Data Visualisation(590V) - Project Report
Submitted by : Ankita Mehta
Webpage Link

Dataset Overview

The dataset taken for this project has 48, 842 records with 15 features. It is the adult dataset with size 5.3 MB, also knows as the ”Census Income” dataset hosted by the UCI Machine Learning Repository. Here is the link from where it has been downloaded : https://archive.ics.uci.edu/ml/datasets/Adult/

Dataset has following fields:

1. Age : Ordinal attribute denoting age of the person - Range from 17 to 90
2. Workclass : Nominal attribute denoting the working class of a person with values - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. Fnlwgt : Ordinal attribute denoting the survey weight - Range from 12285 to 1490400
4. Education : Nominal attribute denoting the education level of a person with values - Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. Education-num : Ordinal attribute denoting the educational number given to each education level - Range from 1 to 16.
6. Marital-status : Nominal attribute denoting the marital status of a person with values - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. Occupation : Nominal attribute denoting occupation level of a person with values - ech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. Relationship : Nominal attribute denoting the relationship status of a person with values - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. Race : Nominal attribute denoting race of a person with values - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. Sex : Nominal attribute having two values - male or female.
11. Capital-gain : Ordinal attribute denoting the capital gain contributed by the person - Range from 0 to 99999
12. Capital-loss : Ordinal attribute denoting the capital loss caused by the person - Range from 0 to 4356
13. Hours-per-week : Ordinal attribute denoting the total number of hours worked per week by the person - Range from 1 to 99
14. Native-country : Nominal attribute denoting the country of origin of the person with values - United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad &Tobago, Peru, Hong, Holand-Netherlands.
15. Income : Nominal attribute having two values $> 50k$, $< 50k$

Various points that can be noticed here are as follows:

- Maximum records in this dataset are from United States.
- There are many missing values in the dataset, denoted by ?.

- The Nominal values in the dataset has to be changed to ordinal to perform various Machine Learning techniques like classification, clustering, PCA.
- It can also be noticed that educational-num and education attribute are giving the same information i.e the level of education of the person. For e.g: education-num 6 has been given to education level -10th and so on.
- fnlwgt attribute can be ignored here, because it is not related anyhow to the other attributes.

Interest

I am interested in this dataset because of the following reasons:

- Firstly, the attributes/fields present in this dataset are related to daily life which makes it very easy to understand and work upon.
- Secondly, this dataset has a very concrete goal i.e, whether the salary of the person is $> 50k$ or $< 50k$ and it has lot of records, so unsupervised and supervised machine learning techniques can be easily explored. Also it has 15 attributes, so Feature Selection and PCA can also be considered.
- Thirdly, it has a geospatial attribute - Native Country, so Chloropleth visualisation can be explored for this project.

Various hypothesis that I answered about this dataset through my analysis are:

- What are the attributes upon which salary of a person is most-commonly dependent. Is the salary been affected by their race or marital-status etc. ?

  Answer to this hypothesis is that salary of a person is very much influenced by his sex, marital-status, education and occupation whereas race, number of hours worked and his age are not the influential factors while analysing the salary of an individual.
- How are the factors like sex, education etc related to each other ?

  Answer to this hypothesis has been given by using parallel coordinates visualisation technique - which is education and educational-num attributes are highly related to each other as they are conveying the same information. Whereas other attributes are not directly related to each other but collectively, say sex and education, are creating a great impact on the salary of an individual.
- Which classifier has highest performance over this dataset : KNN , SVM etc.

  This hypothesis is done to evaluate the analysis done about the dataset. It has been shown that an individual's income is highly impacted by the factors given by predicting this attribute using various classifiers. Random Forest performed the best on this dataset by predicting the salary of a person with $85 percent$ accuracy and acquiring 90 percent area under the Roc curve.
- How does Principal Component Analysis performed on this dataset ?

  As a part of evaluation of analysis, dimensions of given data are reduced to two. This has been done dynamically by choosing the attributes to include for PCA. It can be shown that persons salary is most affected by his education, marital-status, sex and occupation.

Analytics

Data is cleaned before processing in following ways :

- Firstly, Since dataset has around 48k records and around 3500 records has ? as missing values. So, these records can be deleted straightforward.
- Secondly, Nominal values needs to be changed to ordinal values to apply Machine Learning techniques and to make Parallel coordinates Visualisation. For an eg: $> 50k$ and $< 50k$ can be changed to 1 and 0 respectively.

Various analytical techniques that are applied here :

- Statistical information like Minimum/Maximum/Average are calculated for ordinal attributes in the dataset.
- Classification techniques like SVM, KNN, Random Forest, Naive Bayes and Decision Trees have been applied and their performance haven been evaluated.
- Dimensionality Reduction techniques like PCA has been explored.

Visualizations

Various Visualizations that will be explored for this dataset are :

1. **Table** having dataset information.
2. **Grouped Barchart** showing the division of records according to the specific attributes for salary $> 50k$ and $< 50k$. (For example : $> 50k$ and $< 50k$ salary distribution for both male and female - sex attribute)
3. **Scatterplot** showing the relationship between nominal attributes(X-axis) and the ordinal attributes(Y-axis) with Income as the color attribute and also for PCA analysis
4. **Pie Chart** having percentage of samples according to various attributes.
5. **Choropleth** showing the world map with distribution of records in various countries.
6. **Parallel Coordinates** is used for showing the correlation between various features of this dataset.
7. **Line chart** is used for comparing the performance of various classifiers.
8. **Bar chart** is for comparing the performance of various classifiers.

Interactions

Selection, Probing, Drop-down menu , Buttons, Checkboxes have been used wherever necessary.

Story

This project has explored the Adult income dataset by telling what are the main factors affecting the salary of a person. It has been done by analysing the dataset using various plots viz: grouped bar chart, scatterplot, pie-chart. And then it has also been seen that if these attributes are related to each other or not. Then at the end, all the analysis done has been evaluated by applying various Machine Learning classifiers. It has been done to check which classifier performs the best or whether our analysis done is correct or not. It may be possible that salary of a person is not at all dependent on his/her education, sex , marital-status and we have done wrong predictions. As a part of evaluation, PCA has also been applied on the dataset, which is nothing but a dimensionality reduction technique. Using that one can see which all attributes when combined together can cluster the dataset into two parts effectively.

Implementation

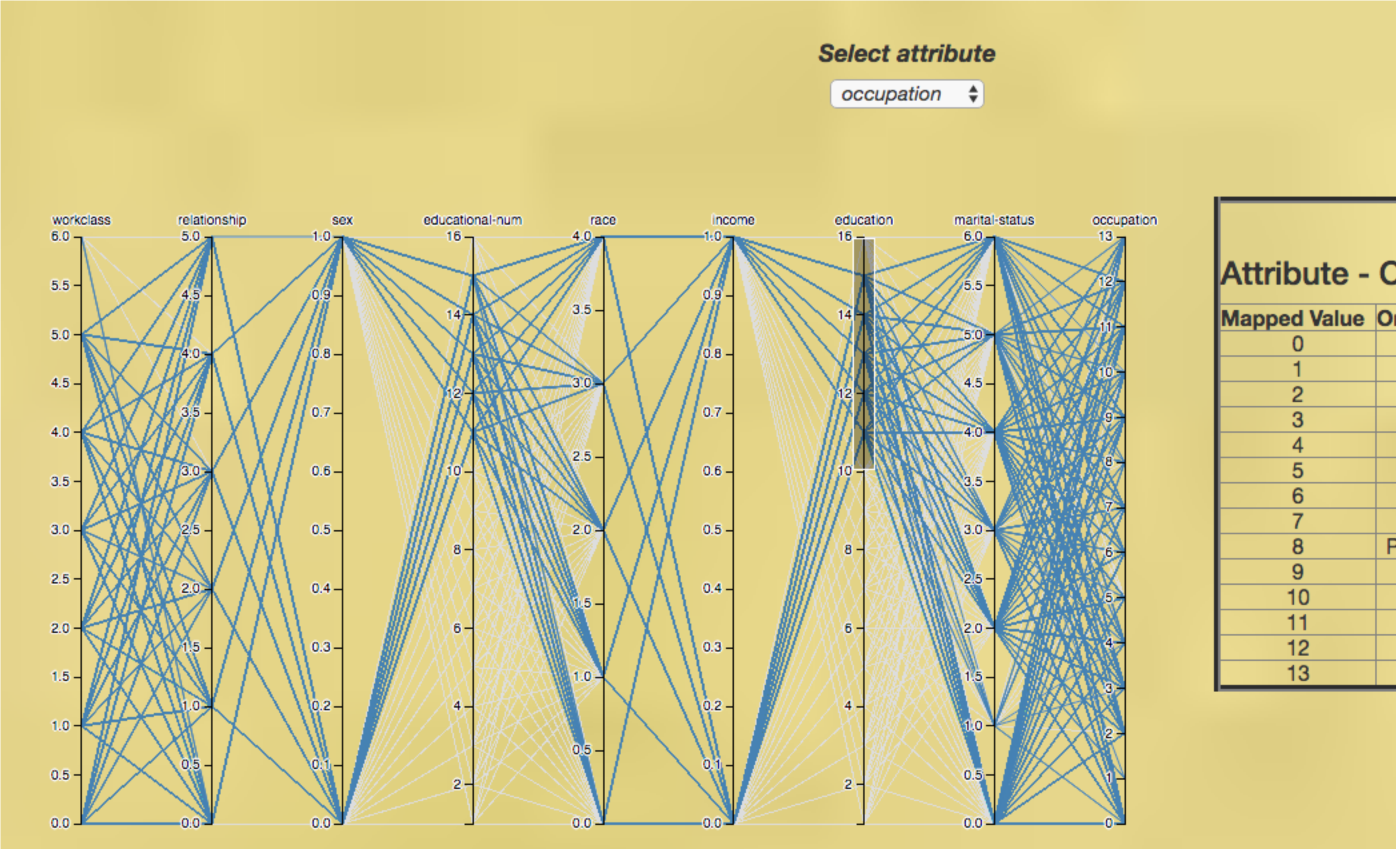
Since dataset is bit large, loading webpage will take around 30-40sec. The last plot loaded will be Scatterplot for PCA. As you can see, there are four sections in the webpage:

- Firstly, the dataset has been explained - what are the factors that can affect the salary of an adult. Like - education, relationship, marital-status, sex, origin, race etc. Tables and world map have been made to describe the dataset.
- Then three plots viz: Grouped bar chart, Pie chart and scatterplot have been made for analysing the affect of various factors described above. All these charts are interacting. (On clicking any pie/bar of pie chart and grouped bar chart, respective area will be displayed in all the charts.) Drop-down menu has been provided to select the X and Y-axis for the charts.

  Please have some patience while interacting with the charts, as they will take some seconds to filter.
- In the next section, Parallel coordinates visualisation has been made which shows the correlation amongst the attributes described in 1st section. For this, nominal attributes have been mapped to the ordinal values which can be seen in the table. Attributes can be selected from the drop down menu.

  Parallel coordinates are interactive in the sense that you can change the order of attributes by dragging and dropping. Specific points can be selected using the cursor over that axis.

You can find the visualisation like this after selection -



Select attribute

occupation

Attribute - C

Mapped Value	Or
0	
1	
2	
3	
4	
5	
6	
7	
8	F
9	
10	
11	
12	
13	

Interacting with this visualisation will take some time as filtering needs to be done over the large dataset.

- Last section is for evaluating the analysis done using classification techniques like SVM, Random Forest, Decision Trees, Naive Bayes, KNN. Two plots viz: bar chart and line chart has been made for their performance analysis.
- As a part of evaluation, scatterplot has been made to perform PCA dynamically. With the given checkboxes, you can select the attributes and check which set of attributes can cluster the dataset into two parts effectively.

Previous work and WebPage Link

The page referenced for the analysis of this dataset : Previous work

Link to the webpage for this project can be found at : Project Link