**ML Assignment-1**

**Problem Statement**

For this assignment, we examined the Census Income dataset available at the UC Irvine Machine Learning Repository. We aim to predict whether an individual’s income will be greater than $50,000 per year based on several attributes from the census data.

**Exploratory Analysis**

The Dataset

The Census Income dataset has 48,842 entries. Each entry is distributed in 15 variables. Out of which 6 are numeric and 9 are categorical. Each entry contains the following information about an individual:

1. **age**​: The age of an individual ranges from 17 to 90.
2. **workclass**​: The employment status of an individual

Private, Self­emp­not­inc, Self­emp­inc, Federal­gov, Local­gov, State­gov,Without­pay, Never­worked.

1. **fnlwgt​:** Final weight. In other words, this is the number of people the census believes the entry represents.
2. **education​:** The highest level of education achieved by an individual.

Bachelors, Some­college, 11th, HS­grad, Prof­school, Assoc­acdm, Assoc­voc,

9th, 7th­8th, 12th, Masters, 1st­4th, 10th, Doctorate, 5th­6th, Preschool.

1. **education­num**​: the highest level of education achieved in numerical form. (Ranges from 1 to 16)
2. **marital­status**​: Marital status of an individual. Married­civ­spouse corresponds to a civilian spouse while Married­AF­spouse is a spouse in the Armed Forces.

Married­civ­spouse, Divorced, Never­married, Separated, Widowed,

Married­spouse­absent, Married­AF­spouse.

1. **occupation**​: The type of occupation of an individual

Tech­support, Craft­repair, Other­service, Sales, Exec­managerial,

Prof­specialty, Handlers­cleaners, Machine­op­inspct, Adm­clerical,

Farming­fishing, Transport­moving, Priv­house­serv, Protective­serv,

Armed­Forces.

1. **relationship**​: Represents what this individual is relative to others. For example, an

individual could be a Husband. Each entry only has one relationship attribute and is

somewhat redundant with marital status.

Wife, Own­child, Husband, Not­in­family, Other­relative, Unmarried.

1. **race​:** Descriptions of an individual’s race.

White, Asian­Pac­Islander, Amer­Indian­Eskimo, Other, Black.

1. **sex​:** Gender of the individual

Female, Male

1. **capital­gain**​: Capital gains for an individual

Ranges from 0 to 99999

1. **capital­loss​:** Capital loss for an individual

Ranges from 0 to 4356

1. **hours­per­week​:** The hours an individual has reported to work per week

Ranges from 1 to 99

1. **native­country​:** Country of origin for an individual

United­States, Cambodia, England, Puerto­Rico, Canada, Germany,

Outlying­US(Guam­USVI­etc), India, Japan, Greece, South, China, Cuba, Iran,

Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal,

Ireland, France, Dominican­Republic, Laos, Ecuador, Taiwan, Haiti, Columbia,

Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El­Salvador,

Trinadad&Tobago, Peru, Hong, Holand­Netherlands.

1. **income​:** The target variable, which predicts if the income is higher or lower than 50K$.

<=50K 24720

<=50K. 12435

>50K 7841

>50K. 3846

**Machine Learning Flow**

This is the general flow for solving a classification problem with missing values and other preprocessing steps:

**a. Data Cleaning:**

* **Handling Missing Values**:
  + Replace missing values (?) with either:
    - The **most frequent value** (for categorical columns).
    - The **mean or median** (for numerical columns).
  + Optionally, you could remove rows with too many missing values or use advanced techniques like **KNN imputation**.
* **Encoding Categorical Variables**:
  + Convert categorical variables into numerical format.

**b. Data Preprocessing:**

* **Feature Scaling**: Normalize or standardize numerical features so that they are on a similar scale.
* **Train-Test Split**: Split the dataset into training and test sets, typically in an 80-20 or 70-30 ratio.

**c. Modeling:**

* Choose appropriate algorithms for the task, such as:
  + **Logistic Regression** (for binary classification).
  + **Decision Trees** or **Random Forests** (for non-linear relationships).
* **Model Training**: Train the model using the training dataset.

**d. Evaluation:**

* Evaluate the model using metrics like **accuracy** and **precision.**
* If it's a regression task, use **mean squared error (MSE)** or **R-squared**.