**PROJECT3**

Data: - census\_income.csv

**Introduction:**

The 1994 census database contains a wide range of demographic and employment data gathered during the US Census that year. It includes information such as age, occupation, education level, marital status, race, sex, hours worked per week, native country, and whether an individual's income exceeds $50,000 annually. This dataset is a valuable tool for predictive modeling, especially in machine learning, to create accurate models predicting income levels based on demographic and socio-economic factors. Researchers and data scientists can use this dataset to analyze correlations between different attributes and income, providing valuable insights into socio-economic trends and predictive analytics.

**Descriptive Statistics:**

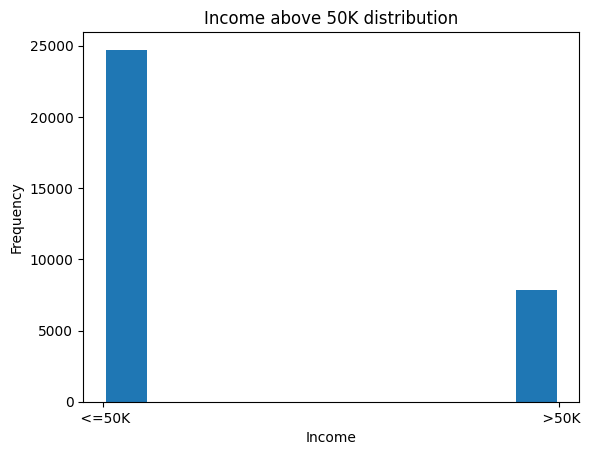
1. From the descriptive statistics, we could analyze that number of hours people worked for a week is 40 on an average.

2. The data consists of people between the age group from 17 to 90.

3. The average of income\_above\_50k is 0.2 which means that there are very a few people with inomce above 50k.

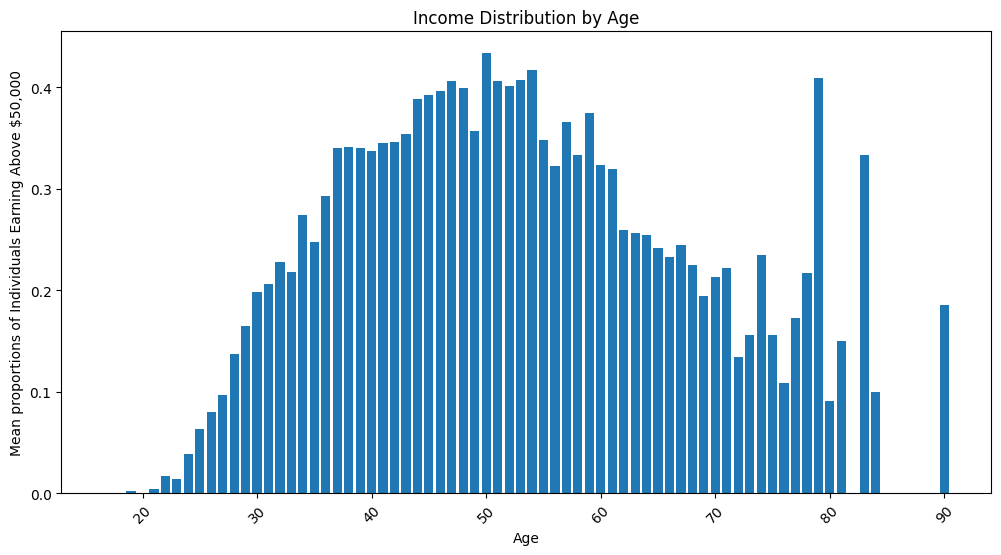
4. The education level is across the 1 to 16 levels. That means the given data consists of people from different educational background.

**Histogram Plot**

****

From the above Histogram, It is clearly known that about 8000 people are with income above 50K and about 24000 people are with below 50k income.

**Bar Graph**

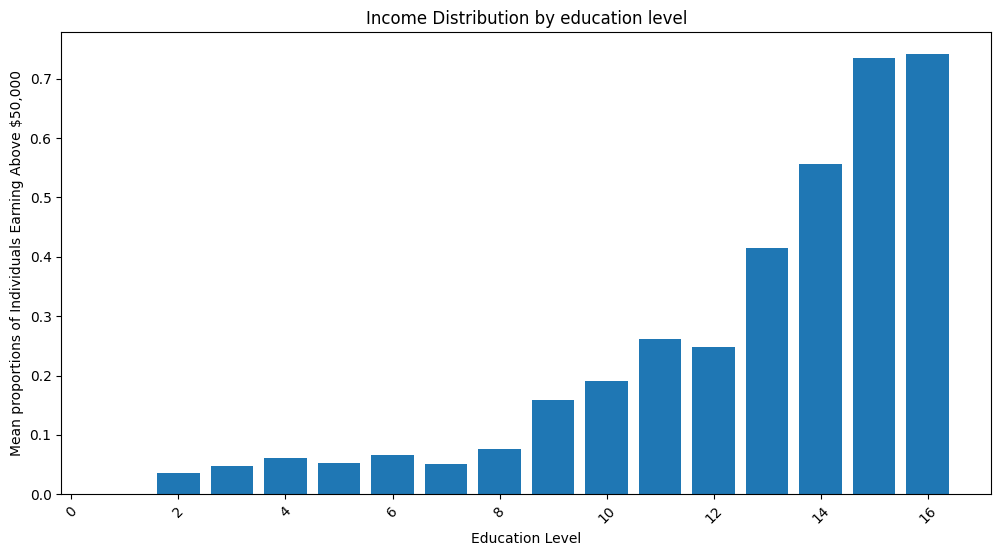


1. As individuals age, their earnings typically rise until they reach their 40s or 50s. According to the graph, the percentage of individuals earning over $50,000 steadily increases until around the age of 40. After that point, it stabilizes and may even experience a slight decline.

2. Young adults have a relatively low proportion of individuals earning over $50,000.

3. The graph illustrates that less than 20% of individuals aged between 20 and 30 earn more than $50,000. This can be attributed to various factors, such as lower wages for entry-level positions and the fact that many young adults are still pursuing their education.

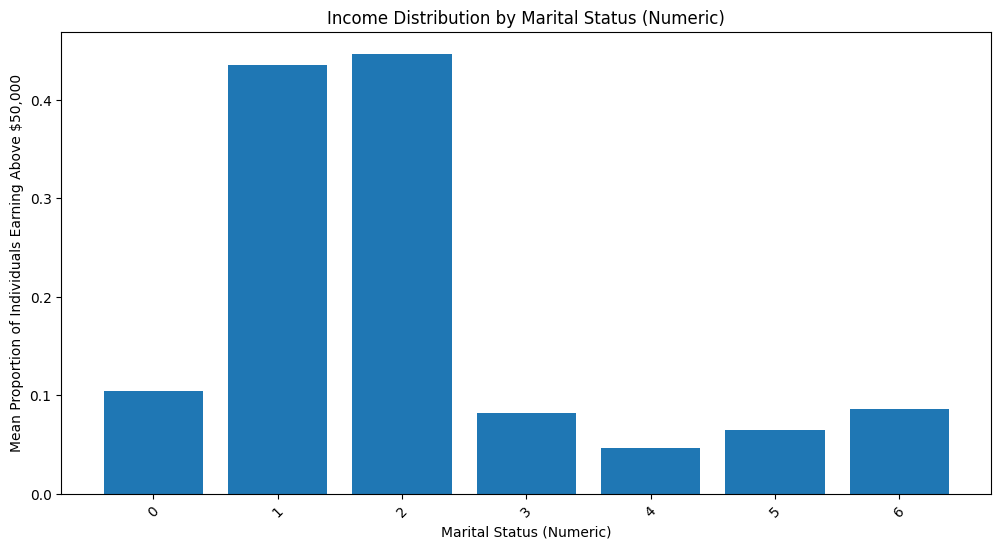
**Bar Plot**



1. The graph clearly illustrates that as education level increases, the likelihood of earning more than $50,000 per year also increases. The y-axis represents the proportion of people, while the x-axis represents education level. Starting from a low point, the line steadily rises, indicating that education plays a significant role in determining income levels.

2. According to the graph, individuals with a professional degree have the highest probability of earning more than $50,000. The line on the graph reaches its peak at the education level associated with a professional degree, highlighting the strong correlation between professional degrees and higher incomes.

**Bar Plot**



1. No one who is married to a civilian spouse (Married-civ-spouse) earns more than $50,000.

2. All divorced individuals (Divorced) earn more than $50,000.

3. A significant number of people who have never been married (Never-married) earn more than $50,000.

4. The income distribution for the other marital statuses (Separated, Widowed, Married-spouse-absent, Married-AF-spouse) is inconclusive due to data limitations on the y-axis.

**Classifier Predictions and Accuracy and Interpretation**

**1. Classifier Predictions**

KNN Classifier: Predicted [1] (income > $50k)

Decision Tree Classifier: Predicted [1] (income > $50k)

Random Forest Classifier: Predicted [1] (income > $50k)

Neural Network Classifier: Predicted [1] (income > $50k)

All the classifiers have agreed on predicting an income greater than $50k, showing a unanimous decision on the classification outcome for the specific instance(s).

**2. Prediction Scores (Accuracy)**

K Nearest Neighbors (KNN): Prediction score (accuracy) = 0.2295786758383491

Decision Tree: Prediction score (accuracy) = 0.23547475740081072

Random Forest: Prediction score (accuracy) = 0.23215821152192606

Neural Network: Prediction score (accuracy) = 0.22626212995946443

The prediction scores, ranging from around 22.6% to 23.5%, indicate the accuracy of each classifier in determining whether the income exceeds $50k or not in a test dataset. These scores reflect the proportion of correct predictions made by each model.

**3. Interpretation**

The unanimous forecast ([1]) from all classifiers indicates their agreement in categorizing income based on the provided input features.

The prediction scores offer valuable information about the overall effectiveness of each classifier, with accuracy rates varying between approximately 22.6% and 23.5%. Although these accuracies may seem relatively low, they demonstrate the models' capability to accurately classify income categories using the test dataset.

**Conclusion**

The analysis of the 1994 census data highlights significant correlations between demographic factors and income levels. Individuals with higher education levels and older age tend to have a higher probability of earning more than $50,000 annually, as illustrated in the graphs. Marital status also plays a role in income distribution, with divorced and never-married individuals having a higher percentage of higher earners compared to those married to civilian spouses. While classifier predictions focus on incomes above $50,000, the prediction scores indicate a moderate accuracy range of 22.6% to 23.5%, suggesting room for model enhancement. These results emphasize the intricate relationship between socio-economic variables and income outcomes, providing valuable insights for further research and policy development.