CSE 5243 - Data Mining Homework 1 Report

-Srividhya Chandrasekharan (chandrasekharan.12)

INTRODUCTION:

This report documents the various steps carried out in the process of finding ‘k’ data records that are closest or most similar to a given example. Also, it contains an extensive exploratory data analysis of the Income dataset and results of other analysis tasks.

Input Datasets:

- Iris.csv
- Income.csv

Output file:

- output.csv

EXPLORATORY ANALYSIS OF INCOME DATASET:

Income dataset contains 520 records and it can be accessed here . The dataset has 16 attributes wherein the 16th attribute is the target class label {“<=50K”,”>50K”}. The following data visualizations were created in the hope of identifying some interesting pattern(s).

The following is a pie chart depicting the portion of the dataset that falls into each of the income categories. 420 records have the class label ‘<=50K’ while, only 100 records have the class label ‘>50K’. Clearly, the number of data records in each income category is unequal and, there are more records under the ‘<=50K’ income class.

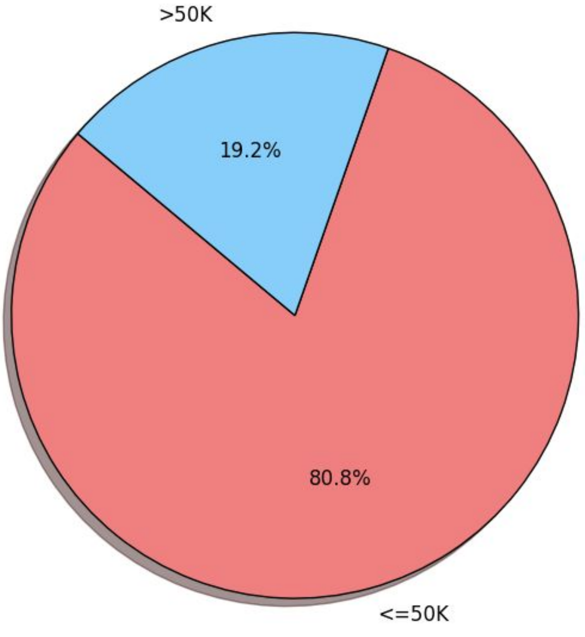


Figure 1 : Pie chart of classes

The Income dataset has 5 numeric attributes namely - age, fnlwght, capital gain, capital loss and hours per week. Histograms were plotted for each of these 5 attributes in the hope of finding trends in the values.

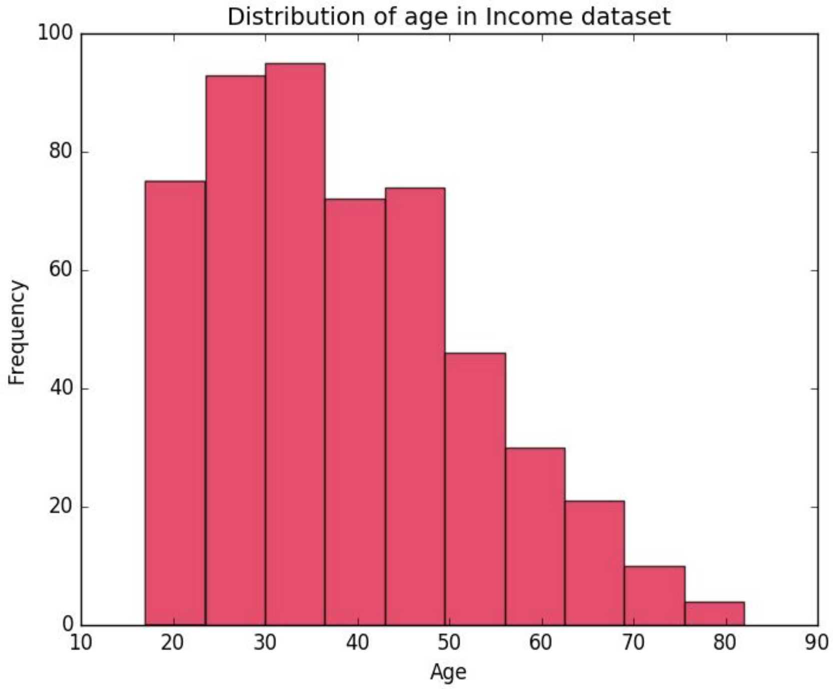


Figure 2 : Histogram of age attribute

The minimum age encountered in the dataset is 17 and the maximum age is 82. More than 50% of the data records fall in the age range 20-50.

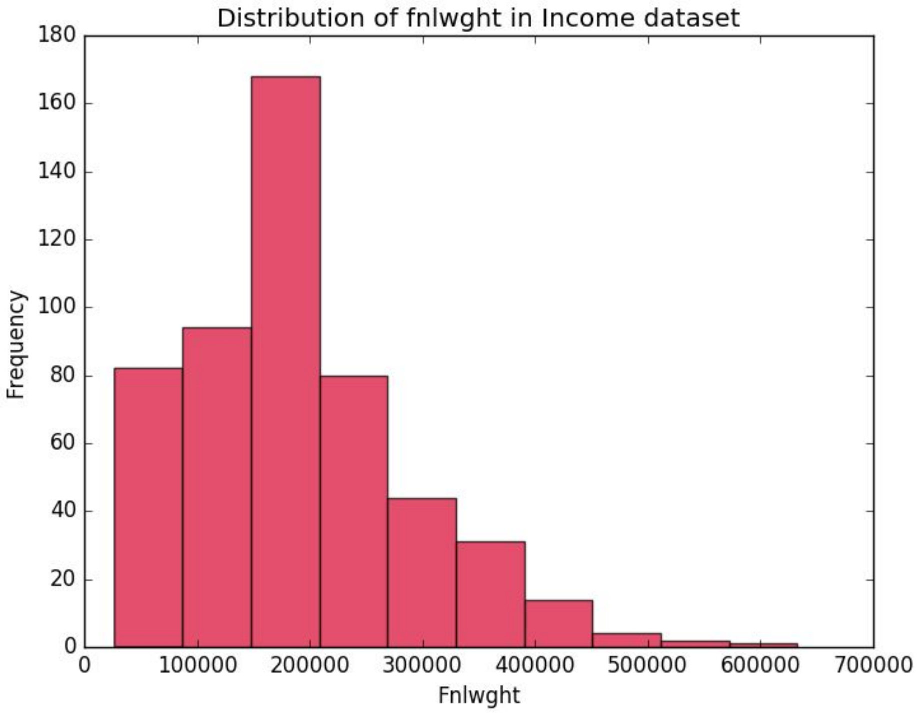


Figure 3 : Histogram of fnlwght attribute

The attribute fnlwght has examples in the dataset with values ranging from 26880 and 632613. From the histogram it's clear, that the number of examples in the datasets for very high value of fnlwght are less. While, majority of the data examples have their fnlwght values in the range [10000,20000].

The following are the histograms of Capital Gain and Capital Loss attributes of the Income dataset.

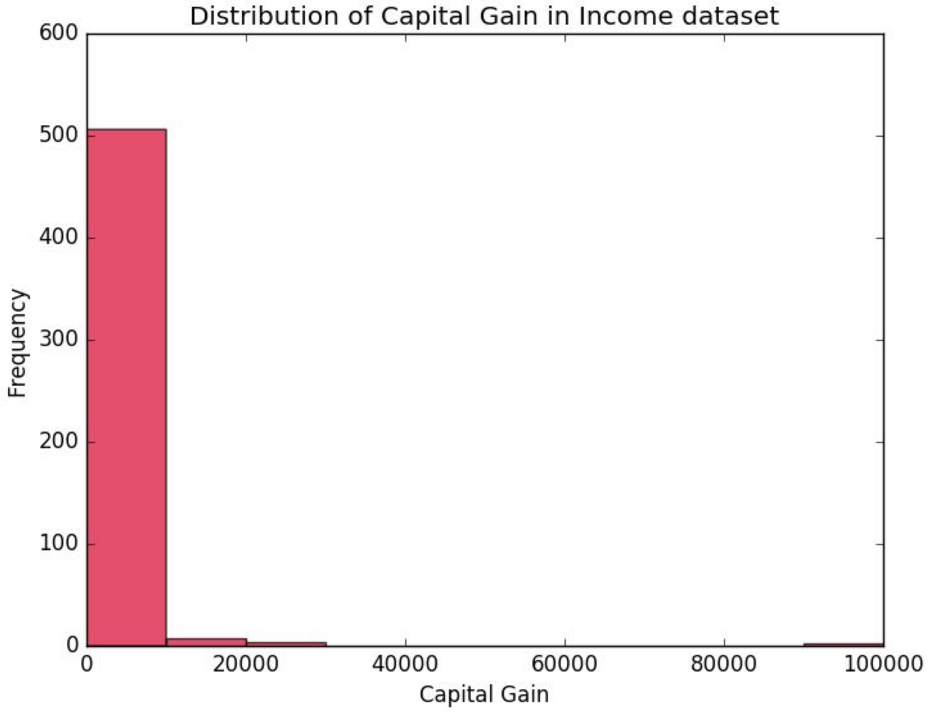


Figure 4 : Histogram of Capital Gain attribute

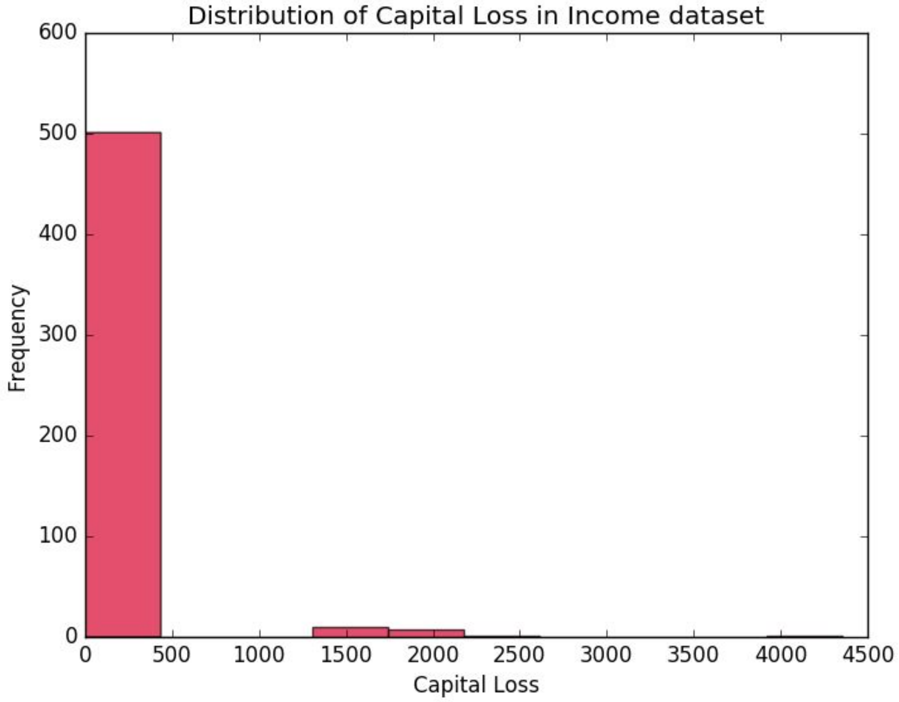


Figure 5 : Histogram of Capital Loss attribute

From Figures 4&5, it's obvious that the Capital Gain and Capital Loss have the value 0 for more than 90% of the data records in the Income dataset. Only a small portion of the dataset contain a non-zero value for these two attributes and out of this portion, only very few records have a very high value. It is interesting to note that no record has non-zero values for both the capital gain and loss attributes.

The following visualization (Figure 6) is a histogram of the attribute - Hours per week. Again, it's evident that majority of the training examples work 40 hours per week. Some people work less than 10 hours per week and some have clocked hours greater than 60 per week with the highest being 99 hours!

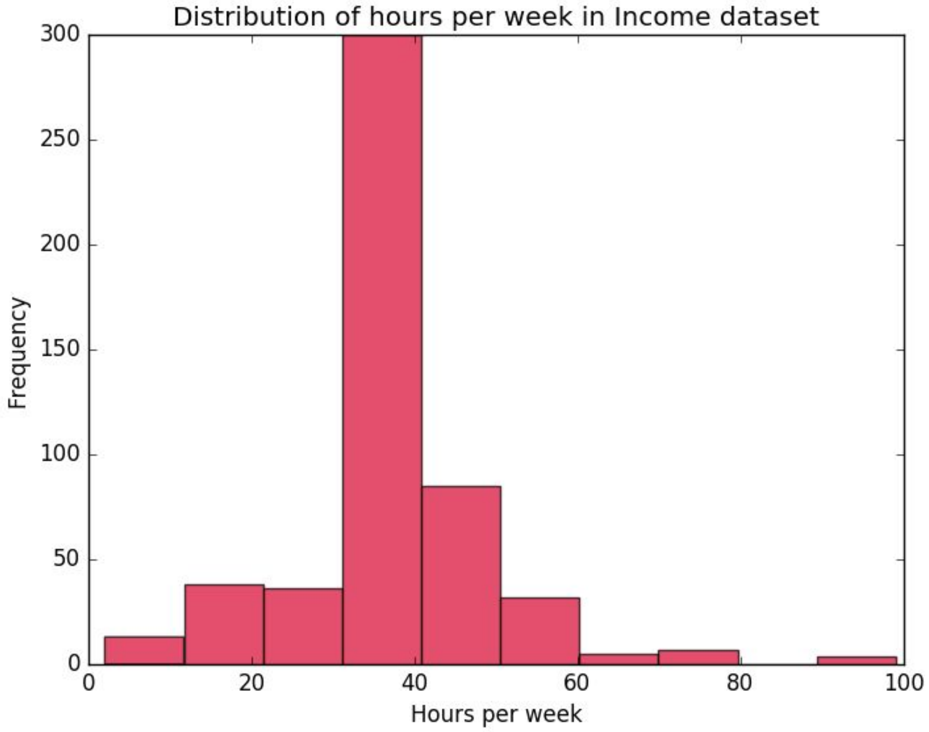


Figure6:Histogram of Hours per week attribute

The following visualization (Figure 7) is a stacked bar graph that was created to analyze the number of Male and Female data records in each of the income classes. After analysis, the number of females are less than the number of males in the Income dataset.

- Out of the 100 records which fall under the >50K income class, 19% are female and the remaining 81% are male.
- Out of all the records labelled with <=50K income class, 60% are Male and 40% are female.

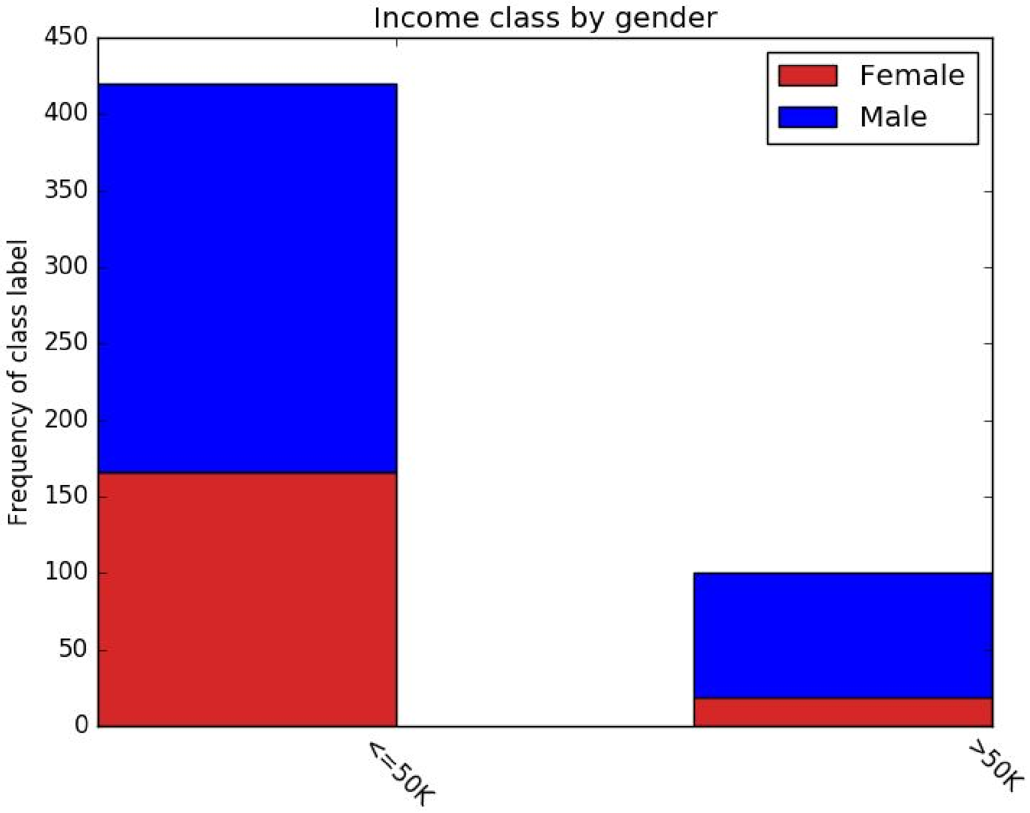


Figure 7: Gender vs Income class

PROGRAM DESCRIPTION:

General points

- Programming language used - Python
- Each data record was assigned a transaction ID for convenience.
- Program will be able to handle Iris and Income datasets and this was made possible by using metadata. Metadata contains the information about the attributes. This information is useful when handling the attribute values for comparisons. Based on the input .csv file, metadata is correctly chosen.
  - ➔ Metadata for Income Dataset: [[0, 'tid'], [1, 'ID'], [2, 'num'], [3, 'cat'], [4, 'num'], [5, 'cat'], [6, 'num_cat'], [7, 'cat'], [8, 'cat'], [9, 'cat'], [10, 'cat'], [11, 'cat'], [12, 'num'], [13, 'num'], [14, 'num'], [15, 'cat'], [16, 'class']]
  - ➔ Metadata for Iris Dataset: [[0, 'tid'], [1, 'num'], [2, 'num'],[3, 'num'], [4, 'num'], [5, 'class']]

tid - transaction ID
num - numeric attribute
cat - categoric attribute
num_cat - numeric categorical attribute
class - class label

Handling missing data

- Income dataset had 33 missing values. This was indicated by '?'.
- Iris dataset did not have any missing values.
- Instead of deleting records with incomplete data, reasonable values for attributes missing data were computed.
- When the value of a categorical attribute was missing, the '?' was replaced by the mode of the attribute. The **mode** is the value that appears most often in a set of data.
- For example: The attribute 'workclass' had some missing values. The records which were missing values for 'workclass' were assigned the mode of 'workclass' attribute - 'Private'. Similarly,, the mode of the 'native_country' attribute - United-States was assigned in place of '?' to those records which were missing native_country value.
- The mean of an attribute was assigned in place of '?' for numeric attributes.

Data Preprocessing

- Numeric attributes were normalized using min-max normalization so as to bring all the values to a common scale [0,1].

$$v' = \frac{v - min}{max - min}\,(max_{new} - min_{new}) + min_{new}$$

- By normalizing attributes, we make sure that different attributes contribute equally to the distance measures.
- No transformations were applied on the dataset.
- No outliers were detected.

Distance measures used

- Two distance measures were used in the program - Euclidean and Cosine similarity.
- Program has a variable **ch** **-** which is used to set the distance metric used. When **ch** is set to 1, Euclidean distance is used and when **ch** is set to 2, cosine similarity is used. By default, **ch** is set to 1.

Computing Similarity

- Program has a variable **k** **-** which is used to set the number of closest number of records to be returned for the given example. By default, **k** is set to 5.
- Equal weight is assigned to numeric and categorical attributes.

$$weight = \frac{1}{total\ number\ of\ attributes}$$

- For each data record in the Income dataset, similarity measures are calculated with all the other records in the dataset. The top 5 records are selected and written into output.csv file.

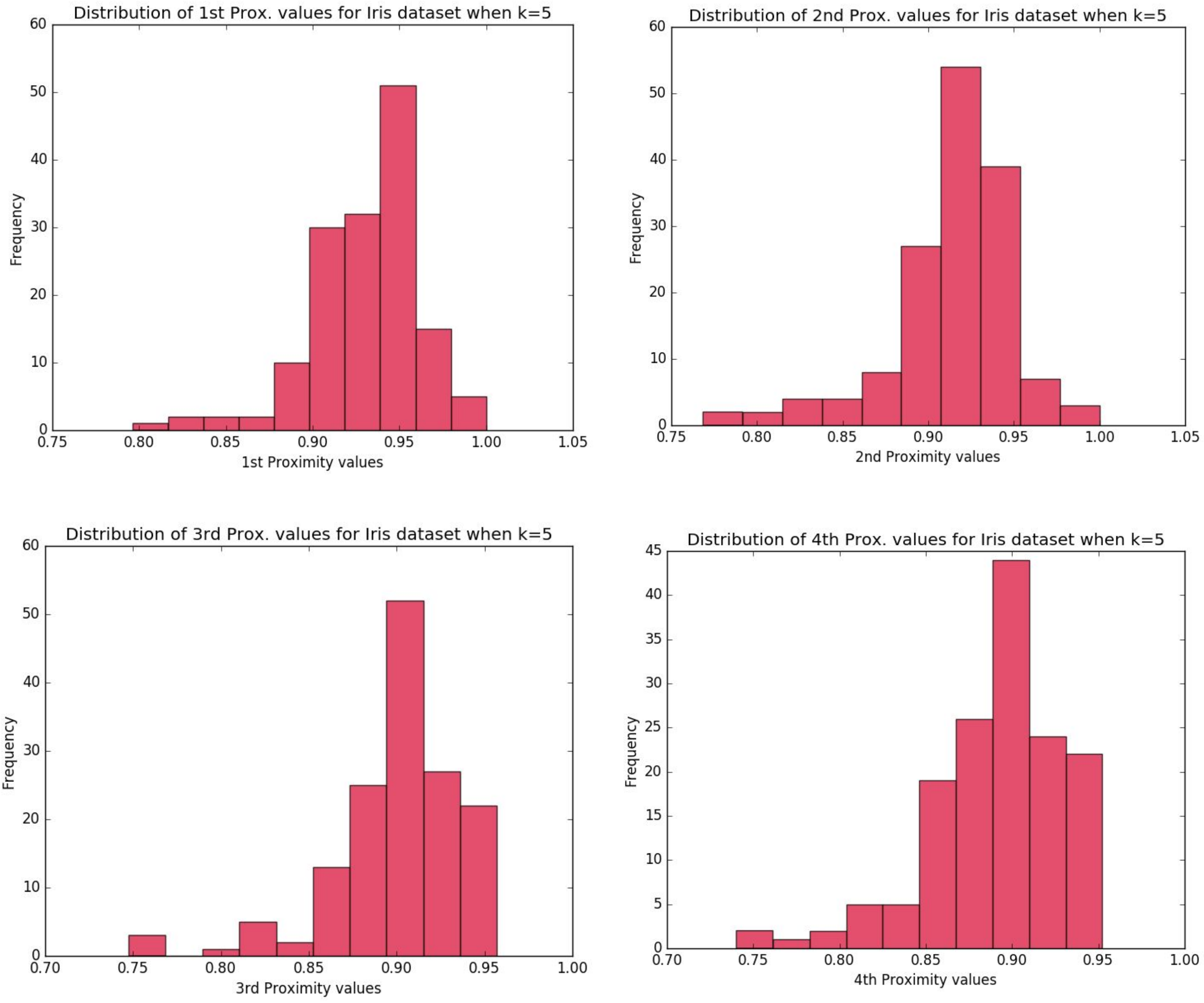
$$Similarity\ measure = \frac{1}{1 + distance}$$

- Each data record analyzed is split into a list of numerical and categorical values and these lists are used for comparisons.

- The list of numeric attributes are used to calculate numeric score (distance).
- Each of the categorical attribute values of the 2 records in hand are individually compared value by value.
- The numerical and categorical scores are combined together based on the weights and the similarity measure is calculated.

ANALYSIS OF RESULTS:

★ The following visualizations were produced in the hope of finding patterns in the k proximity values computed for the Iris dataset for k=5 using **Euclidean distance** measure.



Distribution of 1st Prox. values for Iris dataset when k=5

Distribution of 2nd Prox. values for Iris dataset when k=5

Distribution of 3rd Prox. values for Iris dataset when k=5

Distribution of 4th Prox. values for Iris dataset when k=5

Figures 8-11: Distribution of proximities

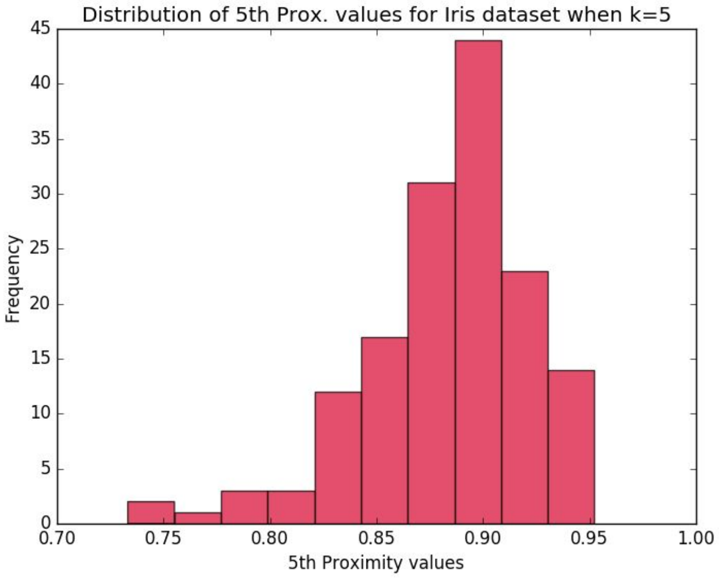


Figure 12: Distribution of 5th proximity for Iris dataset using Euclidean distance

On observing these histograms, we can see that the 3rd and 4th proximity distributions are similar to each other.

- ★ The following visualizations were produced in the hope of finding patterns in the k proximity values computed for the Iris dataset for k=5 using the **Cosine similarity** measure.

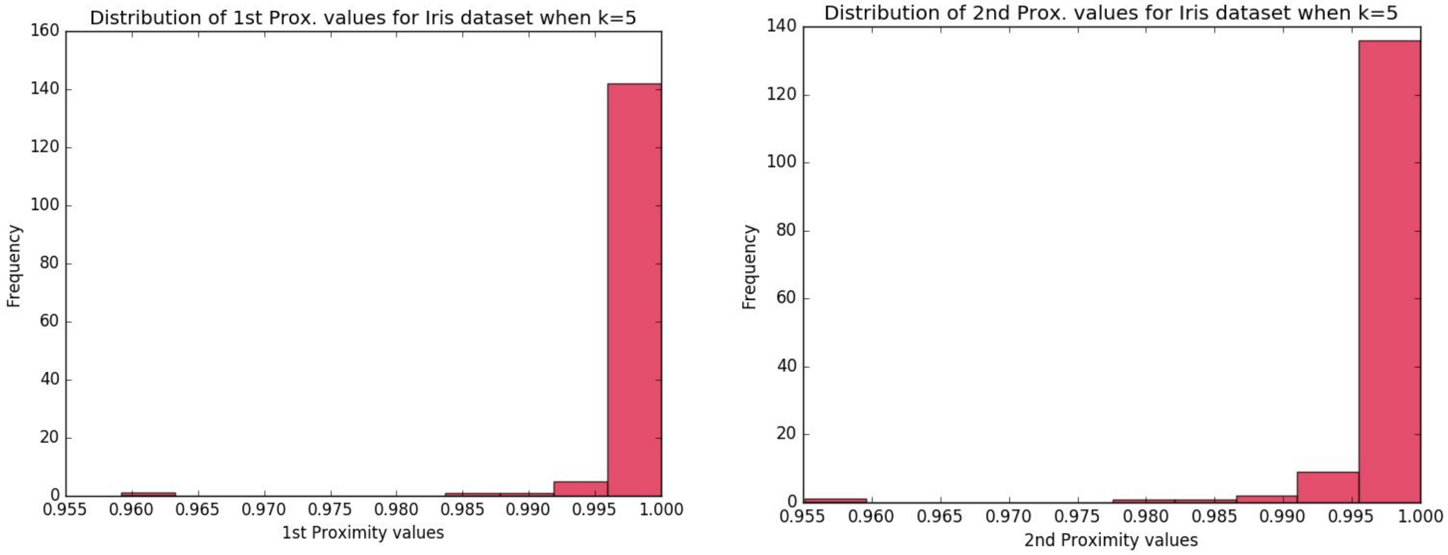
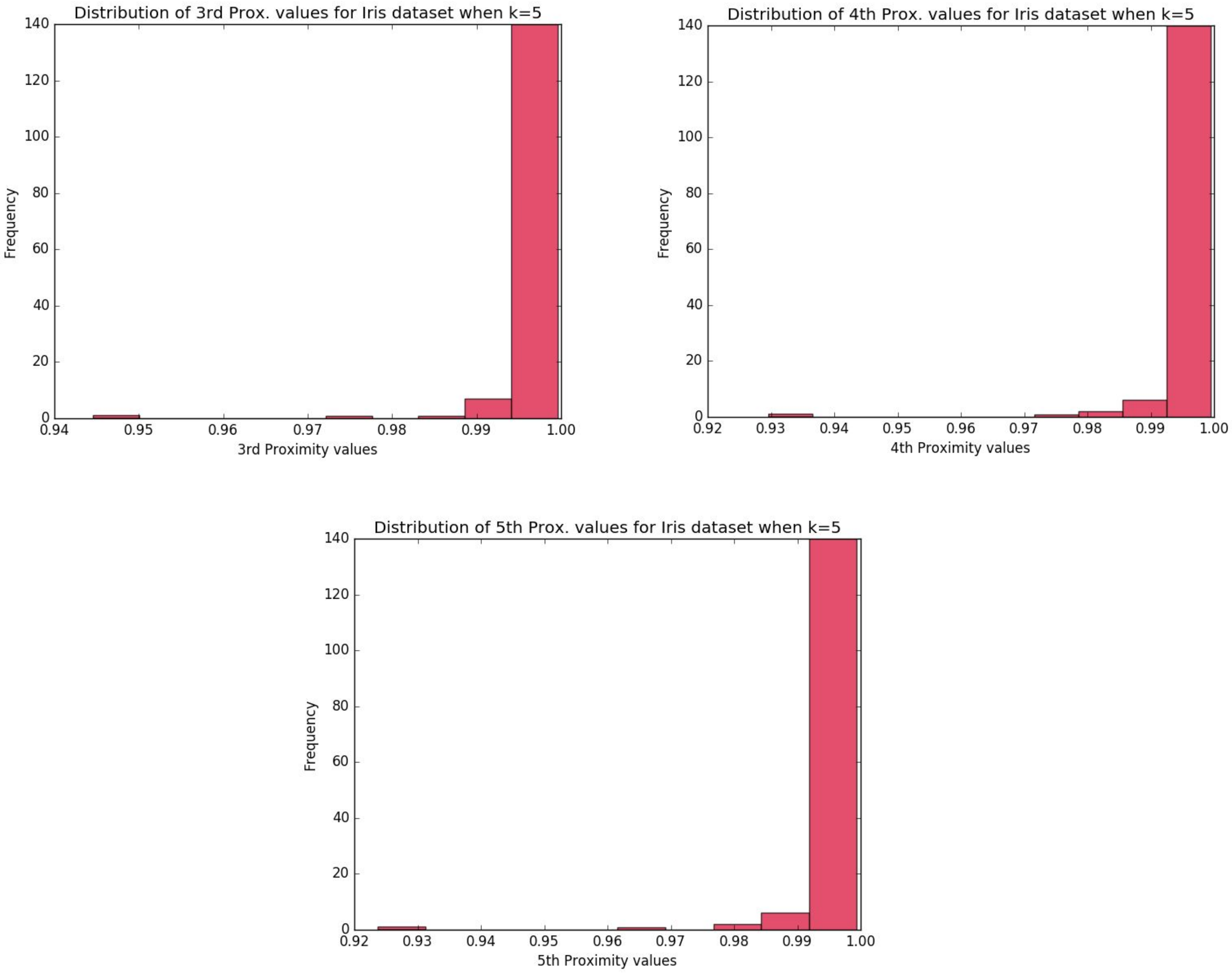


Figure 12-13: Distribution of 1st and 2nd proximities for Iris dataset using Cosine similarity



Figures 15-17: Distribution of 3rd, 4th and 5th proximities for Iris dataset using Cosine similarity

The above histograms show that the proximity measures follow a logarithmic distribution for the Iris dataset using Cosine similarity.

Though the similarity measures of record i and j are the same, i and j may not be ranked the same while looking at the k closest records.

Also, there isn't one record that's closest to all records.

The following visualization(Figure 18) is a table encompassing the means of each of the proximity distributions got by running the program for Iris dataset setting k=5 using

Euclidean and Cosine Similarity measures. We can see that the means of proximity distributions are different for Euclidean and cosine similarity measures.

Proximity Distribution i	**Cosine Similarity**	**Euclidean Distance**
1	0.99849	0.93286
2	0.997716	0.91352
3	0.997164	0.90031
4	0.99648	0.8914
5	0.995977	0.88394

Figure 18 - Table of means

CONCLUSION

Analysis of the Iris and Income datasets were carried out and the above visualizations and techniques were documented.