**Classification of 1994 Census Income Data**

**Problem Statement**: To build a model that will predict if the income of any individual in the US is greater than or less than USD 50K based on the data available about that individual.

**Data Set Description**: This Census Income dataset (adult\_income.csv) was collected by Barry Becker in 1994.This data set will help you understand how the income of a person varies depending on various factors such as the education background, occupation, marital status, geography, age, number of working hours/week, etc.

Here’s a list of the independent or predictor variables used to predict whether an individual earns more than USD 50K or not:

* age
* workclass
* fnlwgt
* education
* education-num
* marital-status
* occupation
* relationship
* race
* gender
* capital-gain
* capital-loss
* hours-per-week
* native-country

The dependent variable is the “income” that represents the level of income. This is a categorical variable and thus it can only take two values:

1. <=50k
2. >=50k

**Questions**

1. Import the dataset (**adult\_income.csv**) and do a quick exploration on the data
2. Perform Data cleaning – remove redundant variables, treat missing value and outliers.
3. Perform a detailed exploratory data analysis (EDA) on the dataset to understand variables and their interactions (relationship with target variable)
4. Build classification model using different algorithms (Logistic regression, Decision tree, Random Forest, KNN etc.)
5. Compare performance of different models using best model evaluation metrics and choose the best classification model for this problem.