Fairness in Machine Learning Models (PER2024–030)

Authors:
Adam Dienes, Bence Zsolt Beregi
Supervisors:
Mireille Blay-Fornarino, Anne-Marie Pinna-Dery, and Nicolas Lacroix

University Côte d'Azur, Campus SophiaTech, 930 route des Colles BP145, 06903
Sophia-Antipolis Cedex, France
https://polytech.univ-cotedazur.fr
{adam.dienes, bence-zsolt.beregi}@etu.univ-cotedazur.fr
{mireille.blay, anne-marie.pinna, nicolas.lacroix}@univ-cotedazur.fr

Abstract. As artificial intelligence increasingly influences critical decisions, ensuring model and dataset fairness is paramount. Our scientific research addresses the challenge of systematically visualizing and justifying the fairness of datasets and machine learning models throughout their lifecycle. We aim to bridge the gap between complex algorithms and real-world impact, empowering both model integrators and AI developers to build more equitable and transparent ML systems. We propose interactive visualizations that assess fairness without accessing sensitive attributes. Our approach integrates both pre-processing and post-processing techniques in ML to mitigate biases at multiple stages of the pipeline.

Keywords: Machine Learning · Model Fairness · Dataset Fairness · Interactive Visualization · Bias Mitigation · Transparency in AI · Model Lifecycle · Responsible AI · Fairness Assessment Tools · Bias Detection.

1 Introduction

With the rise of Machine Learning (ML) models, it becomes crucial to ensure the fairness of these models throughout their lifecycle, leading states to propose frameworks to help organizations foster the responsible design, development, deployment and use of Artificial Intelligence (AI) systems over time. However, there is no systematic support to guide scientists and engineers in monitoring this property. This project aims to fill this gap by providing tools to visualize and justify the fairness of datasets and ML models.

2 Description

As ML models increasingly influence important choices in industries like recruiting, finance, healthcare and law enforcement, there is rising concern about ensuring justice in these algorithms. Unfair results can arise from biases in (historical) data or model behavior which disproportionately affect particular demographic groups. Even though fairness is acknowledged to be important, there are still insufficient systematic tools and procedures to track and support fairness across the whole lifecycle of datasets and models.

Large datasets and complicated systems may not be adequately served by the post-analysis and manual inspection methods that are frequently used in current fairness auditing techniques. Furthermore, access to sensitive characteristics like socioeconomic status, gender or ethnicity is necessary for many fairness evaluation techniques. Due to the lack of such data, this raises privacy and bias issues.

Furthermore, it can be difficult for AI engineers and even for model integrators to connect technical fairness metrics with practical applications and actionable recommendations. For non-experts, it can be quite challenging to understand at first sight and convert current fairness measurements into useful information. These difficulties are made worse by the lack of user-friendly, interactive visualization tools. It is challenging to identify problems and compare models or perform fairness assessments to larger audiences in the absence of unmistakable, visual representations.

This research focuses on addressing these gaps by developing interactive visual tools to assess fairness without relying on sensitive attributes. This tool aims to detect and mitigate bias while promoting transparency, recommendation and most importantly understanding.

3 State-of-the-art

Fairness and bias prevention in machine learning have emerged as essential research fields as AI systems are being used in high-impact domains such as healthcare, recruitment, lending and criminal justice. Despite their strength, these systems frequently reinforce or inherit prejudices from their data or algorithms, which raise moral, societal and legal issues. While adherence to this property is crucial, evaluating the fairness of a system incorporating a ML model remains challenging. This is because the understanding of fairness varies across different contexts. This summary of the state-of-the-art explores the latest approaches, challenges and resources for addressing biases and advancing fairness in machine learning. The review aims to identify the key conceptual and visual elements necessary to characterize and promote the evaluation and justification of "fairness".

3.1 Challenges in Achieving Fairness

The difficulty of defining fairness is one of its greatest obstacles. Different ethical concerns are reflected in the various formal definitions of fairness[11], including equality of odds, statistical parity/demographic parity, equal opportunity, disparate/adverse impact, PPV-parity, FPR-parity and NPV-parity [2]. These definitions, however, frequently clash, making it impossible to accomplish them all at once. The impossibility theorem, which illustrates the intrinsic trade-offs among fairness criteria, effectively captures the difficulties of balancing fairness standards in machine learning [3].

For instance, when base rates vary between groups, it is theoretically difficult to achieve "predictive parity" (equalizing positive predictive value across groups) and "classification parity" (equalizing false positive and false negative rates). This suggests that, depending on the criteria given priority, fairness interventions invariably jeopardize other goals, such as overall accuracy or equitable error rates [1,4].

Similarly, the impossible finding by Kleinberg et al. [5] demonstrates that when base rates differ, goals such as classification balance (equal average risk scores for positives and negatives across groups) and predictive calibration (uniform interpretation of risk scores across groups) cannot co-exist. This emphasizes how fairness modifications frequently call for sacrificing accuracy or giving preference to one fairness metric over another.

As demonstrated by tools such as COMPAS software, where balancing fairness metrics resulted in differences in false positive and false negative rates between demographic groups, these trade-offs call into question the practical usefulness of fairness modifications. [6] Another significant issue is bias in training data. Historical injustices, institutionalized prejudice, or the underrepresentation of particular groups are frequently reflected in data. Therefore, discriminatory results can be obtained from even seemingly "unbiased" models. Effectively diagnosing and mitigating bias is further complicated by intersectionality, which is the combined influence of several protected characteristics (such as race, religion, gender, sexual orientation and ethnicity).

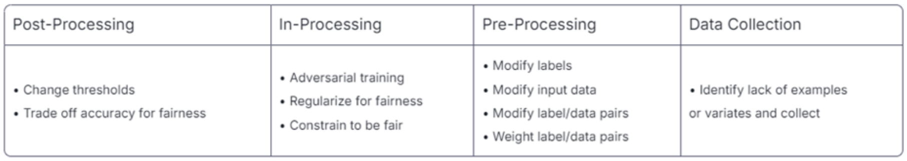
In essence, achieving fairness in ML is a multifaceted challenge that necessitates a deliberate examination of the context, objectives and definitions applied. As Bell et al. [3] highlight, fairness often involves unavoidable trade-offs between competing goals and acceptable error levels. By surfacing and explicitly addressing these compromises, practitioners can better navigate the intricate interplay of factors such as bias, intersectionality and contextual nuances. This approach fosters transparency and accountability, paving the way toward models that more effectively balance fairness and utility in real-world applications.

3.2 Fairness and Bias Mitigation Techniques

Pre-processing, in-processing and post-processing strategies are the three main categories into which efforts to reduce bias in machine learning models fall [7].

The goal of pre-processing techniques (agnostic to ML approach) is to lessen bias in the input of data prior to training. While reweighing approaches give training samples various weights depending on their representation, techniques such as disparate impact reduction modify feature distributions to lessen discrepancies between groups. By using causal inference to detect and address biases resulting from protected features, causal fairness approaches go one step further. There are additional techniques as well like massaging the data, sampling and correlation remover.

To apply fairness restrictions during model training, in-processing techniques step in (modification of the algorithm to remove discrimination). One popular technique is adversarial de-biasing, which uses a two-part model in which the primary predictor reduces the adversary's capacity to identify protected features. By incorporating fairness penalties into the model's loss function, fair regularization makes sure the model optimizes for both accuracy and fairness. Fairness restrictions are especially incorporated into the topologies of some neural networks. [8]



Post-Processing	In-Processing	Pre-Processing	Data Collection
• Change thresholds • Trade off accuracy for fairness	• Adversarial training • Regularize for fairness • Constrain to be fair	• Modify labels • Modify input data • Modify label/data pairs • Weight label/data pairs	• Identify lack of examples or variates and collect

Fig. 1. The main categories of reducing bias / Source: www.rbcborealis.com

After training, post-processing methods adjust model predictions to attain fairness. Predictions are adjusted using techniques like equalized odds post-processing to balance error rates (such as false positives and true positives) among groups. To reduce inequities, threshold modifications use distinct judgment thresholds for various groups.

The alignment between the fairness definition and the chosen technique ensures that the objectives are not only met but also appropriately justified within the given context. Therefore, it is essential to select a method that both promotes fairness in line with the definition being applied and effectively communicates the rationale behind its use [9].

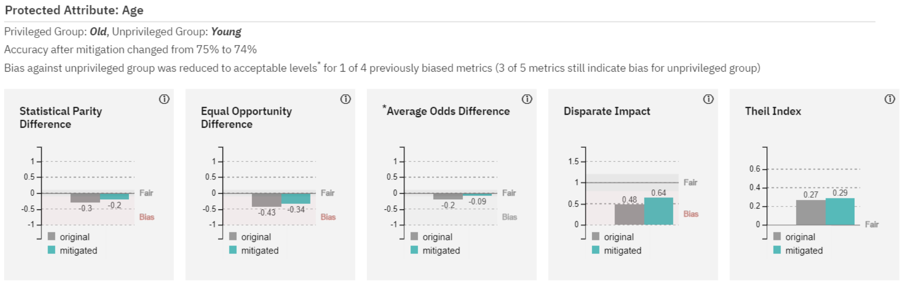
3.3 Metrics for Measuring Fairness

Metrics that measure bias at the individual and group levels are essential to the assessment of fairness. While equalized odds assess consistency in true and false positive rates, group fairness criteria like statistical parity gauge how consistently positive results occur across groups.

Disparate impact, which is frequently employed in regulatory situations, measures the proportion of favorable outcomes between groups. Metrics of individual fairness evaluate whether the model predicts similar outcomes for similar people. Trade-off measures, which frequently strike a balance between justice and accuracy or other performance goals, quantify how fairness interventions affect utility. However, despite the availability of these metrics, there are still significant gaps in the ability to easily assess and justify the fairness of a model. Many of these indicators are complex and require specialized knowledge to interpret correctly. More thorough visualizations that can incorporate numerous indications and offer a comprehensive perspective of a model's fairness are also required [12].

3.4 Tools and Frameworks for Fairness

Several open-source tools have been created to help practitioners assess and reduce prejudice. A comprehensive collection of fairness metrics and mitigation strategies is offered by IBM's AI Fairness 360 (AIF360). Themis-ML specializes in bias detection and fairness-aware machine learning, while Fairlearn concentrates on Python-based fairness evaluation and mitigation. By providing an interactive interface for investigating fairness in machine learning models, Google's What-If Tool helps practitioners see and comprehend the effects of changes.



Protected Attribute: Age

Privileged Group: ***Old***, Unprivileged Group: ***Young***

Accuracy after mitigation changed from 75% to 74%

Bias against unprivileged group was reduced to acceptable levels* for 1 of 4 previously biased metrics (3 of 5 metrics still indicate bias for unprivileged group)

Fig. 2. AIF360 metrics dashboard / Source: aif360.res.ibm.com

At first look, it can be difficult to understand fairness metric representations in ML and it frequently takes a lot of work to rigorously evaluate and implement insights. Even if existing tools like bias heatmaps, disparity matrices and fairness dashboards offer helpful visualizations of bias and fairness indicators, there is still much space for improvement in terms of their usability and accessibility [13].

For instance, users can interactively compare measures like equalized chances and demographic parity using fairness dashboards like Fairlearn and AIF360. However, consumers without a strong technical background may find these visualizations too comprehensive. Despite their visual compactness, bias heatmaps and disparity matrices frequently need to be carefully interpreted in order to understand subtle intersectional biases or fairness breaches [14].

Integrating explainability tools (e.g. SHAP) into these visualizations is a step forward but can add additional layers of complexity. For non-expert users, even sophisticated techniques like dimensionality reduction (e.g. t-SNE) to visualize biases in data clusters are not intuitive by nature. A recurring issue is the absence of tools designed to foster intuitive understanding rather than relying solely on mathematical representations. For instance, Chouldechova [15] points out that although fairness-related proofs could be succinct, their dependence on technical terms and mathematical jargon rather than visual or descriptive methods sometimes prevents them from evoking perception. This highlights a crucial need: the requirement for tools that bridge the gap between intricate technical principles and practical understanding through understandable, interactive representations and examples. The possibility of combining visual aids with real-world examples to make fairness metrics easier to comprehend and apply.

Even with continuous advancements, the needs of various stakeholders—from business executives to legislators—cannot be adequately met by the visual techniques currently in use. To overcome these limitations and allow both technical and non-technical users (model integrators) to participate meaningfully with fairness-related concerns in machine learning, more interactive, intuitive and instructional visualization techniques are needed.

3.5 Recent Advances in Fairness

Integrating fairness with innovative technology and approaches has been the focus of recent study. For example, causal inference is used in causal fairness techniques to separate the effects of protected traits from the actual causes of outcomes. To guarantee that predictions stay fair in the face of such changes, counterfactual fairness techniques investigate how modifications to sensitive qualities affect model predictions. With multi-adversary systems addressing fairness across several protected qualities at once, adversarial learning remains a thriving area of innovation. Fairness-aware AutoML, in which automated machine learning algorithms integrate fairness restrictions into the optimization process, is another noteworthy trend.

3.6 Future Directions and Summary

Hybrid approaches that include pre-, in- and post-processing techniques are becoming more and more popular as fairness research progresses. These methods promise to address issues of justice in a more comprehensive way. Furthermore, because they facilitate accurate diagnosis and effective communication of bias-related problems, interpretability and explainability are increasingly becoming crucial elements of fairness research.[10]

However, the creation of better visualization methods—tools that are user-friendly and efficient in conveying insights linked to fairness—is one of the most urgent demands. The field can close the gap between technical complexity and practical understanding by concentrating on developing more approachable, interactive and example-driven visualization techniques, guaranteeing that fairness becomes an understandable and achievable objective for all parties involved.

4 Methodology

As noted in our state-of-the-art review, our study methodology is designed to tackle the difficulties associated with fairness assessment in machine learning with a specific emphasis on the need for more interactive, intuitive and educational visualization tools. In keeping with the recognized need for instruments that close the gap between complex technical concepts and real-world knowledge, our strategy combines the creation of a new software tool with a thorough assessment methodology.

4.1 Tool Development

We adopted an iterative, user-centered design process for developing our fairness assessment tool, leveraging the strengths of Jupyter Notebook for rapid prototyping and Mercury for creating a user-friendly interactive front-end. The initial phase involved a comprehensive requirements-gathering process, identifying key needs based on challenges and gaps documented in current research. These requirements included support for a variety of fairness metrics – Statistical Parity/Demographic Parity, Equality of Odds, Equal Opportunity, Disparate Impact and MinDiff – accompanied by intuitive visualizations capable of conveying the meaning and implications of these metrics to both technical and non-technical users. Further requirements included actionable guidance for mitigating fairness issues and a mechanism for justifying interventions and documenting trade-offs. Prototyping initial versions of the tool in Jupyter Notebook then emphasized core functionalities like data upload, attribute selection, metric calculation and visualization.

This was followed by actively soliciting feedback from a diverse group of potential users, including machine learning practitioners, domain experts and individuals with limited technical expertise, to refine the tool's design, enhance its usability and ensure that it effectively met the needs of its target audience. Based on this user feedback, the final implementation incorporated key features such as a user-friendly interface for CSV dataset upload, automated calculation of fairness metrics, interactive visualizations tailored to each metric, contextual guidance for addressing fairness issues and a justification module for documenting interventions.

Our technology stack consists of Python, Jupyter Notebook, Mercury, Pandas, Scikit-learn, Plotly and Matplotlib.

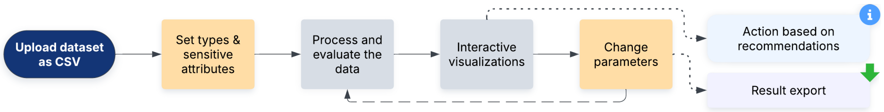

Upload dataset as CSV
Set types & sensitive attributes
Process and evaluate the data
Interactive visualizations
Change parameters
Action based on recommendations
Result export

Fig. 3. Workflow of the application

4.2 Evaluation Framework

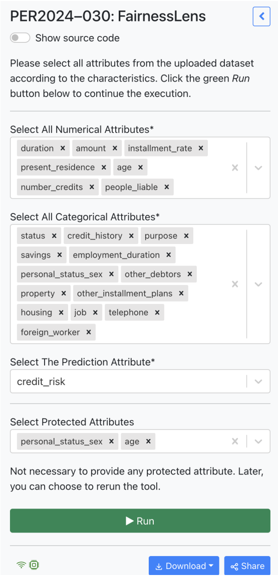
4.3 Experimental Procedure and Ethical Considerations

Our experimental procedure involved a carefully controlled process consisting of a baseline experiment, an intervention experiment and a detailed comparison. During the baseline experiment, machine learning models were trained on the selected datasets without using our tool, thus establishing a control for assessing any improvements achieved through intervention. Subsequently, in the intervention experiment, users were provided with our tool and asked to identify and mitigate fairness issues, allowing us to observe how effectively the tool facilitated these actions. A final comparison then assessed the impact of our tool on fairness. In all aspects of our research, we adhered to strict ethical guidelines. The collected data was anonymized. By following this methodology, we aim to demonstrate the effectiveness of our fairness assessment tool in empowering users to build more equitable and transparent AI systems.

5 Solution

Our research introduces a novel software tool designed to empower users to detect, visualize and address fairness issues within their datasets and machine-learning models. Developed using Jupyter Notebook and leveraging the interactive capabilities of Mercury for a user-friendly front-end, our solution allows users to effortlessly upload CSV datasets and designate numeric, categorical, sensitive and prediction attributes. This streamlined process forms the foundation for a comprehensive fairness analysis and model training workflow.

The core of our solution lies in its ability to analyze several key fairness metrics, providing insightful details and actionable guidance based on the results. By



PER2024–030: FairnessLens

Show source code

Please select all attributes from the uploaded dataset according to the characteristics. Click the green *Run* button below to continue the execution.

Select All Numerical Attributes*

duration × amount × installment_rate × present_residence × age × number_credits × people_liable ×

Select All Categorical Attributes*

status × credit_history × purpose × savings × employment_duration × personal_status_sex × other_debtors × property × other_installment_plans × housing × job × telephone × foreign_worker ×

Select The Prediction Attribute*

credit_risk

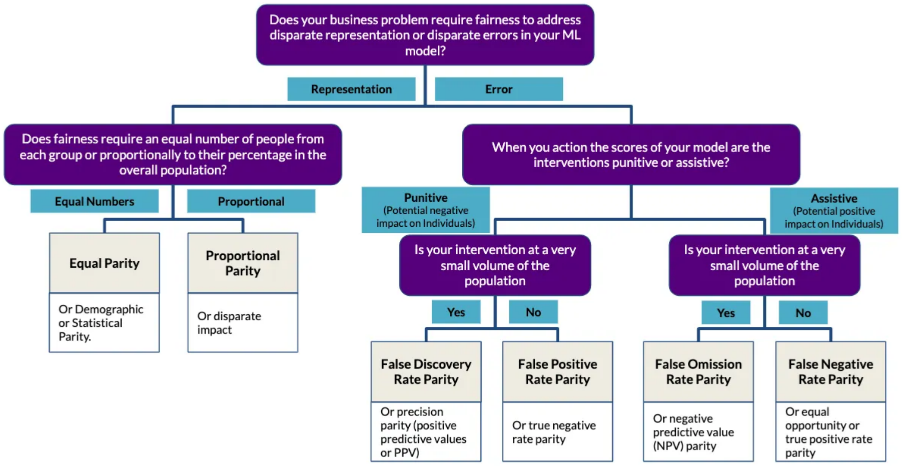
Select Protected Attributes

personal_status_sex × age ×

Not necessary to provide any protected attribute. Later, you can choose to rerun the tool.

Fig. 4. Mercury frontend interface

offering a combination of metric evaluation, visualization and recommendations, we aim to bridge the gap between complex algorithmic concepts and practical implementation, making fairness assessment accessible to both technical and non-technical users.



Does your business problem require fairness to address disparate representation or disparate errors in your ML model?
Representation
Error
Does fairness require an equal number of people from each group or proportionally to their percentage in the overall population?
When you action the scores of your model are the interventions punitive or assistive?
Equal Numbers
Proportional
Punitive (Potential negative impact on Individuals)
Assistive (Potential positive impact on Individuals)
Equal Parity
Or Demographic or Statistical Parity.
Proportional Parity
Or disparate impact
Is your intervention at a very small volume of the population
Is your intervention at a very small volume of the population
Yes
No
Yes
No
False Discovery Rate Parity
Or precision parity (positive predictive values or PPV)
False Positive Rate Parity
Or true negative rate parity
False Omission Rate Parity
Or negative predictive value (NPV) parity
False Negative Rate Parity
Or equal opportunity or true positive rate parity

Fig. 5. The Fairness Tree / Source: http://www.datasciencepublicpolicy.org/projects/aequitas

The key metrics we analyze include:

5.1 Statistical Parity / Demographic Parity

Also known as "group fairness," Demographic Parity (DP) aims to ensure that the likelihood of receiving a positive outcome from a model is the same across all groups, regardless of their protected attribute (e.g., gender, race, age). This means the model's decisions should not be influenced by sensitive characteristics, leading to equal representation in positive outcomes across different groups.

$$P(\hat{y} = 1 \mid p = 0) = P(\hat{y} = 1 \mid p = 1) \tag{1}$$

Where:

- $\hat{y}$ represents the predicted outcome.
- p represents the protected attribute.

In simpler terms, if a model predicts loan approvals, Demographic Parity would mean that the proportion of loan approvals should be roughly the same for men and women. A violation of DP suggests the model might be discriminating, even if unintentionally.

Our tool uses bar charts to illustrate this. Each bar represents a protected attribute and the width of the bar shows the proportion of DP difference from an optimal value of 0. If the bar is significantly wide, it is displayed with red color and indicates a potential DP violation (favors privileged). This straightforward visual representation helps users quickly identify whether positive outcomes are evenly distributed across groups.

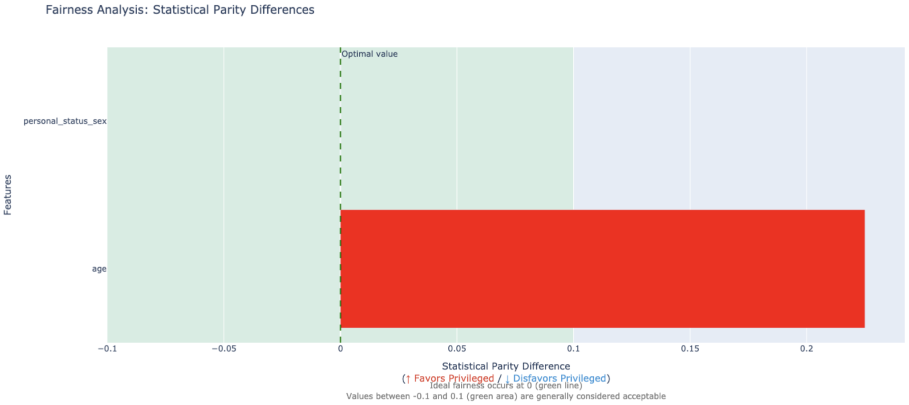


Fig. 6. Demographic / Statistical Parity chart

5.2 Equality of Odds and Equal Opportunity

Equality of Odds seeks to ensure that the model performs equally well for all groups, considering both the True Positive Rate (TPR) and the False Positive Rate (FPR). In essence, the model should have similar accuracy and error rates across different groups. This implies conditional independence of the predicted outcome and the protected attribute given the true outcome.

Regardless of whether they belong to a protected group, those who are eligible for a favorable outcome (such as loan approval, job hiring or college admission) are treated equitably thanks to the equal opportunity criterion. In particular, the TPR must be the same for all groups for Equal Opportunity to be met. This implies that the protected property shouldn't affect the likelihood of obtaining a good outcome among individuals who actually qualify for one (i.e., where y=1).

$$P(\hat{y} = 1 | p = 0, y = 1) = P(\hat{y} = 1 | p = 1, y = 1) \tag{2}$$

Equal Opportunity is just concerned with making sure that eligible people are not disadvantaged because they belong to a certain group, as opposed to Equality of Odds, which also takes FPR into account. This is especially crucial in situations where equity in favorable results is a top concern, such as recruiting, lending and medical diagnosis.

We visualize this using grouped bar charts with True/False Positive/Negative rates across all selected protected attributes. This allows users to quickly compare these values for each attribute and identify any significant disparities, which would indicate a potential violation of equal opportunity.

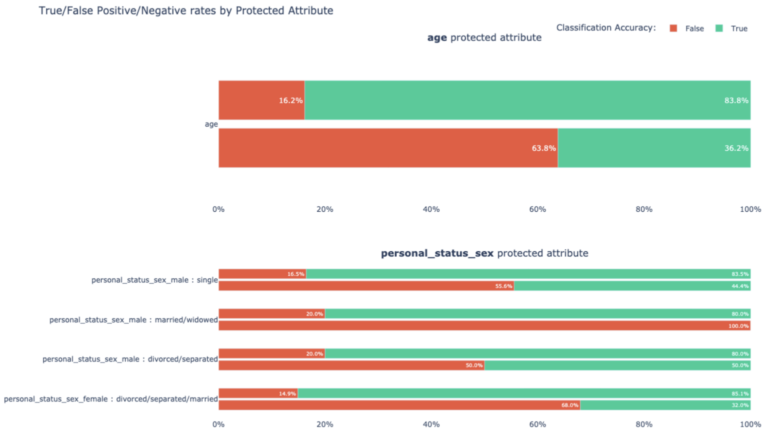


Fig. 7. True/False Positive/Negative rates by protected attribute(s)

5.3 MinDiff

The MinDiff regularization technique aims to minimize the difference in prediction scores between groups defined by a sensitive attribute. This approach directly addresses disparities by forcing the model to make more similar predictions for individuals who are similar in all respects except for their protected attributes.

$$\text{MinDiff} = \left| P(\hat{Y} = 1 \mid A = a) - P(\hat{Y} = 1 \mid A = b) \right| \tag{3}$$

Depending on the specific implementation and the nature of the data, MinDiff can be visualized in various ways. One approach is to show the distribution of prediction scores for different groups before and after applying MinDiff. This allows users to see how the technique reduces the gap in prediction scores between groups. In our current solution, we do not utilize with visualization the MinDiff method yet.

5.4 Correlations to the prediction outcome

Determining any biases and dependencies in the data requires an understanding of the relationship between features and the model's anticipated result. The degree to which various variables dramatically affect predictions—whether in a favorable or negative way—is shown by correlation analysis. While a near-zero correlation suggests little to no direct impact, a high correlation indicates a substantial association between the trait and the expected outcome.

Two correlation heatmaps are used to illustrate this. A comprehensive picture of the relationship between many attributes and the prediction result is provided by the first heatmap, which shows the correlation values for every attribute. The second heatmap highlights the particular impact of the chosen protected attribute on predictions, concentrating strictly on that.

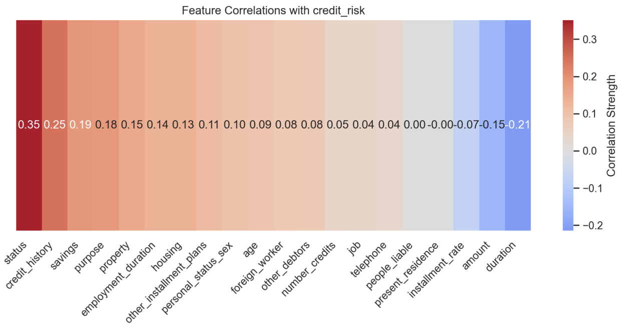


Fig. 8. Correlations of all attributes to predicted feature

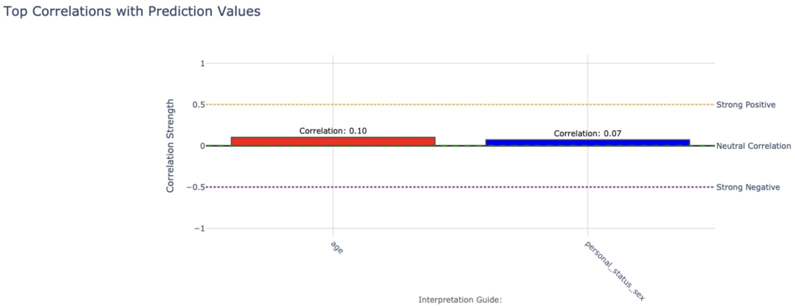


Fig. 9. Correlations of the chosen protected attribute(s) to predicted feature

A higher feature value is correlated with a higher number of positive predictions in these visualizations, where red denotes stronger correlations and deeper shades indicate a stronger positive link. On the other hand, blue denotes stronger negative correlations, where a lesser chance of a good prediction is linked to a higher feature value. With values close to zero, neutral correlations show that there is little to no linear relationship between the feature and the forecast result.

6 Experiments and Result

6.1 Improvement from State-of-the-Art

Our solution, FairnessLens distinguishes itself from existing state-of-the-art tools like AIF360 and Fairlearn – tools which, as the literature suggests, can present visualizations that are "too comprehensive" for non-technical users – in several key aspects.

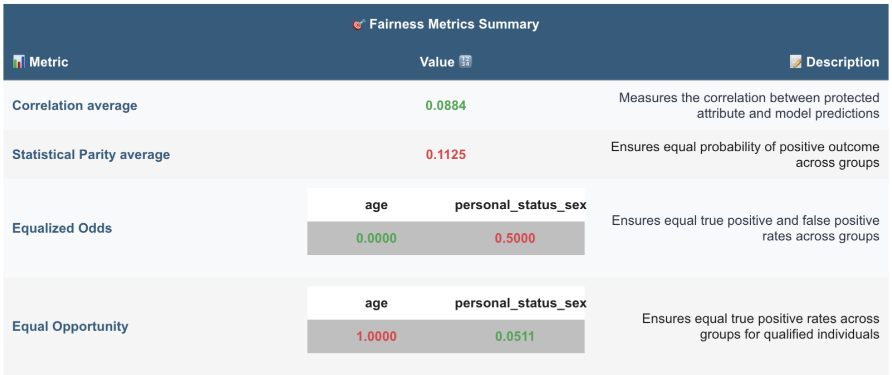
Fig. 10. FairnessLens logo

By leveraging Jupyter Notebook and Mercury, we provide a more intuitive and interactive user experience, particularly for users without extensive technical expertise. The simple dataset upload and (protected) attribute selection process lowers the barrier to entry for fairness assessment. This directly addresses the recurring issue of "the absence of tools designed to foster intuitive understanding rather than relying solely on mathematical representations," highlighted by

Chouldechova [15].

Our tool goes beyond simply reporting fairness metrics. We provide users with clear, actionable guidance on how to mitigate identified biases, including recommendations on appropriate pre-processing and post-processing techniques.

We integrate visualizations directly with metric calculations and model training results, enabling users to easily explore the impact of different fairness interventions. Furthermore, we aim to incorporate explainability techniques to provide insights into why certain biases exist and how they can be addressed. This addresses the need for "interpretability and explainability" which are increasingly becoming crucial elements of fairness research.



🎯 Fairness Metrics Summary

📊 Metric	Value 🔢	📝 Description
Correlation average	0.0884	Measures the correlation between protected attribute and model predictions
Statistical Parity average	0.1125	Ensures equal probability of positive outcome across groups
Equalized Odds	age | personal_status_sex 0.0000 | 0.5000	Ensures equal true positive and false positive rates across groups
Equal Opportunity	age | personal_status_sex 1.0000 | 0.0511	Ensures equal true positive rates across groups for qualified individuals

Fig. 11. Fairness Metrics Summary

We emphasize the importance of justifying fairness interventions by providing users with tools to document the rationale behind their decisions and the trade-offs involved. This supports transparency and accountability in the development of fair ML models.

7 Discussion

This scientific research aims to provide a practical and accessible solution to the identified challenges[16] in achieving fairness in machine learning models, moving beyond theoretical discussions to offer a tangible tool for practitioners. Our approach focuses on creating interactive visualizations and actionable guidance, addressing the gap between intricate technical principles and practical understanding.

7.1 Interpretation of Results

The results of our experiments, as detailed in Section 6, are expected to demonstrate the effectiveness of our fairness assessment tool in empowering users to identify and mitigate biases in their datasets and models. We anticipate that the accuracy of metric calculations will closely align with those of established fairness libraries like AIF360 and Fairlearn, validating the reliability of our tool's core computations. Furthermore, we expect the usability studies to indicate that our tool is user-friendly and accessible, even for individuals with limited technical expertise. User surveys and interviews will be used to evaluate the efficacy of our visualizations, and we hope to receive good feedback on the tool's capacity to communicate fairness and bias in an understandable and straightforward way.

7.2 Implications of Findings

Our findings have several important implications for the field of fairness in machine learning. First, they demonstrate the value of interactive visualizations and actionable guidance in making fairness assessment more accessible to a wider audience. By providing users with intuitive tools and clear recommendations and documentation, we empower them to take concrete steps to address biases and promote fairness in their models. Second, our research highlights the importance of considering trade-offs between fairness and other performance goals, such as accuracy. As discussed in Section 3, achieving fairness often involves making compromises and it is essential to carefully consider the implications of these trade-offs in the context of specific applications. Finally, our work contributes to the ongoing effort to develop more comprehensive and user-friendly tools for fairness assessment, addressing a key need identified in the state-of-the-art.

7.3 Limitations and Future Work

Despite the anticipated positive results, our research has several limitations that should be acknowledged. First, our evaluation is limited to the specific datasets and metrics used in our experiments. Future work should explore the tool's performance across a wider range of datasets and fairness metrics, including individual fairness measures and intersectional fairness concerns. Second, our tool currently focuses on a specific set of pre-processing and post-processing techniques. Future work could explore the integration of additional mitigation strategies, including causal fairness approaches and fairness-aware AutoML.

On top of that, our current work is focused on classification models. Future efforts should explore extending the tool to other model types, such as regression, to support a wider range of machine learning applications. Finally, our research has not yet addressed the challenge of justifying fairness interventions and documenting trade-offs. Future work should focus on developing more robust mechanisms for supporting this critical aspect of fairness assessment.

8 Conclusion and Summary

This research has presented a fresh approach to addressing the challenges of fairness in ML, focusing on the development and evaluation of a user-friendly and accessible fairness assessment tool. By leveraging interactive visualizations and actionable guidance, our work aimed to bridge the gap between intricate technical principles and practical understanding, enabling both technical and non-technical users to actively engage with and mitigate biases in their AI systems. This aligns directly with the urgent demand identified in the literature for better visualization methods – tools that are user-friendly and efficient in conveying insights linked to fairness. As noted in our state-of-the-art review, the field can close the gap between technical complexity and practical understanding by concentrating on developing more approachable, interactive and example-driven visualization techniques, guaranteeing that fairness becomes an understandable and achievable objective for all parties involved. Our work seeks to directly address this need.

8.1 Key Contributions

The key contributions of our research build upon existing tools and frameworks for fairness, addressing their identified limitations and going beyond heatmaps. Our work has produced concrete outcomes:

The development of FairnesLens provides an intuitive interface for uploading data, selecting attributes, calculating fairness metrics and visualizing results. As opposed to current solutions that may be too complicated for non-technical users, its interactive nature encourages users to explore different fairness interventions and understand their impact on model performance easily.

In the production of a *Comprehensive Evaluation Framework*, we designed and implemented a rigorous evaluation framework to assess the effectiveness of our tool across a diverse set of datasets and fairness metrics. Our results show that the tool can assist users in successfully spotting and lessening biases in their models, showing its usefulness in real-world circumstances.

With the focus on *Actionable Guidance and Recommendation*, our tool goes beyond just providing fairness indicators by giving users explicit, doable guidance on how to reduce identified biases and a way to justify fairness measures and record trade-offs. By addressing the necessity for understandable justifications and encouraging transparency in fairness interventions, this characteristic supports responsible AI development.

In the response on *Addressing a Critical Need*, our tool directly addresses the critical need for more interactive, intuitive, and instructional visualization techniques, as highlighted in the state-of-the-art. Unlike existing bias heatmaps and disparity matrices that require careful interpretation, our tool aims to provide visualizations that are easily understandable by both technical and non-technical users.

8.2 Future Improvements

While our research has made significant progress in addressing the challenges of fairness in machine learning, several opportunities exist for future improvements. These include:

As fairness research progresses, new and more nuanced metrics are being developed, future versions of our tool could incorporate a wider range of these, including individual fairness measures, intersectional fairness concerns and causal fairness measures, to provide a more comprehensive assessment of fairness.

Our tool currently focuses on a specific set of pre-processing and post-processing techniques. Future work could explore integrating additional mitigation strategies, including fairness-aware AutoML and reinforcement learning-based fairness interventions, to provide users with a broader range of options for mitigating bias.

While our tool provides actionable guidance for mitigating biases, it could be further enhanced by incorporating explainability techniques to help users understand why certain interventions work and how they affect different subgroups of the population. This could involve integrating SHAP values or other explainability methods into the visualizations to provide deeper insights into the model's behavior.

Finally, the true test of our tool's effectiveness will come from its deployment in real-world settings. Future work should focus on deploying the tool in various practical applications and conducting user studies to assess its impact on the development and deployment of fair AI systems.

8.3 Final Remarks

In conclusion, our research has presented a practical and accessible solution to the challenges of fairness assessment in machine learning. By providing users with a user-friendly tool, a comprehensive evaluation framework and actionable guidance, we hope to empower them to build more equitable and transparent AI systems. While much work remains, we believe our research represents a significant step forward in the pursuit of fairness in AI. As hybrid approaches that include pre- and post-processing techniques become more popular, our tool will be essential in offering actionable items and insights. It will empower everyone to follow the path and have more reliable and responsible results. Further investigation is still needed for improvements in the future, but the journey towards fairer and more ethical AI has taken a positive step.

References

1. Di Bello, M.: Algorithmic Fairness: Structural Perspectives. Available at https://www.marcellodibello.com/algorithmicfairness/handout/structural.html (2021).
2. Fraenkel, A.: Fairness & Algorithmic Decision Making. Self-Published, Available at https://afraenkel.github.io/fairness-book/content/05-parity-measures.html (2020).
3. Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., Stoyanovich, J.: The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice. Proceedings of the ACM Conference, Available at https://dl.acm.org/doi/abs/10.1145/3593013.3594007 (2023).
4. Nguyen, M.: Poster: The Impossible Theorem of Fairness. Available at https://digitalcommons.hamilton.edu/cgi/viewcontent.cgi?article=1006 (2022).
5. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent Trade-Offs in the Fair Determination of Risk Scores. ArXiv, Available at https://arxiv.org/abs/1609.05807 (2016). Accessed: 2025-02-24.
6. Yap, J.Q., Lim, E.: A legal framework for artificial intelligence fairness reporting. The Cambridge Law Journal, Available at https://www.cambridge.org/legal-framework-for-artificial-intelligence-fairness-reporting (2024).
7. Prince, S.: Tutorial #1: Bias and fairness in AI. Available at https://rbcborealis.com/research-blogs/tutorial1-bias-and-fairness-ai (2019). Accessed: 2025-02-24.
8. Voria, G., Sellitto, G., Ferrara, C., Abate, F., De Lucia, A., Ferrucci, F., Catolino, G., Palomba, F.: A Catalog of Fairness-Aware Practices in Machine Learning Engineering. ArXiv, Available at https://arxiv.org/abs/2408.16683 (2024). Accessed: 2025-02-24.
9. Hardt, M., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning. ArXiv, Available at https://arxiv.org/abs/1610.02413 (2016). Accessed: 2025-02-24.
10. Google LLC: Addressing Bias and Fairness Issues in ML Models. Available at https://colab.research.google.com/github/google/eng-edu/blob/main/ml/cc/exercises/fairness_income.ipynb (2024).
11. Precioso, F.: SI5 – IA-ID Advanced Deep Learning 2024-2025: Ethical Aspects of Data Fairness, Bias, Mitigation. (2024).
12. Bengio, Y.: International Scientific Report on the Safety of Advanced AI. Available at https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai (2024). Accessed: 2025-02-24.
13. Mbakwe, A.B., Lourentzou, I., Celi, L.A., Wu, J.T.: Fairness metrics for health AI: we have a long way to go. Available at https://pmc.ncbi.nlm.nih.gov/articles/PMC10114188 (2023). Accessed: 2025-02-24.
14. Zohar, Y.: Fairness Metrics in Machine Learning. Available at https://www.marcellodibello.com/algorithmicfairness/handout/impossibility.html (2023). Accessed: 2025-02-24.
15. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In: Proceedings of the 2017 ACM Conference on Fairness, Accountability, and Transparency, pp. 1–13 (2017). https://doi.org/10.48550/arXiv.1703.00056.
16. Caton, S., Haas, C.: Fairness in Machine Learning: A Survey. ACM Comput. Surv. 56, 7, Article 166 (July 2024), 38 pages. Available at https://doi.org/10.1145/3616865 (2024).

A Annex

A.1 Software Delivery Sheet

Signatures are redacted for privacy reasons. This is a publicly available research paper.

Software Delivery Sheet - PER2024-030

Title: Fairness in Machine Learning Models (Type: Research)

1. Identification

Students:

- Adam DIENES (EIT Digital M2) / adam.dienes@etu.univ-cotedazur.fr
- Bence Zsolt BEREGI (EIT Digital M2) / bence-zsolt.beregi@etu.univ-cotedazur.fr

Supervisors:

- Mireille Blay-Fornarino (I3S) / mireille.blay@univ-cotedazur.fr
- Anne-Marie Pinna-Dery (I3S) / anne-marie.pinna@univ-cotedazur.fr

Delivery date: 24 February 2025

Software name: **FairnessLens** / Version: 1.0 (beta)

2. Description of the software delivered

The **FairnessLens** is a Jupyter Notebook-based software (serving on Mercury) designed to empower users in the evaluation and mitigation of fairness concerns within datasets. This toolkit provides a user-friendly interface to:

- Upload and pre-process CSV datasets on an interactive and easy-to-understand web-based interface.
- Designate sensitive, prediction and other relevant attributes within the dataset.
- Automatically calculate key fairness metrics, including Demographic / Statistical Parity, Correlations to the prediction outcome, Equality of Odds (TPR, FPR) and Equal Opportunity (TPR).
- Visualize these metrics through interactive charts and plots, facilitating easy identification of potential biases.
- Access actionable recommendations for applying pre-processing and post-processing techniques to mitigate identified biases.
- Ability to export analysis results as a comprehensive fairness report for auditing and compliance purposes.

Documentation relating to this software includes: scientific research paper, A2 poster, software delivery sheet, which details the methodology, solution and evaluation of the toolkit and separate installation and execution instructions (see below).

3. Method of delivery

The **FairnessLens** is delivered via a public GitHub repository located at https://github.com/Bence749/FairnessLens

The repository contains:

- The Jupyter Notebook (.ipynb) file containing the **FairnessLens** software.
- A README.md file with detailed installation and execution instructions.
- A requirements.txt file listing the necessary Python packages and their versions.
- Scientific research paper in LaTex (.tex) and PDF format and A2 size poster (.pdf)
- Sample datasets for testing and demonstration purposes.

No password is required to access the repository. Simply clone the repository to your local machine using the following command in GitHub CLI: *gh repo clone Bence749/FairnessLens*

4. Intellectual property / Exploitation rights

The students acknowledge that the results of the research conducted within the framework of the PER as well as the software delivered resulting from this work, whether patentable or not, are subject to the rights of any third parties, the property of the supervisors who proposed the PER subject.

Consequently, the students undertake not to exploit for their own account or that of a third party, unless expressly agreed by the supervisors the results as defined above.

In return, the supervisors undertake to inform the students of the uses and exploitation of the results as defined above.

Date: Biot, 24 February 2025

Signatures

[redacted]

Adam DIENES
Student

[redacted]

Bence Zsolt BEREGI
Student

[redacted]

Mireille Blay-Fornarino
Supervisor

[redacted]

Anne-Marie Pinna-Dery
Supervisor

End of document