Evaluation of Fairness in Machine Learning Models using the UCI Adult Dataset

Lucas Sena[1,2], **Javam Machado**[1,2]

[1]Laboratório de Sistemas e Banco de Dados (LSBD)
Departamento de Computação (DC)
Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brasil

[2]Mestrado e Doutorado em Ciência da Computação (MDCC)
Departamento de Computação (DC)
Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brasil

{lucas.sena,javam.machado}@lsbd.ufc.br

Abstract. *This paper presents a comprehensive analysis of fairness in machine learning models using the UCI Adult Dataset. The study focuses on mitigating biases related to sensitive attributes such as race and gender by reducing the dimensionality of the dataset. We evaluated the performance and fairness of three popular machine learning models—Logistic Regression, Random Forest, and Gradient Boosting—both with and without including sensitive features. The results indicate that while performance metrics remain stable, the fairness metrics reveal significant insights, underscoring the necessity of considering fairness alongside performance in machine learning applications.*

1. Introduction

The increasing application of machine learning models in business activities has led to significant advancements in various areas, such as sentiment analysis and audio classification [Chaves et al. 2022, Sena et al. 2022]. Despite these advancements, there are scenarios where the deployment of machine learning models requires careful attention to prevent potential biases that could lead to discrimination and adverse effects for the company [Barocas et al. 2023]. Ensuring fairness and mitigating bias in machine learning models is essential to maintain business operations' integrity and ethical standards [Stoyanovich et al. 2020].

The focus on fairness in machine learning has gained prominence, with new techniques being developed to detect and mitigate biases in these models. This ensures that machine learning applications remain ethical and reliable, fostering trust and equitable outcomes across different demographics [Mehrabi et al. 2021, Žliobaitė 2017, Caton and Haas 2024].

Recent studies have highlighted the importance of addressing bias in machine learning models, particularly in applications involving sensitive attributes. One such approach is the Protected Attribute Suppression System (PASS), which mitigates bias in face recognition by reducing the encoding of protected attributes like gender and skin tone without requiring end-to-end retraining of the entire model [Dhar et al. 2021]. Another significant work [Girhepuje 2023], which investigates bias reduction using Ensemble Learning on the UCI Adult dataset, focusing on gender bias in wage prediction and employing Kullback-Leibler (KL) divergence to measure bias.

In this work, we compare the fairness of models trained with and without protected attributes using the UCI Adult Dataset, a widely used benchmark for evaluating fairness in machine learning models. By analyzing the impact of including or excluding sensitive attributes such as race and gender, we aim to identify biases and assess their influence on model performance. Our findings contribute to the broader discourse on creating fairer and more transparent machine learning systems, aligning technological advancements with ethical standards in business operations.

2. Theoretical Foundation

2.1. Demographic Parity

Demographic Parity is achieved when the probability of receiving a favorable outcome is the same, regardless of whether an individual belongs to a privileged or unprivileged group. Formally, this is expressed as:

$$P(\hat{Y} = +|G = unprivileged) = P(\hat{Y} = +|G = privileged).$$

This metric is widely used in fairness research for classification tasks [Barocas et al. 2023], as it measures whether the outcomes are distributed equally between different groups. However, caution must be exercised, as achieving perfect Demographic Parity could result in reverse discrimination [Koumeri et al. 2023].

Example: Imagine a loan approval system that uses machine learning to decide whether a person qualifies for a loan. Demographic Parity would be achieved if the approval rate is the same for two groups, such as men and women. For instance, if 70% of men and 70% of women receive loan approvals, the system would satisfy Demographic Parity. However, if one group consistently receives a higher approval rate (e.g., 80% for men and 60% for women), this would indicate a violation of Demographic Parity, highlighting potential bias.

2.2. Equalized Odds

Equalized Odds ensure that the model's prediction accuracy, whether a true positive or false positive, is consistent across privileged and unprivileged groups, conditional on the actual outcome [Hardt et al. 2016]. Formally, it is defined as:

$$P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y),$$

for all outcomes y. This fairness metric has gained prominence due to its robustness in ensuring equitable outcomes across different demographics [Mehrabi et al. 2021].

Example: Consider a hiring model that predicts whether a candidate should be hired based on their qualifications. Equalized Odds would be satisfied if, given that a candidate is qualified, the likelihood of being correctly predicted as "hired" is the same for all demographic groups. For example, if the model predicts that 85% of qualified men and 85% of qualified women are hired and the false positive rates (incorrect predictions of being hired when unqualified) are also the same, the model would meet the Equalized Odds criterion. Disparities in these rates across groups would indicate bias in the model's decision-making process.

3. Related Work

The issue of fairness in machine learning models has garnered increasing attention in the research community, with various approaches and methodologies being proposed to mitigate biases and promote more equitable decisions [Caton and Haas 2024].

[Dhar et al. 2021] propose the Protected Attribute Suppression System (PASS) to reduce bias in face recognition by suppressing the encoding of protected attributes such as gender and skin tone. Their method operates on face descriptors from pre-trained networks, achieving high verification accuracy without the need for end-to-end retraining. Unlike Dhar et al.'s study, our study focuses on evaluating discrimination in the UCI Adult Dataset. We analyze bias using fairness metrics such as Demographic Parity Difference and Equalized Odds Difference, providing a detailed investigation into bias mitigation strategies within this specific context.

Several studies have addressed bias in machine learning models concerning sensitive attributes. Girhepuje's work [Girhepuje 2023] examines gender bias in wage prediction using Ensemble Learning on the UCI Adult dataset, revealing significant disparities and higher bias in tree-based models. In contrast, our study evaluates discrimination in the UCI Adult Dataset by analyzing two fairness metrics, Demographic Parity Difference, and Equalized Odds Difference, specifically investigating biases related to race and gender for a comprehensive analysis of bias mitigation.

4. Methodology

4.1. Dataset

The UCI Adult dataset, also known as the "Census Income" dataset, is a popular benchmark for machine learning. It comes from the 1994 U.S. Census and is used to predict if a person's income exceeds $50,000 per year based on demographic information such as age, education, occupation, race, and gender. The dataset contains 48,842 entries with 14 attributes. The binary target variable indicates whether the income is less than or equal to $50,000 or greater than $50,000. The UCI Adult dataset is extensively used for classification tasks, fairness analysis to evaluate and mitigate biases, and benchmarking different machine learning algorithms.

4.2. Experimental Design

The primary objective of our experiments was to assess the impact of including sensitive features (such as race and sex) on the performance and fairness of machine learning models. To achieve this, we conducted the following experiments for each dataset:

- **Model Training without Sensitive Features:** In this setup, we trained the models using all available features except the sensitive ones (e.g., race and sex). This approach evaluates the model's performance and fairness without directly considering sensitive attributes.
- **Model Training with Sensitive Features:** In this setup, we included the sensitive features in the training process. This approach allows us to assess the impact of sensitive attributes on the model's predictions and identify potential biases.

4.3. Models and Metrics

We selected three popular machine learning models for our experiments: **Logistic Regression**, **Random Forest**, and **Gradient Boosting**. These models were chosen due to their widespread use in various applications and their different approaches to handling data.

The performance of each model was evaluated using the most commonly utilized metrics, **Accuracy** and **F1 Score**. Additionally, we assessed the fairness of the models using **Demographic Parity Difference** and **Equalized Odds Difference**.

- **Demographic Parity Difference**: Measures the difference in positive outcome rates between groups. It ranges from -1 to 1, where a value of 0 indicates perfect fairness. Values different from zero indicate a disparity, with values further from zero indicating greater disparity.
- **Equalized Odds Difference**: Measures the difference between true and false positive rates between groups. It also ranges from -1 to 1, where 0 indicates perfect fairness. Values different from zero suggest that the model's predictions are not equally accurate for different groups, with values further from zero indicating greater inequality.

By comparing the results of models trained with and without sensitive features, we aim to highlight the importance of considering fairness in machine learning applications and provide insights into how sensitive attributes can impact model performance and bias.

5. Experiments and results

5.1. Performance Metrics

Table 1 shows the performance metrics for Logistic Regression, Random Forest, and Gradient Boosting models trained with and without sensitive features on the UCI Adult Dataset. The inclusion of sensitive features had a minimal impact on the performance metrics.

Table 1. Performance Metrics for UCI Adult Dataset (Bold indicates the best result)

Model	Accuracy	F1 Score
Without Sensitive Features		
Logistic Regression	85.22%	0.6585
Random Forest	**83.62**%	**0.6378**
Gradient Boosting	86.34%	0.6742
With Sensitive Features		
Logistic Regression	**85.23**%	**0.6588**
Random Forest	83.45%	0.6351
Gradient Boosting	86.34%	0.6742

The results indicate that the performance metrics, such as accuracy and F1 score, remain relatively stable regardless of whether sensitive features are included. For instance, the accuracy of the Logistic Regression model is 85.22% without sensitive features and

85.23% with sensitive features, demonstrating minimal change. Similar trends are observed for the Gradient Boosting model, which maintains an accuracy of 86.34% and an F1 score of 0.6742 in both cases. The Random Forest model shows a slight decrease in accuracy from 83.62% to 83.45% and a small reduction in the F1 score from 0.6378 to 0.6351 when sensitive features are included, indicating a minimal impact overall.

5.2. Fairness Metrics

Table 2 presents the fairness metrics for the same models. We report the Demographic Parity Difference and Equalized Odds Difference for both race and sex sensitive features.

Table 2. Fairness Metrics for UCI Adult Dataset (Bold indicates the greatest disparity)

Model	Metric	Race	Sex
Without Sensitive Features			
Logistic Regression	Demographic Parity Difference	0.0528	0.0193
	Equalized Odds Difference	0.1008	0.0599
Random Forest	Demographic Parity Difference	0.0100	0.0008
	Equalized Odds Difference	**-0.0349**	0.0111
Gradient Boosting	Demographic Parity Difference	0.0599	0.0127
	Equalized Odds Difference	0.1186	0.0422
With Sensitive Features			
Logistic Regression	Demographic Parity Difference	0.0528	0.0193
	Equalized Odds Difference	0.1008	0.0599
Random Forest	Demographic Parity Difference	**0.0158**	**0.0121**
	Equalized Odds Difference	-0.0110	**0.0216**
Gradient Boosting	Demographic Parity Difference	0.0599	0.0127
	Equalized Odds Difference	0.1186	0.0422

The results for the Random Forest model showed a slight increase in disparity when sensitive features were included, which was not entirely unexpected. The Demographic Parity Difference increased from 0.0100 to 0.0158 for race and from 0.0008 to 0.0121 for sex, indicating a more pronounced disparity for sex. Similarly, the Equalized Odds Difference for sex increased from 0.0111 to 0.0216 with the inclusion of sensitive features. These values, ranging from -1 to 1, are close to 0 and thus indicate relatively fair models. However, the increases observed do suggest some bias, though not necessarily significant. The variations between scenarios (with and without sensitive features) are small but relevant, showing that including sensitive features can introduce or amplify bias. The variation between models is also noteworthy, as the Random Forest model showed more sensitivity to including sensitive features compared to Logistic Regression and Gradient Boosting, which remained consistent regardless of the features included.

6. Discussion and Contributions

Including sensitive features had a minimal impact on performance metrics but revealed significant variations in fairness metrics. For the UCI Adult Dataset, the Random Forest model showed increased disparity when sensitive features were included, while Logistic

Regression and Gradient Boosting remained consistent. This indicates that including sensitive features can sometimes lead to increased bias, particularly in models like Random Forest, which may handle interactions between features differently.

These findings underscore the importance of considering fairness alongside performance in machine learning, particularly in applications with real-world consequences. Our results suggest that while some models handle sensitive features without significant bias, others, like Random Forest, require further scrutiny and potential bias mitigation strategies.

6.1. Contributions

Our study extends existing research on fairness in machine learning by evaluating multiple models on real-world datasets, specifically the UCI Adult Dataset. We provide insights into how fairness metrics are affected by the inclusion of sensitive attributes and emphasize the necessity of ongoing evaluation of fairness and performance metrics in developing ethical machine learning systems.

6.2. Limitations and Future Work

The primary limitation of this study is the reliance on a single dataset, which may restrict the generalizability of our findings. Future research should explore additional datasets from domains like healthcare and finance to validate the observed fairness metrics across different contexts. Further exploration of complex models such as deep neural networks is necessary to understand how fairness can be ensured in more advanced applications. Practical case studies implementing fairness interventions would provide valuable contributions to the field.

7. Conclusion

Our study highlights the critical importance of evaluating and addressing fairness in machine learning models, especially those that can significantly impact individuals' lives. The UCI Adult Dataset results showed that including sensitive features can sometimes lead to increased disparity, as seen in the Random Forest model. Conversely, the Logistic Regression and Gradient Boosting models maintained consistent fairness metrics regardless of whether sensitive features were included. These findings emphasize that removing sensitive features alone is insufficient to guarantee fairness. The disparities observed in the Random Forest model underscore the necessity of balancing performance and fairness, as certain models may inherently handle sensitive features differently. This illustrates that simply excluding sensitive features is not a comprehensive solution. Therefore, our results stress the need for ongoing evaluation and refinement of fairness interventions in machine learning models, encouraging the development of effective and ethically sound systems. Future work should continue to explore diverse datasets and refine fairness interventions to ensure equitable outcomes across different demographic groups.

Acknowledgment

This research was partially funded by CNPq/Brazil under grant number 316729/2021-3.

References

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38.

Chaves, I. C., Martins, A. D. F., Praciano, F. D., Brito, F. T., Monteiro, J. M., and Machado, J. C. (2022). Bpa: A multilingual sentiment analysis approach based on bilstm. In *ICEIS (1)*, pages 553–560.

Dhar, P., Gleason, J., Roy, A., Castillo, C. D., and Chellappa, R. (2021). Pass: protected attribute suppression system for mitigating bias in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15087–15096.

Girhepuje, S. (2023). Identifying and examining machine learning biases on adult dataset. *arXiv preprint arXiv:2310.09373*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Sena, L. B., Praciano, F. D., Chaves, I. C., Brito, F. T., Neto, E. R. D., Monteiro, J. M., and Machado, J. C. (2022). Audio-mc: A general framework for multi-context audio classification. In *ICEIS (1)*, pages 374–383.

Stoyanovich, J., Howe, B., and Jagadish, H. V. (2020). Responsible data management. *Proceedings of the VLDB Endowment*, 13(12).

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089.