#### Predict Income Levels

Rubiyah Biamin, K2222445D

**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 diferente patterns and concepts which led to prediction of future events, not easy to obtain.

Data Mining is the extraction of hidden information from large database. Classifation is a data mining task of predicting the value of a categorical variable by building a model based on one or more numerical and/or categorical variables (predictors or attributes). Classication mining function is used to gain a deeper understanding of the database structure. There are various classification techniques like decision tree induction, Bayesian networks, lazy classifier and rule based classifier.

In this Project, we will use a number of different supervised algorithms to precisely predict individuals’ income using the Census data set collected from the UCI machine learning repository. We will then choose the best candidate algorithm from preliminary results and further optimize this algorithm to best model the data. Our goal with this implementation is to build a model that accurately predicts the income levels.

We will also 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.

Figure 1

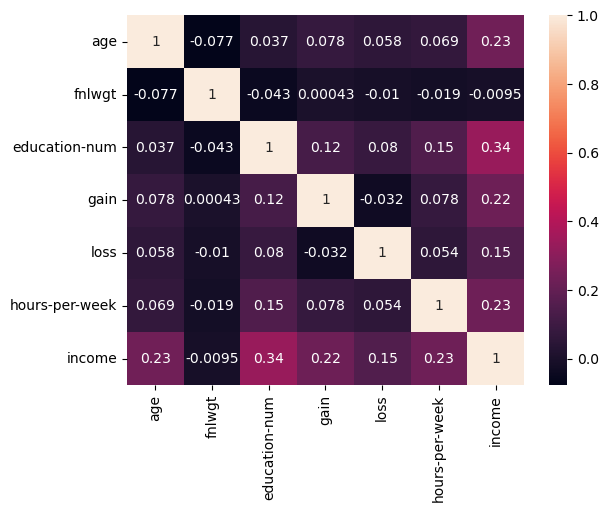
|  |  |  |
| --- | --- | --- |
| ID | Attribute Name | Type |
| F1 | age | continuous |
| F2 | workclass | categorical |
| F3 | fnlwgt | continuous |
| F4 | education | categorical |
| F5 | education-num | continuous |
| F6 | marital-status | categorical |
| F7 | occupation | categorical |
| F8 | relationship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

**A** **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.

**Figure 2**



Relatively education has the highest correlation +0.34 with income

Gain, age and hours per week are also positively correlated with income with a correlation coefficient of around 0.20.

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

A group of lines with numbers

Description automatically generated with medium confidence

A colorful pie chart with numbers

Description automatically generated

A graph with red and blue dots

Description automatically generated

**Key Findings**

The minimum age is 17 and the maximum is90 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.

**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 nmakes the data suitable for machine
* Perform fit and Transform

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Performance Evaluation**

A blue and white chart with numbers

Description automatically generated**Key Findings and Insights**

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

**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.

2. Clustering Analysis:

KMeans lustering revealed distinct groups within the dataset, with the silhouette score indicating reasonable cluster separation.

Further exploration of the charateristics defining these clusters could provide valuable insighrts into demographic and socioeconomic patterns.

1. Demographic Patterns:

Exploratory Data Analysis highlighted the diverse distribution of demographic features, with variations in age, education, and occupation.

Correlation analysis identified positive correlations between certain features and higher income levels.

**CONCLUSION**