US Census Income Prediction (1994 – 1995)

PROJECT OVERVIEW:

This data set comprises weighted census data from the US Census Bureau's current population surveys conducted between 1994 and 1995.
Our task is to predict if a person makes over $50,000 a year or not based on several employment and demographic-related attributes like age, marital status, citizenship, stock dividends, family status, Migration status, income, etc.

DATA DESCRIPTION:

1. The dataset was collected from the *UCI Machine Learning Repository* - https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29
2. There are approximately *1,99,523* instances in the data file and *99,762* in the test file (*2,99,285* instances), with *42* employment and demographic-related attributes.
3. The dataset contains both Numeric and Categorical variables.

COLUMN NAME	DESCRIPTION	DATATYPE
age	Age of the worker	Numeric
class_worker	Class of worker	Categorical
det_ind_code	Industry code	Numeric
det_occ_code	Occupation code	Numeric
hs_college	Enrolled in educational institution	Categorial
education	Level of education	Categorial
wage_per_hour	Wage per hour	Numeric
major_ind_code	Major industry code	Categorial
major_occ_code	Major occupation code	Categorial
hisp_origin	Hispanic origin	Categorial
sex	Sex	Categorial
region_prev_res	Region of previous residence	Categorial
stock_dividends	Dividends from stocks	Numeric
det_hh_fam_stat	Detailed household and family status	Categorial
det_hh_summ	Detailed household summary in household	Categorial
union_member	Member of a labor union	Categorial
mig_chg_msa	Migration code - change in MSA	Categorial
unemp_reason	Reason for unemployment	Categorial
mig_chg_reg	Migration code - change in region	Categorial

full_or_part_emp	Full- or part-time employment status	Categorial
capital_losses	Capital losses	Numeric
state_prev_res	State of previous residence	Categorial
mig_move_reg	Migration code - move within region	Categorial
	Tax filer status	Categorial
tax_filer_stat	Live in this house one year ago	Categorial
mig_same	Migration - previous residence in sunbelt	Categorial

APPROACH:

To predict if a person makes over $50,000 a year or not, we will follow the following approach:

1. **Understand the Problem:**
   In this approach, we will try to understand the business problem or do a sort of pre-analysis where-in we will understand which variables are essential and ask questions to perform the analysis.

2. **Data Preparation and Cleaning:**
   This step will clean the data by handling missing values, performing string manipulations, removing duplicates, and converting them into appropriate data types. In short, we will improve the data quality for our analysis.

3. **Exploratory Data Analysis:**
   This step will discover patterns and anomalies present in our dataset. We also understand if there is any correlation between any variables.

4. **Feature Scaling and Feature Engineering:**
   In this step, we will transform the raw data into suitable features for machine learning. We will perform "*one-hot*" and "*dummy encoding"* to convert the categorical features into numeric attributes.
   Then, we will scale the data using "*Normalization*" and "*Standardization*" techniques.
   Next, we will perform "*Principal Component Analysis*" (PCA) to remove the set of highly correlated variables and keep only those variables that convey most of the information.

5. **Applying Machine Learning Models:**
   We will divide the dataset into training, testing, and validation dataset.
   Since it is a classification problem, we will apply *Logistic Regression*, *Decision trees*, *Random Forest*, *Naïve Bayes*, *K Nearest Neighbors*, and *Support Vector Machine* algorithms.

6. **Model Performance and Evaluation:**
   For evaluating the model performance, we will use *Confusion Matrix*, *AUC* and *ROC* curve, *Precision*, and *Recall.*