Census Income Project

support@intellipaat.com

+91-7022374614

US: 1-800-216-8930 (Toll-Free)

In this project, you are going to work on the The "Census Income" data set from the UCI Machine Learning Repository that contains the income information for over 48,000 individuals taken from the 1994 US census.

For more details about this dataset, you can refer to the following link: https://archive.ics.uci.edu/ml/datasets/census+income

Problem Statement:

In this project, initially you need to preprocess the data and then develop an understanding of different features of the data by performing exploratory analysis and creating visualizations.Further, after having sufficient knowledge about the attributes you will perform a predictive task of classification to predict whether an individual makes over 50K a year or less,by using different Machine Learning Algorithms.

Tasks to be done:

1. **Data Preprocessing:**
   a) Replace all the missing values with NA.
   b) Remove all the rows that contain NA values.

2. **Data Manipulation:**
   a) Extract the “education” column and store it in “census_ed” .
   b) Extract all the columns from “age” to “relationship” and store it in “census_seq”.
   c) Extract the column number “5”, “8”, “11” and store it in “census_col”.
   d) Extract all the male employees who work in state-gov and store it in “male_gov”.
   e) Extract all the 39 year olds who either have a bachelor's degree or who are native of the United States and store the result in “census_us”.
   f) Extract 200 random rows from the “census” data frame and store it in “census_200”.
   g) Get the count of different levels of the “workclass” column.
   h) Calculate the mean of the “capital.gain” column grouped according to “workclass”.
   i) Create a separate dataframe with the details of males and females from the census data that has income more than 50,000.
   j) Calculate the percentage of people from the United States who are private employees and earn less than 50,000 annually.
   k) Calculate the percentage of married people in the census data.
   l) Calculate the percentage of high school graduates earning more than 50,000 annually.

3. **Linear Regression:**
   a) Build a simple linear regression model as follows:

- Divide the dataset into training and test sets in 70:30 ratio.
- Build a linear model on the test set where the dependent variable is “hours.per.week” and the independent variable is “education.num”.
- Predict the values on the train set and find the error in prediction.
- Find the root-mean-square error (RMSE).

4. Logistic Regression:

a) **Build a simple logistic regression model as follows:**

- Divide the dataset into training and test sets in 65:35 ratio.
- Build a logistic regression model where the dependent variable is “X”(yearly income) and the independent variable is “occupation”.
- Predict the values on the test set.
- Build a confusion matrix and find the accuracy.

b)**Build a multiple logistic regression model as follows:**

- Divide the dataset into training and test sets in 80:20 ratio.
- Build a logistic regression model where the dependent variable is “X”(yearly income) and independent variables are “age”, “workclass”, and “education”.
- Predict the values on the test set.
- Build a confusion matrix and find the accuracy.

5. Decision Tree:

a) Build a decision tree model as follows:

- Divide the dataset into training and test sets in 70:30 ratio.
- Build a decision tree model where the dependent variable is “X”(Yearly Income) and the rest of the variables as independent variables.
- Predict the values on the test set.
- Build a confusion matrix and calculate the accuracy.

6. Random Forest:

a) Build a random forest model as follows:

- Divide the dataset into training and test sets in 80:20 ratio.
- Build a random forest model where the dependent variable is “X”(Yearly Income) and the rest of the variables as independent variables and number of trees as 300.
- Predict values on the test set

- Build a confusion matrix and calculate the accuracy

7. For this problem, use the population dataset, and perform the following:

1. EDA on the time series to find trends and seasonality.
2. Forecast the population on the given dataset for the next 6 months.