CSE 5243 – Introduction to Data Mining

Homework 1

Rajarshi Biswas, Sayam Ganguly

{biswas.91, ganguly.32}@osu.edu

The Ohio State University

Table of Contents

[1 Exploratory analysis of the Income Dataset 3](#_Toc473727228)

[1.1 Overview of the income Dataset 3](#_Toc473727229)

[1.2 Analysis 4](#_Toc473727230)

[1.2.1 Age 4](#_Toc473727231)

[1.2.2 Work class 5](#_Toc473727232)

[1.2.3 Education 6](#_Toc473727233)

[1.2.4 Marital status 6](#_Toc473727234)

[1.2.5 Occupation 7](#_Toc473727235)

[1.2.6 race 8](#_Toc473727236)

[1.2.7 Gender 9](#_Toc473727237)

[1.2.8 Capital loss and Capital gain. 9](#_Toc473727238)

[1.2.9 Hours per week 10](#_Toc473727239)

[1.2.10 Native country 11](#_Toc473727240)

[2. Preprocessing 11](#_Toc473727241)

[2.1 Missing Values 11](#_Toc473727242)

[2.2 Outliers 12](#_Toc473727243)

[2.3 Attributes Transformation 12](#_Toc473727244)

[2.4 Normalization 14](#_Toc473727245)

[2.5 Proximity function selection 14](#_Toc473727246)

[3 Analysis of the result 14](#_Toc473727247)

[3.1 Distribution of proximities: 14](#_Toc473727248)

[3.2 Distribution of proximities for each class 15](#_Toc473727249)

[3.3 Closest neighbor 15](#_Toc473727250)

[3.4 Observation of results when change proximity measure 16](#_Toc473727251)

[4. Language/ Tools Used 16](#_Toc473727252)

# 1 Exploratory analysis of the Income Dataset

The Income dataset is an extraction from the 1994 census database. From the income point of view, the dataset is categorized between two classes. One class consist of individuals having more than 50K income and the other class’ income is less than or equal to 50K. In our exploratory data analysis, we try to find the relationship of different attributes of the dataset to one income class.

## 1.1 Overview of the income Dataset

The income dataset has total 520 entries. Below the names, type, and a brief description of each of the attributes are mentioned. In the next section we will present our thorough analysis.

* ID: Identification number for a person in the dataset; Nominal.
* Age: Age of the person; Continuous and ratio.
* Work class: Indicates the work sector; Categorical.
* fnlwgt: Meaning is ambiguous; Continuous.
* Education: Highest education of a person; Categorical.
* Education-num: Education category. Continuous.
* Marital-status: Marital status of an individual; Categorical.
* Occupation: Occupation of a person; Categorical.
* Relationship: Relationship status of a person; Categorical.
* Race: Race of a person; Categorical.
* Gender: indicates whether a person is male or female; Categorical.
* capital-gain: continuous.
* capital-loss: continuous.
* Hours-per-week: Hours per week a works; Continuous.
* Native-country: Native country of an individual in the dataset; Categorical.
* Class: Income class of an individual distinguished between two classes – less than or equal to 50K income and more than 50K Income

We observed that 28 rows have both work class and occupation values as ‘?’. We have not considered those rows in our exploratory data analysis.

## 1.2 Analysis

### 1.2.1 Age

Fig 1.1 scatter plot below shows the age group distribution of individuals in the dataset.

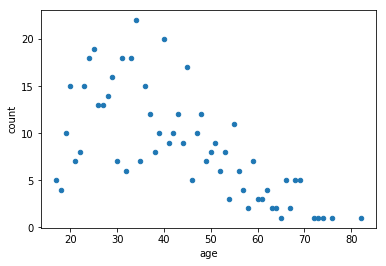


Fig 1.1

From this scatter plot it is evident that age group distribution is dense between age 22 and 55. We see there are outliers above 70. Furthermore, there are a small set of people with age below 20. This pattern indicates the typical working age range.

In addition, in Fig 1.2, we show the income class distribution in different age groups. We can infer from this pie charts that with more experience, probability of earning more than 50K is higher.

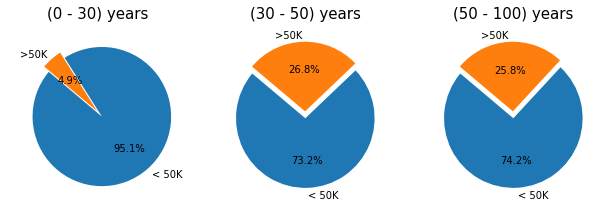


Fig 1.2

### 1.2.2 Work class

Fig 1.3 shows the distribution of individuals working in different categories. Based on this we broadly categorized the work class into private, government and self-employed. In addition, we plot Fig 1.4 that indicates the total number of people earning more than 50K in each of the work class categories. We can see the private sector is dominating here.

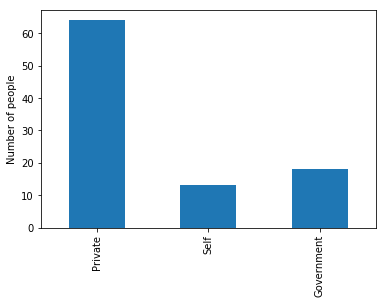
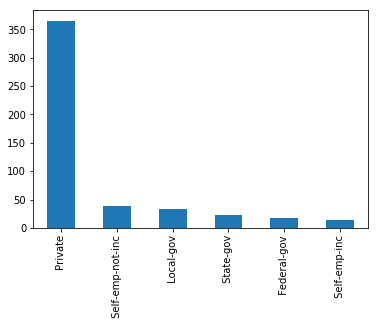


Fig 1.3 Fig 1.4

However, to infer which work class has the more potential of a higher income we plot Fig 1.5. This pie chart shows that self and government employees have much more potential of earning more than private employees.

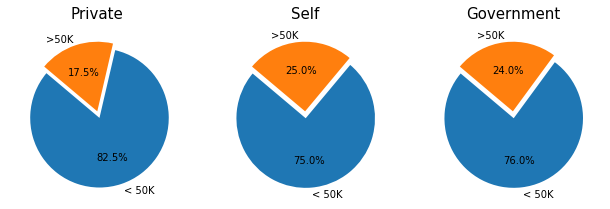


Fig 1.5

### 1.2.3 Education

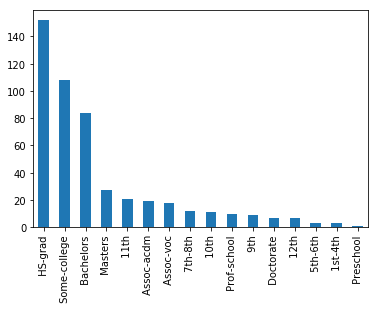


Fig 1.6 shows the distribution of highest educational qualification. From this bar plot, it is evident that HS-grad and College education predominates the dataset.

In addition, we plot Fig 1.7 below to find the relation of earning more with having an advance degree. These plots suggest that an advance educational qualification leads to higher income.

Fig 1.6

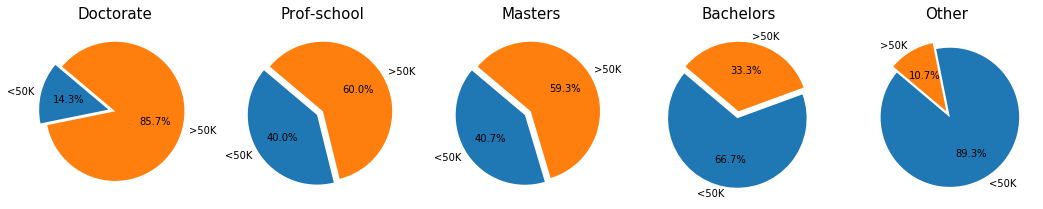


Fig 1.7

### 1.2.4 Marital status

The distribution of marital status in Fig 1.8 shows the dataset comprised of equal number of married and not married individuals. Interestingly, the data also reveals a pattern in Fig 1.9 that married individuals are earning more than others.

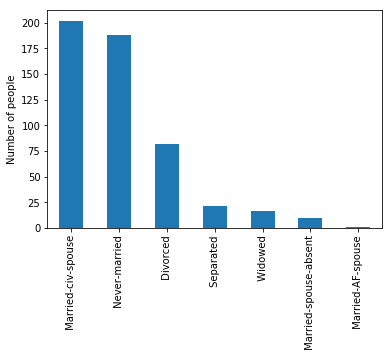
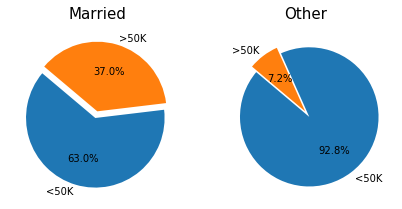
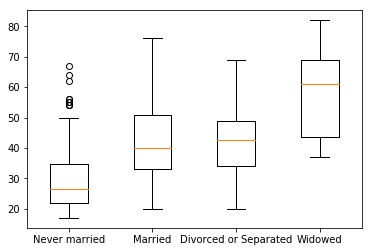
. 

Fig 1.8 Fig 1.9



Then we try to find a relation between marital status and the age of an individual. Fig 1.10 shows a clear trend of marital status corresponding to the age. Widowed, divorced, and separated people tend to be elder when compared to the other classes.

Not surprisingly, individuals who have never married falls into the smaller age group.

Fig 1.10

### 1.2.5 Occupation

We plot in Fig 1.11 the number of people having different occupation, in addition, Fig 1.12 depicts the percentage of people earning more than 50K in each occupation category.

### 

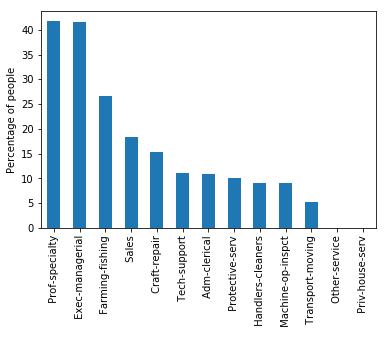
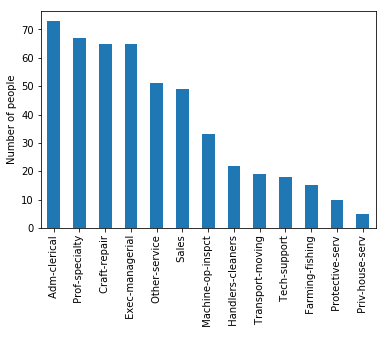


Fig 1.11 Fig 1.12

We can see from the above plots that although the number of people having jobs like Adm-clerical and Craft repair are predominant in our dataset, when it comes to earning more, jobs like prof-specialty and exec-managerial are more rewarding.

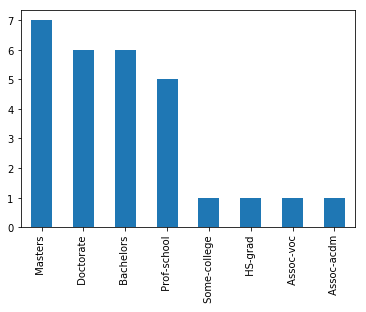


Fig 1.13

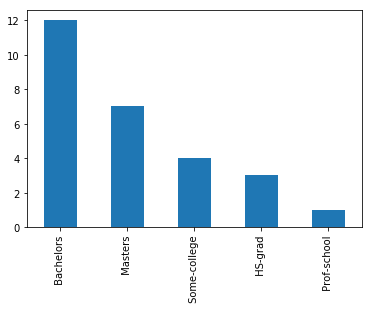
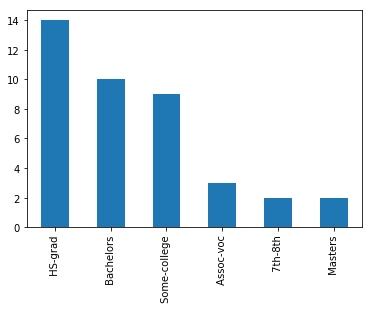
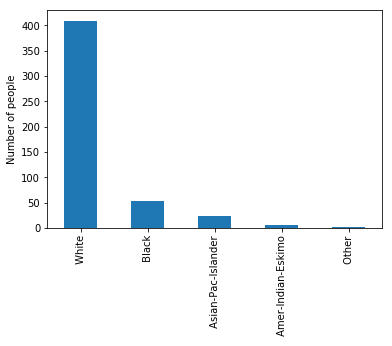
 

Fig 1.14 Fig 1.15

To compare the relationship of educational qualification of people working in the most rewarding jobs to a lesser rewarding one, we plot the above three graphs. Fig 1.13, 1.14, and 1.15 show that the educational qualification distribution of workers in Prof-specialty, Exec-managerial, and Craft-repair occupation respectively. The evident trend here is jobs that are more rewarding demands higher educational qualification.

### 1.2.6 race



We plot the bar chart of race in Fig 1.16. The data set is majorly dominated by the white race.

Fig 1.16

### 1.2.7 Gender

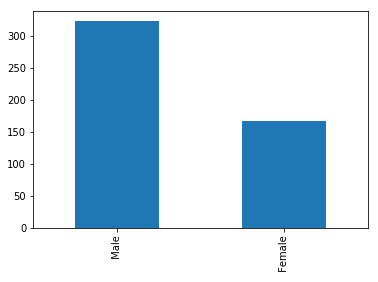
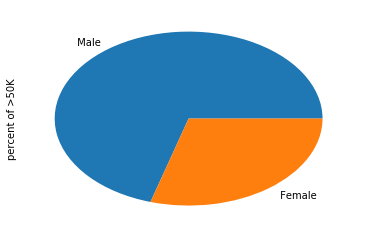
 

Fig 1.17 Fig 1.18

We plot the gender distribution in the dataset in figure 1.17. We see from the plot that majority of the sample in the dataset is male. In addition, we plot in Fig 1.18 the percentage of each gender category that earns more than 50K. Here, we see that males dominate the higher income class.

### 1.2.8 Capital loss and Capital gain.

Below we show the summaries of capital gain and capital loss. From the summaries we can observe that there are a lot of 0s in both the columns. It may be an indication of 0 capital gain or 0 capital loss.

count 492.000000

mean 69.528455

std 377.638974

min 0.000000

25% 0.000000

50% 0.000000

75% 0.000000

max 4356.000000

Capital loss

count 492.000000

mean 1085.599593

std 6979.369258

min 0.000000

25% 0.000000

50% 0.000000

75% 0.000000

max 99999.000000

Capital Gain

### 1.2.9 Hours per week

We plot the number of people working same hours per week in Fig 1.19. We can infer form this scatter plot that almost half of the people in our income dataset works for 40 hours a week.

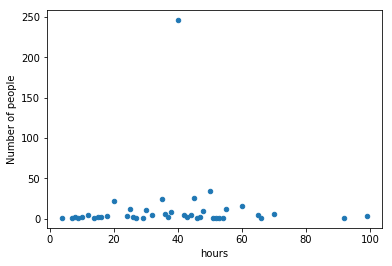


Fig 1.19

Then, we generate a bar plot, Fig 1.20, to find the number of people earning more than 50K for different work hour week. We see that mostly people who work greater than or equal to 40 hours a week tend to earn more.

In addition, Fig 1.21 shows the education qualification distribution of people who have more than 40 hours work week and earn more than 50K. Not surprisingly this class is the most educated.

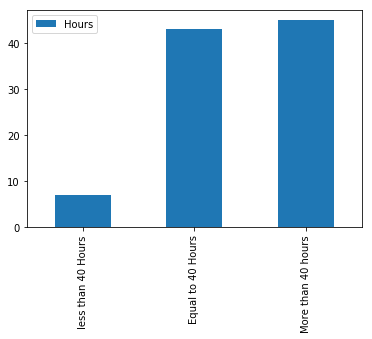
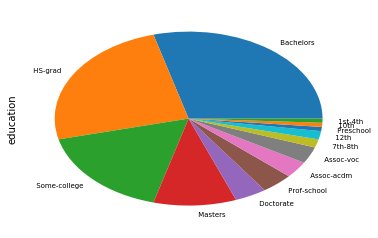
 

Fig 1.20 Fig 1.21

### 1.2.10 Native country

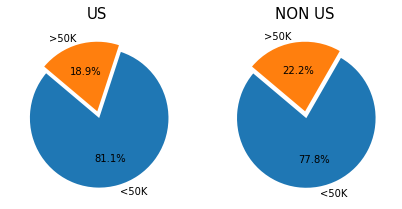
 

Fig 1.22 Fig 1.23

From Fig 1.22, we see the United states dominates the dataset. We broadly categorized the native country in US and NON US classes. We also plot in Fig 1.23 the income classes of US and NON US citizens. This pie chart indicates that the earning potential is same no matter of the native country.

# 2. Preprocessing

## 2.1 Missing Values

* We have 520 records in our datasets and among them there are 62 total missing values which are identified by the ‘?’ character.
* The fields ‘workclass’ and ‘occupation’ have 28 missing values each and they belong to the same rows. It follows from the fact that if a person’s occupation is not known then It’s not possible to determine the work class. As the dataset pertains to a person’s income so not knowing either of occupation and work class makes these records irrelevant to our analysis and we have ignored these records.
* The rest of the missing values belong to the ‘native\_country’ field. From the Bar plot below in Fig 2.1, it can be seen that majority of our records belong to the US and a small portion is from other country. Hence we have considered these missing value records in the “Non-US” category.

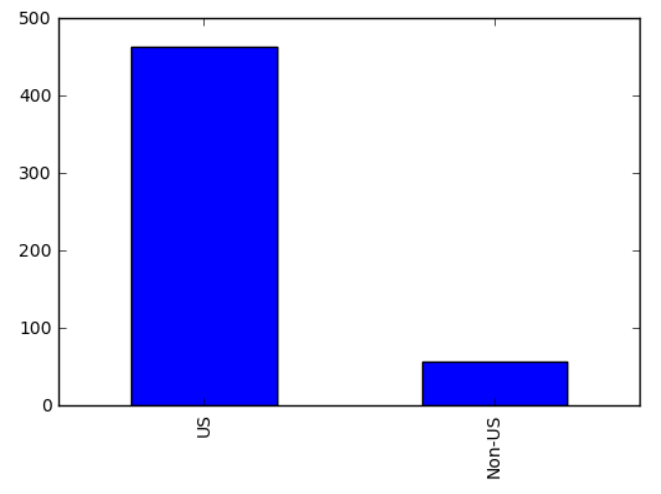


Fig 2.1

## 2.2 Outliers

There are some outliers in ‘age’ field with values >70. We have however ignored these records for simplicity.

## 2.3 Attributes Transformation

* The first feature of the dataset “ID” is a nominal attribute identifying each record which should not have any impact on the income. Also there is another feature “fnlgwt” whose purpose is ambiguous. We have removed these features from our proximity analysis.
* From Fig 1.4 depicts that “workclass” can be broadly categorized into 3 subsets – “Private”, “Self” and “Government”. We have applied this transformation to this attribute and then we have assgined unique numerical values (1, 2, and 3) to each of the categories.

1 Private 365

2 Govt 75

2 Self 52

* Next we concentrate on the two features “education” and “education\_cat”. A closer look at the dataset reveals that the “education\_cat” field is just a numerical representation of the “education” field and they are redundant. Hence, we have dropped the “education” field and rename “education\_cat” to “education” which will serve our purpose for numerical analysis. Furthermore, we have grouped the data in 6 categories to simplify analysis.

3 College 278

4 Bachelors 84

1 School 74

2 Associate 39

5 Masters 28

6 Doctorate 17

* Then we have transformed the feature “marital\_status” into two major categories – Married and Not-Married. Numerical values have then been assigned to each of the categories to facilitate numerical analysis.

1 Married 307

2 Not Married 213

* Next we look at “occupation” feature in the dataset and have mapped all the values into numerical categories.
* We have not considered “relationship” feature in our data analysis as we are already considering “Marrital\_status” both of which seem to be representing attributes that are somewhat redundant. Hence we have dropped this feature in our analysis.
* From the bar plot of “race” we see that the predominant value for the feature is “white” and others constitute a small fraction. Hence we decided to categorize race in two classes “white” and “non white”.

1 White 433

2 Non-White 87

* For the feature “gender” it is obvious that it can be categorized in two groups – “Male” and “Female”

1 Male 335

2 Female 185

* We considered features “capital\_gain” and “capital\_loss” and both of the features have asymmetric values which is evident from the fact that most of the records have a zero value. Hence, we have categorized it as a binary field where if there is some value then w e are assigning it a value of 1 and 0 otherwise.
* Next we have considered a major feature “age”. Although it seems that it is a continuous attribute we choose to categorize it into different intervals. The rationale behind this comes from Fig 1.1 where it can be seen that the number of records belonging to each age range has a pattern where the region above 25 and up to 55 is dense compared to other regions and the other regions on both sides are equally dense. Hence we have chosen to group the records in intervals below 30, 30 – 50 and above 50.
* Finally we categorize “hours\_per”week” into 3 intervals – 40 hours/week, less than 40 hours/week and more than 40 hours/week. Fig 1.19 shows the distribution of number of people working in different hours per week and we can see that with these 3 groups we can distribute them uniformly.

## 2.4 Normalization

We have categorized all of the features in the Income dataset hence applying normalization to theme will not be appropriate and we have not done that. Instead we normalized all the four features in the Iris dataset to scale all of them between 0 and 1 by applying min-max normalization.

## 2.5 Proximity function selection

We applied Euclidean and Minkowski distance measure for the Iris dataset. This selection was motivated by the fact that all the features are in the continuous space and these distance measures would be an ideal choice for that.

For the Income dataset the first proximity measure that we applied was the Cosine Similarity. The Income dataset has many attributes that are categorical and through our transformation we have even converted some of the continuous features into categorical space, hence Cosine Similarity was chosen as our first measure. Next, we also wanted to explore how Euclidean performed in case of categorical data and that led us to implement Euclidean as our second proximity measure.

# 3 Analysis of the result

## 3.1 Distribution of proximities:

Fig 3.1 shows the proximity distribution (Euclidean distance) of the Iris dataset. From the figure we observe that, when K = 1, that is the distances of first nearest neighbor for each row are lesser when compared to the distances when K = 10, that is the distances to 10th nearest neighbor. The graph clearly shows that the distance to the neighbors increases as we increase K.

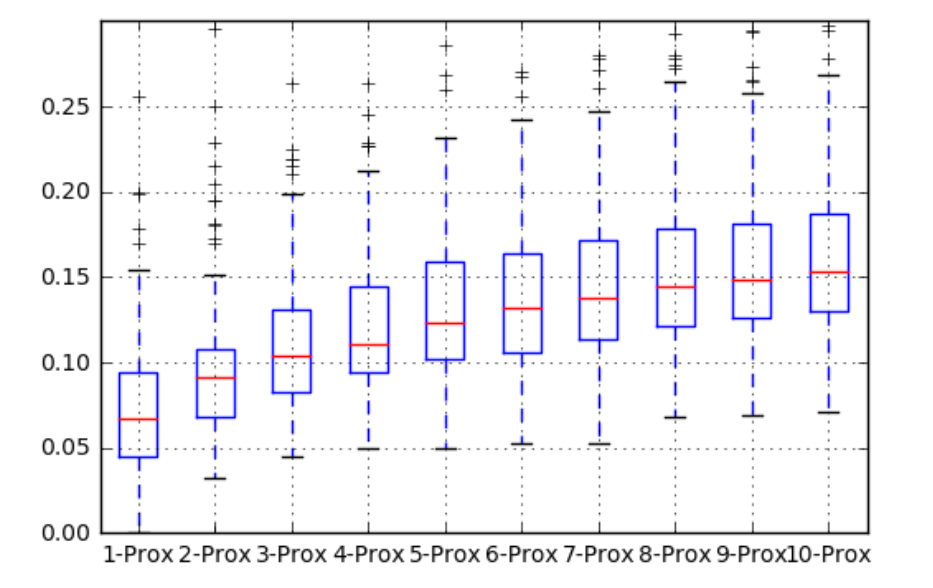


Fig 3.1

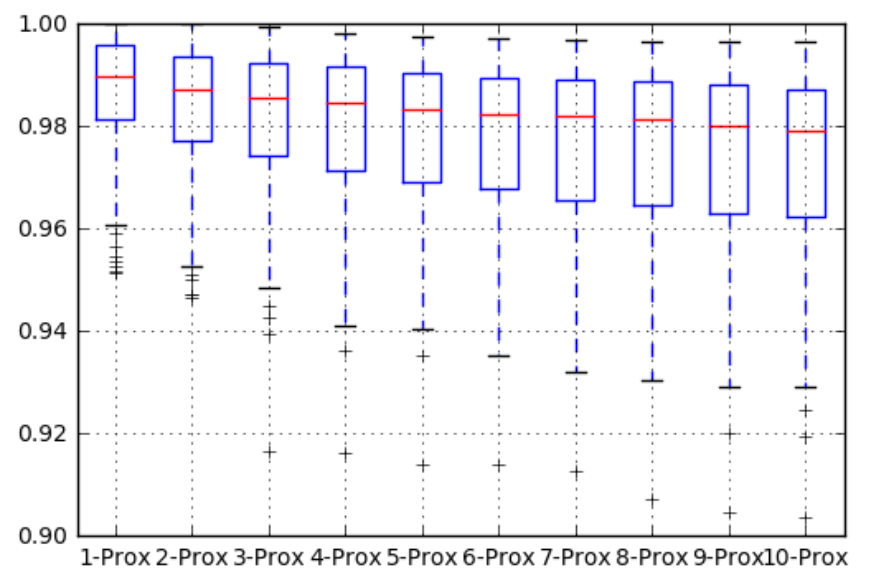


Fig 3.2

We also analyze the proximity distribution (cosine similarity) of the income dataset in Fig 3.2. A value of 1 in cosine similarity indicates that two rows are similar. We observe that with increasing K the value of cosine similarity gets smaller. On other words, with increasing K, the Kth neighbors of each row become more dissimilar.

## 3.2 Distribution of proximities for each class

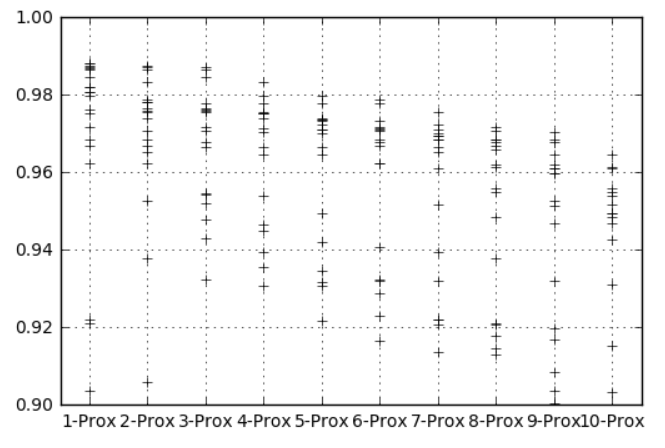
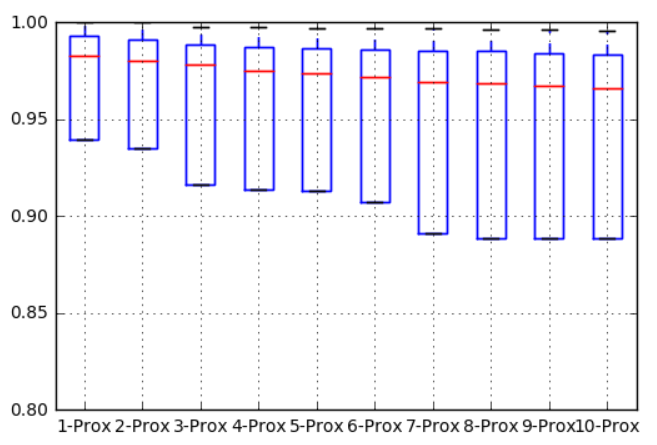
 

Fig 3.3 Fig 3.4

Fig 3.3 shows the cosine proximity distribution of income class greater or equal to 50K. Moreover, Fig 3.4 depicts the cosine proximity of income class less than 50K.

## 3.3 Closest neighbor

We analyzed from the income dataset that 26th record is the closest (1st proximity distance) to the largest number of other examples.

In the Iris dataset, 47th record is the closest to the largest number of other examples.

## 3.4 Observation of results when change proximity measure

We observed that in the income dataset cosine similarity measure seems to outperform Euclidean measure. We anticipate that this is because of the categorical nature of the income dataset.

# 4. Language/ Tools Used

* Python
* Numpy
* Pandas
* Jupyter
* Collaboration using GIT