Local-vs-global models for fairness-aware classification

By

Gul Jabeen

M.Sc in Internet Technologies and Information Systems (ITIS)
Leibniz University Hannover
Hannover, Germany

Project Supervisor: Prof. Eirini Ntoutsi
Approval Professor: Prof. Dr. Wolfgang Nejdl

l3s Research Center
Appelstrasse 4
30167 Hannover, Germany
Email:gul.jabeen@stud.uni-hannover.de

February 05,2019

Contents

1 Introduction

In the past few years, data analysis has become an important part of the world. Besides this, the growth of data, storage for new technologies and processed data is rapidly increasing. Machine learning helps to extract and store the data, discover patterns and concepts in the data which generalized future results.

Machine learning provides different techniques to extract information from raw data in the databases and data sets. This process contains several steps: selection, preprocessing, transformation of learning algorithms, data mining techniques and evaluation. There are two main fields of machine learning which are supervised learning and unsupervised learning. In supervised learning, the labels are known whereas, in unsupervised learning, labels are not known. The algorithms developed for classification are decision trees, naïve Bayes classifier, K-nearest neighbour algorithm, Logistic Regression, Support Vector Machine (SVM) whereas in clustering are k-means, DB-scan, agglomerative clustering and many more.

The objective of my report is to detect and remove the errors to increase the quality of data. To partition the data into a pre-defined number of clusters using K-means or K-modes then train a model for making income level predictions using classification algorithms. We will apply a global classifier against clustering to improve the quality of classification. To design algorithms that make fair predictions which avoid disparate impact and unfair results. Sensitive attributes must be examined like race and gender.

2 Dataset

2.1 Description of dataset

The dataset used in this project is the census adult dataset [1] which contains 49,000 records and a binary label determining whether a person's salary is greater than 50K a year. In the given dataset, a class label of <50K contains 76% of the records in comparison to >50k which contains 26% of the records.

There are 14 attributes consisting of 6 numeric, 2 binary and 6 non-numeric attributes (Table 1). The employment attribute contains the following types of employers such as self-employed or federal. Occupation attribute specifies the employment types such as farming or managerial. The education attribute consists of the highest level of education like high school graduate or doctorate. Education number is a numeric form of education attribute from range 1 to 13. The relationship attribute has categories involving unmarried, married or separated.

Other non-numeric attributes are country of residence, gender and race. The numeric attributes are age, hours worked per week, capital gain, capital loss and a survey weight attribute which is a demographic score assigned to an individual based on information such as the area of residence and type of employment.

2.2 Data set understanding

Table (1) describes a percentage of all features of the adult dataset along with the number of missing values. In the work class attribute, the percentage of private is higher than other features of the attribute. The relationship attribute percentage of the husband is higher with respect to other features of the relationship attribute.

Similarly, the race has a higher percentage of white people, and the native country contains 89% of United States features. In salary attribute, ≤ 50 has a higher percentage than ≥ 50.

According to Table (2), the standard deviation in the numerical data indicates that there is a significant number of values in each attribute and there are no missing values in numeric data.

Whereas, Figure (1) defines the range of each attribute in the data set for Table (2) by using box plots which explains the range of each numeric attributes.

Table 1: Description of features: Adult dataset

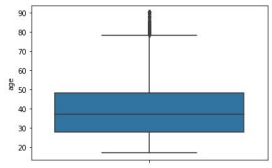
Attributes	**Values**	**Missing Values**
Work Class	Private(69.70%), Self-emp-not-inc(7.80%), Local-gov(6.43%), (5.64%), State-gov (3.99%), Self-emp-inc (3.43%), Federal-gov (2.95%), Without-pay (0.04%), Never-worked (0.02%)	1836
Education	HS-grad (32.25%), Some-college (22.39%), Bachelors (16.45%), Masters (5.29%), Assoc-voc (4.24%), 11th (3.61%), Assoc-acdm (3.28%), 10th(2.87%), 7th-8th (1.98%), Prof-school (1.77%), 9th (1.58%), 12th (1.33%), Doctorate (1.27%), 5th-6th (1.02%), 1st-4th (0.52%), Preschool (0.16%)	0
Relationship	Husband (40.52%), Not-in-family (25.51%) Own-child (15.56%),Unmarried (10.58%), Wife (4.82%), Other-relative (3.01%)	0
Race	White (85.43%), Black (9.59%), Asian-Pac-Islander (3.19%) Amer-Indian-Eskimo (0.96%), Other (0.83%)	0
Marital Status	Married-civ-spouse (45.99 %), Never-married (32.81%), Divorced (13.65%), Separated (3.15%), Widowed (3.05%), Married-spouse-absent(1.28%), Married-AF-spouse (0.07%	0
Occupation	Prof-specialty (12.71 %), Craft-repair (12.59 %), Exec-managerial (12.49 %),Adm-clerical (11.58 %), Sales (11.21 %), Other-service (10.12 %), Machine-op-inspct (6.15 %), ? (5.66 %), Transport-moving (4.90 %),Handlers-cleaners (4.21 %), Farming-fishing (3.05 %), Tech-support (2.85 %), Protective-serv (1.99 %), Priv-house-serv (0.46 %), Armed-Forces (0.03 %	1843
Native-Country	42 categories- United states(89.59 %)	583
Salary	<=50K (75.92%), >50K (24.08%)	0
Gender	Male (66.92 %), Female (33.08 %)	0

Table 2: Description of Numeric features: Adult dataset

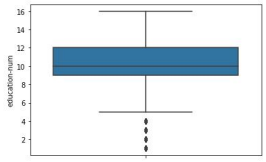
Attributes	**Mean**	**Median**	**Std. Dev**	**Range**	**Missing Values**
Age	38.58	37	13.64	17-90	0
Hours-per-week	40.44	40	12.35	1-99	0
Edu -Num	10.08	10	2.57	1-16	0
Capital gain	1078	0	7385	0-99999	0
Capital loss	87.3	0	403	0-4356	0

Figure 1: Box Plots of Numeric attributes

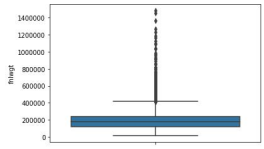
(a) Age



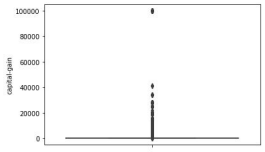
(b) Education Number



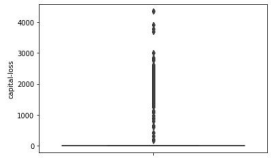
(c) fnwgt



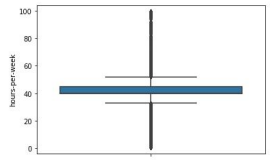
(d) Capital Gain



(e) Capital loss



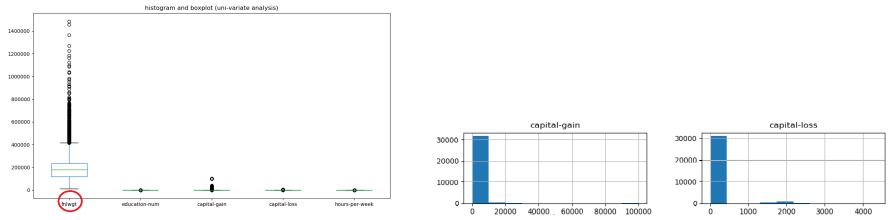
(f) Hours per week



2.3 Data cleaning

According to Figure (2a), the survey weight attribute has 21,000 unique values out of 31,000 instances, which may suggest that this attribute may not be significantly predictive.

Whereas Figure (2b), indicates that in the capital gain and capital loss attributes, mostly values are equal to 0 or either $\geq$ zero. Therefore, we can delete the survey attribute and also we can set the capital gain to 0 and 1. If a value is equal to 0, then set 0 otherwise 1.



(a) Boxplot of numerical data

(b) Histogram of numerical data

Figure 2: Verification of data quality

Figure (3) shows that an education number is a numeric form of education attribute in the dataset. The highest educational value is 14 whereas 1 is the lowest education value. Therefore, we can remove education attribute from the dataset.

education	education-num
Bachelors	13
Bachelors	13
HS-grad	9
11th	7
Bachelors	13
Masters	14
9th	5
HS-grad	9
Masters	14

Figure 3: Education Number is numeric form of education attribute

3 Data Preprocessing/ Data Cleaning

Data cleaning is the process of detecting and removing missing values and inconsistencies from a data set in order to improve its quality.

As shown in Table (3), there are missing values in the following attributes; work class, occupation and native country. We need to remove these missing values from the data set in order to get correct results.

Attributes	**Missing Values**
age	0
workclass	1836
education	0
education-num	0
marital-status	0
occupation	1843
relationship	0
race	0
gender	0
capital-gain	0
capital-loss	0
hours-per-week	0
native-country	583
salary	0

Table 3: Data Processing

After getting all missing values in the dataset, we use dropna function to remove all instances with missing values. Table (4) shows the dataset values after cleaning.

Data Set	**Values**
Before Cleaning	(32561, 14)
After Cleaning	(30162, 14)

Table 4: Before and after Cleaning of data

4 Feature transformation to apply K-means clustering algorithm

Feature transformation (FT) is a transformation of non-numeric features into numeric features in a dataset. We convert all features into numeric form because k-means doesn't work on non-numeric and categorical data. It cant take the mean of non-numeric data. This approach helped us to apply k-means. Also, after getting results, we reverse the transformation to non-numeric features to interpret the results. That's the reason this approach was highly effective.

In Figure (4), we have converted all the features into numeric features for applying the k-means algorithm. That will give accurate and better results.

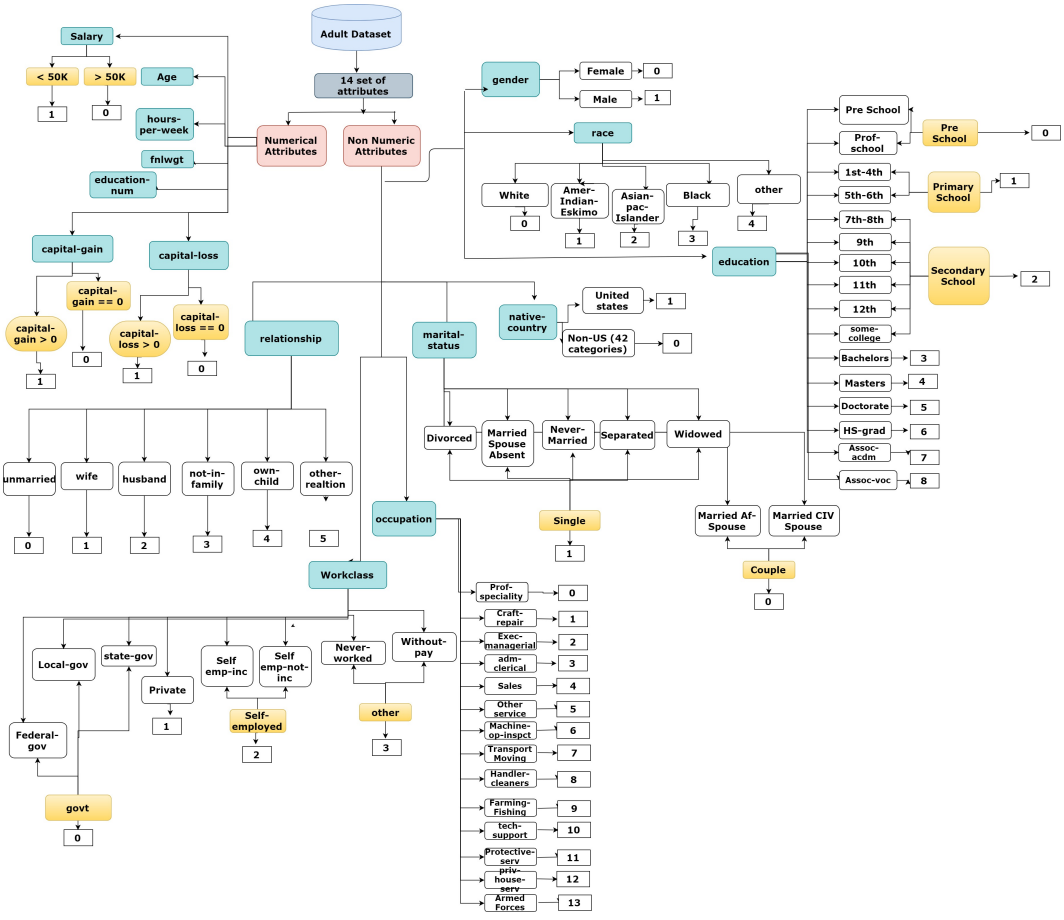


Figure 4: Conversion of Categorical data/non-numeric data into numeric data

5 Data Analysis

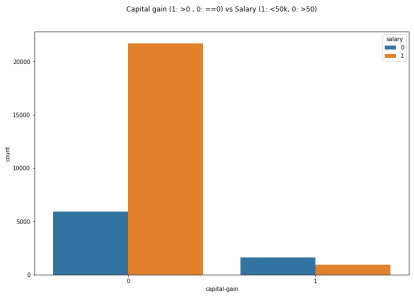
In this section, we have explained two types of analysis, univariate analysis and bivariate analysis.

5.1 Uni-Variate Analysis

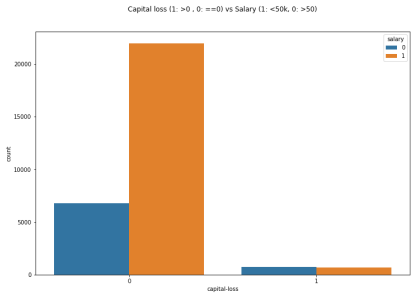
Univariate analysis has only one variable which doesn't deal with causes or relationships, and its major purpose is to describe the data in terms of a single feature. In all the figures below, class 0 shows salary > 50 whereas class 1 shows salary ≤ 50k.

In Figures (a) & (b), capital gain and capital loss contain 0 and 1 features. In Figure (a), capital gain (0) has a higher percentage of salary ≤ 50k compared to capital gain (1) contains fewer features but it has a higher percentage of salary > 50k.

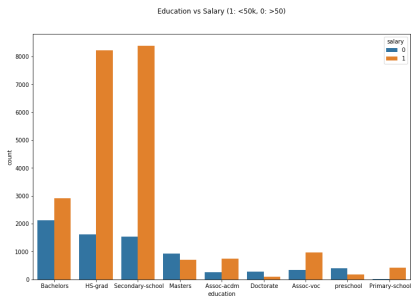
In Figure (c), education types like bachelors, HS-grad, and secondary school have mostly salary > 50k. Although in Figure (d), males on an average earn more than females.



(a) Capital Gain



(b) Capital Loss

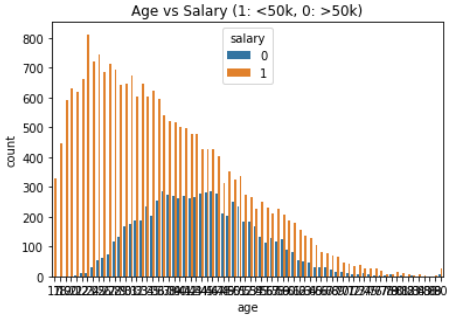


(c) Education

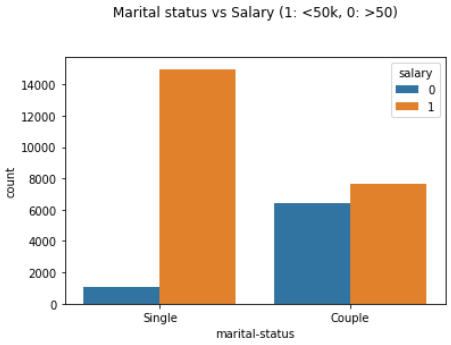
(d) Gender

Figure (a) shows younger people are paid less than older people whereas Figure (b) presents married citizens with a spouse have a higher chance of earning more than those who are unmarried/divorced/widowed/separated.

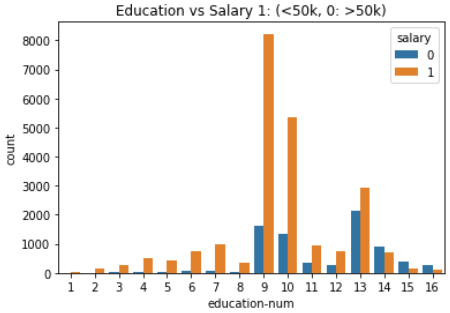
Figure (c) demonstrates that higher education can lead to higher income in most cases. Figure (d) indicates that mostly dataset is populated with the US population.



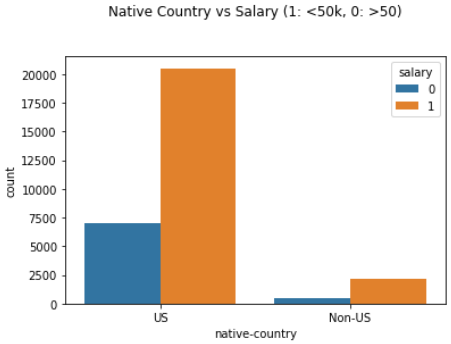
(a) Age



(b) Marital status



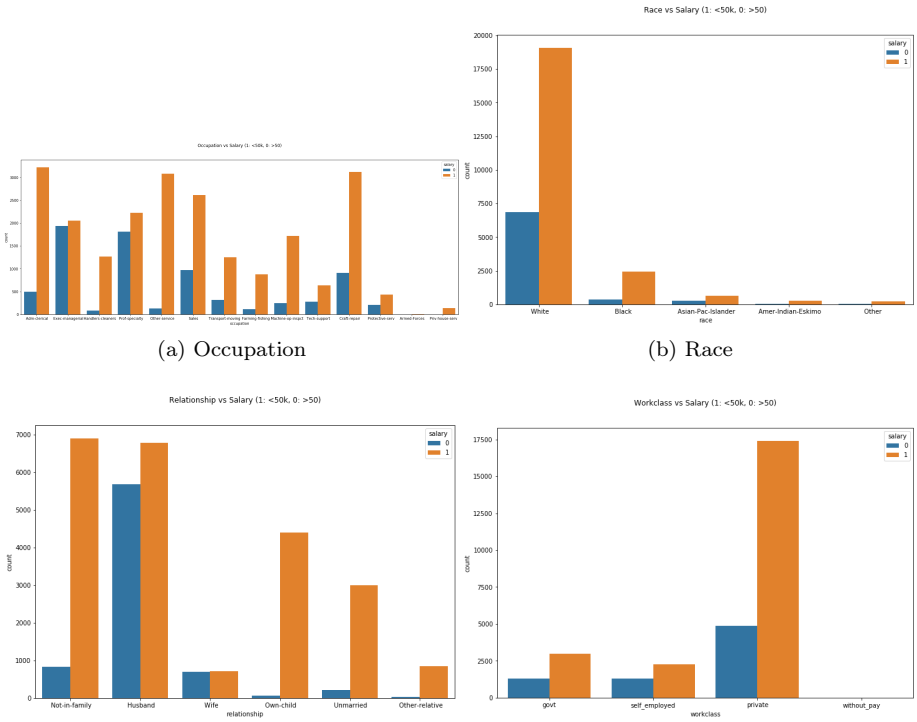
(c) Education



(d) Native Country

Figure (a) indicates exec-managerial, prof-speciality and sales have a higher income than other occupation types. Figure (b) presents Asian-Pacific-Islanders and white are two races that have the highest average income.

Figure (c) shows that husbands are getting paid more than other relationship categories whereas 'own-child' type of relationship is getting the least salary. Figure (d) indicates the higher salary income rate is higher on private compared to other work class attribute types.



(a) Occupation

(b) Race

(c) Relationship

(d) Work Class

Figure 5: Uni-variate Analysis of all attributes vs Salary

5.2 Bi-Variate Analysis

Bivariate analysis is the analysis of two variables which explores a relationship or association between two variables. Even if there exists an association and the strength of this association, or even if there are differences between two variables and the significance of these differences.

Below Figure (6) shows the correlation between attributes.

— Salary and Marital-status are correlated (0.45)
— Relationship and Marital-status are correlated (0.35)

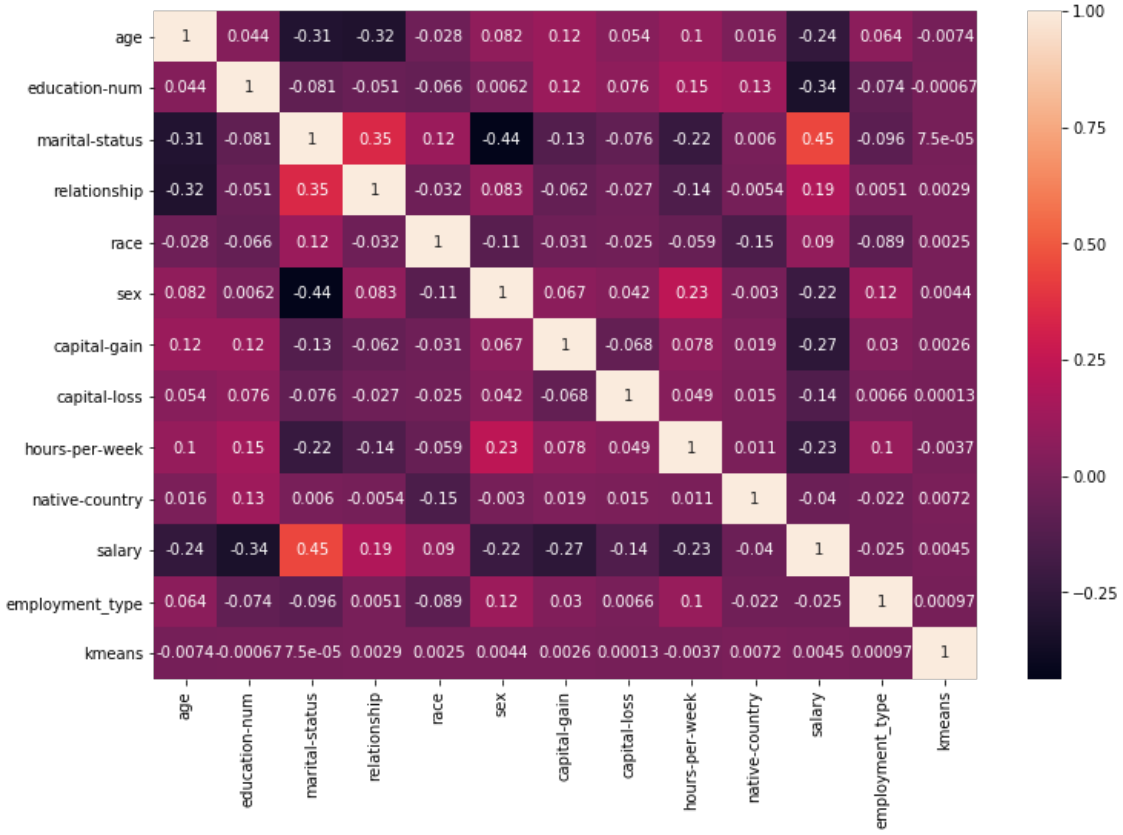


Figure 6: Correlation Matrix of all attributes of an adult dataset

6 Clustering

In this section, first, we explained three methods for the selection of a number of clusters (k) for applying k-means clustering. Second, we have applied PCA (Principal Component Analysis) before and after k-means in order to check the performance of our algorithm in both terms. Third, we have done some cluster analysis. Finally, we showed results after applying clustering algorithms by comparison of k-means and k-modes clustering algorithms.

6.1 K-Means Algorithm

K-Means Algorithm is the type of partitioning clustering algorithm. The initial K value is known in the K-mean clustering to get a cluster from database D of n points. the k-means algorithm has four steps for implementation:

— Randomly picks k-objects as cluster centers which can be represented as c1, . . . , ck.
— Assign the rest of the points to their closest cluster centers.
— Update the center of each cluster based on the new point assignments.
— Repeat until convergence.

6.1.1 Selection of K, the number of clusters

1. **Method 1: Elbow Method [2]**

   Elbow method runs the k-means clustering algorithm on each value of k. For each k, it calculates the sum of squared errors (SSE). Then, draws a line chart of the SSE on each value of k, if a line chart makes an arm with an elbow then that's the best value to be chosen for K.

   SSE should be small, and the SSE tends to decrease toward 0 as we increase k. The goal is to select a small value of k which have a low SSE, and the elbow represents where we start to have weakened returns by increasing k.

But Sometimes we get such results as shown in Figure (7) in which elbow is not clear that's why we used dendrogram and silhouette coefficient to select K.

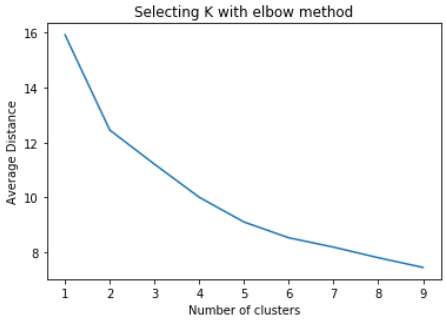


Figure 7: Elbow Method

2. **Method 2: Dendrogram**[4]

Hierarchical clustering is the composition of many clusters on each layer.It makes a cluster tree called dendrogram which represents data, where each node or group is associated with two or more groups. The groups are nested and organized as a tree where each node in the cluster tree contains a group of similar data.

As shown in Figure (8), the two blue clusters are the largest clusters in the data set, whereas three clusters (green,red,teal) are quite separated from each other. Then each cluster is composed of more sub-clusters. Therefore two or three clusters could be the best selection for K.

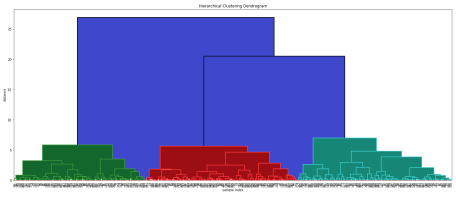


Figure 8: Dendrogram: Hierarchical Clustering

3. **Method 3 : Silhouette Coefficient**[3]

Silhouette analysis is a way to measure how close each point in a cluster is to the points in its neighboring clusters. It's a neat way to find out the optimum value for k during k-means clustering. Silhouette values lie in the range of [-1, 1]. A value of +1 indicates that the sample is far away from its neighboring cluster, and very close to the cluster its assigned. Similarly, the value of -1 indicates that the point is close to its neighboring cluster than to the cluster it's assigned. And, a value of 0 means it's at the boundary of the distance between the two clusters. Value of +1 is idea and -1 is least preferred. Hence, higher the value better is the cluster configuration.[3]

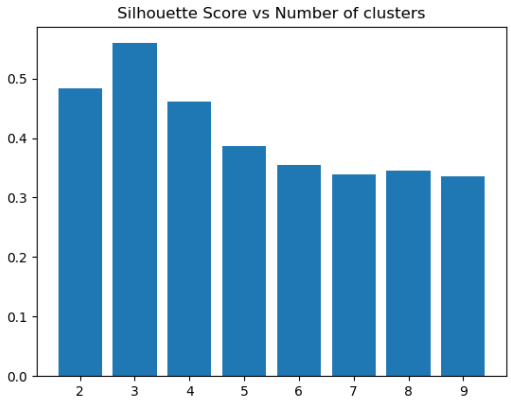


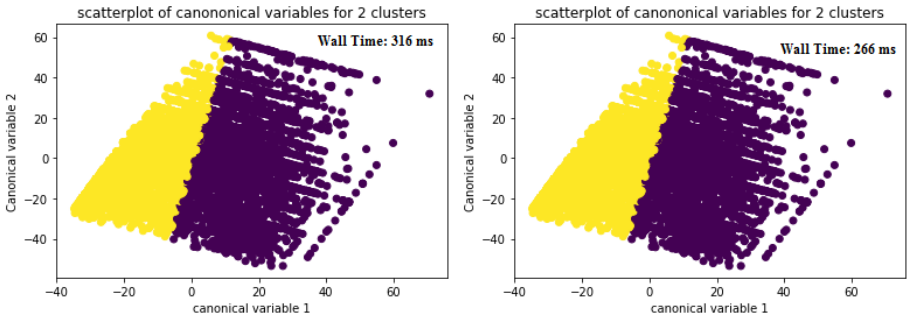
Figure 9: Silhouette Coefficient

6.1.2 Applying PCA before and after K means algorithm

PCA (principal component analysis) is the statistical method for dimension reduction. In this section, we explain the benefits of applying dimension reduction before and after clustering algorithms.

Figure 10(a) shows the computing cost of applying PCA after k-means is 316s compared to the computation cost of applying PCA before k means clustering is 266s as shown in Figure 10(b).

PCA (Principal Component Analysis) should apply before a clustering algorithm (such as k-means). It improves the clustering results along with performance. It helps to reduce noise in the dataset before applying k-means. Besides this, it also reduces the dimensions and decreases the computation cost. On the other hand, its performance depends upon the distribution of the dataset and correlation features. So if you want to cluster the data on the bases of many features, then using PCA before clustering is very reasonable and time-saving.



(a) Applying PCA after K means

(b) Applying PCA before K means

Figure 10: Applying PCA before and after K means algorithm

6.1.3 Cluster Analysis

1. **Total no. of features in each cluster**
   Table (5) shows the division of clusters after applying k means clustering algorithm. Cluster 0 contains 16,283 features considering that cluster 1 consists of 13879 features.

Clusters	Total no. of features
Cluster 0	16283
Cluster 1	13879

Table 5: Cluster Description

2. **Cluster description**
   Table (6) gives a description of each cluster. The top represents the most occurring attribute type in each cluster and frequency shows the number of features with respect to each attribute type in each cluster.

Attribute	Clusters	Top	Frequency
Occupation	1	Exec-managerial	2428
	0	Adm-clerical	2242
Education	1	HS-grad	4427
	0	Secondary School	5881
Age	1	40s	6540
	0	20s	8211
Marital status	1	Couple	16293
	0	single	10725
Relationship	1	Husband	7808
	0	Husband	4655
Race	1	White	12098
	0	white	13835
Gender	1	Male	10016
	0	Male	10364
Native-Country	1	US	12741
	0	US	14763
WorkClass	1	private	8976
	0	private	13310

Attribute	Clusters	mean	std	min	25%	50%	75%
Hours-per-week	1	43	12	1	40	40	50
	0	38	10	1	35	40	40

Table 6: Cluster Description

3. **Visualization of clusters with respect to all features**

In this section, we have compared each cluster with all features of the adult dataset. We used histogram for the analysis of our clusters with respect to each attribute.

(a) **Clusters vs Age**

Cluster 0: Young people
Cluster 1: Old people

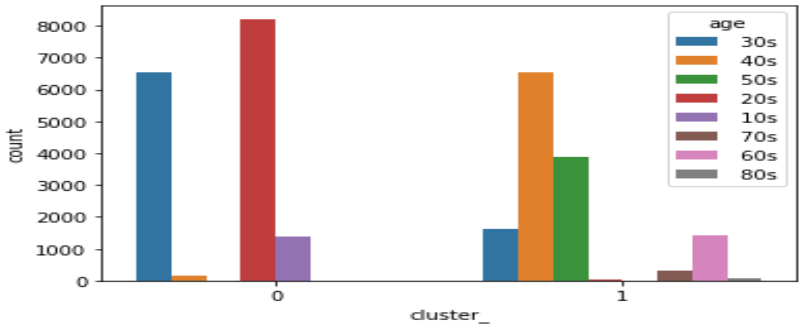


Figure 11: Cluster vs Age

(b) **Clusters vs Education Number**

Cluster 0: High Education Rate
Cluster 1: Medium Education Rate

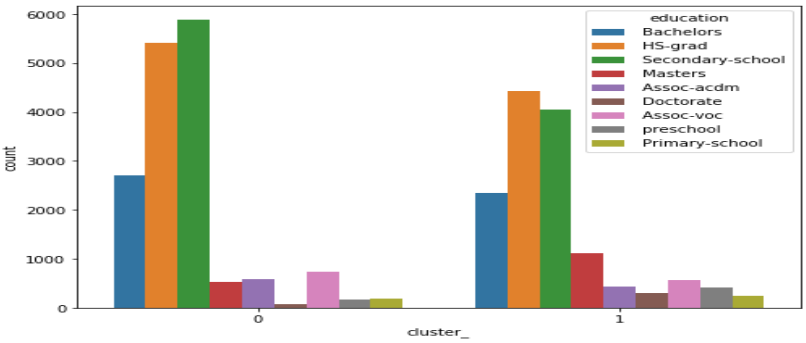


Figure 12: Cluster vs Education

(c) **Clusters vs Gender**

Cluster 0: High rate of male population and few women participation
Cluster 1: High rate of male population.

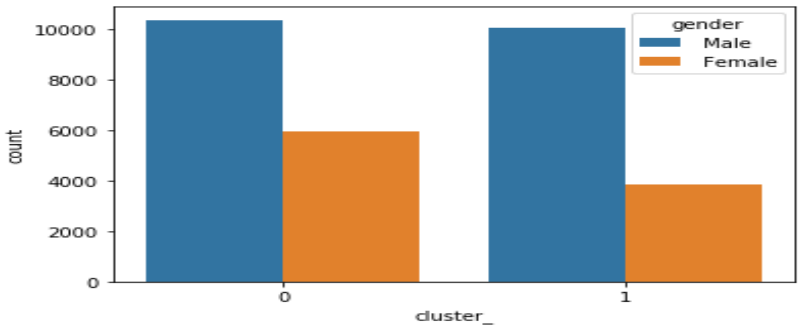


Figure 13: Cluster vs Gender

(d) **Clusters vs Marital status**

Cluster 0: High rate of single People
Cluster 1: High rate of couples

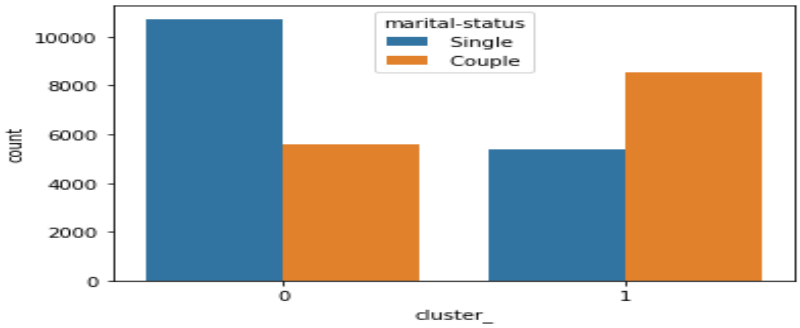


Figure 14: Cluster vs Marital status

(e) **Clusters vs Native Country**

Cluster 0: Highly populated with US-based people whereas Non-US participation is really low.
Cluster 1: Least participation of non-US people

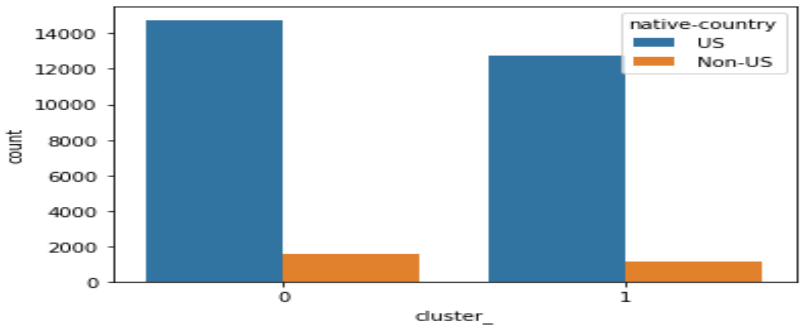


Figure 15: Cluster vs Native Country

(f) **Clusters vs Occupation**

Cluster 0: Second high occupation rate (Max. Exce-managerial, Prof Speciality, Craft repair)
Cluster 1: Moderate occupation rate (Max. Exce-managerial)

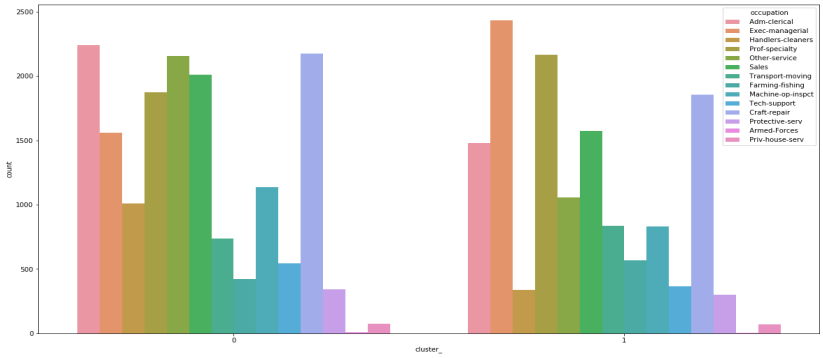


Figure 16: Cluster vs Occupation

(g) **Clusters vs Race**

Cluster 0: White people and few black people
Cluster 1: Overly white people society

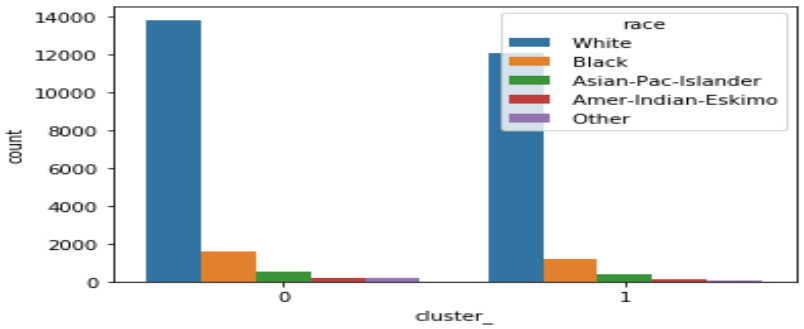


Figure 17: Cluster vs Race

(h) **Clusters vs Relationship**

Cluster 0: Moderate population of each category
Cluster 1: Mostly husbands

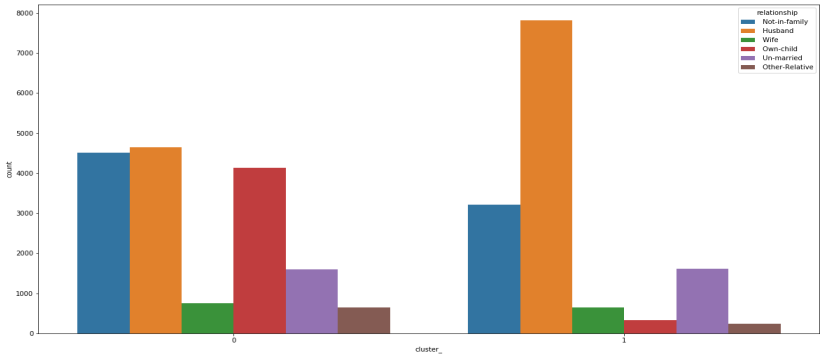


Figure 18: Cluster vs Relationship

(i) **Clusters vs Salary**

Cluster 0: Lower salaries (less than 50K)
Cluster 1: Lower and greater salaries both (less or greater than 50K)

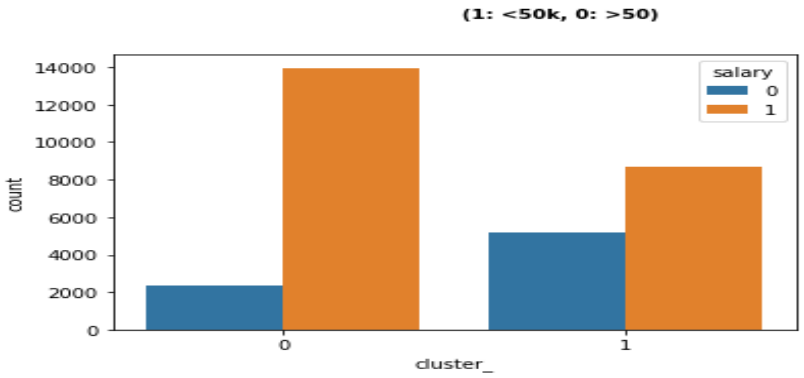


Figure 19: Cluster vs Salary

(j) **Clusters vs Workclass**

Cluster 0: Private class
Cluster 1: Few self-employed but mostly private class

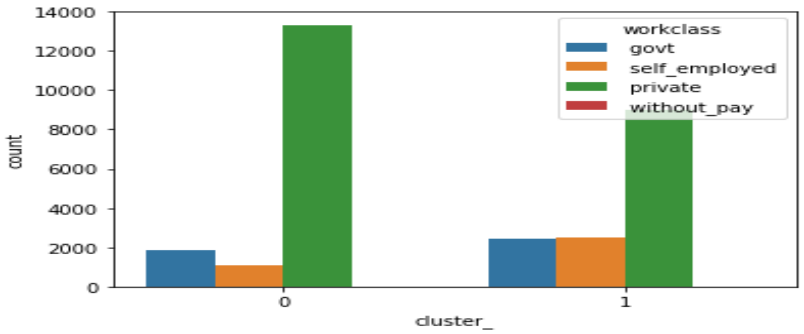


Figure 20: Cluster vs Work Class

6.1.4 K Means Limitations and their Solutions

K means performance can be improved by removing all these limitations. Following are the solutions for k means clustering w.r.t to each limitation:

1. **K means algorithm doesn't work on categorical data**

   Solution 1: Feature Engineering

   We can map data to 0/1 values but it cannot generate quality clusters for high dimensional data.

   Solution 2: K Mode Algorithm

   K-mode is the extension to the k-means by replacing the mean of clusters with modes. K-mode clustering algorithm can handle the categorical data. This technique provides better accuracy and effective results for the categorical data. It finds the dissimilarity measure for categorical objects and modes instead of their means which helps to find similar objects that make clusters.

2. **Always need to specify K**

   Solution1: Silhouette Coefficient
   Solution2: Apply other algorithms of clustering like dbscan, Spectral clustering

3. **Unable to handle noisy data**

   Solution: Apply PCA

   It is a method to apply PCA (Principal Component Analysis) before a clustering algorithm like kmeans. It improves the clustering results in practice such as noise reduction and dimension reduction.

4. **Not suitable to discover the cluster with non-convex shapes**

6.2 K-Mode Clustering

K-modes work well on a large data set which contains categorical data. In order to check the quality of my results on K means by using the feature engineering technique, we have applied the k-mode Algorithm to compare my result with a k-means algorithm to check whether my feature transformation was good or not.

6.2.1 Comparison of Cluster's Description of K means and Kmode

In Table (6), we have compared top features of both clustering algorithms k-means and k-mode which shows quite similar results.

Attribute	Clusters	Top (K-means)	Top (K-modes)
Occupation	1	Exec-managerial	Craft Repair
	0	Adm-clerical	Adm-clerical
Education	1	HS-grad	HS-grad
	0	Secondary School	Secondary School
Age	1	40s	30s
	0	20s	20s
Marital status	1	Couple	Couple
	0	Single	Single
Relationship	1	Husband	Husband
	0	Husband	Not-in-family
Race	1	White	White
	0	White	White
Gender	1	Male	Male
	0	Male	Female
Native-Country	1	US	US
	0	US	US
WorkClass	1	Private	Private
	0	Private	Private

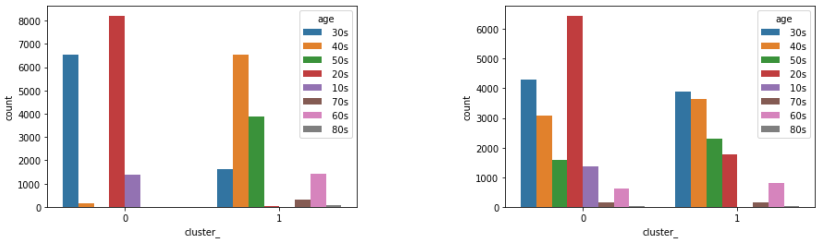
Table 7: Comparsion of Clusterś Description of Kmodes and K means

6.2.2 Visualization Comparison of K-means and K-modes

In this section, we have visualized compared both K-means and K-modes clustering algorithms. The results were quite similar for both algorithms as shown in the figures below.

1. **Age**

   Cluster 0: Young people
   Cluster 1: Old people

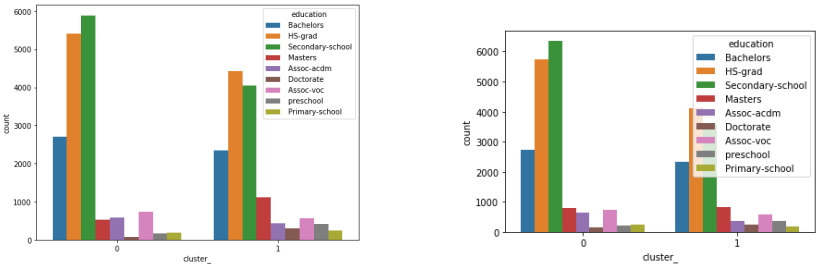


(a) K-means-Age (b) K-modes-Age

Figure 21: Visualization Comparison of K-means and K-modes w.r.t Age attribute

2. **Education**

   Cluster 0: High education rate
   Cluster 1: Medium education rate

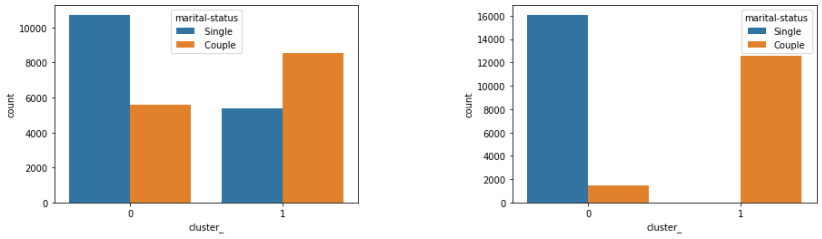


(a) K-mean-Education (b) K-modes-Education

Figure 22: Visualization Comparison of K-means and K-modes w.r.t Education attribute

3. **Marital-Status**

Cluster 0: Mostly single people
Cluster 1: Mostly couples

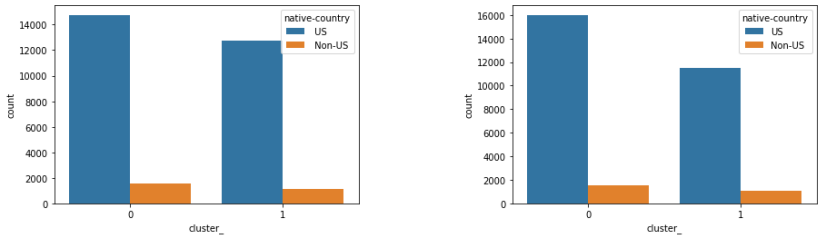


(a) K-means-Marital status (b) K-modes-Marital status

Figure 23: Visualization Comparison of K-means and K-modes w.r.t Marital Status attribute

4. **Native-Country**

Cluster 0: Highly populated with US based people whereas Non-US participation is really low.
Cluster 1: Least participation of non-US people

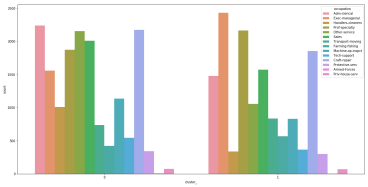


(a) K-means-Native Country (b) K-modes-Native Country

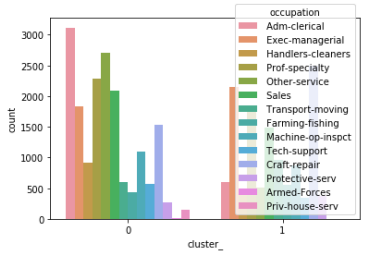
Figure 24: Visualization Comparison of K-means and K-modes w.r.t Native-Country attribute

5. **Occupation**

Cluster 0: Second high occupation rate (Max. Exce-managerial, Prof-Speciality, Craft repair)
Cluster 1: Moderate occupation rate (Max. Exce-managerial)



(a) K-means-Occupation

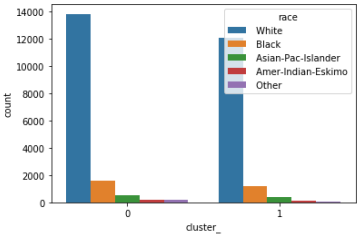


(b) K-modes-Occupation

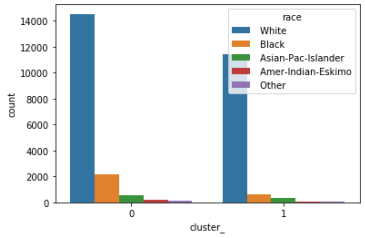
Figure 25: Visualization Comparison of K-means and K-modes w.r.t Occupation attribute

6. **Race**

Cluster 0: White people and few black people
Cluster 1: Overly White people society



(a) K-means-Race

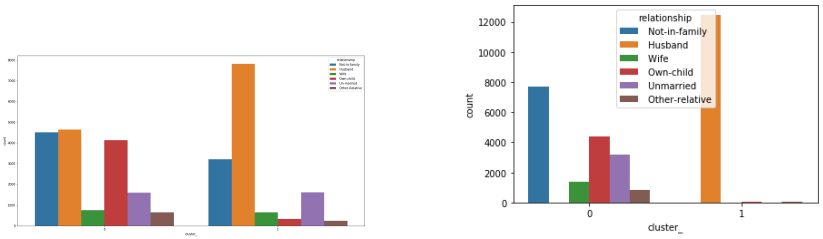


(b) K-modes-Race

Figure 26: Visualization Comparison of K-means and K-modes w.r.t Race attribute

7. **Relationship**

Cluster 0: (No husbands) high rate of 'not in the family', 'own-child', and "unmarried"
Cluster 1: Mostly husbands

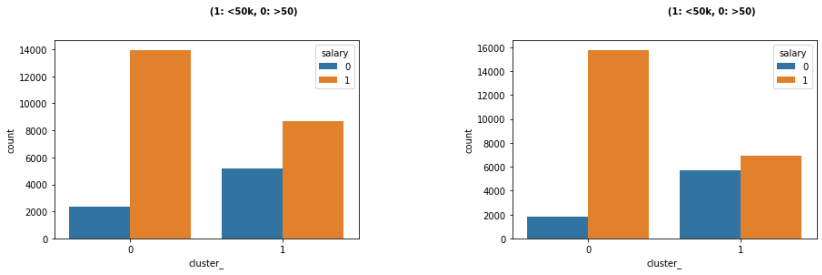


(a) K-means-Relationship (b) K-modes-Relationship

Figure 27: Visualization Comparison of K-means and K-modes w.r.t Relationship attribute

8. **Salary**

Cluster 0: Low salaries (less than 50K)
Cluster 1: High salaries (greater than 50K)

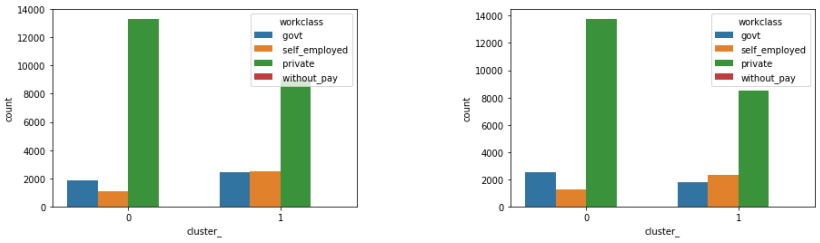


(a) K-mean-Salary (b) K-modes-Salary

Figure 28: Visualization Comparison of K-means and K-modes w.r.t Salary attribute

9. **Work-Class**

Cluster 0: Private class
Cluster 1: Few Self employed but mostly private class



(a) K-means-WorkClass

(b) K-modes-WorkClass

Figure 29: Visualization Comparison of K-means and K-modes w.r.t Work-Class attribute

7 Labeling the Clusters

After applying k means and k modes, we finally got the comparison of two clusters. Table (8) describes the results of both clusters 0 and 1. These results are generated by comparing k-means, and k-mode clustering algorithm. Cluster 1 and 0 contain instances (people) of (mainly) the following properties:

Features	**Cluster1**	**Cluster 0**
Age	Old people	Young people
Gender	High rate of male population	Both (male and female)
Native-Country	US-populated mostly	US-populated mostly
Marital Status	Mostly couple	Mostly single
Relationship	Mostly husbands	(No husbands) high rate of 'not in the family', 'own-child', and 'unmarried'
Salary	High salaries	Low salaries
Education-num	Low education rate	High education rate
Gender	Male-populated	Male-populated

Table 8: Comparison of Clusters

According to the following characteristics, we renamed clusters as

- Cluster 1 as higher_salary_couples
- Cluster 0 as lower_salary_single

8 Combining K means and Classification

In this section, first, we have applied global classifier (logistic regression) on the adult dataset. Second, we have applied global classifier against k-means clustering in order to improve the performance of classification.

8.1 Global classification model

Here are the results after applying a logistic regression algorithm which managed to achieve an F1 score of 0.57. This is an average score.

	Logistic Regression
Accuracy	0.82
Confusion Matrix	Cor. Pred: 6464+1010 = 7474 Incor. Pred :1200+375= 1575
Precision	0.81
Recall	0.82
F1-score	0.550960

Table 9: Global Classifier Performance Measures [7]

In the following figure, our correct predictions are equal to 7474 compared to incorrect predictions of 1575.

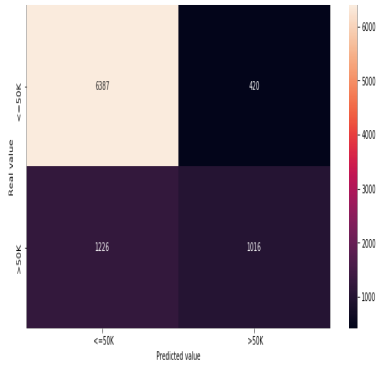


Figure 30: Confusion matrix

In below Figure (31) the features that seem to contribute most positively to have an income of more than $50K are Capital Gain, Education-Num and Gender, while the features that contribute most negatively are Marital Status and Relationship.

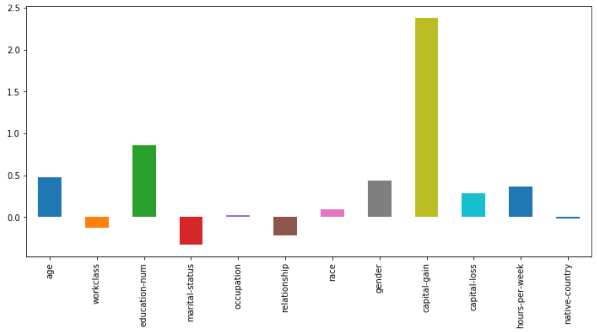


Figure 31: Logistic Regression Classification Results [6]

8.2 Local classification models

In order to improve classification, we will apply k means first and get the clusters then apply global classifier against k-means. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

In this local classification model, the independent variable is a cluster which has been extracted by applying k means algorithm in section 6.1 whereas remaining attributes will be dependent variables. This approach gives better performance measures compared to the global classification model.

8.2.1 Local Classification model 1

1. **Performance Measures**

   The F1 score has improved to 0.93 which is near to best value 1 compared to global classification model which managed an F1-score near 0.6.

	LR-classifier against clustering
Accuracy	0.99
Confusion Matrix	Cor. Pred: 4667+3818 = 8485 Incor. Pred : 344+220= 564
Precision	0.99
Recall	0.99
F1-score	0.93

Table 10: Performance Measures of Local Classifier 1 [7]

*notes*To quote from Scikit Learn: The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier to not label a sample as positive if it is negative. The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.The F-beta score weights the recall more than the precision by a factor of beta. beta = 1.0 means recall and precision are equally important. The support is the number of occurrences of each class in y_test.

2. **Summary of Results**

As you can see managed to achieve an F1 score of 0.93 which show the classification worked well compared to global classification model which managed to achieve F1 of 0.5, and the explained features that seems to contribute most positively (to cluster 1) to are old people, working as private and their salaries are greater than 50k, while the features that contribute most negatively (in cluster 1) are young people and fewer salaries.

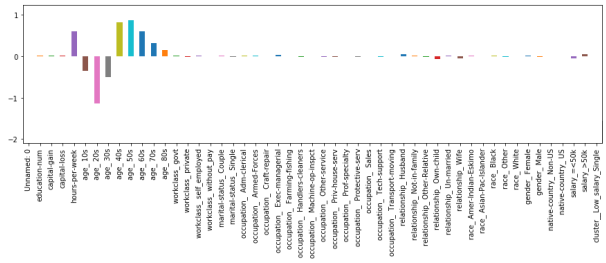


Figure 32: Summary of Results for Cluster1 as higher_salary_couples

8.2.2 Local Classification model 2

1. **Performance Measures**

The F1 score of local classification model 2 is the same as the local classification model 2. It has improved to 0.93 which is near to the best value 1 compared to global classification model which managed an F1-score near 0.6.

	LR-classifier against clustering
Accuracy	0.99
Confusion Matrix	Cor. Pred: 4667+3818 = 8485 Incor. Pred : 344+220= 564
Precision	0.99
Recall	0.99
F1-score	0.93

Table 11: Performance Measures of Local Classifier 2 [7]

*notes*To quote from Scikit Learn: The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives. The precision is intuitively the ability of the classifier to not label a sample as positive if it is negative. The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0.The F-beta score weights the recall more than the precision by a factor of beta. beta = 1.0 means recall and precision are equally important. The support is the number of occurrences of each class in y_test.

2. **Summary of results** Whereas cluster 0 name as low_salary single mostly contains single young people with fewer salaries. As you can see in the below results.

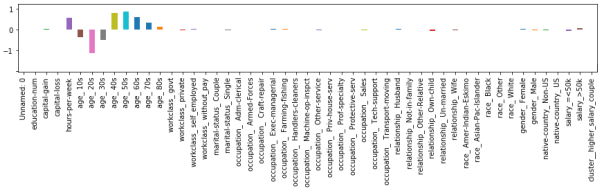
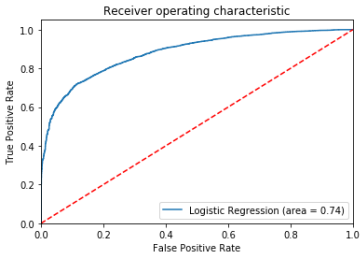


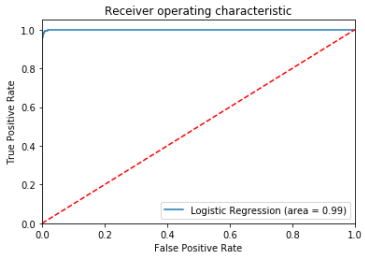
Figure 33: Summary of Results for Cluster 0 as lower_salary_single

8.2.3 ROC Curve

The receiver operating characteristic (ROC) curve is a tool which helps to identify the performance of our binary classifier. The dotted line shows the ROC curve of a random classifier. A good classifier stays distant from that line as possible (toward the top-left corner).



(a) LR Classifier ROC Curve



(b) LR Classifier against K-Means ROC Curve

9 Fairness

Machine learning models are used in an everyday routine now like recommendation systems, decision makings and evaluations. These models are impacting on our society because it could affect others; the decision shouldn't be biased against someone. These matters should be concerned with machine learning models. After applying models, it is necessary to check whether our predictions are right or not and without any discrimination or injustice.

9.1 Remove Sensitive attributes from the dataset

Let's train the classifier which predicts whether the salary is >50k or $\leq$ 50k. Such kind of datasets can help to decide the loan, insurance and other financial aspects. .[8]

— **Parse the data into three datasets:**

Before training the model, parse the dataset into three datasets:

Features: X
Targets: yϵ\{income>50K,income$\leq$50K\}
Sensitive attributes:$z_{race}\epsilon$\{black,white\} and$z_{sex}\epsilon$\{male,female\}

Features are the attributes for the prediction like age, education, occupation. The target contains salary attribute whereas sensitive attributes are race and gender.

— the dataset of sensitive attributes race and gender has not used by features dataset which has used for training the model. [8]

— **Results after training a model**

ROC AUC: 0.91
Accuracy: 85.1%

— **Basic classifier performs pretty well!**

9.2 Qualitative Model Fairness

Let's start the investigation on the fairness of our classifier by evaluating test data which shows the distribution of (Income >50k) on given sensitive attributes.

The figure shows that race and sex do affect the fairness of the model. As a race (left plot) and gender (right plot), the blue prediction distributions have a larger peak at the low end of probability range. This means if the person is

female or black then its higher chance of salary is less than 50k which is unfair compared to when someone is white or male.

The results of qualitative measures give us a clear picture of the unfairness in the model with respect to race and gender as our model favors more to race (=white)and gender (=male) for a high-income level. Therefore we need to make a prediction which can bring fairness in the results.

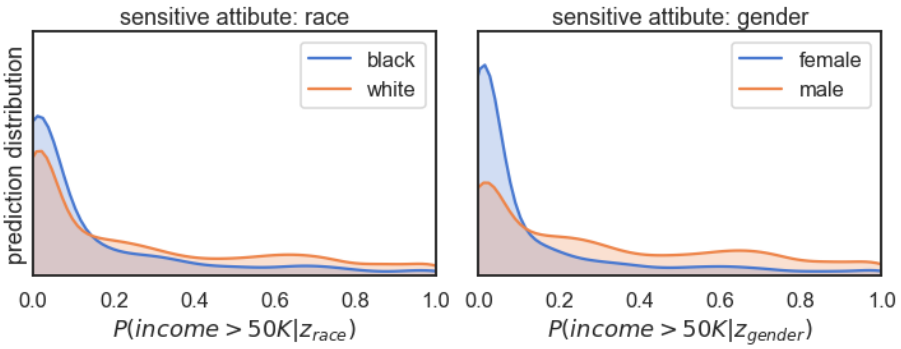


Figure 34: Qualitative Model Fairness

9.3 Quantitative Model Fairness

In to order to find quantitative fairness of the classifier. We used 80% rule to quantify the unfairness on a group of people of a sensitive attribute. Also, Zafar et al. has used the same technique of p% rule in their paper Fairness Constraints: Mechanisms for Fair Classification, which can be used to quantify fairness of a classifier.[9] This rule states as: A binary class prediction $\{\hat{y}\} \in \{0, 1\}$ has made by classifier, given a binary sensitive attribute $z \in \{0, 1\}$ satisfies the p%-rule if the following inequality holds:

$$min(\frac{P(\hat{y} = 1|z = 1)}{P(\hat{y} = 1|z = 0)}, \frac{P(\hat{y} = 1|z = 0))}{P(\hat{y} = 1|z = 1))} \geq \frac{p}{100} \quad [9]$$

The following equation[1] that the ratio between the probability of a positive outcome given the sensitive attribute being valid. And the same probability is sensitive attribute being false is no less than p:100. So, when a classifier is completely fair in a model, it will meet the expectations of a 100%-rule whereas, when it is entirely unfair it satisfies a 0%-rule.

In determining the fairness of our classifier, we will follow the 80% rule. The model is fair when it satisfies at least an 80%-rule.

The classifier satisfies the following %p-rules:

— **given attribute race; 45%-rule**
  According to p% rule, 45% is less than 80% which shows race is biased in the dataset.
— **given attribute gender; 34%-rule**
  According to p% rule, 34% is less than 80% which shows gender is also biased in the dataset.

This satisfies that the trained classifier is unfair in making its predictions. Whereas gender is more unfair in the dataset which shows males are the priority than females likewise for white over black people.

Such results can bring unfairness in predictions that's why we need to make our model fair in order to get fair decisions in machine learning.

9.4 Fighting a Bias

The Classifier is still giving unfair results after removing the sensitive attributes like gender and race. It's still providing biased results against gender and race. Therefore that's not enough.

We found the same unfair results in an adult dataset. The adult dataset gathered in 1994, but income inequality is still a problem nowadays. In the dataset, mostly white males and white women are getting higher income concerning black people which has categorised in lower pay.

The predictive model can quickly learn the bias through other attributes in the dataset like education level, occupation and relationship. Therefore the model will end up with unfair results even after removing sensitive attributes.

9.5 Removing the Bias from the Dataset

There are two ways that can help us to remove the bias from the dataset.

— De-bias the dataset by separating sensitive attributes from the dataset as shown in section 9.1.
— Constrain the model so that it is forced into making fairer predictions. Adversarial training procedure can be used for this approach [10]

9.6 Adversarial networks and GAN systems

Bias problem can be resolved by taking an idea of Good Flew Seminal paper [10], which was published in 1994. In which they introduced the GAN system (Generative Adversarial Networks) which has two neural networks; a generative model and an adversarial classifier. These two networks are competing with each other in a zero-sum game. In the game, the generative model focuses on producing samples that are identical from real data, on the other side, the adversarial classifier tries to match whether the samples came from the generative model or from the real data.

By taking inspiration from the GAN system, we can train the classifier by using pivotal property on the predictive model.

9.7 Adversarial training procedure

In Adversarial training procedure, the system has two neural networks same as GAN system with some amendments. First, we use the generative model as a predictive model to make predictions for the second network. It generates the predictions $\hat{y}$ based on x instead of data. Secondly, the adversarial classifier doesn't distinguish real from generated data whereas it predicts the sensitive attributes values $\hat{z}\epsilon\hat{Z}$ from the anticipated $\hat{y}$ of the classifier as shown in Figure (35)

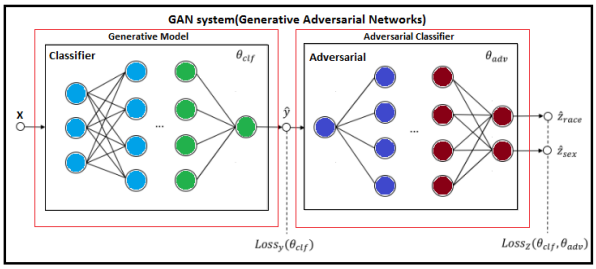


Figure 35: Adversarial training procedure[11]

a) First neural network: Predictive Model

The predictive model has the following objectives;

— It makes the best possible income level predictions without using sensitive attributes. The objective function is:

$$min_{clf}[Loss_y(\theta_{clf}) - \lambda Loss_Z(\theta_{clf}, \theta_{adv})] \quad [12]$$

— It minimizes its own prediction losses while maximizing that of the adversarial whereas increasing λ gives fair predictions while sacrificing prediction accuracy.

b) Second neural network: Adversarial Classifier

The second neural network is Adversarial which has an objective to predict race and gender based on income level predictions of the classifier. It minimizes the prediction loss without concerning accuracy.

$$min_{\theta adv}[Loss_Z(\theta_{clf}, \theta_{adv})]. \quad [12]$$

After pre-training the neural networks, we will start adversarial training on the given predictions produced by the predictive model. We have simultaneously trained both networks for 165 iterations while tracking the performance of the classifier on the test data.

The λ values, that tune fairness versus accuracy, are set to λ_{race}=130 andλ_{gender}=30. During training, we heuristically found that these fixed values give fair results while satisfying the p%-rule values. It is harder to implement fairness for racial attributes than for gender.

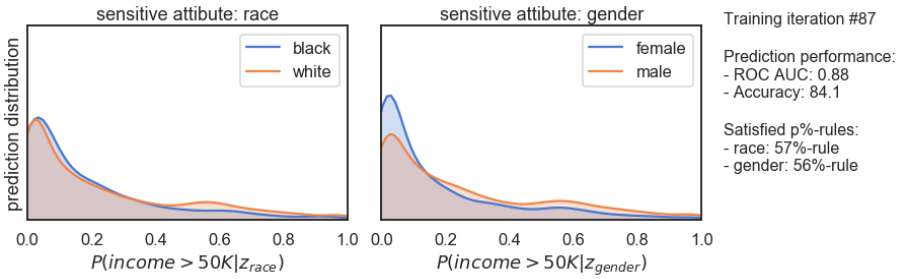


Figure 36: Adversarial Model Fairness at initial iteration [12]

In the above figure when iteration is at 87, it still didn't satisfy the p% rule. But by increasing λ the result got better, and it gave fair results and also satisfied the p% rule.
Whereas the predictions are very much the same as observed for the previously trained classifier: both high in bias and prediction performance.

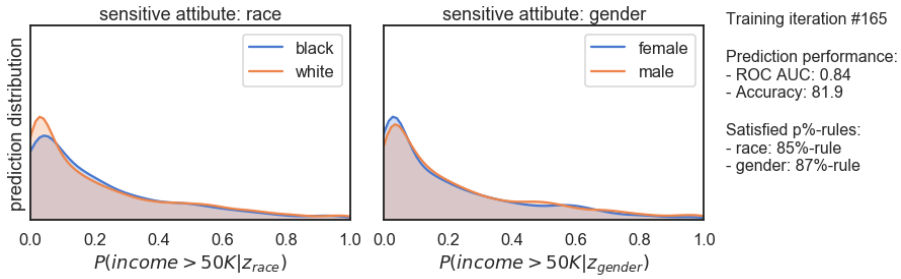


Figure 37: Adversarial Model Fairness at iteration 165 [12]

In the above figure, as the λ increases, we see that the predictions have gradually become more and fairer while prediction performance is slightly declining. At 165th iteration, the classifier satisfies the 80%-rule for both sensitive attributes while achieving a ROC AUC 0.88 and an accuracy of 84%.

Therefore, it seems that the adversarial network training on adult data set works quite well. After sacrificing only 7% of prediction performance, we got a classifier which produces fair results with respect to race and sex.

10 Conclusion

In this report, we have presented several models and how to implement them on an adult dataset, addressing several issues like feature transformation, PCA, global classification model and how to improve classification accuracy by local classification models.

During data understanding phase of the project some of the attributes were found to be a broadly single value such as the country (with 'United States')and capital gain and loss (with 0) in over ninety per cent of instances. The survey weight attribute has removed from the dataset. The education number and education level attributes were found to be similar, and a correlation also identified between the marital status and relationship attributes.

We could clearly see that Local Classifier can make a better model than the Global Classifier. Fairness is an essential factor in machine learning algorithms. But removing sensitive attributes from the dataset is not enough to bring fairness in results and predictions. Fairness towards prediction acquires clever techniques like Adversarial Training. Fair predictions can affect the cost of the algorithm but its a small price to pay.

References

[1] UCI Adult Dataset: https://archive.ics.uci.edu/ml/datasets/census+income

[2] Trupti M. Kodinariya, Dr Prashant R. Makwana: Review on determining the number of Cluster in K-Means Clustering(2013)

[3] Peter J.Rousseeuw: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis

[4] Cluster Analysis:https://www.sciencedirect.com/science/article/pii/B9780123850225000154

[5] P.Prabhu, N.Anbazhagan: Improving the Performance of K-Means Clustering For High Dimensional Data Set

[6] Hannes Wettig, Peter Grunwald: On Discriminative Bayesian Network Classifiers and Logistic Regression

[7] Performance Measures: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.ht

[8] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi: Fairness Beyond Disparate Treatment Disparate Impact: Learning Classification without Disparate Mistreatment

[9] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi: Fairness Constraints: Mechanisms for Fair Classification

[10] Ian J. Goodfellow: Generative Adversarial Networks (2014)

[11] Brian Hu Zhang, Blake Lemoine, Margaret Mitchell: Mitigating Unwanted Biases with Adversarial Learning

[12] David Madras, Elliot Creager, Toniann Pitassi, Richard Zemel: Learning Adversarially Fair and Transferable Representations