Data Science Project

Final Project on “Census Income” Dataset

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.

Census Income Dataset:

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband
3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband
5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife
6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife
7	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family
8	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband
9	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family

Lab Environment: RStudio

Domain: Social

Tasks to be done:

1.Data Preprocessing:

a)Replace all the missing values with NA.
b)Remove all the rows that contain NA values.
c)Remove all whitespaces from the columns.

2.Data Manipulation:

Your task is to perform data manipulation to analyze the data set using various functions from the **dplyr** package.

Questions:

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 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 “capital.gain” column grouped according to “workclass”.

3.Data Visualization:

a) Build a bar-plot for the “relationship” column and fill the bars according to the “race” column.

i. Set x-axis label to ‘Categories of Relationships’
ii. Set y-axis label to ‘Count of Categories’
iii. Fill the bars according to “sex”
iv. Set the position of the bars to “dodge”
v. Set the title of plot to be ‘Distribution of Relationships by Sex”

b) Build a Histogram for the “age” column with number of bins equal to 50.

i) Fill the bars of the histogram according to yearly income column i.e., “X”
ii)Set the title of the plot to "Distribution of Age".
iii)Set the legend title to "Yearly income".
iv) Set the theme of the plot to black and white.

c)Build a scatter-plot between “capital.gain” and “hours.per.week”. Map “capital.gain” on the x-axis and “hours.per.week” on the y-axis.

i)Set the transparency of the points to 40% and size as 2.
ii)Set the color of the points according to the “X” (yearly income) column.
iii)Set the x-axis label to “Capital Gain”, y-axis label to “Hours per Week”, title to “Capital Gain vs Hours per Week by Income”, and legend label to “Yearly Income”.

d) Build a box-plot between “education” and “age” column.Map “education” on the x-axis and “age” on the y-axis.

i) Fill the box-plots according to the “sex” column.
ii)Set the title to "Box-Plot of age by Education and Sex".

4. Linear Regression:

a) Build a simple linear regression model as follows:

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

5.Logistic Regression:

a)Build a simple logistic regression model as follows:

i)Divide the dataset into training and test sets in 65:35 ratio.
ii)Build a logistic regression model where the dependent variable is “X”(yearly income) and independent variable is “occupation”.
iii)Predict the values on the test set.
iv)Plot accuracy vs cut-off and pick an ideal value for cut-off.

v)Build a confusion matrix and find the accuracy.

vi)Plot the ROC curve and find the auc(Area Under Curve).

b)Build a multiple logistic regression model as follows:

i)Divide the dataset into training and test sets in 80:20 ratio.

ii)Build a logistic regression model where the dependent variable is “X”(yearly income) and independent variables are “age”, “workclass”, and “education”.

iii)Predict the values on the test set.

iv)Plot accuracy vs cut-off and pick an ideal value for cut-off.

v)Build a confusion matrix and find the accuracy.

vi)Plot the ROC curve and calculate the auc(Area Under Curve).

6.Decision Tree:

a)Build a decision tree model as follows:

i) Divide the dataset into training and test sets in 70:30 ratio.

ii)Build a decision tree model where the dependent variable is “X”(Yearly Income) and the rest of the variables as independent variables.

iii)Plot the decision tree.

iv)Predict the values on the test set.

v)Build a confusion matrix and calculate the accuracy.

7.Random Forest:

a)Build a random forest model as follows:

i)Divide the dataset into training and test sets in 80:20 ratio.

ii)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.

iii)Predict values on the test set

iv)Build a confusion matrix and calculate the accuracy