probability of predicting a real negative will be positive. Figure 14 below shows the ROC curve of the logistic regression algorithm.

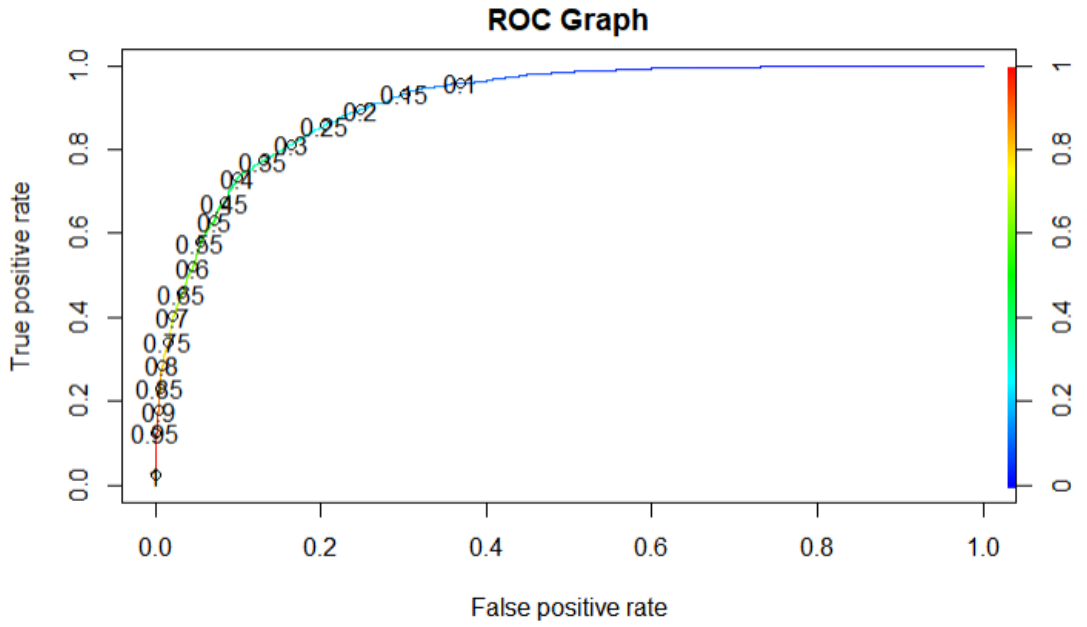

Figure 14 ROC Curve Graph

3.4.2. Support Vector Machine

The Support Vector Machine can be used to classify the dataset into two groups, which is fit for the adult dataset. The support vector machine technique seeks out an N-dimensional hyperplane that clearly divides the input points into different categories. In a multidimensional hyperplane, an SVM model represents many classes. To reduce inaccuracy, SVM will iteratively build the hyperplane. SVM aims to divide the datasets into categories to find a maximum marginal hyperplane (MMH). (Jain, 2020) In this section, the SVM is used to predict the income level. Then, the accuracy of the algorithm will be shown and discussed.

Figure 15 shows the confusion matrix for SVM. SVM operates reasonably well when there is a distinct margin of separation between classes, which is why it is effective. SVM is more practical than the algorithms reproduced in the result reproduction section in high-dimensional space.

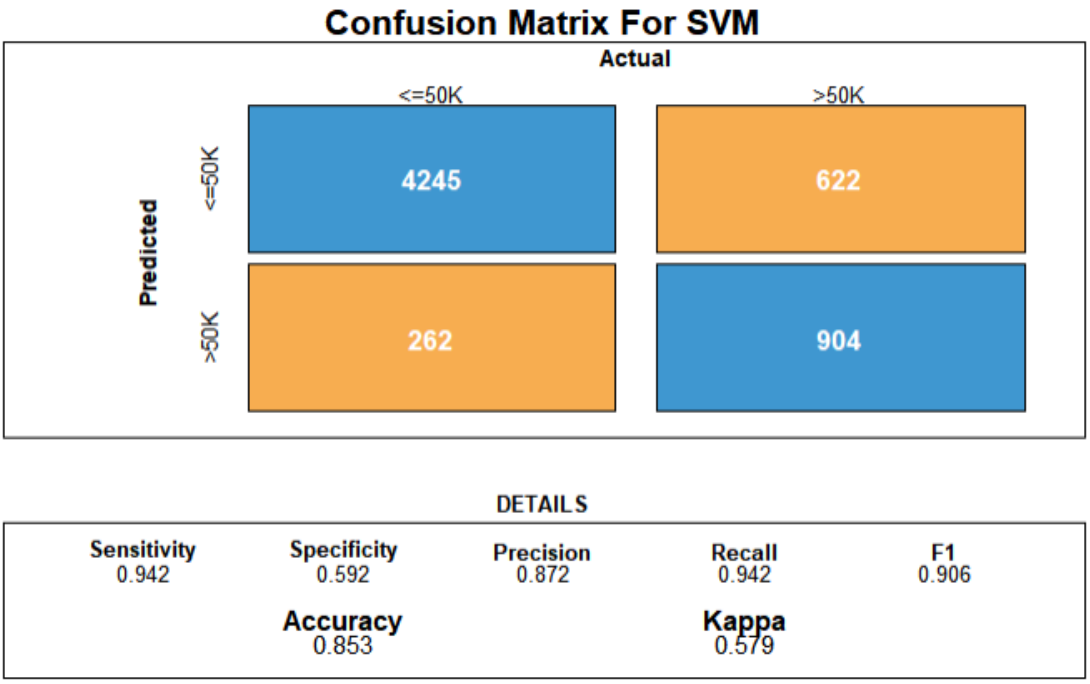

Figure 15 Confusion Matrix of SVM

4. Comparison

The random forest classifier and its subsequent results are reproduced from the first paper. According to the second paper, three more algorithms are implemented: Naive Bayesian, Zero R, and k -means. Finally, two new algorithms, logistic regression and SVM, are introduced for predicting the dataset. In this section, a comparison is made with these perspectives: accuracy, run time complexity, and space complexity.

Table 1 Comparison of Algorithms

Algorithm	Accuracy
Random Forest	83.8%
Naïve Bayesian	81.7%
K-means Clustering	63.4%
Zero R	74.1%
Logistic Regression	85.4%
SVM	84.8%

For accuracy, table 1 shows that logistic regression has the highest accuracy, then SVM, random forest, and naive bayesian have pretty good accuracy. However, the K-means and Zero R have low accuracy because of the reason that is mentioned in the above sections. The K-means algorithm does a fantastic job at capturing the structure of data clusters if they resemble spheres. This suggests that when the clusters contain a complex geometric system, K-means is ineffective at categorizing the data.

Table 2 Time and Space Complexity for Each Algorithm

	Time complexity	Space complexity
Random Forest	O(k * m)	O(p * m)
Naive Bayesian	O(d * c)	O(d * c)
K-means	O(n * K * I * d)	O((n + K) * d)
Zero R	O(n)	O(n)
Logistic regression	O(d)	O(d)
SVM	O(s * d)	O(s)

n = number of points

d = dimensions

k = depth of the tree

m = number of decision trees

p = number of nodes in a tree

c = number of classes

K = number of clusters

I = number of iterations

s = number of support vectors

Table 2 shows that Zero R and Logistic regression have good efficiency, and K-means have low efficiency and enormous space complexity. Zero R, Logistic regression, and SVM have relatively small space complexity.

5. Conclusion

In this project, the adult dataset is analysised, and exploratory data analysis is done to visualise the data, and data cleaning is performed for future algorithm implementation and result validation. Two papers' result are reproduced and got similar results. The first paper introduced a random forest classifier and ranked the attributes according to the importance score. The second paper compared four different classifier algorithms' accuracy. Also, two different algorithms has been inroduced, logistic regression and SVM, both representing a relatively high accuracy. The fitted classifier model's results indicate that factors such as marital status, capital gains, education, age, and work hours account for a significant portion of the difference between low and high-income levels. Of six algorithms, logistic regression performs the best for classification and prediction, it has the highest accuracy, high efficiency, and low space complexity. K-means performs the worst, it has the lowest accuracy, low efficiency, and high space complexity. Although K-means brings unsupervised learning for this dataset, the principle of the working may not suitable for the high-demensional dataset. Looking ahead, other data mining methods like deep learning and association can be used in conjunction with the current study. It may be expanded to include other classification algorithms. There are still many social problems can be handled by applying the model, with the deepening of the study about the income level prediction, some visionary application directions will be created.

6. References

Bekena, S. M. (2017). *Using decision tree classifier to predict income levels*.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Deepajothi, S., & Selvarajan, S. (2012). A comparative study of classification techniques on adult data set. *International Journal of Engineering Research & Technology (IJERT)*, *1*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer series in statistics. *Springer New York*.

jain, apurv. (2020, September 25). Support Vector Machines(S.V.M)—Hyperplane and Margins. *Medium*. https://medium.com/@apurvjain37/support-vector-machines-s-v-m-hyperplane-and-margins-ee2f083381b4

K-means clustering. (2022). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=K-means_clustering&oldid=1127979478

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, *3*(22), 41–46.

Tong Wu. (2022). *Electric Bike Performance Enhancement – Security System*. The University of Manchester.

UCI Machine Learning Repository: Adult Data Set. (n.d.). Retrieved 19 December 2022, from https://archive.ics.uci.edu/ml/datasets/Adult

Variable importance plot (mean decrease accuracy and mean decrease Gini). (2020, April 1). [Figure]. Figshare; PLOS ONE. https://doi.org/10.1371/journal.pone.0230799.g002