From Data to Decision

Investigating Bias Amplification in Decision-Making Algorithms

Elena Mihalache[1]

Supervisor(s): Sarah Carter[1]**, Jie Yang**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Elena Mihalache
Final project course: CSE3000 Research Project
Thesis committee: Sarah Carter, Jie Yang, Stefan Buijsman, Marcus Specht

Abstract

This research investigates how biases in datasets influence the outputs of decision-making algorithms, specifically whether these biases are merely reflected or further amplified by the algorithms. Using the Adult/Census Income dataset [1] from the UCI Machine Learning Repository[2], the research explores biases through the lens of three machine learning models: Logistic Regression, Decision Tree, and Random Forest. The analysis reveals that all models exhibit varying degrees of bias, dependent on the fairness metrics applied: Demographic Parity, Disparate Impact, Equal Opportunity, Equalized Odds. It has been found that higher accuracy does not necessarily equate to fairness. The findings emphasize the complex nature of algorithmic bias and the challenge of achieving fairness in automated decision-making systems. This research contributes to the understanding of bias amplification in algorithms and underscores the need for continued efforts to develop fairer decision-making systems in various sectors.

Key terms: bias amplification, decision-making algorithms, fairness metrics

1 Introduction

In the context of computer systems, bias refers to "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others" [1]. In today's digital age, there is an increasing reliance on automated decision-making systems in diverse sectors such as, but not limited to, housing [2], employment [3], social welfare [4, 5], healthcare [6] and justice [7–10]. Despite the benefits of automated systems, recent instances have highlighted that algorithms can replicate or even amplify societal biases present in their training datasets [11, 12]. For instance, in the housing market, there are ethnicity biases in AI-based mortgage application approval systems that lead to applicants of certain ethnicities to be declined mortgages that would otherwise be approved to white applicants [2]. Prior research also indicates that bias is amplified when using AI-based decision support systems [2], or when using an algorithm to allocate limited medical resources [6]. These biases can have far-reaching consequences, especially when these systems are used to make crucial decisions that affect human lives.

Previous work in the area has laid a strong foundation for understanding and measuring algorithmic bias [1, 6, 13]. Significant research has been focused on identifying and mitigating biases in specific cases, but a comparative analysis between the biases inherent in datasets and those present in the algorithm outcome has not been carried out extensively.

The core of this research focuses on exploring how the bias present in the data used by decision-making algorithms compares to the bias in the outputs of these algorithms. Specifically, the research seeks to determine whether these algorithms only reflect the existing biases in the data or if they exacerbate them further. The main question guiding this research is : *"How does the amplification of bias in decision-making algorithms compare to the inherent biases present within their training datasets?"*. This main question is supported by several sub-questions aimed at addressing the various dimensions of algorithmic bias:

- *What characteristics of datasets most significantly contribute to bias in algorithms?*
- *How do different algorithms respond to sensitive or biased data?*
- *How do different fairness metrics compare when applied to the same decision-making system?*
- *Which fairness metrics are most effective in identifying bias amplification?*

In addressing these questions, the research aims to show how bias is replicated and potentially amplified by algorithms. By systematically comparing the biases in datasets with those in algorithmic outputs, the research will provide empirical evidence on the extent of bias amplification by decision-making algorithms.

The paper is structured as follows: Section 2 reviews the related work, providing context on bias within decision-making algorithms and outlining sensitive attributes, types of bias and fairness metrics. Section 3 presents the methodology, detailing the dataset selection, preprocessing steps, model training process and the bias measurement approach. Section 4 describes the experimental setup and preprocessed dataset analysis, revealing insights into data distribution and inherent biases. Section 5 reports the results, focusing on model training outcomes and bias measurement using fairness metrics. Section 6 discusses responsible research practices, including ethical considerations and reproducibility. Section 7 engages in a comprehensive discussion and outlines limitations. Finally, Section 8 concludes the research, summarizing findings and suggesting directions for future work.

2 Related Work

This section delves into prior research focusing on bias within decision-making algorithms. It outlines historically sensitive attributes that have been misused in decision systems, contributing to unfair outcomes, and discusses the various types of biases that can occur in data-driven environments and the different approaches to measuring fairness.

2.1 Sensitive Attributes

Protected or sensitive attributes, also referred to as variables, are disproportionately less likely to be positively classified [14]. Haeri and Zweig [15, p.1] observe that "historically, sensitive attributes of individuals were exploited to abuse the rights of individuals, leading to unfair decisions". As a consequence, legal frameworks like the Fair Housing Act (FHA)

[1] Adult/Census Income dataset available at: https://archive.ics.uci.edu/dataset/2/adult

[2] UCI Machine Learning Repository available at: https://archive.ics.uci.edu/

and the Equal Credit Opportunity Act (ECOA) [16] have been developed to prevent discrimination and to include provisions that prohibit the use of certain sensitive attributes such as race, sex, national origin, and others in decision-making systems that could lead to discriminatory outcomes.

2.2 Types of Bias

Castelnovo et al. [17, p.3] observe that "most sources of bias in data-driven decision making lie in the data itself and in the way in which they are collected". Mehbrabi et al. [18] present different types of biases in data that might result in biased algorithmic outcomes when used by machine learning algorithms, such as, but not limited to:

- **Representation Bias**: occurs when the data does not accurately represent the real-world demographics.
- **Aggregation Bias**: occurs when conclusions about individuals are derived from aggregated data, which might not hold true at the individual level.

2.3 Measuring Fairness

Berk et al. [19, p.1] have highlighted it is "impossible to maximize accuracy and fairness at the same time and impossible simultaneously to satisfy all kinds of fairness", indicating an expected variance in performance across different fairness metrics.

For the purpose of this research, group fairness metrics, meaning treating different groups equally [18], were considered. Caton and Haas [14] categorize these metrics into **parity-based**, **confusion matrix-based**, and **calibration-based** types. For this study, *Demographic Parity* and *Disparate Impact* were selected as parity-based metrics, while *Equal Opportunity* and *Equalized Odds* were chosen from the confusion matrix-based group.

Parity-based metrics

Parity-based metrics consider the predicted positive rates across different groups [14].

Demographic Parity defines fairness as "an equal probability of being classified with the positive label" [14, p.7] with the following formula:

$$P(\hat{Y} = 1 | G = \text{priv.}) = P(\hat{Y} = 1 | G = \text{unpriv.})$$

where P denotes probability, $\hat{Y}$ is the predicted outcome, and G represents the group, indicating whether an individual belongs to the privileged (priv.) or unprivileged (unpriv.) group.

For this research, Demographic Parity Difference was employed as the difference in the probability of positive outcomes between the privileged and unprivileged groups. A value of 0 indicates perfect parity.

Disparate Impact "considers the ratio between unprivileged and privileged groups" [14, p.7] with the following formula:

$$\frac{P(\hat{Y} = 1 | G = \text{unpriv.})}{P(\hat{Y} = 1 | G = \text{priv.})}$$

where P denotes probability, $\hat{Y}$ is the predicted outcome, and G represents the group, indicating whether an individual belongs to the privileged (priv.) or unprivileged (unpriv.) group.

For this research, Disparate Impact Ratio was employed as the proportion of positive outcomes for the unprivileged group to that of the privileged group. A value of 1 indicates no disparate impact, and values less than 0.8 are often used as a threshold to indicate potential discrimination [14]. Conversely, a value greater than 1.25 indicates that the proportion of positive outcomes for the unprivileged group is significantly higher than for the privileged group.

A shortcoming of Demographic Parity and Disparate Impact is that a fully accurate classifier might still be deemed unfair if the proportion of actual positive outcomes differ significantly between groups [20].

Confusion matrix-based metrics

Confusion matrix-based metrics "consider additional aspects such as True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR)" [14, p.8].

Equal Opportunity considers "potential differences in the groups being compared" [14, p.8] by promoting that the TPR is the same across different groups, with the following formula:

$$P(\hat{Y} = 1 | Y = 1, G = \text{unpriv.}) = P(\hat{Y} = 1 | Y = 1, G = \text{priv.})$$

where P denotes probability, $\hat{Y}$ is the predicted outcome, Y is the actual outcome, and G represents the group, indicating whether an individual belongs to the privileged (priv.) or unprivileged (unpriv.) group.

For this research, Equal Opportunity Difference was employed as the difference in TPR between unprivileged and privileged groups. A value of 0 indicates that both groups have the same TPR.

Equalized Odds is similar to Equal Opportunity, but also considers FPR, with the following formula:

$$P(\hat{Y} = 1 | Y = 1, G = \text{priv.}) = P(\hat{Y} = 1 | Y = 1, G = \text{unpriv.})$$
$$\text{and}$$
$$P(\hat{Y} = 1 | Y = 0, G = \text{priv.}) = P(\hat{Y} = 1 | Y = 0, G = \text{unpriv.})$$

where P denotes probability, $\hat{Y}$ is the predicted outcome, Y is the actual outcome, and G represents the group, indicating whether an individual belongs to the privileged (priv.) or unprivileged (unpriv.) group.

Equalized Odds was developed to overcome the disadvantages of Demogrphic Parity and Disparate Impact [20].

For this research, Equalized Odds Difference was employed as a measure of the largest difference in TPR and FPR between the privileged and unprivileged groups. A value of 0 indicates that both groups have the same TPR, TNR, FPR, and FNR.

3 Methodology

This research employed a quantitative analysis approach to compare the level of bias in the dataset and the outputs of decision-making algorithms. The experiment was carried out in three stages:

1. **Dataset Selection and Preprocessing**: a dataset was selected and several dataset cleaning and preprocessing

operations were performed, and the results were analysed.

2. **Model Training**: various machine learning algorithms were trained on the preprocessed data and the performance was evaluated using standard metrics.
3. **Bias Measurement**: fairness metrics were employed to measure bias and assess amplification from training set to prediction for each of the chosen algorithms.

The analysis was conducted using Python[3], with the help of libraries such as pandas[4] and numpy[5] for data manipulation, scikit-learn[6] for machine learning model implementation and fairlearn[7] for fairness metrics computations. The results were plotted using matplotlib[8] and seaborn [9].

4 Experimental Setup

This section describes the selection and preprocessing of the dataset, and the analysis performed on the resulted data.

4.1 Dataset Selection and Preprocessing

The publicly available Adult/Census Income dataset, sourced from the UCI Machine Learning Repository, was chosen due to its common use in fairness-related studies [18]. It contains information from the 1994 U.S. Census Bureau database by Ronny Kohavi and Barry Becker [21] regarding characteristics of people such as age, education, occupation, race and sex, with the aim of determining whether a person makes over $50K a year.

The dataset initially contains 32,561 entries in the training set and 16,281 entries in the test set, with 15 features—6 numerical and 9 categorical. The cleaning and preprocessing steps included, in this order:

- Dropping the 'education' column due to its redundancy with 'education-num', and removing the 'fnlwgt' column as it was not useful for the analysis.
- For entries where the individual had never worked (as determined by the 'workclass' column), setting the 'occupation' to 'None' (previously a missing value represented by '?') for logical consistency.
- Replacing missing values (represented by '?') in 'workclass', 'occupation', and 'native-country' columns with NaN (Not a Number).
- Identifying and removing 24 duplicate entries and the remaining 2,399 entries with missing values (represented by NaN after previous steps) from the training set.
- Implementing one-hot encoding for categorical features and scaling numerical features to ensure consistent data input formats for machine learning models.
- Converting the target variable 'income' to a binary numeric value, where 1 is for individuals with an income above $50K and 0 is for individuals with an income below $50K.
- Applying the same preprocessing operations on the test set and ensuring that it is appended with columns to match the training set.

The aforementioned operations resulted in 30,146 entries with 90 features in the training set and 15,055 entries with 90 features in the test set, as can be seen in Table 1.

Dataset	Stage	Entries	Features
Training Set	Before Preprocessing	32,561	15
	After Preprocessing	30,146	90
Test Set	Before Preprocessing	16,281	15
	After Preprocessing	15,055	90

Table 1: Dataset Before and After Preprocessing for Both Training and Test Sets.

4.2 Preprocessed Dataset Analysis

Analysis performed on the preprocessed data revealed:

- The dataset is imbalanced concerning the income variable, with a larger proportion of individuals earning below $50K (Appendix A.1).
- The age distribution is right-skewed, indicating a higher concentration of younger individuals in the dataset. Adults (26-45) and middle-aged (46-65) groups dominate the workforce in both income categories. The young (17-25) and seniors (66-99) show significantly fewer high-income earners (Appendix A.2).
- Most individuals have an education level of high school graduation (9-10 years), some college or bachelor's (12-13 years), and few reach higher education levels (14-16 years). Education levels ≥ 13 generally correspond to higher income (Appendix A.3).
- Capital gain is highly concentrated at zero with very few entries showing higher gains (Appendix A.4).
- Similar to capital gains, most individuals report no capital loss, with a very small number showing higher losses (Appendix A.4).
- The distribution of hours worked per week is strongly peaked around 30-40 hours, typical of full-time employment. Higher income earners are more likely to work more hours, with significant counts in the 40-60 hour range (Appendix A.5).
- 'White' dominates both income categories, with other races showing considerably fewer counts (Figure 1).
- 'Male' dominates the higher income category compared to 'Female' (Figure 2 and Table 5).
- Proportionally, income above $50K is significantly less frequent among minority groups (Table 4, Table 5, Figure 1 and Figure 2).

[3] Python available at: https://www.python.org/

[4] pandas available at: https://pypi.org/project/pandas/

[5] numpy available at: https://numpy.org

[6] scikit-learn available at: https://scikit-learn.org/stable/

[7] fairlearn available at: https://fairlearn.org/

[8] matplotlib available at: https://matplotlib.org/

[9] seaborn available at: https://seaborn.pydata.org/

- Individuals in the ’Private’ sector constitute the majority (Appendix A.6).
- Married individuals show a higher rate of earning above $50K. (Appendix A.7).
- Single or individuals who have never been married predominantly earn below $50K (Appendix A.7).

The dataset exhibits a clear representation bias, as it shows an over-representation of certain demographic groups, which does not accurately reflect the general population. For instance, the dataset indicates a much higher proportion of individuals identified as ’White’ at 85.97%, compared to ’Black’ at 9.35%, ’Asian-Pac-Islander’ at 2.97%, ’Amer-Indian-Eskimo’ at 0.95%, and ’Other’ at 0.77%. (Figure 1 and Table 3). Additionally, there is a significant gender imbalance, with ’Male’ individuals constituting 67.57% of the dataset, while ’Female’ individuals make up only 32.43%. (Figure 2 and Table 2).

Sex	Percentage in Training Set (%)
Male	67.57
Female	32.43

Table 2: Percentage of Males Compared to Females in the Training Set.

Race	Percentage in Training Set (%)
White	85.97
Black	9.35
Asian-Pac-Islander	2.97
Amer-Indian-Eskimo	0.95
Other	0.77

Table 3: Percentage of Each Race in the Training Set.

Race	≤50K (%)	>50K (%)
Amer-Indian-Eskimo	88.11	11.89
Asian-Pac-Islander	72.26	27.74
Black	87.01	12.99
Other	90.90	9.10
White	73.62	26.38

Table 4: Percentage Income Distribution by Race.

Sex	≤50K (%)	>50K (%)
Female	88.62	11.38
Male	68.61	31.39

Table 5: Percentage Income Distribution by Sex.

Furthermore, grouping individuals into broader categories such as ’Amer-Indian-Eskimo’ and ’Asian-Pac-Islander’ can likely introduce aggregation bias, as it masks the diversity and potential disparities within these broadly defined groups. Races and ethnicities are inherently diverse with significant cultural, economic, and social differences and can have significantly different employment rates or average incomes.

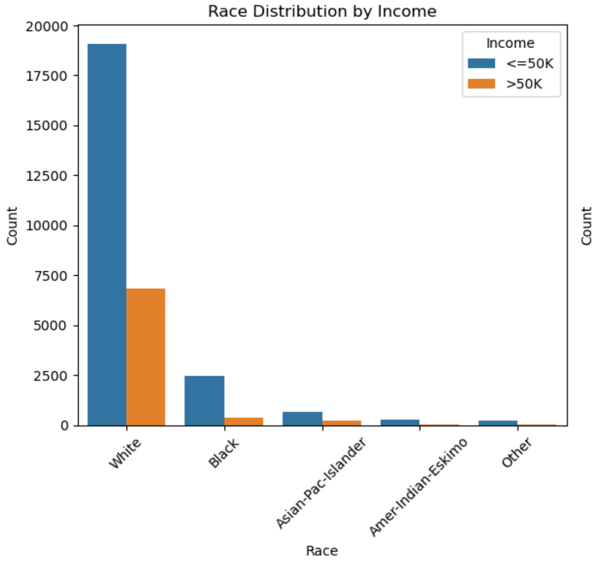


Figure 1: Count Income Distribution by Race.

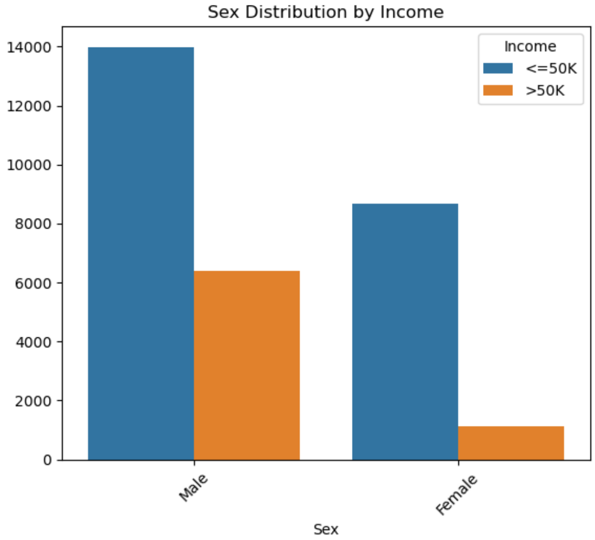


Figure 2: Count Income Distribution by Sex.

5 Results

This section outlines the training of the machine learning models and explains how different fairness metrics were used to measure and compare biases in the model outputs, based on sensitive attributes.

5.1 Model Training

Three machine learning models were trained to predict income levels: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF). The models were selected for their suitability in handling sparse data resulting from one-hot encoding and their ability to model both linear and non-linear relationships. The overall accuracies of the models can be found in Table 6.

The hyperparameters were not specifically tuned for the models in the experiment. The classifiers were initialized with their default hyperparameters, with only minor adjustments for Logistic Regression (max_iter=1000) and setting the random state for the Decision Tree and Random Forest classifiers (random_state=42). The reason for the maximum iterations adjustment was to ensure convergence, while for random state it was to ensure reproducibility.

Model	Accuracy (%)
LR	84.74
RF	84.07
DT	81.13

Table 6: Overall Accuracies of the Chosen Classfiers.

Figure 3 illustrates the performance of the same classifiers across 'Male' and 'Female' groups. The results indicate that all classifiers achieve higher accuracy for females compared to males, with Logistic Regression slightly outperforming the others. Specifically, Logistic Regression achieves an accuracy of 93% for females and 81% for males, indicating a considerable performance gap. Decision Tree, although less accurate overall, follows a similar trend with a more pronounced disparity between sexes. Random Forest maintains consistent performance close to Logistic Regression for females and marginally better than Decision Tree for males.

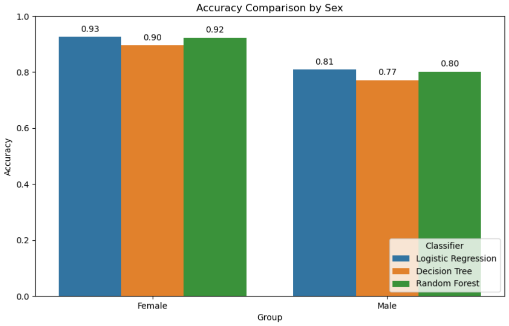


Figure 3: Accuracy Comparison of Classifiers across Different Sex Groups.

Figure 4 illustrates the performance of the aforementioned classifiers across different racial groups. While Logistic Regression consistently maintains high accuracy across all groups, the Decision Tree classifier exhibits significant variability, particularly underperforming for the 'Amer-Indian-Eskimo' and 'Asian-Pac-Islander' groups. In contrast, Random Forest demonstrates relatively stable performance, outperforming Decision Tree in almost all cases and occasionally surpassing Logistic Regression, particularly for the 'Black' and 'Other' racial groups.

5.2 Bias Measurement

After preprocessing and training the classifiers on the dataset, the fairness metrics presented in Section 2.3 were employed to quantitatively assess fairness in the algorithmic outcomes. The sensitive attributes of 'sex' and 'race' were retained in the analysis, the specific choice being motivated by the information provided in Section 2.1.

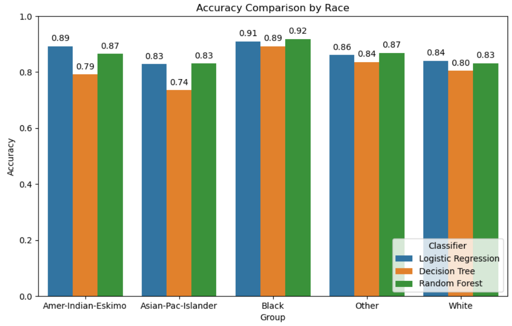


Figure 4: Accuracy Comparison of Classifiers across Different Racial Groups.

Sensitive Attribute 'sex'

Figure 5 shows the metrics values with 'sex' as the selected protected attribute, where 'sex_Female' is the unprivileged group and 'sex_Male' is the privileged group.

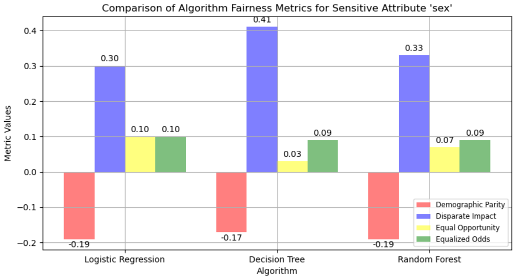


Figure 5: Comparison of Algorithm Fairness Metrics for Sensitive Attribute 'sex'.

Negative values for Demographic Parity Difference across all models suggest that females are less likely to receive favourable outcomes compared to males. This aligns with the representation bias observed in the dataset, where males are more likely to be in the higher income category. This finding points to a significant disparity that persists across the different algorithms.

Values significantly less than 1 for Disparate Impact Ratio (ranging from 0.30 to 0.41) indicate a strong disparity against females. According to legal standards, a value below 0.8 typically signifies potential discrimination. All three models exhibit values well below this threshold, highlighting a severe fairness issue.

The Equal Opportunity Difference values for the three models (ranging from 0.03 to 0.10) suggest that there are discrepancies in the likelihood of correctly identifying positive outcomes for each sex. This metric reveals that the models are less likely to correctly identify positive cases for females compared to males.

The Equalized Odds Difference values for all models (approximately 0.09 to 0.10) indicate that there are noticeable differences in how each model predicts positive outcomes and errors for males and females. This suggests that these models are not equally accurate across different sexes, thereby reinforcing existing biases.

Sensitive Attribute ’race’

Figure 6 shows the Demographic Parity Difference values with ’race’ as the selected protected attribute across all classifiers. For each race comparison pair, the first group is considered unprivileged, while the second is privileged.

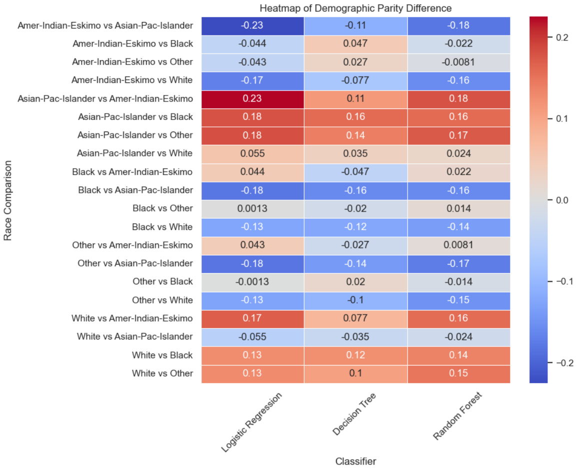


Figure 6: Demographic Parity Differences by Race Across Classifiers.

The largest disparities exist when comparing the ’Amer-Indian-Eskimo’ race with other races, particularly with ’Asian-Pac-Islander’, where the Logistic Regression model shows a large negative value of -0.23, indicating a substantially higher positive rate for the ’Asian-Pac-Islander’ group. This group receives higher values than all others, followed by ’White’.

For the ’Black’ vs. ’White’ and ’Black’ vs. ’Asian-Pac-Islander’ comparisons, the Demographic Parity Difference is consistently negative across all classifiers, suggesting the ’White’ and ’Asian-Pac-Islander’ groups tend to have a higher positive rate.

Figure 7 shows the Disparate Impact Ratio values with ’race’ as the selected protected attribute across all classifiers. For each race comparison pair, the first group is considered unprivileged, while the second is privileged.

Similar to Demographic Parity Difference, the largest disparities are observed when comparing ’Amer-Indian-Eskimo’ with ’Asian-Pac-Islander’. The Logistic Regression model shows a very high ratio of 5.8, indicating the ’Amer-Indian-Eskimo’ group has a much lower positive rate compared to ’Asian-Pac-Islander’. A comparable pattern is seen for ’White’ vs. ’Amer-Indian-Eskimo’, where the Logistic Regression model has a ratio of 4.6, suggesting the ’White’ group has a significantly higher positive rate.

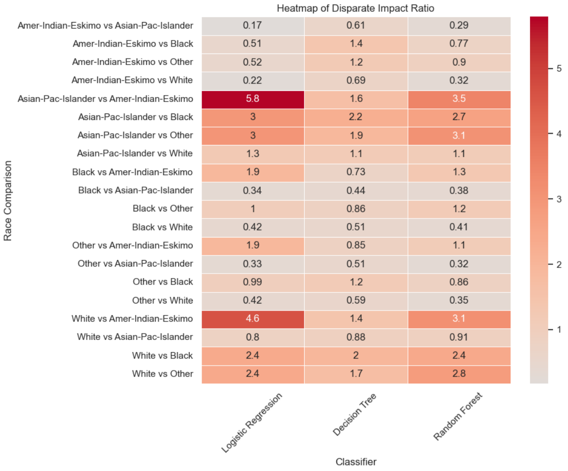


Figure 7: Disparate Impact Ratios by Race Across Classifiers.

Several race comparisons involving ’Amer-Indian-Eskimo’ as the unprivileged group exhibit Disparate Impact ratios well below 0.8 across all classifiers, indicating a significantly lower positive rate compared to other races like ’Asian-Pac-Islander’ (0.17), ’Black’ (0.51), and ’White’ (0.22) with Logistic Regression.

Figure 8 shows the Equal Opportunity Difference values with ’race’ as the selected protected attribute across all classifiers. For each race comparison pair, the first group is considered unprivileged, while the second is privileged.

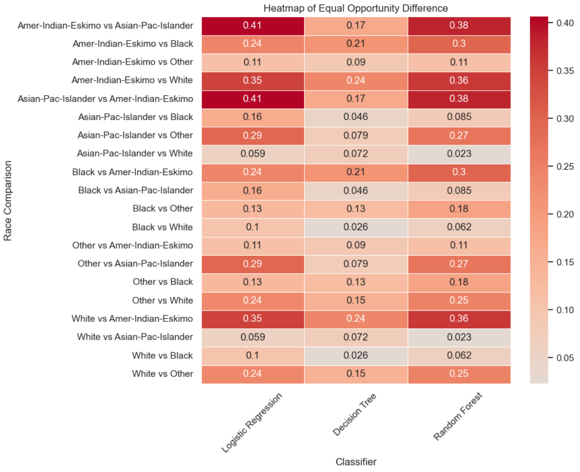


Figure 8: Equal Opportunity Differences by Race Across Classifiers.

The largest disparities are observed when comparing ’Amer-Indian-Eskimo’ with ’Asian-Pac-Islander’, with Equalized Odds Difference values of 0.41 and 0.38 for Logistic Regression and Random Forest, respectively. This suggests significant differences in true and false positive rates between these two groups. Comparisons involving the ’Amer-Indian-Eskimo’ group generally exhibit higher Equalized Odds differences across all classifiers, indicating potential biases in the model’s predictions for this group.

The Random Forest classifier tends to have slightly higher Equalized Odds Difference values compared to Logistic Regression and Decision Tree for certain race comparisons, such as 'Amer-Indian-Eskimo' vs. 'Asian-Pac-Islander' (0.38) and 'Amer-Indian-Eskimo' vs. 'Black' (0.3). Comparisons involving the 'Black' race group generally exhibit moderate Equalized Odds differences, with values ranging from 0.062 ('White' vs. 'Black') to 0.3 ('Amer-Indian-Eskimo' vs. 'Black') for the Random Forest classifier.

Figure 9 shows the Equal Opportunity Difference values with 'race' as the selected protected attribute across all classifiers. For each race comparison pair, the first group is considered unprivileged, while the second is privileged.

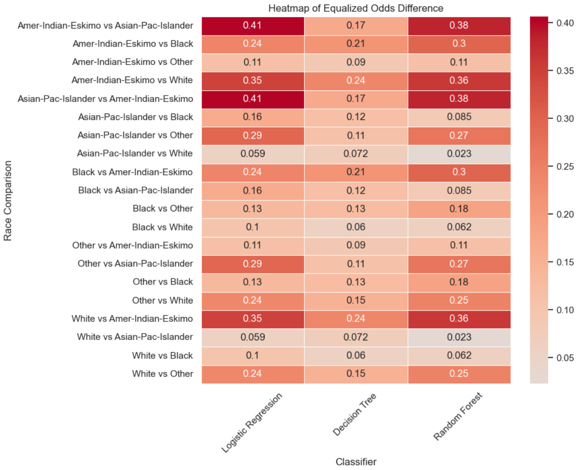


Figure 9: Equalized Odds Differences by Race Across Classifiers.

Equal Opportunity Difference has much of the same values as Equalized Odds Difference for Logistic Regression and Random Forest, meaning that in those cases the maximum between the TPR and the FPR differences was the former. This suggests that the TPR difference is the more dominant factor compared to the FPR difference when these metrics are applied, meaning that ensuring that true positives are equal across groups is more challenging than balancing false positives. Some differences exist in a number of cases for Decision Tree where the values are slightly lower, for example, when 'Asian-Pac-Islander' is taken as the unprivileged group.

6 Responsible Research

By identifying biases in algorithms, this research aims to contribute positively towards the development of fairer automated decision-making systems. All data used were anonymized and publicly available, adhering to privacy norms and ethical research standards. However, it is worth noting that sensitive data often includes information about vulnerable populations who might not have the power to consent or withdraw their data.

To support the reproducibility of the results, all methods and procedures were documented in detail. The machine learning models, data preprocessing steps, and fairness metrics used are explicitly described to enable others to replicate the research. The software and libraries used in the analysis are open-source.

7 Discussion

How does the amplification of bias in decision-making algorithms compare to the inherent biases present within their training datasets? Similar to prior work on analysing bias in prediction tasks [13], the research demonstrates that decision-making algorithms can indeed amplify inherent biases present in their training datasets. All three models, Logistic Regression, Decision Tree, and Random Forest, showed varying degrees of bias amplification depending on the fairness metric used: Demographic Parity, Disparate Impact, Equal Opportunity, Equalized Odds.

What characteristics of datasets most significantly contribute to bias in algorithms? The imbalances and representation biases in the dataset significantly contribute to bias in the algorithm outputs. The Adult/Census Income dataset used in this study exhibited clear representation biases, with certain demographic groups (e.g., 'White' and 'Male') being overrepresented. Additionally, aggregation bias is also present, where racial classifications could obscure meaningful variations within groups like 'Asian-Pac-Islander' and 'Amer-Indian-Eskimo', potentially leading to misleading conclusions about the extent and nature of bias. This imbalance was reflected and, in some cases, amplified by the algorithms. For instance, the Demographic Parity Difference and Disparate Impact Ratio revealed that females and some minority races were less likely to receive favourable outcomes compared to their 'Male' and 'White' counterparts. However, it is noteworthy that outcomes for the 'Asian-Pac-Islander' group often appeared more favourable compared to other groups, even the over-represented 'White' category. This is likely due to the fact that in the training set, 'Asian-Pac-Islander' is the only racial group with a distribution of incomes similar to the 'White' group, in fact with a slightly higher percentage of high income earners (Table 4).

How do different algorithms respond to sensitive or biased data? Regarding how different algorithms respond to biased data, Logistic Regression exhibited a consistent performance across various demographic groups but showed signs of bias amplification, particularly in terms of Disparate Impact and Demographic Parity. Decision Tree displayed the most variability in performance, suggesting a sensitivity to the nuances in the training data. Random Forest generally provided more stable outcomes across different groups, but still reflected underlying biases in the dataset.

How do different fairness metrics compare when applied to the same decision-making system? The comparative effectiveness of fairness metrics such as Demographic Parity, Disparate Impact, Equal Opportunity, and Equalized Odds was evaluated to assess how each metric highlights different facets of fairness within the same decision-making systems. Demographic Parity and Disparate Impact, focusing on outcome-based measures, were useful in identifying broad disparities in treatment across groups but did not account for the correctness of model predictions. Equal Opportunity and Equalized Odds, which consider both the rate of positive predictions and the accuracy of these predictions, provided a more nuanced view that captured discrepancies in algorithmic performance that affect fairness. This comparison highlights that, while no single metric can fully account for fairness, a combination of outcome-based and error-based metrics can provide a more comprehensive understanding of how biases manifest in algorithmic decisions.

Which fairness metrics are most effective in identifying bias amplification? To identify bias amplification effectively, certain fairness metrics were more insightful than others in specific cases. Demographic Parity and Disparate Impact can indicate if one group is systematically favoured over another in terms of positive outcomes, but do not capture whether the decisions are correct or equally accurate across groups. Equalized Odds and Equal Opportunity consider the correctness of the predictions (TPR and FPR), revealing biases in how accurately the classifiers perform for different groups. As a consequence, Equalized Odds and Equal Opportunity captured additional layers of bias not seen with Demographic Parity and Disparate Impact alone. For example, the 'Amer-Indian-Eskimo' vs. 'Black' and 'White' vs. 'Asian-Pac-Islander' groups both received scores of approximately -0.02 for Demographic Parity in the Random Forest outcomes, suggesting almost perfect parity. However, when computing the Equalized Odds for the same pairs and classifier, the latter pair got a score of approximately 0.02, indicating almost no difference, the same as Demographic Parity suggested, while the former pair got a score of approximately 0.3, signifying considerable disparities. This example clearly illustrates how Equalized Odds can capture biases in TPR or FPR that Demographic Parity might overlook.

Although the fairness metrics values seem to point to the disadvantage of 'Female' individuals, the accuracies for this group are significantly higher than those for 'Male' individuals across all classifiers. Having higher accuracy might seem like an advantage at first sight, but this is likely due to the fact that outcomes for 'Female' individuals are more homogenously distributed. In this context, this means that a higher proportion of 'Female' individuals are in the below $50K income bracket and, as a consequence, the classifiers can predict a lower income with higher confidence for this group. This finding reinforces the fact that higher accuracy does not equate to fairness.

While Kamiran and Žliobaitė [22] attribute the higher average annual income of males compared to females to the fact that females typically work fewer hours per week than males, it is important to recognize that the number of hours an individual works each week could itself be a manifestation of systemic biases. Societal pressures and norms often dictate that women take on more significant caregiving responsibilities, which can limit their availability for full-time work [23]. Therefore, the observed differences in working hours should be considered a proxy for underlying systemic biases rather than an independent variable. It is, however, not the case that females are paid less because they work fewer hours, as the Centre for Data Ethics and Innovation (CDEI) [24] used conditional parity with working hours as a risk factor to demonstrate that within the same bins of hours worked per week, women still receive lower income scores than men. This indicates that women working the same hours as men are still less likely to be high earners. Similar patterns are observed when grouping individuals by race, where disadvantaged demographics are less likely to be high earners, even if they work the same number of hours.

A significant limitation in the research is the reductionist binary and oversimplified definition of categories such as sex and race. The dataset's categorization into broad groups (e.g., 'Male' and 'Female', 'White' and 'Black') can mask underlying disparities within these groups [12]. This can also turn into aggregation bias that can contribute to misleading conclusions about the extent and nature of bias.

Another limitation is the Formalism Trap [25], where the focus on quantitative metrics and formal definitions of fairness may obscure the underlying societal and ethical complexities. Fairness metrics provide a mathematical way to assess bias, but they may not fully capture the nuanced realities of discrimination and inequity. This trap can lead to a false sense of impartiality, where achieving favourable metric values is mistaken for genuine fairness. Algorithms may still perpetuate harmful biases not accounted for by the selected metrics, and the sociotechnical context in which these algorithms operate is often overlooked.

One other limitation is the temporal relevance of the dataset. The Adult/Census Income dataset originates from 1994, and the socio-economic context has significantly evolved since then. Demographic trends, income distribution, and societal attitudes towards sex and race have all changed over the past decades. As a consequence, the biases identified in this research may not accurately reflect current realities, and the findings might have limited applicability to contemporary datasets.

It is also important to acknowledge that some of the race labels used in the dataset, such as 'Amer-Indian-Eskimo,' are considered archaic and inappropriate. These terms are not reflective of the current understanding and respect for the diversity and identity of these groups. Using such outdated terms can perpetuate stereotypes and fail to recognize the preferred and more accurate terms. For example, 'Amer-Indian-Eskimo' can be more appropriately referred to as 'Indigenous Peoples' or by specific tribal affiliations if known. Similarly, 'Asian-Pac-Islander' could be separated into 'Asian' and 'Pacific Islander' to better represent these distinct groups [26].

8 Conclusions and Future Work

This research investigated the extent to which decision-making algorithms amplify biases inherent in their training datasets. Using the Adult/Census Income dataset and applying machine learning models like Logistic Regression, Decision Tree, and Random Forest, it was found that all models displayed bias to varying degrees based on different fairness metrics such as Demographic Parity, Disparate Impact, Equal Opportunity, and Equalized Odds.

The findings indicate that achieving higher accuracy in predictions does not ensure fairness, as all models demonstrated bias amplification. The varying degrees of bias amplification observed across the different models suggest that a one-size-fits-all approach to fairness is inadequate. Algorithm designers should address the specific biases associated with the datasets and the chosen models. Moreover, the amplification of biases by decision-making systems also highlights the necessity for robust policy and regulatory frameworks. These should require transparency in algorithmic processes and regular audits to ensure compliance with fairness standards.

Significant disparities were also revealed in how different demographic groups are treated by algorithms, which can have profound implications for affected individuals. For instance, the observed biases against certain racial and gender groups call for a reevaluation of the deployment of such algorithms in critical areas like employment, housing, and justice, where biased decisions can exacerbate existing social inequalities.

Future research can include longitudinal studies that track the impact of algorithmic decisions over time, thus revealing long-term biases and their consequences. Additionally, deploying algorithms in real-world settings and continuously monitoring their performance can help identify and address biases that may not be evident in controlled environments.

On top of that, further work can also focus on developing new fairness metrics that can adapt to changing societal norms and specific application contexts.

It is also worth considering that the issue of fairness in automated decision-making is one that requires interdisciplinary research. Collaborations between computer scientists, ethicists, sociologists, and policymakers can provide diverse perspectives and solutions that are grounded in both technical feasibility and ethical considerations.

All in all, this research highlights the intricate relationship between dataset biases and their amplification through decision-making algorithms. The findings underscore that higher accuracy does not guarantee fairness, as biases can persist or even intensify within algorithmic outputs. It is imperative for developers and policymakers to prioritize fairness alongside accuracy in the development and deployment of these systems. Future research should continue to explore and address these complexities, ensuring that automated decision-making tools contribute positively to society by promoting equity and reducing bias.

References

[1] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on information systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996.

[2] L. Zou and W. Khern-am nuai, "Ai and housing discrimination: the case of mortgage applications," *AI and Ethics*, vol. 3, no. 4, 2022.

[3] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, 2016.

[4] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions," in *Conference on Fairness, Accountability and Transparency*, pp. 134–148, PMLR, 2018.

[5] R. Shroff, "Predictive analytics for city agencies: Lessons from children's services," *Big data*, vol. 5, no. 3, pp. 189–196, 2017.

[6] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, "The measure and mismeasure of fairness," *The Journal of Machine Learning Research*, vol. 24, no. 1, 2023.

[7] M. C. Cohen, S. Dahan, W. Khern-am nuai, H. Shimao, and J. Touboul, "The use of ai in legal systems: Determining independent contractor vs. employee status," *SSRN Electronic Journal*, 2022.

[8] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of data and analytics*, pp. 254–264, Auerbach Publications, 2022.

[9] R. Berk, *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media, 2012.

[10] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[11] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[12] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, pp. 77–91, PMLR, 2018.

[13] P. Garg, J. Villasenor, and V. Foggo, "Fairness metrics: A comparative analysis," in *2020 IEEE international conference on big data (Big Data)*, IEEE, 2020.

[14] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Computing Surveys*, 2020.

[15] M. A. Haeri and K. A. Zweig, "The crucial role of sensitive attributes in fair classification," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 2993–3002, IEEE, 2020.

[16] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ACM, 2019.

[17] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I. G. Penco, and A. C. Cosentini, "A clarification of the nuances in the fairness metrics landscape," *Scientific Reports*, vol. 12, no. 1, p. 4209, 2022.

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[19] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.

[20] D. Pessach and E. Shmueli, “A review on fairness in machine learning,” *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[21] R. Kohavi, “Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid,” in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[22] F. Kamiran and I. Žliobaitė, “Explainable and non-explainable discrimination in classification,” in *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases*, pp. 155–170, Springer, 2013.

[23] National Partnership for Women & Families, “The female face of family caregiving.” https://nationalpartnership.org/wp-content/uploads/2023/02/female-face-family-caregiving.pdf, Feb. 2023. (accessed June 22, 2024).

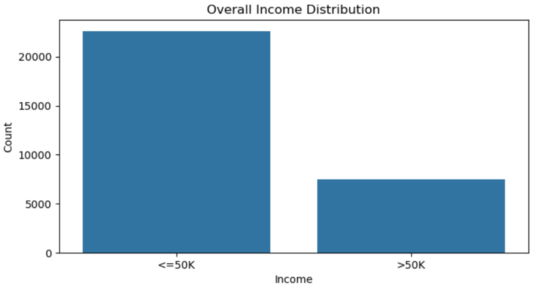
[24] Centre for Data Ethics and Innovation (CDEI), “Machine Learning Bias Mitigation.” https://cdeiuk.github.io/bias-mitigation/. (accessed June 22, 2024).

[25] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, “Fairness and abstraction in sociotechnical systems,” in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68, 2019.

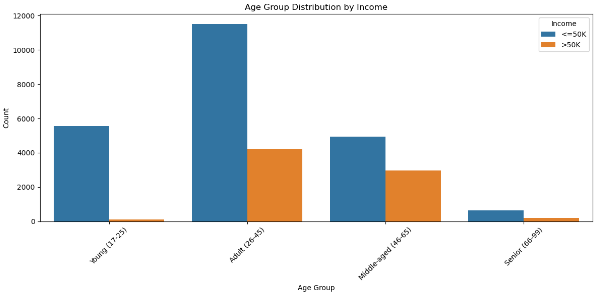
[26] U.S. Census Bureau, “Detailed races and ethnicities in the u.s. and puerto rico: 2020 census.” https://www.census.gov/library/visualizations/interactive/detailed-race-ethnicities-2020-census.html, 2023. (accessed June 23, 2024).

A Preprocessed Dataset Analysis Plots

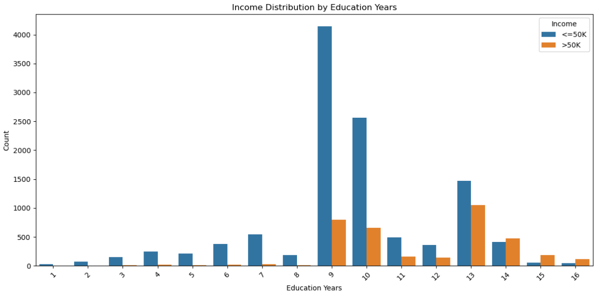
A.1 Overall Income Distribution Plot



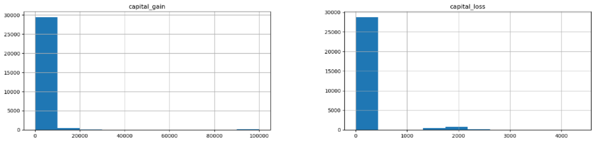
A.2 Age Distribution Plot



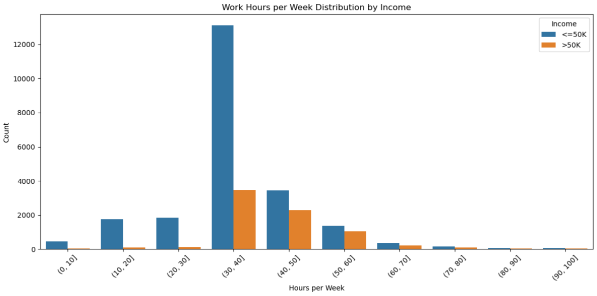
A.3 Education Years Plot



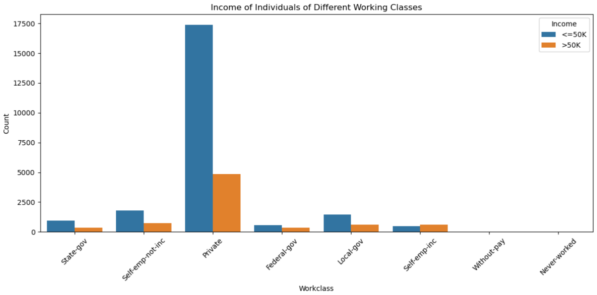
A.4 Capital Gain and Loss Plots



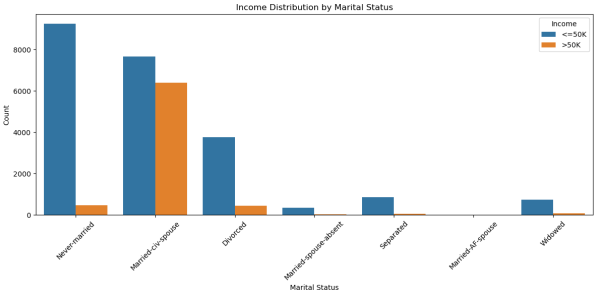
A.5 Working Hours per Week Plot



A.6 Working Classes Plot



A.7 Marital Status Plot


Income Distribution by Marital Status
Income
🟦 <=50K
🟧 >50K
Count
Never-married
Married-civ-spouse
Divorced
Married-spouse-absent
Separated
Married-AF-spouse
Widowed
Marital Status