ADULT INCOME CENSUS DATASET PROJECT PLANNING

BACKGROUND

This machine learning project examines the potentially significant effects that the census income data has on the yearly income of an individual. This project consists of two primary components, performing data cleaning techniques on the dataset, and applying machine learning models on the cleaned data. By performing these machine learning techniques we aim to successfully predict that an income exceeds the value of $50,000 annually or not.

GOALS

To predict if an individual has an annual income above a set limit or not. For this project, the limit is set to $50,000 per year. The machine learning model we aim to use for the dataset is the classification model. The classification model is defined as the model which uses machine learning techniques to predict the outcome by labeling classes based on the various inputs given to the model.

DATA DESCRIPTION

The dataset was obtained from the university of california Irvine machine learning repository. This dataset which is named as the adult dataset was extracted by Barry Becker from the 1994 census database. A set of pre conditions were applied to obtain the data to retrieve clean records. The data description is as follows:

Dataset Characteristics: Multivariate, the dataset consists of two or more variable quantities which are observed

Number of instances: 48842, the number of rows in the dataset

Area: Social, the domain of the dataset comes from a social background perspective

Attribute Characteristics: Categorical, Integer. The attributes of the dataset take the form of a categorical data and integer based data.

Number of attributes: 14, the number of columns in the dataset

Date Donated: 1996-05-01

Associated Tasks: Classification, the machine learning model to be applied on this dataset is the classification model for optimal results

Missing Values: Yes, the dataset consists of missing values which needs to be cleaned

Number of web hits: 2183850

ATTRIBUTE INFORMATION

The attributes of the dataset are defined as the columns in the dataset. Below is an explanation of each attribute by first stating the type of the attribute and what it signifies.

- **age**: The age of an individual
  - Continuous variable with an integer value of above 0
- **workclass**: The employment status of an individual
  - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
- **fnlwgt**: The number of people an entry represents
- **education**: The level of education of an individual
  - 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**: The level of education in numerical form
  - Continuous variable
- **marital-status**: Marital status of an individual
  - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
- **occupation**: Occupation of an individual
  - Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
- **relationship**: What the relationship status of an individual is
  - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
- **race**: Race of an individual
  - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
- **sex**: Biological sex of an individual
  - Female, Male
- **capital-gain**: The capital gain of an individual
  - Continuous variable with an integer value of 0 or above
- **capital-loss**: The capital loss of an individual
  - Continuous variable with an integer value of 0 or above
- **hours-per-week**: The number of hours an individual has worked in a week
  - Continuous
- **native-country**: Country of origin for an individual
  - 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, Trinidad Tobago, Peru, Hong, Holland-Netherlands

PROJECT PLAN

We will be using different types of classification algorithms to classify our output in order to minimize the error and improve the accuracy. Once the results are obtained we will then compare the results from the different classification algorithms to determine the most accurate classification technique for predicting the yearly income of an individual. For implementation, we will be dividing our data into two sets, training data and testing data. This will be done to train the model and validate the trained model against the test data to check our model. We will be using RMSE (Root mean squared error) and other testing algorithms to check the accuracy of our model.