ALY 6140: Python and Analytics Systems Technology

Module 3 - Capstone Project Proposal

Harsh Anadkat

Owais Baig

Rohith Joginapally

Sudindra Karkera

Instructor: Prof. Zhi He

Northeastern University

3rd May 2024

Introduction

Our project's main objective is to use the Adult dataset to thoroughly examine the complex relationship between social demographics and income levels. We aim to provide important insights into the variables driving income inequality as well as the ability of individual characteristics to predict income levels. We use a combination of machine learning models and exploratory data analysis (EDA) techniques to do this.

Goals of the Project:

1. **Recognize How Social Demographics Affect Income:** Finding the social and demographic factors that have the most effects on income levels is our main objective. Our goal is to find trends and connections that clarify how variables like age, education, occupation, gender, race, and region affect income differences.

2. **Predictive Modeling of Income Levels:** Our goal is to create models that can accurately anticipate income levels by utilizing individual variables. This entails evaluating the effectiveness of machine learning algorithms in estimating income thresholds as well as investigating the predictive potential of various characteristics.

Questions to Investigate:

1. Which social and demographic variables have the most effects on income levels?
2. Can we use individual attributes to properly forecast income levels?
3. What role do occupation and level of education have in the differences in income?
4. Do racial and gender inequalities in income levels translate into meaningful differences?
5. Is it possible to pinpoint demographic groupings that have a higher propensity to earn more than a given threshold?
6. How does the geographical location of an individual affect their income level, and can this be used to improve income prediction models?

Methods Used in Analysis:

1. **Exploratory Data Analysis (EDA):** To learn more about the distribution, correlation, and trends in the dataset, we first perform EDA. This entails producing summary statistics, looking at correlations between data, and creating visualizations (such as box plots and histograms).

2. **Feature engineering:** We handle missing values, scale numerical features, and encode categorical variables as part of the preprocessing step of the data. By ensuring that features are in a format that is appropriate for machine learning algorithms, this phase gets the data ready for modeling.

3. **Machine Learning Models:** To create predictive models for income levels, we use a variety of machine learning algorithms, including logistic regression, decision tree classifiers, random forest classifiers, and support vector classifiers. F1-score, accuracy, precision, recall, and other performance metrics will be used to train, validate, and assess these models.

Exploratory Data Analysis (EDA):

To get a better understanding of the variables affecting income levels, the study starts by loading the Adult dataset and then moving on to data extraction, cleansing, and visualization.
Data Extraction: The Adult dataset includes data on the demographics and socioeconomic characteristics of individuals, including age, workclass, education, marital status, occupation, race, sex, hours worked per week, income, and native country.

Data Cleanup:

1. Missing Values Handling: The dataset initially contains missing values represented as '?' in some columns. These missing values are replaced with 'NA' and subsequently removed from the dataset to ensure data integrity.

```
# Checking for missing values
missing_counts = adult_data.isna().sum()

print(missing_counts)
```

```
age                0
workclass       1836
fnlwgt             0
education          0
education.num      0
marital.status     0
occupation      1843
relationship       0
race               0
sex                0
capital.gain       0
capital.loss       0
hours.per.week     0
native.country   583
income             0
dtype: int64
```

```
# Checking for missing values
missing_counts = adult_data.isna().sum()

print(missing_counts)
```

```
age               0
workclass         0
fnlwgt            0
education         0
education.num     0
marital.status    0
occupation        0
relationship      0
race              0
sex               0
capital.gain      0
capital.loss      0
hours.per.week    0
native.country    0
income            0
dtype: int64
```

2. Encoding Target Variable: The target variable 'income' is converted into binary form ('<=50K' and '>50K') for classification purposes.

```
# Converting the Target variable to binary
adult_data['income'] = adult_data['income'].map({'>50K': 1, '<=50K': 0})

# Check the first few rows to confirm the changes
print(adult_data['income'].head())
```

```
0    0
1    0
2    0
3    0
4    0
Name: income, dtype: int64
```

Data Visualizations:

1. Histograms for Numerical Data Analysis

We used histograms to graphically investigate the distributions of important numerical variables in our analysis of the Adult dataset. Understanding the spread and central tendencies of these variables and how they affect income levels is made easier by this graphical portrayal. Using Seaborn's `histplot` function, we generated histograms with 30 bins for each numerical variable. This bin size allows for a detailed view of the distribution while avoiding excessive granularity.
The histograms are organized in a grid layout with 3 rows and 2 columns, ensuring a clear and systematic presentation of the data. Each histogram is labeled with the corresponding variable name for easy identification.

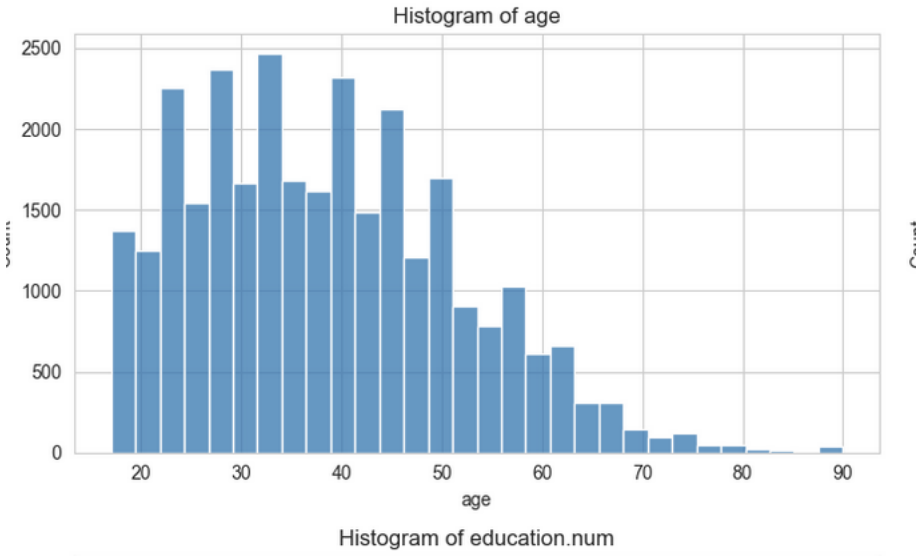
Numerical Variables Examined

We selected the following numerical columns for histogram analysis:

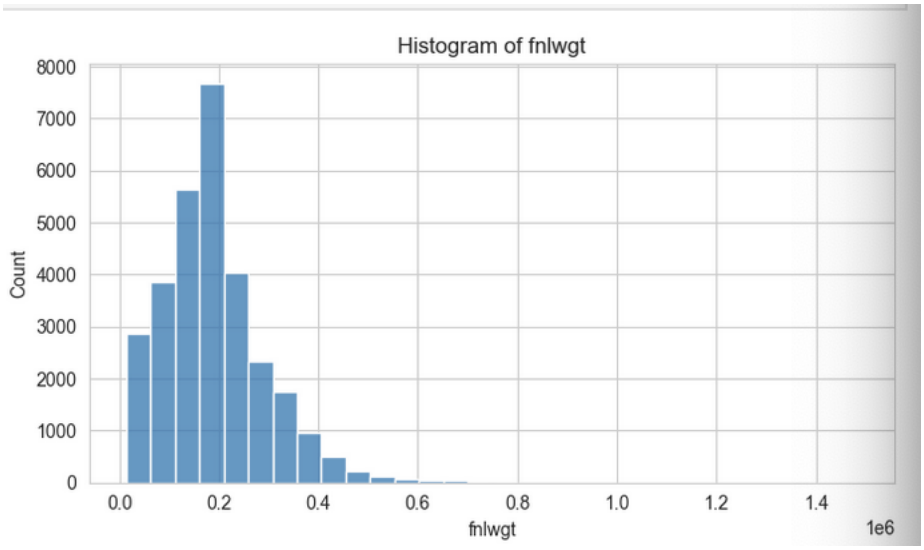
- Age
- Final Weight (fnlwgt)
- Education Number (education.num)
- Capital Gain
- Capital Loss
- Hours Per Week

Insights and Interpretation

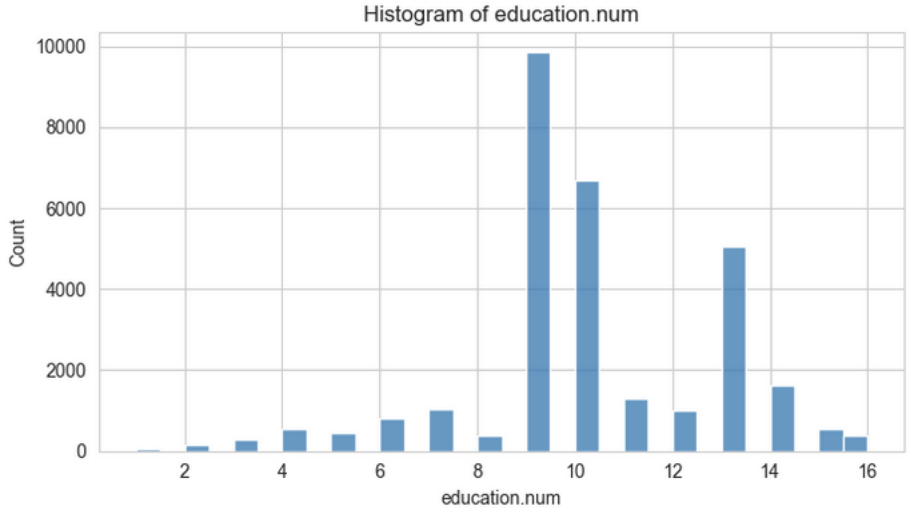
1. **Age Distribution**: The age histogram sheds light on the age distribution of the dataset's participants. We see a surge in the middle age group, and as people get older, the frequency gradually declines. This distribution is essential to comprehending the potential effects of age on income levels.



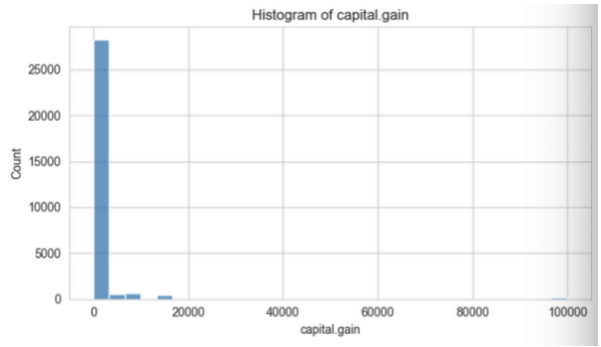
2. **Final Weight Distribution**: The fnlwgt histogram reveals the distribution of final weights assigned to individuals, which reflects sample biases or population representation in the dataset.

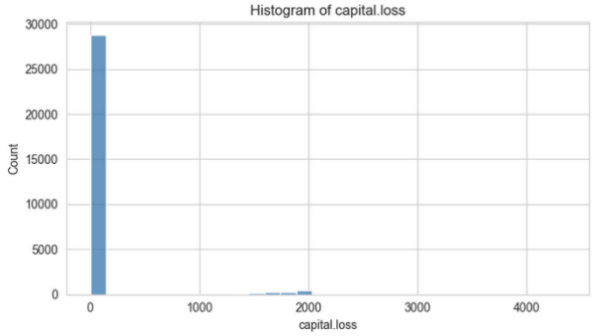


3. **Education Impact**: The distribution of numerically encoded education levels is shown by the education number histogram. Understanding people's educational backgrounds and any potential relationships to income is made easier by this analysis

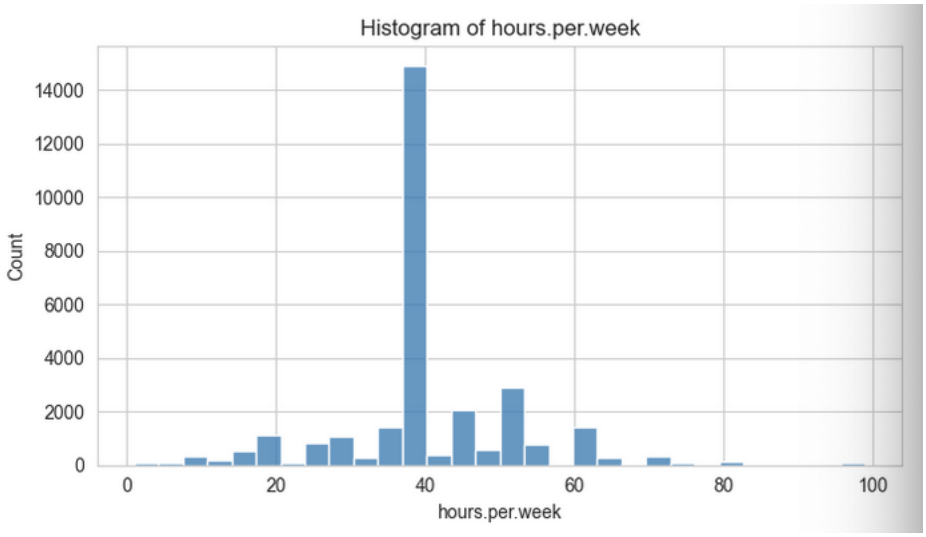


4. **Financial Metrics (Capital Gain/Loss)**: The distribution of individual financial profits and losses is displayed in the capital gain and loss histograms. These measures are essential forevaluating the dataset's wealth accumulation trends and financial standing.





5. **Work Hours**: The hours per week histogram offers insights into the distribution of work hours among individuals. This variable's distribution can shed light on the relationship between work commitment and income levels.



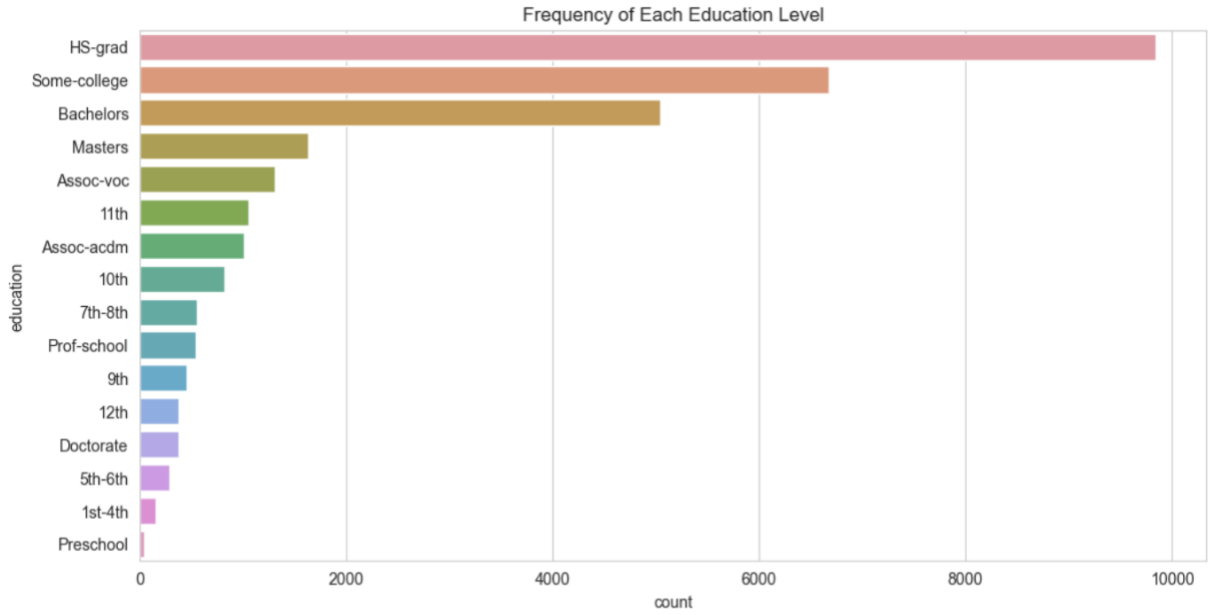
2. Bar Chart for Categorical Data Analysis

We used a bar chart to evaluate categorical data in our investigation of the Adult dataset, with a particular emphasis on the frequency distribution of education levels among people. This graphic depiction provides insightful information about the population's educational background within the dataset and how it might affect income levels. We made a horizontal bar chart to show the frequency of each education level using Seaborn's countplot function. The bars show the total number of people connected to each educational group arranged from highest to lowest frequency.

The horizontal arrangement of the bar chart improves readability and facilitates a quick comparison of the frequency of education levels. The corresponding education level is indicated on each bar, giving an understandable and straightforward representation of the distribution.

Insights and Interpretation

1. **Education Level Distribution**: The bar chart highlights the distribution of education levels within the dataset. By examining the heights of the bars, we can identify the most prevalent education categories and their relative frequencies.



Based on the bar chart, we can observe the following:

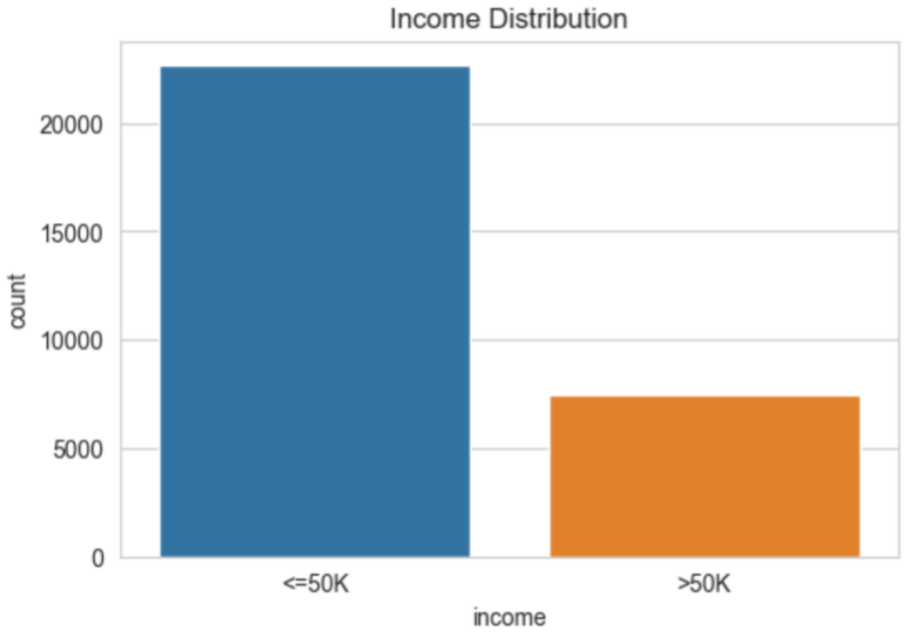
- **The Most Common Education Level**: The education level with the highest frequency is Bachelors. This may indicate that a larger percentage of the adult population has at least a bachelor's degree compared to other education levels.

- **Lower Education Levels**: The bar chart shows a clear downward trend in the frequency of adults with lower education levels, from Doctorate to some high school graduates. This may indicate that there is a negative correlation between education level and income or occupation.

- **Unusual Category**: The "Other" category appears to be quite high in frequency compared to other education levels. This could be due to data entry errors or an influx of people with no education or unspecified education levels.

- **Possible Gaps**: There could be a gap in the education level distribution. For example, there are fewer people with some college education (1 or 2 years of college) and high school graduates (some high school, no high school diploma) compared to the other education levels.

2. **Income Distribution Bar Chart:** The bar chart shows the distribution of income levels in the `adult_data` dataset.



Here are some observations based on the chart:

- **Two Major Income Categories**: The chart shows two major income categories: <=50K and >50K. This suggests that the income distribution is bimodal, with many people earning less than or equal to $50,000 and a smaller number of people earning more than $50,000.

- **Income Gap**: The gap between the two income categories is significant, indicating a large income disparity between the two groups. This could be due to various factors, such as education level, occupation, or geographical location.

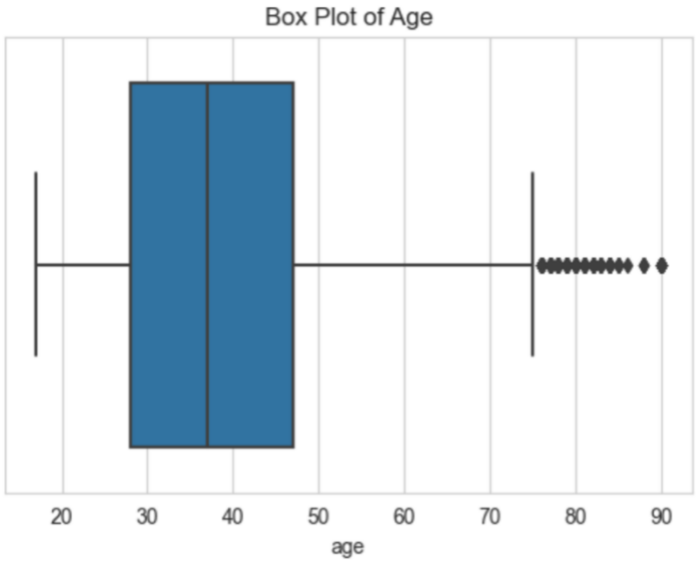
- **Frequency of Income Levels**: The chart shows that the frequency of people earning <=50K is higher than those earning >50K. This could indicate that a larger percentage of the population earns a lower income compared to a higher income.

- **Possible Outliers**: There may be some outliers in the dataset, as there are a few people who earn significantly more than the rest of the population. This could be due to various factors, such as high-paying jobs or inheritance.

3. Box Plot Analysis for Age Distribution:

\A box plot was employed in our research of the Adult dataset to investigate the age distribution of the subjects. Understanding age-related patterns within the dataset is made easier by this graphical depiction, which provides insights into the central tendency, spread, and existence of outliers in the age variable. We made a box plot that concentrated on the age distribution using Seaborn's boxplot function. The box plot shows important statistical metrics, such as the potential outliers (data points outside the whiskers), quartiles (box margins), and median (box center line).

The interquartile range (IQR), which shows the range of age values among people, is shown by the height of the box. As a visual representation of the dispersion of the data, the whiskers stretch out to 1.5 times the IQR from the quartiles.



Insights and Interpretation:

- **Central Tendency**: The median line in the box plot represents the central tendency of ages within the dataset. It indicates the age value that divides the dataset into two equal parts, providing a measure of the dataset's central age value.

- **Spread of Ages**: The length of the box (interquartile range) and the whiskers (1.5 times IQR) demonstrate the spread of age values. A wider box and longer whiskers suggest greater variability in age among individuals.

- **Outlier Detection**: Any data points beyond the whiskers are considered potential outliers. These outliers, if present, could indicate unusual age values that deviate significantly from the typical age distribution in the dataset.

Significance of Results:

- The EDA reveals insights into the demographic composition and income distribution within the Adult dataset.

- Understanding the age distribution, education levels, and income disparities provides a foundation for further analysis and modeling.

- Visualizations such as histograms, bar charts, and box plots enhance the understanding of data patterns and help in identifying potential trends and outliers.

Overall, the EDA component sets the stage for subsequent analysis and modeling, providing valuable insights into the dataset's characteristics and key variables influencing income levels.

Predictive Models:

Data Preprocessing:

The dataset was first preprocessed to suit the requirements of machine learning models. This included:

- Encoding Categorical Variables: Categorical features were transformed using one-hot encoding, allowing models to better interpret these features.

- Scaling Numerical Variables: Numerical data was standardized to have a mean of zero and a standard deviation of one, which helps in normalizing the range of data features and improving the performance of several algorithms.

- Data Splitting: The data was split into training (80%) and testing (20%) sets to enable model validation.

Models Deployed

Four different classification models were evaluated:

1. **Logistic Regression:** A statistical model that estimates the probabilities using a logistic function.

2. **Decision Tree Classifier:** A model that partitions the data into subsets based on feature values, creating a tree-like model of decisions.

3. **Random Forest Classifier:** An ensemble of decision trees, designed to improve on the predictive capability of a single tree by averaging multiple trees that individually consider random subsets of features and samples.

4. **Support Vector Machine (SVM):** A powerful classifier that works by finding the best hyperplane that divides the dataset into classes.

Model Performance:

The performance of the models was assessed based on accuracy, precision, recall, and F1-score. Here are the results:

1. **Logistic Regression:**

```
Results for Logistic Regression:
              precision    recall  f1-score   support

       <=50K       0.88      0.93      0.90      4976
        >50K       0.72      0.58      0.64      1537

    accuracy                           0.85      6513
   macro avg       0.80      0.76      0.77      6513
weighted avg       0.84      0.85      0.84      6513
```

- **Performance:** The model achieved an overall accuracy of 85%, indicating a strong ability to distinguish between individuals earning <=50K and >50K. The model demonstrated high precision (88%) and recall (93%) for predicting individuals earning <=50K, suggesting it is particularly effective at identifying lower income brackets.

- **Challenges with Higher Income Prediction:** For predicting individuals earning >50K, the model showed lower precision (72%) and recall (58%). This indicates a challenge in correctly identifying higher earners, as the model tends to miss a significant portion of this group.

2. **Decision Tree:**

```
Results for Decision Tree:
              precision    recall  f1-score   support

       <=50K       0.88      0.87      0.88      4976
        >50K       0.60      0.62      0.61      1537

    accuracy                           0.81      6513
   macro avg       0.74      0.75      0.74      6513
weighted avg       0.82      0.81      0.81      6513
```

- **Performance:** The Decision Tree model demonstrated good performance for individuals earning <=50K, with a precision of 88% and recall of 87%. This indicates a high level of

accuracy in identifying lower-income individuals, reflected in an f1-score of 88%.

- **Challenges with Higher Income Bracket:** For individuals earning >50K, the model's precision (60%) and recall (62%) were notably lower, indicating difficulties in accurately identifying higher earners. The f1-score of 61% suggests there is room for improvement in this category.

3. Random Forest:

Results for Random Forest:

	precision	recall	f1-score	support
<=50K	0.89	0.93	0.91	4976
>50K	0.72	0.61	0.66	1537
...				
accuracy			0.85	6513
macro avg	0.81	0.76	0.78	6513
weighted avg	0.85	0.85	0.85	6513

- **Performance:** The Random Forest model excelled in predicting individuals earning <=50K, achieving a precision of 89% and a recall of 93%. This resulted in a high f1-score of 91%, indicating robust identification of lower-income individuals.

- **Moderate Success with Higher Income Bracket:** For those earning >50K, the model displayed a precision of 72% and a recall of 61%. The f1-score of 66% reflects a moderate level of accuracy, suggesting some challenges in consistently identifying higher earners accurately.

4. Support Vector Machine:

Results for SVM:

	precision	recall	f1-score	support
<=50K	0.87	0.94	0.91	4976
>50K	0.74	0.56	0.64	1537
accuracy			0.85	6513
macro avg	0.81	0.75	0.77	6513
weighted avg	0.84	0.85	0.84	6513

- **Performance:** The SVM model showed strong performance in identifying individuals earning <=50K, with a recall of 94% and a precision of 87%. The high recall indicates the model's effectiveness in capturing the majority of the lower-income group, achieving an f1-score of 91%**.**
- **Challenges in Higher Income Predictions:** For those earning >50K, the SVM model achieved a precision of 74% but a lower recall of 56%. The f1-score of 64% reflects challenges in accurately identifying a significant portion of higher earners, suggesting a potential area for model tuning to improve performance.

Model Comparison:

1. **Balanced Performance Across Classes:** Random Forest and SVM generally performed better in terms of balancing recall and precision across both income classes, unlike the Decision Tree which showed significant disparities, particularly struggling with the higher income bracket. Both Random Forest and SVM maintained a good balance, making them superior to the Decision Tree and slightly better than Logistic Regression in overall effectiveness.

2. **Overall Accuracy and F1-Scores:** All models demonstrated similar overall accuracies around 85%. However, the Random Forest stood out with slightly better f1-scores, especially for the higher income bracket (>50K), where its f1-score of 0.66 surpassed that of SVM (0.64), Logistic Regression (0.64), and Decision Tree (0.61).

3. **Handling of Higher Income Bracket:** The Random Forest model not only offered competitive performance in identifying lower income individuals but also showed relatively better results in identifying higher earners compared to other models. This is crucial as the higher income bracket generally had lower recall and precision, making Random Forest's performance particularly valuable.

Conclusion:

The analysis demonstrated that while individual models have their unique strengths, the Random Forest classifier provided the most reliable and robust predictions for this particular dataset. It effectively captured the complexity and variability of the data, making it the recommended model for further development and deployment in predicting income levels based on demographic and employment factors.

Future Recommendations:

Model Tuning and Experimentation: Further tuning of the Random Forest parameters could potentially enhance its performance. Additionally, experimenting with advanced ensemble techniques or neural networks might yield improvements, especially in handling the underrepresented higher income bracket.

Feature Engineering: There may be an opportunity to explore more sophisticated feature engineering techniques, such as interaction terms between features or more complex encodings for categorical variables, to capture more nuanced patterns in the data.

References

- Bentéjac, C. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 1937–1967. Retrieved from https://link.springer.com/article/10.1007/s10462-020-09896-5
- DeMaris, A. (2013). Logistic Regression. *Converting Data into Evidence*, 115–136. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4614-7792-1_7
- Mucherino, A. (2009). k-Nearest Neighbor Classification. *Data Mining in Agriculture*, 83–106. Retrieved from https://link.springer.com/chapter/10.1007/978-0-387-88615-2_4