**Analysis of the Census Income Dataset**

**Data Definition**

| **Name** | **Data Type** | **Description** | **Values** |
| --- | --- | --- | --- |
| age | Integer | Age of the participant | 17-90 |
| workclass | Text | The working class is a socioeconomic term that describes people who earn a living through wage or salary-based work | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.  **Missing values:** 1836 ? ,963 blanks |
| fnlwgt | Integer | In the US Adult Income Dataset, fnlwgt stands for **"final weight"** and is a continuous, positive integer that represents the number of people represented by a row in the data. | No missing values  Range of value : 13862-1024535 |
| education | String | Education of participant | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.no missing values |
| education-num | Integer | Education Level | 1-16, no missing values |
| marital-status | String | Marital status of participants | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. |
| occupation | String | Occupation of participants | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.Missing values- 1843 ?, blanks 966 |
| relationship | String | the state of being connected by blood or marriage. | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. |
| race | String | Race generally refers to a group of people who have in common some visible physical traits, such as skin color, hair texture, facial features, and eye formation. | No blanks |
| sex | String | Sex of an individual | Male - 29554 Female - 15255 No blanks |
| capital-gain | integer | Taxable capital gain is generally equal to the value that you receive when you sell or exchange a capital asset minus your "basis" in the asset. | Range 0- 99999  0 values - 44808  No blanks |
| capital-loss | integer | A capital loss is a loss incurred when a capital asset is sold for less than the price it was purchased for. | Range 0-4356  0 values - 46561  No blanks |
| hours-per-week | integer | The number of hours spent per week. Avg work hours is 40 hours | Range - 1 - 99  No blanks |
| native-country | String | Country of origin | United-States, Cambodia, England, Puerto-Rico, Canada, Germany,Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.19 blanks |
| income | String | money received, especially for work. | <=50K <=50K.  >50K  >50K. |

**Feature variables-** Age,workclass,fnlwgt,education,education-num,marital-status,occupation, relationship, sex, race, capital-gain, capital-loss, hours-per-week, native-country.

**Target variable**- income

**Problem Statement:**

The objective of this machine learning project is to predict whether an individual's income exceeds $50K/year based on demographic and socio-economic data from the dataset. This binary classification problem involves identifying key patterns and correlations in the dataset to accurately classify individuals into two income groups: greater than $50K and less than or equal to $50K.

**ML flow:**

1. **Understanding the problem**:

· Clearly define the problem as a binary classification task where the target variable is income (either > $50K or <= $50K)

· Identify relevant features that might affect income such as age, education, occupation, race, hours worked per week.

2. **Preprocess the data**:

· Handle missing values

· Encode categorical variables

· Standardize numerical values for consistency

· Split the data into training and testing sets

3. **EDA**:

· Visualize the distribution of key features

· Analyze the relationship between the target variable and individual features

· Explore potential correlations between features

4. **Select the model**:

· Choose candidate models for binary classificationconfusion matrix

· Select evaluation metrics (accuracy, precision, recall, etc)

5. **Model training**:

· Train multiple models on the training set

· Cross-validate to assess model performance and avoid overfitting

6. **Model evaluation**:

· Analyze the confusion matrix to understand false positives and false negatives

7. **Model selection and interpretation**:

· Choose the best-performing model based on step 6

Interpret model results and provide insights into which features contribute most to

income prediction