**Toronto Metropolitan University**

**CIND820 XJH**

**Big Data Analytics Project**

**Student name: Farhana Chowdhury**

Student ID: 501069812

**Supervisor: Tamer Abdou**

Date of submission: February 16, 2023

**Abstract**

**Introduction**

Understanding the demographic factors that impact a person’s income is the main objective of this project. Using census data has provided a wide range of demographic information which helps us understand underlying factors behind income levels. UCI Adult Income Dataset (1994) is a very popular dataset that has been used in several anlysis which are of similar kind.

This dataset mostly consist of categorical attributes which facilitated us using multiple classification models to determine income based on provided multivariate dataset. In this classification problem we solve our problem through 2 steps: exploring the data and then building the machine learning model.

We have completed the project by validating our models to identify the most accurate one for the given dataset.

**Dataset**

The analysis has been conducted on the Adult Dataset from UCI Machine Learning Repository. The dataset is also known as Census Income Dataset which is an extraction from 1994 Census database. The entire dataset contains 48,842 records which is split into 2 files – adult.data (main / training datatset) and adult.test (test dataset). A total number of 14 attributes are in the dataset which are a mix of categorical and integer data type.

This project solves a classification problem where the target outcome is to determine whether a person makes <= 50K or >50K/year based on the different demographic info.

**The attribute information are as follows:**

* **Age:** *continuous*
* **Workclass:** (*categorical*) Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
* **fnlwgt:** (*continuous*) final weight, weight on the CPS files prepared by Population Division;
* **Education:** (*Categorical*) Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
* **Education-num**: *continuous*.
* **Marital-status:** (*Categorical*) Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
* **Occupation:** (*Categorical*) Tech-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.
* **Relationship:** (*Categorical*) Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
* **Race:** (*Categorical*) White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
* **Sex:** (*Categorical*) Female, Male.
* **Capital-gain:** *continuous*.
* **Capital-loss:** *continuous*.
* **Hours-per-week:** *continuous*.
* **Native-country:** (*Categorical*) 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.

**Following are a few highlights about the dataset:**

* 8 of the 14 attributes are categorical
* Among the 4 integer attributes some attributes deemed to be not that related to the target variable, such as: fnlwgt, capital-gain and capital-loss which we ended up dropping
* The dataset also seemed to have selection bias as 29,170 instances came from one nationality (US) whereas the rest 19,672 records were taken from 40 nationalities’

**Research Problem**

* Predicting whether a person makes <= 50K or >50K/year

**Methodology**

The entire project has been conducted in Python using Jupyter Notebook.

The first part of our project contains Exploratory Data Analysis (EDA) where we have run extensive explorations to understand the data, clean the data, feature engineering, univariate and bivariate analysis. Using the popular python libraries – Numpy, Pandas, Matplotlib and Seaborn we have conducted the following step by step process for the EDA:

* Creating the dataframe
* Understanding data types
* Dropping irrelevant attributes
* Identifying missing values and replacing them with Mode
* Feature engineering – simplified some categorical attributes and clubbed them in lesser category for better analysis
* Relationship Analysis – correlation heatmap, pairplot, relplot, displot for distribution, countplot, catplot etc.

The second part of the project includes building machine learning models to solve the classification problem. In this part, we have tried to identify accuracy and most important features of the datatset. Using sklearn library we have first created the train-test split on a 70-30 ratio and created the following models to test accuracy:

* Principal Component Analysis (PCA)
* Logistic Regression
* Random Forest
* Decision Tree
* Naïve Bayes Classifier

At the end we also conducted a Cross-validation to identify the best suited model.

**Conclusion:**

We have identified the most accurate model for the dataset to predict whether a person makes <= 50K or >50K/year.

**References:**

UCI Machine Learning Repository: Adult Data Set. (n.d.). Retrieved February 15, 2023, from https://archive.ics.uci.edu/ml/datasets/Adult

Priyanka. (2021, September 2). *Adult UCI dataset analysis with python*. Machine Learning / Deep Learning / AI Tutorials and Projects. Retrieved February 15, 2023, from https://machinemantra.in/adult-uci-dataset-analysis-with-python/

Jieyima. (2018, February 23). *Income classification model*. Kaggle. Retrieved February 15, 2023, from https://www.kaggle.com/code/jieyima/income-classification-model

Lopez Torres, I. (2022, December 29). *Adult UCI Dataset Prediction Analysis*. SSRN. Retrieved February 15, 2023, from https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4307371