Understanding Income Disparities through U.S. Census Data

By

Pandya, Harsh Maheshkumar

Kumar, Prashul

Sheta, Rushank Ghanshyam

hapandya@iu.edu, praskuma@iu.edu, rsheta@iu.edu

Under the guidance of **Prof. Stefan Weiergraeber**

Abstract – This project focuses on evaluating the importance of occupation, levels of education, and demographic characteristics towards income and predict income levels based on these factors. Through the Adult Census Income dataset, this study examines the socioeconomic variables impacting income inequality in the US. Some of the methodologies used in this study are machine learning algorithms, statistical techniques, data visualization, etc. The approach is to validate data quality, handle missing values, perform feature engineering, and to train machine learning models which could predict the income using various demographic attributes.

Keywords – Exploratory Data Analysis (EDA), Data Analysis, Univariate Analysis, Bivariate Analysis, Classification, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, K Nearest neighbors, Naive Bayes.

1. Introduction –

By examining the socioeconomic factors impacting income levels, this project seeks to advance our understanding of income inequality in the US. For this investigation, we make use of the Adult Census Income dataset. There are 32,562 cases and 15 attributes in the Adult Census Income dataset. These elements include demographics, educational history, occupational information, financial parameters, and target income level. Numerous studies indicate that income inequality has significant effects on social mobility, resource accessibility, and the general economic well-being of communities.

First, identifying the factors associated with income disparities can help us inform policy

decisions to reduce inequality and promote economic mobility. Second, understanding the root causes of income gaps can help in addressing systemic issues leading to unequal opportunities and outcomes. Also, understanding the factors influencing income can empower individuals to make informed decisions about their education, career paths, and other life choices.

Machine learning and data analysis offer powerful tools for overcoming these challenges. They enable us to identify complex patterns, handle large datasets, and make predictions. This project will employ a combination of data analysis and machine learning techniques, including data preprocessing, exploratory data analysis, feature engineering, machine learning algorithms, and model evaluation. By applying these techniques, we aim to gain a deeper understanding of the factors contributing to income disparities in the United States and develop robust models for income prediction.

2. Problem Statement -

The project aims to dive into income disparities within the United States using Kaggle's Adult Census Income dataset from the 1994 Census Bureau database. We aim to understand and predict income levels based on a multitude of socio-economic factors. Despite advancements in technology and society, persistent income disparities persist, prompting an investigation into the drivers of these variations. By uncovering the intricate relationships between demographic attributes, educational attainments, occupational affiliations, and income levels, this research seeks to contribute valuable insights for policymakers, economists, and sociologists. Explore the dataset to identify patterns, correlations, and anomalies. Cleanse the data by addressing missing values and inconsistencies. Conduct exploratory data analysis to understand the distribution of variables and their relationships with income. Apply predictive modeling techniques (Logistic Regression with Lasso, Decision Trees, Random Forest) to predict income levels based on socioeconomic attributes. Evaluate model performance and interpret feature importance to understand the key determinants influencing income disparities.

3. Methodology -

- Data Collection and Exploration
- Data Preprocessing
- Exploratory Data Analysis
  1. Univariate Data Analysis
  2. Bi and Multivariate Data Analysis
- Modelling
- Model Evaluation
- Conclusion and Implications

4. Data Collection and Exploration -

This includes steps like Checking for missing values and understanding the five-number summary and statistical summary for each attribute.

Training Data contains attributes like:

- age: Integer values representing the age of individuals.
- workclass: Categorical variable indicating the type of employment or workclass. Contains missing values denoted by "?".

- fnlwgt: Integer values representing final sampling weights.
- education: Categorical variable indicating the highest level of education achieved.
- education.num: Integer values corresponding to the education level.
- marital.status: Categorical variable indicating the marital status of individuals. occupation: Categorical variable denoting the occupation of individuals. Contains missing values denoted by "?".
- relationship: Categorical variable indicating individuals' relationship status.
- race: Categorical variable representing the race of individuals.
- sex: Categorical variable representing the gender of individuals.
- capital.gain: Integer values denoting capital gains.
- capital.loss: Integer values denoting capital losses.
- hours.per.week: Integer values representing the number of hours worked per week.
- native.country: Categorical variable denoting the native country of individuals.
- income: Categorical variable indicating whether an individual's income is greater or less than $50,000 (<=50K or >50K).

5. Data Preprocessing -

Checking for Null and Missing values, we found that there were no Null values in the data. However, the missing values are represented by '?' and three columns namely workclass, occupation and native.country had missing values. We decided to remove the missing values because the total missing values were not that high when compared to the available data. Missing values accounted for 2399 instances out of 32561 total instances or around 7% of the total data.

Feature engineering included adding a new attribute capital calculated from the formula capital.gain - capital.loss and converting the target column i.e income to binary from string type where "<50K", ">50K" are converted to 0 and 1 respectively. Followed by creating dummy variables for categorical variables like marital. status, native.country, etc. Stratified sampling is used to split the processed data into train and test split with 80% or around 26K instances in train data and the remaining 6.5K instances in test data.

6. Univariate Data Analysis -

The Fig 5.1 represents the distribution of Age. It seems to be right-skewed, with more data concentrated from 25 to 50 describing the shape of a distribution and providing insight into the central tendency of the data. Since it is right skewed, it indicates that there are a few observations with very high values, causing the distribution to be pulled in the direction of those higher values.

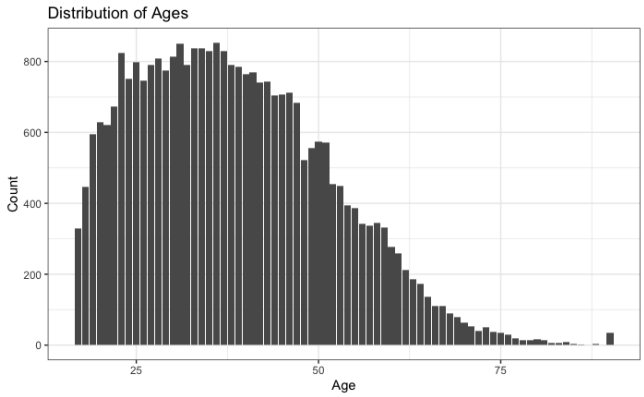


Fig 5.1

The Fig 5.2 represents the count plot of the Education Number. This particular distribution provides insights into the educational diversity of

the dataset. The interpretation of the bar plot suggests that the dataset has a significant number of individuals with the highest education level being "HS grad" (High School Graduate), followed by "Some College" and "Bachelors". As High school education is often considered a baseline or minimum qualification, and a large proportion of the population attains at least a high school diploma, its count is the highest. However, the declining trend in counts from high school to bachelor's suggests a natural progression where fewer individuals attain higher levels of education.

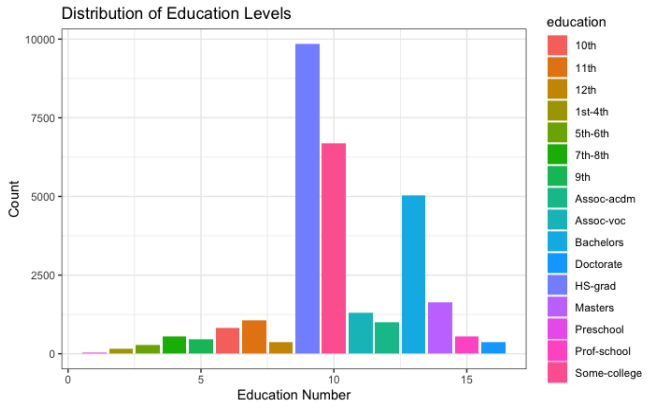


Fig 5.2

For further analysis, we plotted the graph of the Education segment (Fig 5.3). This plot provides a visual representation of the distribution of individuals across the education segments. The bar plot reveals that the education segment "9 to 12" has the highest count, exceeding 15,000 individuals, while the "0 to 4" segment has the lowest count. This distribution suggests a concentration of individuals with education levels corresponding to high school to some college education in the dataset. The lower count in the "0 to 4" segment could be attributed to factors such as a smaller representation of individuals with very basic education levels or potential data collection biases. Overall, the distribution provides insights into the educational diversity of the dataset, with a notable concentration in the middle education segments.

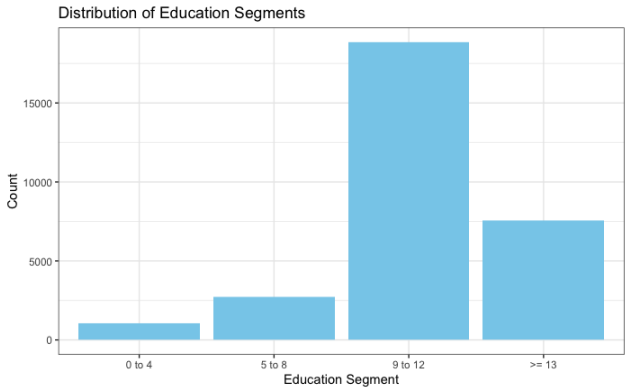


Fig 5.3

From Fig 5.4 The histogram illustrates the distribution of capital values in the dataset, revealing a notable concentration of individuals with lower capital and a comparatively smaller number of individuals with higher capital. This pattern suggests a right-skewed distribution, where the majority of individuals have lower capital, and the frequency gradually decreases as capital values increase, which is an ideal case in the real world as well.

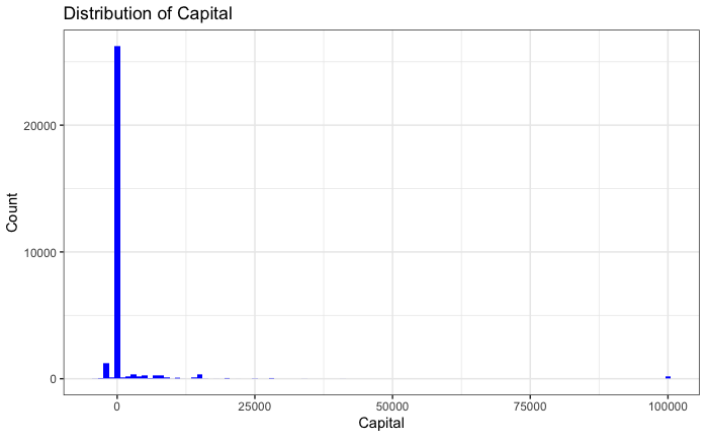


Fig 5.4

6. Bivariate and Multivariate Data Analysis -

The bar chart from Fig 6.1 reveals distinct

patterns in the mean hours worked by gender for specific occupations. For males, executive managerial roles, farming-fishing, and transportation-moving occupations show the highest average hours worked. On the other hand, females tend to have higher average working hours in craft repair, executive-managerial positions, and private house-serving roles.

Fig 6.1

These differences highlight gender-specific trends in the distribution of working hours across various occupational categories. It's essential to consider these patterns when addressing issues related to work-life balance, occupational choices, and potential gender-based disparities in working hours within different job sectors.

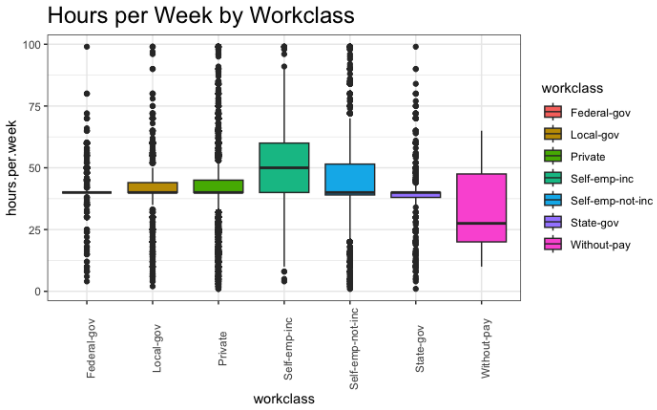


Fig 6.2

The Fig 6.2 represents the boxplot which illustrates the distribution of hours worked per week across different workclasses. Each box represents the interquartile range (IQR), with the horizontal line inside indicating the median hours worked. The chart provides insights into the variability and central tendency of working hours within each workclass. It can be observed that certain workclasses, such as "Without-pay" have limited variability and generally lower median working hours, while others, like "Private" and "Self-emp-inc," exhibit wider distributions with higher median values. This visualization helps us in understanding the overall patterns of weekly working hours across diverse workclasses in the dataset.

Fig 6.3

Fig 6.3 represents the age distribution histogram illustrates that the percentage of individuals with income over 50,000 Dollars tends to increase steadily up to around 40 years old. This could be attributed to individuals advancing in their careers, gaining experience, and securing higher-paying positions during these initial years of their professional lives. However, beyond the age of 40, there is a gradual decline in the percentage of individuals with income exceeding 50,000 Dollars. This decline might be associated

with factors such as career stability, retirement, or a shift towards part-time employment. The visualization provides insights into how age correlates with income levels, highlighting a potential turning point in income distribution around the age of 40.

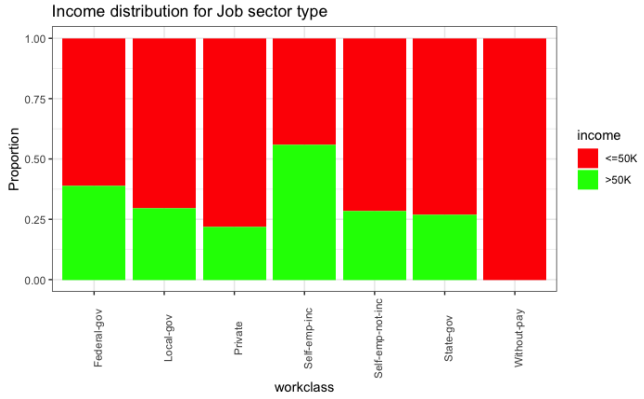


Fig 6.4

The visualization from Fig 6.4 illustrates the distribution of income across different workclass categories. The "Without-pay" category predominantly consists of individuals with income below 50,000 Dollars, indicating that a significant proportion of individuals in this category may not earn a substantial income. In contrast, the "Self-emp-inc" category exhibits a higher proportion of individuals with income over $50,000, suggesting that self-employed individuals in incorporated businesses tend to have higher incomes. This analysis provides insights into the income distribution among different workclass categories, highlighting disparities in earnings.

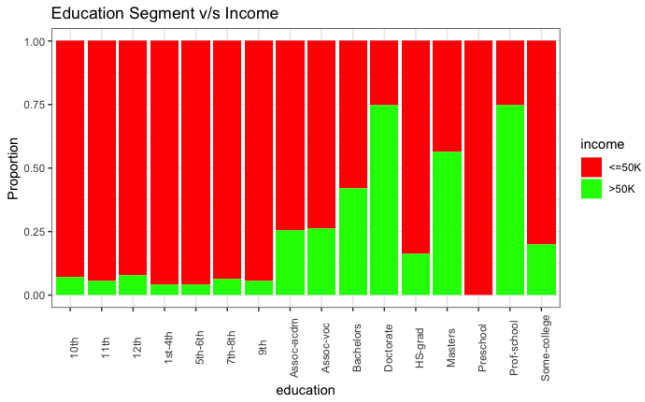


Fig 6.5

The above chart (Fig 6.5) illustrates the income distribution across various education segments, with a focus on the proportion of individuals earning less than or equal to $50,000 and those earning more than 50,000 Dollars. Notably, the lower educational segments such as "Preschool" and education levels from "1st to 10th" predominantly fall within the <=50k income category. On the other hand, the higher educational segments, including "Prof-school," "Doctorate," and "Masters," exhibit a larger proportion of individuals earning above 50,000 Dollars. This pattern suggests a positive correlation between higher education levels and an increased likelihood of earning a higher income, emphasizing the significance of education in income outcomes.

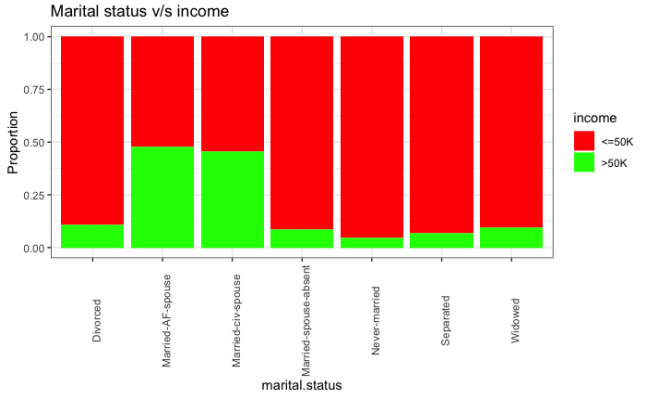


Fig 6.6

The visualization from Fig 6.6 portrays the distribution of income across different marital statuses, emphasizing the proportion of individuals earning less than or equal to 50,000 Dollars and those earning more than 50,000 Dollars. The chart indicates that individuals classified as "Never-married" or "Divorced" have a higher proportion within the <=50k income category. In contrast, married individuals, particularly those categorized as "Married-civ-spouse" as well as "Married-AF-spouse" display a higher proportion in the >=50k income category. This observation suggests a potential correlation between marital status and income, with married individuals tending to have a greater likelihood of earning a higher income.

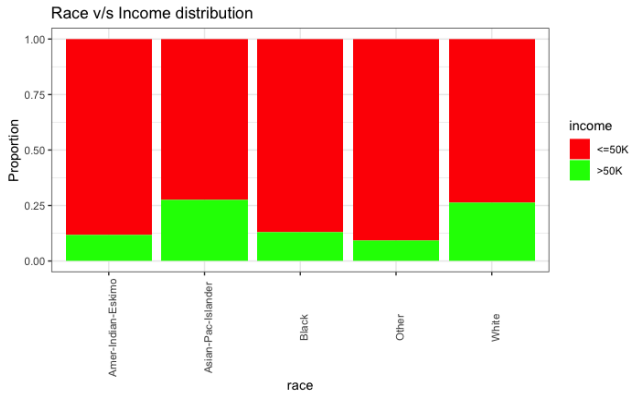


Fig 6.7

The above Fig 6.7 demonstrates the distribution of income across different racial categories, highlighting the proportion of individuals earning less than or equal to 50,000 Dollars and those earning more than 50,000 Dollars. The chart indicates that a higher proportion of individuals from the "Black", "other" racial category falls within the <=50k income category. Conversely, individuals from the "Asian-Pac-Islander" and "White" categories have a relatively higher proportion in the >=50k income category. This observation suggests a potential association between race and income distribution, with variations in income levels across different racial groups.

From Fig 6.8 we The visualization depicts the distribution of income across gender categories, differentiating between individuals earning less than or equal to 50,000 Dollars (red) and those earning more than 50,000 Dollars (green).

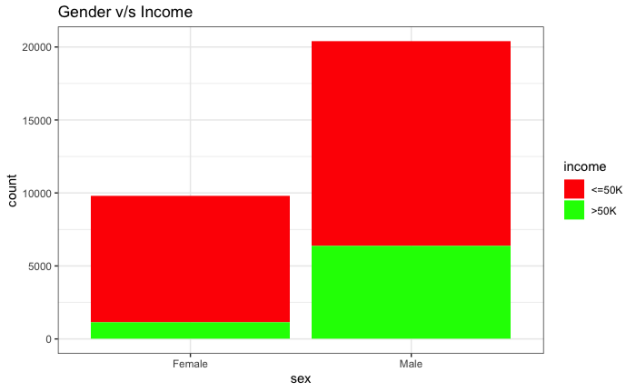


Fig 6.8

The chart indicates that a higher proportion of males falls within the >50k income category compared to females. In contrast, a larger proportion of females is observed in the <=50k income category. This observation suggests a gender-based disparity in income distribution, with a notable difference in the proportion of individuals earning above and below $50,000 based on gender.

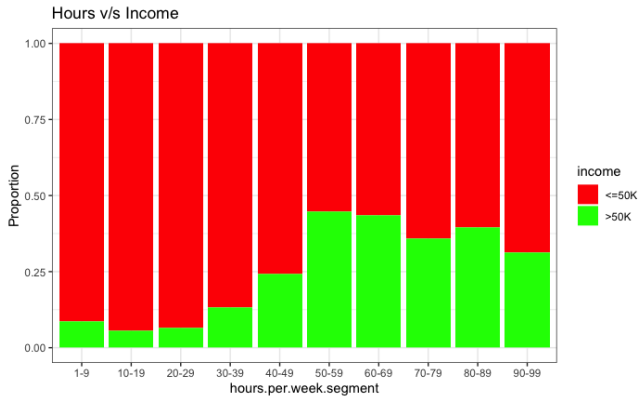


Fig 6.9

The visualization Fig 6.9 illustrates the distribution of income across different segments of hours worked per week, distinguishing between individuals earning less than or equal to 50,000 Dollars (red) and those earning more than 50,000 Dollars (green). The chart shows that the majority of individuals working 1-9 hours per week fall within the <=50k income category, while those working 50-59 hours per week have a more balanced distribution between both income categories. As the number of hours worked per week increases beyond 50, there is a higher proportion of individuals earning above $50,000. However, for those working 90-99 can be considered as exceptions in this case. This suggests a positive correlation between the number of hours worked per week and the likelihood of earning a higher income.

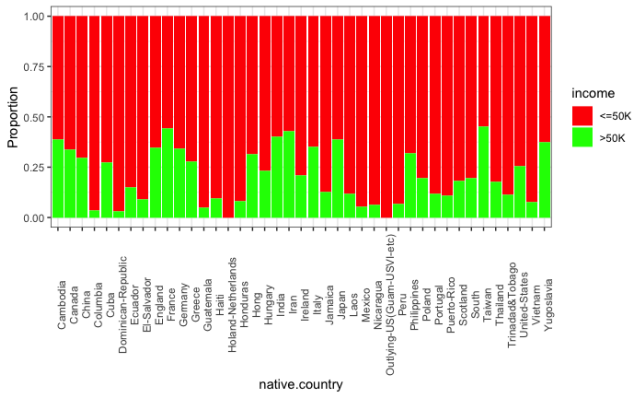


Fig 6.10

The Fig 6.10 showcases the distribution of income across different native countries, with the proportion of individuals earning less than or equal to 50,000 Dollars (red) and those earning more than 50,000 Dollars (green). The chart indicates that the majority of individuals from “Hati” fall within the <=50k income category. In contrast, some other countries, such as India, Taiwan, and Yugoslavia have a higher proportion of individuals in the <=50k income category. This variation suggests differences in income distribution across native countries, possibly influenced by economic factors and job opportunities.

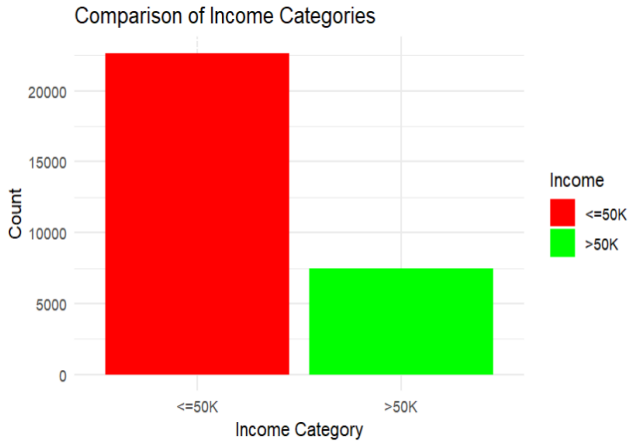


Fig 6.11

From the above Fig 6.11 we can infer that the

dataset exhibits an imbalance in the income categories, with 7,508 individuals earning more than 50,000 Dollars (>50K) and 22,654 individuals earning 50,000 Dollars or less (<=50K). This imbalance can impact the performance of machine learning models, particularly in scenarios where the algorithm may be biased towards the majority class. Thus this will need to be handled accordingly while prediction.

6.2. Statistical Analysis -

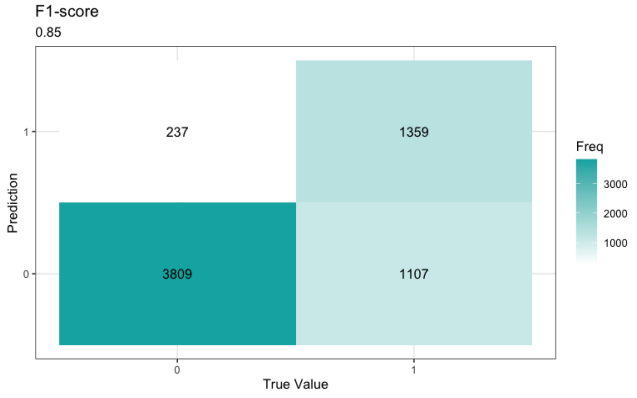
In the statistical analysis conducted, we formulated a null hypothesis (H0) assuming no association between the variables "income" and "sex" in the overall population. Following a rigorous examination, the resulting p-value, which is notably less than the conventional significance level of 0.05, provides compelling evidence to reject the null hypothesis. This suggests a strong indication that gender and income are not independent variables in the broader population. The statistical significance implies that the observed relationship between gender and income is highly unlikely to occur due to random chance alone.

Similarly, in the context of race and income, the obtained result points towards a significant association between the two variables. The small p-value signifies that the observed differences in income levels across various races are unlikely to be attributed to random chance alone. This substantiates the assertion that there is a meaningful and statistically significant relationship between race and income within the population under consideration. These findings contribute valuable insights into the complex interplay of socio-economic factors, shedding light on disparities that warrant further exploration and consideration.

7. Modelling -

A. Logistic regression:

The Logistic Regression model excels within a moderate regularization range, showcasing strong discriminatory power with an AUC of 0.88 and a balanced F1 score of 0.84. Its linear approach efficiently models relationships in the data while leveraging regularization to prevent overfitting. The model demonstrates a commendable ability to distinguish between income classes, attributing its success to its simplicity and effectiveness in capturing linear patterns. The balance between regularization strength and predictive accuracy suggests a robust generalization to new data. Despite its linear nature, Logistic Regression showcases competitive performance.



B. Naive Bayes:

```
F1_Score_nb <- F1_Score(nb_predictions, Y_test)
print(paste("F1 Score: ", F1_Score_nb))
```

```
[1] "F1 Score:  0.609500138083402"
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2323  107
         1 2686 1396

               Accuracy : 0.5711
                 95% CI : (0.559, 0.5832)
    No Information Rate : 0.7692
    P-Value [Acc > NIR] : 1
```

In applying the Naive Bayes classification to the given dataset, the model achieved an F1 Score of approximately 0.62. The confusion matrix reveals insightful performance metrics. The accuracy stands at 57.1%. The model correctly predicted the positive class (income <=50K) , indicating a relatively high accuracy in identifying individuals with lower income. However, a lower accuracy was found for predicting higher-income individuals. As reflected in the F1 Score, the Naive Bayes shows moderate results as it assumes independence between features, which may not fully align with the underlying data distribution.

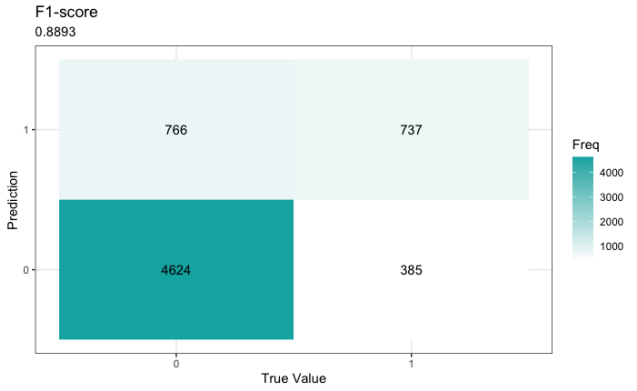
C. KNN:

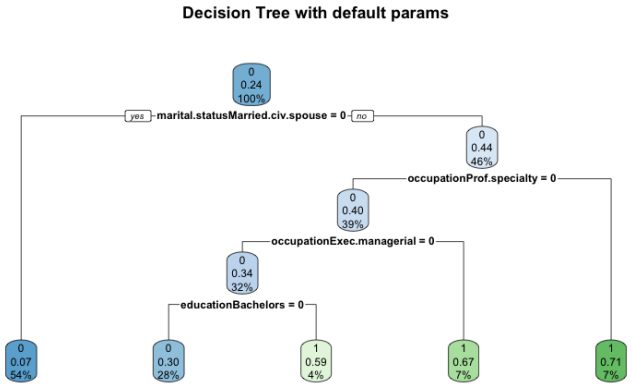
```
[1] "Optimal Number of Neighbors (k): 19"
[1] "Training Accuracy for Optimal k: 0.76221735959153
[1] "Test Accuracy for Optimal k: 0.75291769041769"
[1] "F1 Score for Optimal k: 0.857421355782012"
[1] "Confusion Matrix:"
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4838 1531
         1   78   65
```

The K-Nearest Neighbors (KNN) algorithm was applied to the dataset, and an optimal value for the number of neighbors was determined to be 19, based on maximizing accuracy. The resulting F1 Score for the KNN model was found to be 0.85. The confusion matrix provides insights into the model's performance, revealing an overall accuracy of 75.2%. Notably, the sensitivity, representing the true positive rate, is impressively high at 98.39%, indicating the model's proficiency in correctly identifying individuals with income below or equal to 50K. However, the specificity, representing the true negative rate, is relatively lower, suggesting challenges in accurately predicting higher-income individuals. The KNN model excels in identifying the majority class but faces limitations in capturing the nuances of the minority class for example the False Negatives in KNN are the highest among all of the models.

D. Decision Tree:





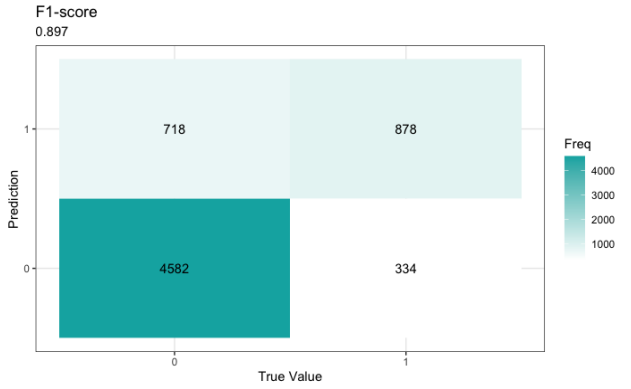
The Decision Tree model reveals crucial features such as marital status, capital-related metrics, and demographics, yielding a high F1 score of 0.889. Its hierarchical structure adeptly captures complex interactions within the data, enabling

accurate predictions based on diverse attributes. This model's strength lies in its ability to create a series of decision rules based on feature importance, allowing a clear understanding of the data's structure. While prone to overfitting in complex datasets, its interpretability and capacity to handle non-linear relationships make it a compelling choice for understanding feature importance and relationships.

	dt_model$variable.importance <dbl>
marital.statusMarried.civ.spouse	1847.3228260
marital.statusNever.married	988.9817116
relationshipNot.in.family	697.3153959
sexMale	588.2682857
age	427.5510456
occupationExec.managerial	356.5111984
occupationProf.specialty	280.4389058
relationshipOwn.child	277.0135923
educationBachelors	149.6649858
educationProf.school	29.6398844

1–10 of 17 rows

E. Random Forest:



Employing ensemble learning, Random Forest emphasizes capital metrics, marital status, and demographics as influential factors. With an F1 score of 0.9024, it effectively aggregates multiple decision trees, mitigating overfitting and enhancing predictive accuracy. The model's ensemble approach harnesses the collective power of diverse decision trees, resulting in robust predictions and a better ability to capture intricate patterns within the data. The feature importance values generated by our Random Forest model provide insights into the relative significance of different features in predicting the target variable. Among the key contributors, "marital status" stands out prominently, indicating that this metric plays a pivotal role in the model's decision-making. These insights aid in understanding the features that hold importance in the predictive accuracy of the Random Forest model and guide further exploration into the underlying patterns in the dataset.

F. XGBoost:

```
training_accuracy
```

```
[1] 0.8653691
```

```
[1] 0.841984
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4561  674
         1  355  922
```

We delved into training and evaluating an XGBoost model for predicting income levels. The goal was to fine-tune the model's hyperparameters and assess its performance on a test dataset. Hyperparameter Tuning: Parameters such as nrounds (number of boosting rounds), max_depth (maximum tree depth), eta (learning rate), and others were optimized through cross-validation to enhance the model's predictive capabilities. F1 score, a balanced metric considering both precision and recall, was computed. Achieving a high F1 score (0.905) indicated the model's strong predictive capabilities. This iterative improvement and feature importance identification aid in building

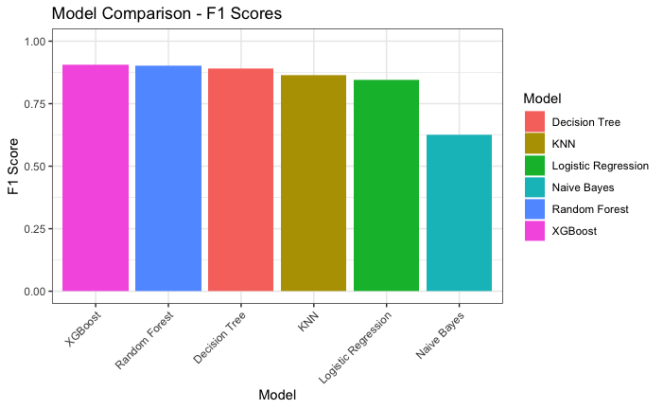
a highly accurate predictive model. The balance between model complexity and accuracy is achieved through careful hyperparameter tuning. Top Features: Education levels (Prof-school, Doctorate, 7th-8th), occupation, relationship status, workclass, and gender emerged as the most influential features.

8. Model Evaluation -

Model	F-1 Score	Train Acc	Test Acc
Logistic Regression	0.854	79.1 %	78.4 %
Naive Bayes	0.6095	57.7 %	57.1 %
KNN	0.865	76.22 %	75.29 %
Decision Tree	0.8893	82.0 %	82.3 %
Random Forest	0.9024	87.5 %	84.4 %
XGBoost	0.9054	86.53 %	84.19 %

As can be seen through the combined model evaluation table - Decision Tree, Random Forest, and XGBoost emerged as strong contenders, demonstrating superior F1 Scores and robust accuracy rates with XGBoost slightly better than the other two models. On comparing Decision Trees v/s Random Forest we can observe that there is slight overfit in Random Forest due to the difference between Train and Test Accuracy, whereas in the Decision Tree even though the Accuracy is less, but the model performs equally better on the test dataset. But with respect to computational efficiency XGBoost's performance was not that great whereas Random Forest performed better in those terms.

9. Conclusion and Future Scope:



The investigation into income disparities in the United States utilizing the Adult Census Income dataset unveiled substantial insights. Analyzing demographic, educational, and occupational attributes highlighted influential factors impacting income levels. The research aimed to predict incomes and evaluate the significance of socio-economic indicators in determining salaries. Leveraging various methodologies and machine learning models—Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors, and Naive Bayes—provided a holistic understanding of income dynamics.

Statistical summaries, exploratory data analysis, and visualizations highlighted intricate relationships between attributes and income levels. Among the models, Logistic Regression revealed a balanced F1 score (0.85) and an AUC of 0.88, showcasing strong discriminatory power with optimized regularization. The Decision Tree model excelled with an F1 score of 0.889, unraveling critical features influencing predictions. Random Forest, through ensemble learning, achieved an F1 score of more than 0.9 which is the highest among all models, emphasizing pivotal metrics such as marital status and demographics. XGBoost, after hyperparameter tuning, obtained a high F1 score of

0.905, revealing influential features like education levels and occupation. However, the Naive Bayes model delivered moderate performance, garnering an F1 score of approximately 0.62, possibly due to its assumption of feature independence, which might not align with the data distribution.

The models' evaluation showcased that Decision Trees, Random Forest, and XGBoost stood out with superior F1 scores and commendable accuracy rates. Despite slight overfitting observed in Random Forest, these models presented robust predictive capabilities. While XGBoost showed the best performance, Random Forest exhibited computational efficiency compared to the rest.

Future research could focus on addressing class imbalances in the dataset to refine model predictions, exploring advanced feature engineering techniques, and evaluating ensemble models to enhance predictive accuracy further. Moreover, investigating the temporal aspect and exploring more recent datasets could offer insights into evolving income disparities over time.

10. References –

1. Data Source - https://www.kaggle.com/datasets/uciml/adult-census-income/data
2. https://www.r-bloggers.com/2021/04/decision-trees-in-r/
3. https://www.analyticsvidhya.com/blog/2016/01/xgboost-algorithm-easy-steps/
4. https://www.qwak.com/post/xgboost-versus-random-forest#:~:text=Random%20Forest%20is%20a%20bagging,and%20determine%20the%20final%20output