➡ FROM DATA TO DECISION:
Investigating Bias Amplification in Decision-Making Algorithms

Author:
Elena Mihalache
e.mihalache@student.tudelft.nl
Supervisors:
Sarah Carter
Jie Yang

⚖️ Introduction

In today's digital age, there is an increasing reliance on automated decision-making systems in diverse sectors such as housing [1], employment [2], healthcare [3] and justice [4]. Recent cases have highlighted that algorithms can replicate or even amplify societal biases present in their training datasets [5], but a comparative analysis between the biases inherent in datasets and those present in the algorithm outcome has not been carried out extensively.

🔍 Research Question

The core of this research focuses on answering this question:
How does the amplification of bias in decision-making algorithms compare to the inherent biases present within their training datasets?

📝 Methodology

This research employed a quantitative analysis approach by carrying out an experiment in three stages:

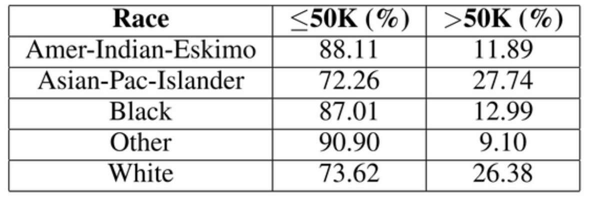
1. **Dataset Selection and Preprocessing:**
   - *Adult/Census Income dataset*
   - *Data cleaning and preprocessing*: removing duplicates, dropping redundant columns, dropping rows with missing values, applying one-hot encoding for categorical variables, scaling numerical features
   - *Result analysis*
2. **Model Training:**
   - *Training various machine learning models to predict whether an individual makes above or below $50K*: Logistic Regression, Decision Tree, Random Forest
   - *Evaluating model performance using standard metrics:* accuracy
3. **Bias Measurement:**
   - *Retaining sensitive attributes*: sex, race
   - *Employing fairness metrics to measure bias:* Demographic Parity, Disparate Impact, Equal Opportunity, Equalized Odds
   - *Assessing amplification from training set to prediction*

🧵 Analysis

Analysis performed on the preprocessed data revealed:

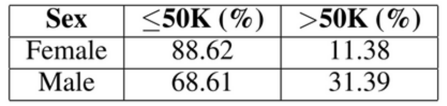
- Imbalance in the income levels, with a larger proportion of individuals earning below $50K.
- Proportionally, income above $50K is significantly less frequent among minority groups (Table 1 and Table 2).

The dataset exhibits a clear **representation bias**, as it shows an over-representation of certain demographic groups ('White' and 'Male'), which does not accurately reflect the general population. The dataset indicates a much higher proportion of individuals identified as 'White' (Table 3) and a significant gender imbalance (Table 4).
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.



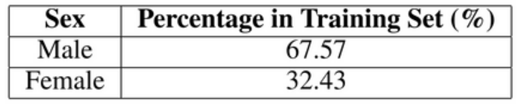
Race	≤50K (%)	>50K (%)
Amer-Indian-Eskimo	88.11	11.89
Asian-Pac-Islander	72.26	27.74
Black	87.01	12.99
Other	90.90	9.10
White	73.62	26.38

Table 1: Percentage Income Distribution by Race.



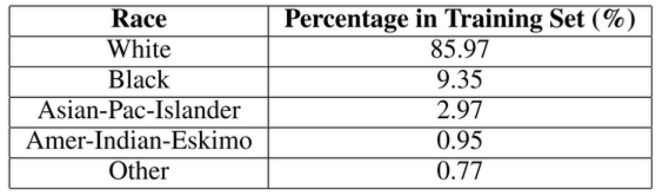
Sex	≤50K (%)	>50K (%)
Female	88.62	11.38
Male	68.61	31.39

Table 2: Percentage Income Distribution by Sex.



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

Table 3: 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 4: Percentage of Each Race in the Training Set.

📊 Results

Model Training
The results indicate that all classifiers achieve higher accuracy for females compared to males. For different racial groups, Logistic Regression and Random Forest maintain stable performance and higher accuracy, while Decision Tree exhibits some variability.
Bias Measurement
Values across all fairness metrics for all classifiers suggest that females are less likely to receive favourable outcomes than males.
Certain racial groups, particularly 'Amer-Indian-Eskimo' and 'Black', consistently receive less favourable outcomes across all fairness metrics for all classifiers compared to 'White' and 'Asian-Pac-Islander' groups.

💡 Discussion

Findings

- Decision-making algorithms can indeed amplify inherent biases present in their training datasets.
- Representation and aggregation biases significantly contribute to biased outcomes. It is, however, interesting to notice that outcomes for the 'Asian-Pac-Islander' group are often more favourable than for other groups, including the over-represented 'White' category, likely because this group has a similar or higher percentage of high-income earners in the training set.
- While outcome-based metrics identify broad disparities, error-based metrics provide a more nuanced view of algorithmic performance, suggesting that a combination of these metrics offers a more comprehensive understanding of biases. In some specific cases, Equalized Odds and Equal Opportunity metrics were more effective at identifying bias amplification, capturing additional layers of bias not seen with Demographic Parity and Disparate Impact.
- Although fairness metrics indicate a disadvantage for 'Female' individuals, their higher classification accuracy is likely due to a more homogeneous distribution of outcomes, with a higher proportion earning below $50K, highlighting that higher accuracy does not equate to fairness.

Limitations

- reductionist binary and oversimplified definition of categories such as sex and race
- Formalism Trap [6]
- temporal relevance of the dataset
- archaic and inappropriate race labels

🏁 Conclusion and Future Work

Future work

- developing robust policy and regulatory frameworks for algorithm designers
- conducting longitudinal studies to track the impact of decision-making algorithms over time
- developing new fairness metrics that can adapt to changing societal norms
- conducting interdisciplinary research to address technical feasibility and ethical considerations for fairness solutions

All in all, decision-making algorithms amplify inherent biases in their training data, demonstrating that higher accuracy does not ensure fairness, thus necessitating tailored fairness approaches, robust policy frameworks, continuous monitoring, and interdisciplinary efforts to mitigate biases and promote equity in automated systems.

REFERENCES
[1] L. Zou and W. Khern-am nuai, "Ai and housing discrimination: the case of mortgage applications," AI and Ethics, vol. 3, no. 4, 2022.
[2] Barocas and A. D. Selbst, "Big data's disparate impact," Calif. L. Rev., vol. 104, 2016.
[3] 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.
[4] 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.
[5] 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.
[6] 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.