#### Prediction of Income Levels

Rubiyah Biamin, K2222445D

*Abstract – For this assignment, we examine the Census Income dataset available at the UC Irvine Machine Learning Repository. We aim to predict whether an individual’s income will be greater than $50,000 per year based on several attributes from the census data.*

**Introduction**

Over the last two decades, humans have grown a lot of dependence on data and information in society and with this advent growth, technologies have evolved for their storage, analysis and processing on a huge scale. The fields of Data Mining and Machine Learning have not only exploited them for knowledge and discovery but also to explore different patterns and concepts which led to prediction of future events, not easy to obtain.

In our first section, we will use a number of different supervised algorithms to precisely predict individuals’ income using the Census data. We will then choose the best candidate algorithm by comparing the models as well as that of others in order to find out what features of significance, what methods are most effective. Our goal with this implementation is to build a model that accurately predicts the income levels.

In the second section, we will use a clustering technique particularly the Kmeans algorithm to reveal segments of the population with similar characteristics.

**A The Dataset**

The data for our study was accessed from the University of California Irvine(UCI) Machine Learning Repository. It was actually extracted by Barry Becker using the 1994 census data. The data set includes figures on 32,561 different records and 14 attributes for 42 nations. The 14 attributes consist of 8 categorical attributes and 6 continuous attributes as shown in Figure 1 below. The binomial label in the dataset is the income level which predicts whether a person earns more than $50,000 dollars per year or not based on the given set of attributes. The last column will be our target variable, ‘income’, and the rest will be the features. (labeled F1 etc in the table below)

Figure 1

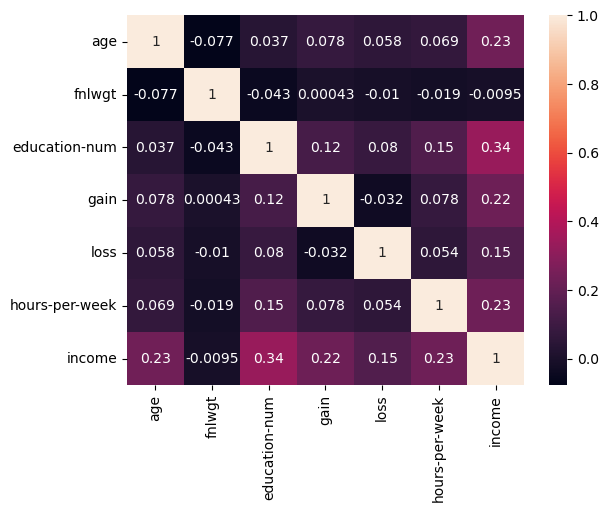
|  |  |  |
| --- | --- | --- |
| ID | Attribute Name | Type |
| F1 | age | continuous |
| F2 | workclass | categorical |
| F3 | fnlwgt | continuous |
| F4 | education | categorical |
| F5 | education-num | continuous |
| F6 | marital-status | categorical |
| F7 | occupation | categorical |
| F8 | relationship | categorial |
| F9 | race | categorical |
| F10 | sex | categorical |
| F11 | capital-gain | continuous |
| F12 | capital-loss | Continuous |
| F13 | Hours-per-week | continuous |
| F14 | Native-country | categorical |

**B** **Feature Study and Selection**

A correlation matrix is shown in Figure 2, in the form of a Heat Map showing Feature-to-Feature and Feature-to-Label Pearson Correlations where all the features are Continuous Variables.

In the heatmap, we observe a positive correlation between education and education-num, indicating that individuals with higher levels of education tend to have a higher education-num. This aligns with our expectations, suggesting that the two variables capture similar information. However, not very strong correlation among variables seen.

**Figure 2**



Relatively education has the highest correlation +0.34 with income

Capital Gain, age and hours per week are also positively correlated with income with a correlation coefficient of around 0.23.

Most of the variables are also positively correlated with each other with highest correlation observed between gain and education, and education and hours worked.

A heatmap of the correlation matrix visually represents the strength and direction of relationships between pairs of features in a dataset. Each cell in the heatmap corresponds to the correlation coefficient between two features. The color intensity indicates the strength and direction of the correlation:

* **Positive Correlation (High values):** A strong positive correlation is indicated by a cell with a color closer to 1.0. This suggests that as one feature increases, the other feature tends to increase as well.
* **Negative Correlation (Low values):** A strong negative correlation is indicated by a cell with a color closer to -1.0. This suggests that as one feature increases, the other feature tends to decrease.
* **Weak or No Correlation (Values close to 0):** Features with values close to 0 indicate a weak or no linear correlation.

**C Data Visualization**

We shall look into the

* Histogram to study the shape of the numeric data
* BoxPlot to have an idea of outliers
* Plotly Pie Charts
* Plotly Scatter Plots
* Countplot for the income variable

A group of blue and white bars

Description automatically generated



The age feature describes the age of the individual. Figure 1 shows the age distribution among the entries in our datset. The age range from 17 to 90 years old wuth the majority of entries between the ages of 25 and 50 years. Becaue there are so many ages being represented, we bucket the entrues into age groups with intervals of tem years to presente the data more concisely as seen in Figure 2. Looking at the graph, we can see that there is a significant amount of variance between the ratio of >50K to <=50K between the age groups. The most interesting ratios to note are those of groups 17-20, 71-80 and 81- 90 where there is almost no chance to have na income of greater than $50,000. The ratio of entries labeled >50K to <=50K for age groups 21-30, 31-40, 41-50, and 51-60 vary significantly as well.

Data visualization has also been done using the Box-and-Whisker plots of all continuous features to clearly understand the measures of central tendency.

A group of lines with numbers

Description automatically generated with medium confidence

A diagram of a graph

Description automatically generatedA graph of a graph with numbers

Description automatically generated with medium confidence

Box and Whisker for ‘Age’ attribute

Box and Whisker for ‘fnlwgt’ attribute

A graph of a number of circles

Description automatically generated with medium confidence

A graph with numbers and a line

Description automatically generated

Box and Whisker for ‘loss’ attribute

Box and Whisker for ‘gain’ attribute

A diagram of a graph

Description automatically generatedA graph with lines and dots

Description automatically generated

Box and Whisker for ‘hours-per-week’ attribute

Box and Whisker for ‘education-num’ attribute

A colorful pie chart with numbers

Description automatically generatedPercentage of education

A graph with red and blue dots

Description automatically generated

**Findings and Insights**

* The minimum age is 17 and the maximum is 90 years, most of the working age group lies between 20-40.
* The minimum hours-per-week is 1 and maximum is 90, with most of the count lying between 30-40
* Outliers observed in almost all the numeric features, these are the extreme values that are presente in the data.

More inferences were made in the Python Notebook illustrating extensive data visualization on different attributes.

For example

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

1. Most of the people who are white receive education for around 9 years and get a salary of <=50k.

2. Very few white people seem to get education for around 13 years and receive a salary of

>50k.

3. All other races do not get that much education and therefore have lower salary.

**D** **Data Preprocessing**

Data must be preprocessed in order to be used in Machine Learning algorithms. This preprocessing phase includes the cleaning and preparing the data.

**Training the MODEL**

* Create X and y object to store the independent variable (X) and dependent variable.
* Perform Standard Scaling to scale the data
* Label Encoding is performed to convert the categorical data into numeric format
* Label Encoder makes the data suitable for machine
* Perform fit and Transform

A screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generatedSupervised Learning:**

We applied three different supervised classification algorithms.

For each algorithm, we evaluated performance using the following metrics:

* Accuracy
* Classification Report (Precision, Recall, F1-score)
* Confusion Matrix

Out of a total of instances present in the dataset, 22 792 instances have been used for training while the rest 9 769 instances have been reserved for testing. After complete evaluation, the model performances are evaluated on the following metrics:

|  |  |
| --- | --- |
| **Classification Algorithm** | **ACCURACY %** |
| **Logistic Regression** | **82.0** |
| **Decision Tree** | **81.1** |
| **Random Forest** | **85.5** |

Confusion Matrix

**Random Forest Classifier (RFC) is chosen for the**

A blue and white chart with numbers

Description automatically generated**Analysis**

Since RFC model uses a number of decision tree classifiers to come up with a mean prediction, it is preferred to applying a single decision tree classifier.

* Random forest classifier(RFC) is chosen because it has a higher accuracy compared to Logistic Regression Classifier. (85% vs 82%)
* RFC is preferred since it reduces the overfitting tendency of decision tree classifier.

All the results are tabulated below:

|  |  |
| --- | --- |
| Training Accuracy | 100% |
| Testing Accuracy | 84.78% |
| Recall | 93% |
| Precision | 88% |
| F1-score | 90% |

According to the obtained metics, it can be concluded that the model is a good fit.

**Unsupervised Learning:**

We applied the K-Means clustering algorithm to uncover segments within the population. The resulting clusters were visualized using a scatter plot.It has been shown that clustering on the Census dataset may not have a clear interpretation or meaningful application, as it is typically used for classification tasks.

A chart with a number of dots

Description automatically generated with medium confidence

K = 3

A diagram of a cluster of colored dots

Description automatically generated

K=4

A diagram of a cluster of dots

Description automatically generated

K=5

Silhouette score = 0.04

A cluster of dots with numbers and lines

Description automatically generated with medium confidence

A silhouette score of 0.04 using KMeans and 0.06 using Agglomerative Clustering technique, in the context of clustering on the Census dataset suggests that the clusters, as defined by the clustering algorithm, are not very well-separated or distinct. The silhouette score measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A score of 0.04 to 0.06 is relatively low, indicating that there is some overlap between clusters, and the objects in the clusters are not well-separated from each other.

|  |  |
| --- | --- |
|  | Silhouette Score |
| KMeans with K= 3,4,5 | 0.04 |
| Agglomerative Clustering | 0.06 |

In the context of the Census dataet, which is commonly used for classification tasks rather than clustering, achieving a high silhouette score might be challenging. The data set is primarily used for predicting income levels (classification task) based on various features such as age, education, occupation etc.

A low silhouette score may suggest that the chosen clustering algorithm or the features used for clustering may not be well suited for identifying clear and distinct groups in the data. Additionally, it could indicate that the inherent structure of the data is not conducive to clustering or that the data may not naturally form well-defined clusters.

**CONCLUSION**

Our analysis aimed to predict income levels and uncover population segments based on demographic features. The findings from supervised learning models provides insights into the effectiveness of different algorithms for income prediction. The K-Means and Agglomerative clustering models shed light on the diversity within the dataset, identifying groups with similar characteristics.

Source Code : <https://github.com/syaq1603/RubiyahBiaminCET182.git>