Assignment 01

Let us consider a public database called the **"Adult"** dataset, hosted on UCI's Machine Learning Repository.1 It contains approximately 32,000 observations concerning different financial parameters related to the US population: age, sex, marital (marital status of the individual), country, income (Boolean variable: whether the person makes more than $50,000 per annum), education (the highest level of education achieved by the individual), occupation, capital gain, etc. We will show that we can explore the data by asking questions like: "Are men more likely to become high-income professionals than women, i.e., to receive an income of over $50,000 per annum?"

Visit the UCI Machine Learning website and search for Adult Data Set
https://archive.ics.uci.edu/ml/datasets/Adult, spend a few moments to understand the dataset. Now, start working with the given dataset as directed:

Hypothesis: - This data set is meant for binary class classification - to predict whether the income of a person exceeds 50K per year based on some census data.

Part A (EDA) :

1. define column names.
2. See if there are any NaNs in the dataframe.
3. Print unique values for the Income column.
4. Convert the <=50Ks into 1 and the >50K into 0.
5. Extract the target variable income into a numpy array and drop it from the dataframe.
6. Let's get some summary statistics on these numerical columns hint dataframe.describe().
7. Find out the distinct count of each categorical column.
8. Now, what insights you can draw from this dataset?
9. Draw Heatmap and find the correlation between the features.

Part A (Models Implementation) :

1. Split the dataset into 60% for training and 40% for testing.
2. Train and fit your model on the given dataset.
3. Apply the following Models:
   Support Vector Machines, Decision Tree, and Random Forest.
4. Apply evaluation matrics on the above models and write which model performs better and why. Justify your answer.