Adult-Census Income Prediction

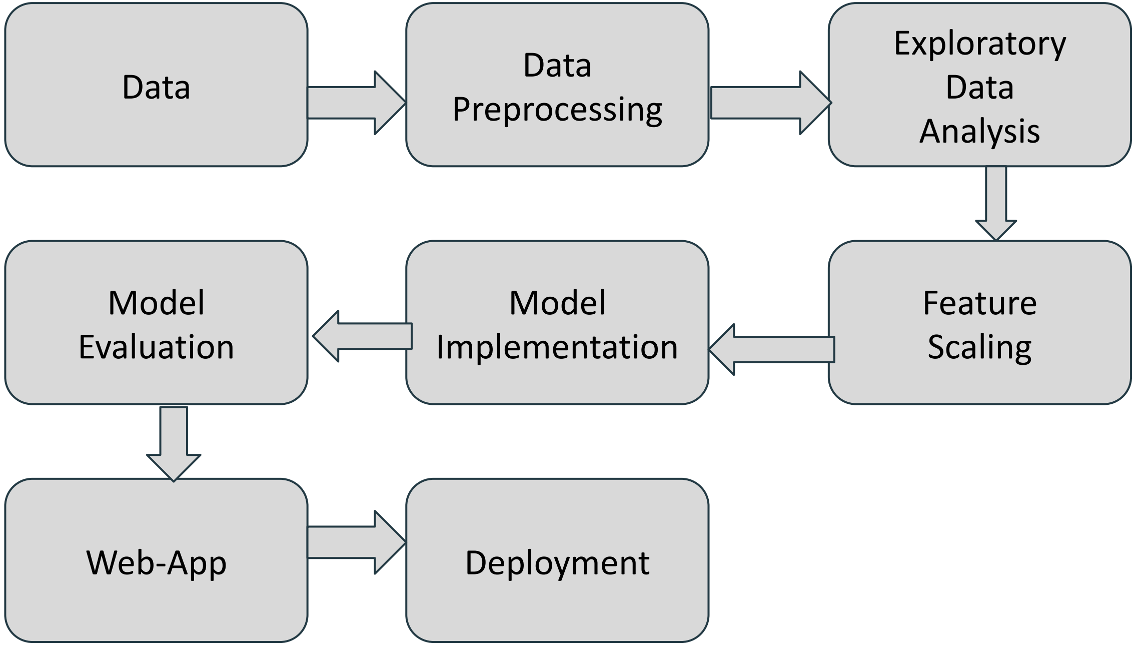
Project Architecture

Domain : Machine Learning

Creator : Meghana Munjeti
Mail-ID : meghameghana297@gmail.com

Date :11/10/2023

Architecture



Data

Data Preprocessing

Exploratory Data Analysis

Feature Scaling

Model Implementation

Model Evaluation

Web-App

Deployment

Architecture Description

Data Preparation

Data Description

The Adult census dataset is a publicly available dataset that contains demographic and socioeconomic information about individuals, such as age, gender, education, occupation, and income. The dataset was collected by the US Census Bureau in 1994 and contains over 48,000 records.

The following is a description of the variables in the Adult census dataset:

- **age:** The age of the individual in years.
- **workclass:** The individual's workclass, which can be one of the following: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt:** The individual's final weight. This is a weight that is used to make the dataset more representative of the US population.
- **education:** The individual's level of education, which can be one of the following: 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 number of years of education that the individual has completed.
- **marital-status:** The individual's marital status, which can be one of the following: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent.

- **occupation:** The individual's occupation, which can be one of the following: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-special, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Armed-Forces.
- **relationship:** The individual's relationship to the person filling out the census form, which can be one of the following: Husband, Wife, Own-child, Unmarried, Other-relative, Not-in-family.
- **race:** The individual's race, which can be one of the following: White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other.
- **sex:** The individual's sex, which can be one of the following: Male, Female.
- **capital-gain:** The individual's capital gains in the past year.
- **capital-loss:** The individual's capital losses in the past year.
- **hours-per-week:** The number of hours per week that the individual works.
- **income:** The individual's annual income. This is the variable that is being predicted in the dataset.

Data Usage

The Adult census dataset can be used for a variety of tasks, including:

- **Predicting income:** The dataset can be used to train a machine learning model to predict the income of an individual based on their demographic and socioeconomic characteristics.
- **Studying poverty:** The dataset can be used to study the factors that contribute to poverty.
- **Developing social programs:** The dataset can be used to develop social programs that are targeted at helping people with low incomes.
- **Conducting research:** The dataset can be used to conduct research on a variety of topics, such as social mobility, economic inequality, and labor markets.

Data Preprocessing:

Data preprocessing is a crucial step in machine learning, involving cleaning, transforming, and normalizing data to make it suitable for model training and analysis.

Data preprocessing steps:

- Handle missing values: The adult census dataset contains missing values which can be handled by removing the rows with missing values or imputing with the mean or median value for the feature.
- Convert categorical features to numerical features: Most of the features in the adult census dataset are categorical, which can be converted to numerical features using techniques such as label encoding.
- Scale features: Scaling features helps improve the performance of machine learning algorithms by transforming the features to have the same mean and standard deviation.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is a process of investigating data to discover patterns, anomalies, and relationships. It is an essential step in any data science project, as it helps us to better understand the data and to identify the best way to approach the problem.

Here are some key EDA findings on the Adult Census dataset:

- **Income distribution:** The distribution of income in the dataset is highly skewed, with a majority of individuals earning less than $50,000 per year.
- **Education and income:** There is a strong correlation between education and income. Individuals with higher levels of education are more likely to earn higher incomes.
- **Occupation and income:** There is also a strong correlation between occupation and income. Individuals in certain occupations, such as Executive/Managerial and Professional/Speciality occupations, are more likely to earn higher incomes.

- **Gender and income:** On average, males earn more than females. However, this gap is narrowing over time.

The following are some EDA techniques that can be used on the Adult Census dataset:

- **Univariate analysis:** This involves analyzing each feature individually to understand its distribution and identify any outliers.
- **Bivariate analysis:** This involves analyzing two features together to identify correlations and relationships.
- **Multivariate analysis:** This involves analyzing three or more features together to identify complex relationships.

Some specific EDA visualizations that can be used on the Adult Census dataset include:

- **Histograms:** Histograms can be used to visualize the distribution of each feature.
- **Scatter plots:** Scatter plots can be used to visualize the relationship between two features.
- **Box plots:** Box plots can be used to compare the distributions of two or more groups.
- **Heatmaps:** Heatmaps can be used to visualize the correlations between all pairs of features.

Model Implementation:

Once the data has been preprocessed and explored, the next step is to implement a machine learning model to predict income. There are many different machine learning algorithms that can be used for this task, such as logistic regression, decision trees, and random forests. The F1-Score is obtained from the Random-Forests is the highest.

Model Evaluation:

Test Dataset is used to evaluate the model. 20% of dataset is separated for tested. Predicted results from the model is compared with the actual data to calculate the model performance.

Deployment:

We have used Flask for web-application.Users can give the values in the respective fields and click on Predict button to display the result. We deployed our model in AWS Cloud Platform.