Journal of Machine Learning Research 25 (2024) 1-7 Submitted 5/24; Revised 9/24; Published 9/24

Aequitas Flow: Streamlining Fair ML Experimentation

Sérgio Jesus[1,2] SERGIO.JESUS@FEEDZAI.COM
Pedro Saleiro[1] PEDRO.SALEIRO@FEEDZAI.COM
Inês Oliveira e Silva[1] INES.SILVA@FEEDZAI.COM
Beatriz M. Jorge[1] BEATRIZ.JORGE@FEEDZAI.COM
Rita P. Ribeiro[2] RPRIBEIRO@FC.UP.PT
João Gama[2] JGAMA@FEP.UP.PT
Pedro Bizarro[1] PEDRO.BIZARRO@FEEDZAI.COM
Rayid Ghani[3] RAYID@CMU.EDU
[1] *Feedzai*
[2] *University of Porto*
[3] *Carnegie Mellon University*

Editor: Sebastian Schelter

arXiv:2405.05809v2 [cs.LG] 30 Oct 2024

Abstract

Aequitas Flow is an open-source framework and toolkit for end-to-end Fair Machine Learning (ML) experimentation, and benchmarking in Python. This package fills integration gaps that exist in other fair ML packages. In addition to the existing audit capabilities in Aequitas, the Aequitas Flow module provides a pipeline for fairness-aware model training, hyperparameter optimization, and evaluation, enabling easy-to-use and rapid experiments and analysis of results. Aimed at ML practitioners and researchers, the framework offers implementations of methods, datasets, metrics, and standard interfaces for these components to improve extensibility. By facilitating the development of fair ML practices, Aequitas Flow hopes to enhance the incorporation of fairness concepts in AI systems making AI systems more robust and fair.

Keywords: Fair machine learning, experimentation, ethical artificial intelligence, open-source framework, python

1 Introduction

Developing Machine Learning (ML) and Artificial Intelligence (AI) systems that result in fairness and equity is a critical topic, especially as such systems get used in high-stakes settings such as hiring (Dastin, 2018), healthcare (Igoe, 2021), criminal justice (Angwin et al., 2016; Chouldechova, 2017), and financial services (Zhang and Zhou, 2019; Bartlett et al., 2019; Jesus et al., 2022). While numerous studies define metrics and properties of algorithmic fairness (Chouldechova, 2017; Calders and Verwer, 2010; Dwork et al., 2012; Feldman et al., 2015; Hardt et al., 2016; Corbett-Davies et al., 2017) and propose methods for fairer models (Fish et al., 2016; Calmon et al., 2017; Zafar et al., 2017; Cotter et al., 2019), gaps in the implementation, user experience. and integration of existing tools hinder end-to-end experimentation (Lee and Singh, 2021) and benchmarking. This makes empirical studies and practical use challenging, scarce, and often limited in scope (Friedler et al., 2019; Lamba et al., 2021), ultimately affecting the adoption of fair ML methods in real-world high-stakes settings.

This paper introduces Aequitas Flow, an open-source framework for reproducible and extensible end-to-end fair ML experimentation that extends Aequitas, our original bias audit toolkit. The goal is to help 1)researchers compare and benchmark new methods they develop against existing methods in a systematic and reproducible manner and 2) practitioners easily evaluate existing bias mitigation methods and deploy ones that best match their goals.

©2024 Sérgio Jesus, Pedro Saleiro, Inês Oliveira e Silva, Beatriz M. Jorge, Rita P. Ribeiro, João Gama, Pedro Bizarro, Rayid Ghani.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v25/24-0677.html.

Table 1: Comparison of packages for training and evaluation of fair ML Methods.

	Packages			
Functionalities	AIF360	Fairlearn	Aequitas	Aequitas Flow
Group fairness metrics	◐	◐	●	●
Pre-processing methods	●	◐	-	●
In-processing methods	◐	●	-	●
Post-processing methods	●	●	-	●
Standardized interfaces for extensibility	◐	◐	-	●
Hyperparameter optimization pipeline	-	-	-	●
Binary classification	●	●	●	●
Regression	●	●	-	-
Model selection	-	-	◐	●
Methods comparison	-	-	-	●
Plotting methods	-	◐	●	●

● exists in package; ◐ partially exists in package; - does not exist in package.

Table 1 compares Aequitas Flow, the latest release of the Aequitas package [1] (Saleiro et al., 2018) to other fair ML packages to highlight some of the key gaps we aimed to fill with this paper.

Fairlearn (Weerts et al., 2023), and AIF360 (Bellamy et al., 2018) are popular fair ML packages to facilitate adoption by offering methods (Feldman et al., 2015; Hardt et al., 2016; Agarwal et al., 2018), fairness metrics (Hardt et al., 2016) and datasets available in the literature (Kohavi, 1996; Angwin et al., 2016; Dua and Graff, 2017; Ding et al., 2021). However, some issues hinder their usability as standard toolkits for fairness studies. First, both lack a defined experimentation pipeline, requiring users to opt for external packages for fundamental tasks, such as dataset splitting and hyperparameter optimization (Schelter et al., 2019). Second, inconsistency in class behavior and implementation force users to customize the code depending on the methods used. For instance, in AIF360's `DisparateImpactRemover` class, most of the parent class methods are not implemented. These issues create a high barrier for users to effectively use the packages. Our work tackles the lack of standardized tools for experimentation with fair ML, with an emphasis on the extensibility of methods, datasets, and metrics, the reproducibility of the experiments, as well as different levels of customization for different user needs.

2 Aequitas Flow

The Aequitas Flow package is a comprehensive framework that integrates the necessary elements for a complete fair ML experiment. These are methods, datasets, and optimization strategies. They can be accessed through a standardized pipeline defined by configuration files, Python dictionaries or instantiated independently. This provides a standardized platform for experimental fairness testing. Figure 1 represents the pipeline's structure, mapping the components and interactions and offering an overview of the fairness experimentation process.

Experiment: The `Experiment` is the main component that orchestrates the workflow within the package. It processes input configurations, which can be provided as either files[2] or Python dictionaries. These specify the methods, datasets, and optimization parameters. The Experiment component initializes and populates the necessary classes, ensuring they interact deterministically throughout the execution process. When an experiment is completed, the results can be analyzed directly within the class with the appropriate methods. A variant of this component allows for

1. `https://github.com/dssg/aequitas`
2. Examples provided in the repository.

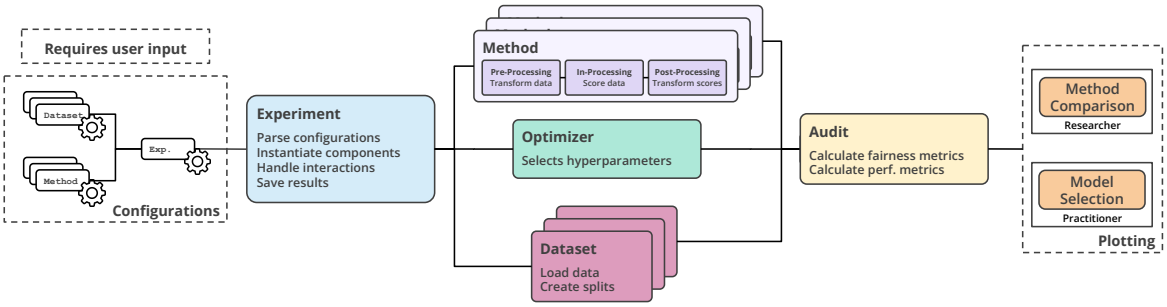


Figure 1: **Diagram of an Experiment in Aequitas Flow.** The user input is passed to the Experiment, which will instantiate the components (Methods, Datasets, and Optimizer) in the pipeline. For each target task (for a researcher or practitioner), different plotting methods can be used to analyze the experimental results.

simplified usage as it only requires the definition of a dataset. This feature is designed to streamline initial experiments and reduce configurations.

```
exp = Experiment(config_file="configs/experiment.yaml")
exp.run()
```

Optimizer: The `Optimizer` component manages hyperparameter selection and model evaluation. It receives the hyperparameter search space of the methods and a split dataset to conduct hyperparameter tuning. It evaluates the performance of models, and stores the resulting artifacts. The component uses Optuna (Akiba et al., 2019) for hyperparameter selection and the bias auditing functionality of Aequitas (Saleiro et al., 2018) for fairness and performance evaluation. This component should only be instantiated by an `Experiment`, to guarantee consistency in input arguments. Several attributes of the hyperparameter optimization can be determined by configurations, such as the number of trials and jobs, the selection algorithm (*e.g.*, random search, grid search), and the random seed.

Datasets: This component has two primary functions: loading the data and generating splits. It maintains information about the prediction target, typing, and sensitive features. The data is stored in a pandas dataframe format (pandas development team, 2020). The framework initially encompasses eleven tabular datasets, including those from the BankAccountFraud (Jesus et al., 2022) and Folktables (Ding et al., 2021). The component also permits user-supplied datasets in CSV or parquet formats with splits based on a column, or randomly.

```
dataset = datasets.FolkTables(variant='ACSIncome')
dataset.load_data()
dataset.create_splits()
dataset.train.X  # return the train feature matrix
```

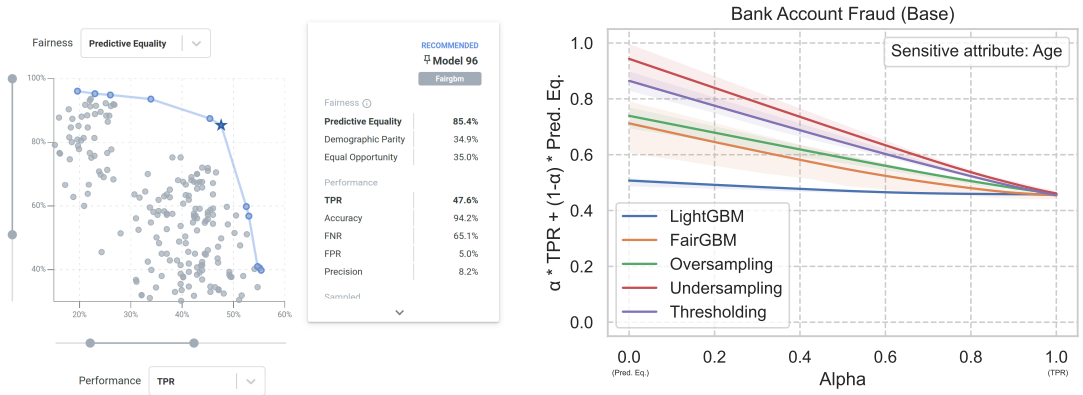
Methods: This group of components handles data processing and creates and adjusts predictions for validation and test sets. Aequitas Flow provides interfaces for the three recognized types of fair ML methods (Caton and Haas, 2023; Mehrabi et al., 2021; Pessach and Shmueli, 2022): pre-processing, in-processing, and post-processing. Pre-processing methods modify the input data, in-processing methods typically directly modify the objective function and generate prediction scores, and post-processing methods adjust these scores or rankings. Additionally, ML classification methods are included in the category of base estimators and function similarly to in-processing methods. The methods adhere to a standardized interface to facilitate calls within the experiment class. In the current version of Aequitas, 15 methods are supported.

```
model = methods.inprocessing.FairGBM()
model.fit(train.X, train.y, train.s)
preds = model.predict_proba(val.X, val.s)
```

Audit: The Aequitas toolkit offers a suite of metrics based on the confusion matrix for the protected groups in the dataset. Users may specify a group as a reference for comparison and select the appropriate fairness metric for their analysis. Experiments leverage the `Audit` class to calculate metrics and disparities when analyzing the produced prediction scores of a model.

```
audit_df = pd.DataFrame({"score": preds, "label": val.y, "group": val.s})
audit = Audit(audit_df)
audit.performance()  # Obtain performance metrics
audit.audit()  # Obtain fairness metrics
```

Plotting: Aequitas Flow provides two workflows based on the goal of the user. The first is around model selection (a), where users can plot the trained models with the desired metrics of fairness and performance in each axis. The Pareto frontier is displayed, with the model with the best fairness-performance trade-off highlighted. The second provides a comparison of methods (b). Confidence intervals for the combined performance and fairness are calculated for each tested method in the trade-offs of these metrics. Additional plotting methods are available for in-depth bias auditing. Figure 2 shows examples of both.



(a) (b)

Figure 2: Plots introduced in Aequitas Flow. Plot (a) is designed for model selection; Plot (b) compares the different tested methods.

3 Conclusion

Aequitas Flow is an open-source framework that makes end-to-end experimentation with fair ML easier through the use of customizable components, namely datasets, methods, metrics, and optimization algorithms. It enhances robustness and reproducibility by addressing the issues of ad-hoc and single-use setups in fair ML experimentation. This can lead to better benchmarking and adoption of fair ML techniques in real world settings. While initially focused on tabular datasets, the framework's flexible interfaces allow adaptation to other data formats, and ongoing updates will incorporate additional implementations, in a welcoming environment to community contributions. Recognizing the challenges associated with responsibly using this framework in real-world applications, we aim to support the widespread adoption of fair ML methodologies and increase their societal impact.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330701. URL https://doi.org/10.1145/3292500.3330701.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2016.

Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. Consumer-Lending Discrimination in the FinTech Era. Technical report, National Bureau of Economic Research, 2019.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, Sep 2010. ISSN 1573-756X. doi: 10.1007/s10618-010-0190-x. URL https://doi.org/10.1007/s10618-010-0190-x.

Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3995–4004, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, aug 2023. ISSN 0360-0300. doi: 10.1145/3616865. URL https://doi.org/10.1145/3616865.

Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, jun 2017. ISSN 2167-6461. doi: 10.1089/big.2016.0047. URL http://www.liebertpub.com/doi/10.1089/big.2016.0047.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098095. URL https://doi.org/10.1145/3097983.3098095.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. Two-player games for efficient non-convex constrained optimization. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 300–332. PMLR, 22–24 Mar 2019. URL https://proceedings.mlr.press/v98/cotter19a.html.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2018.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proc. of the 3rd Innovations in Theoretical Computer Science Conf. on - ITCS '12*, pages 214–226, New York, USA, 2012. ACM Press. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL http://arxiv.org/abs/1104.3913.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL https://doi.org/10.1145/2783258.2783311.

Benjamin Fish, Jeremy Kun, and Ádám Dániel Lelkes. A confidence-based approach for balancing fairness and accuracy. In Sanjay Chawla Venkatasubramanian and Wagner Meira Jr., editors, *Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016*, pages 144–152. SIAM, 2016. doi: 10.1137/1.9781611974348.17. URL https://doi.org/10.1137/1.9781611974348.17.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287589. URL https://doi.org/10.1145/3287560.3287589.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

K Igoe. Algorithmic bias in health care exacerbates social inequities–how to prevent it. *Executive and Continuing Professional Education*, 2021.

Sérgio Jesus, José Pombal, Duarte Alves, André Ferreira Cruz, Pedro Saleiro, Rita P. Ribeiro, João Gama, and Pedro Bizarro. Turning the tables: Biased, imbalanced, dynamic tabular datasets for ML evaluation. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/d9696563856bd350e4e7ac5e5812f23c-Abstract-Datasets_and_Benchmarks.html.

Ron 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*, KDD'96, page 202–207. AAAI Press, 1996.

Hemank Lamba, Kit Rodolfa, and Rayid Ghani. An empirical comparison of bias reduction methods on real-world problems in high-stakes policy settings. *ACM SIGKDD Explorations Newsletter*, 23: 69–85, 05 2021. doi: 10.1145/3468507.3468518.

Michelle Seng Ah Lee and Jat Singh. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445261. URL https://doi.org/10.1145/3411764.3445261.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55 (3), feb 2022. ISSN 0360-0300. doi: 10.1145/3494672. URL https://doi.org/10.1145/3494672.

Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

Sebastian Schelter, Yuxuan He, Jatin Khilnani, and Julia Stoyanovich. Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. *ArXiv*, abs/1911.12587, 2019. URL https://api.semanticscholar.org/CorpusID:208512964.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems, 2023. URL http://jmlr.org/papers/v24/23-0389.html.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/zafar17a.html.

Yukun Zhang and Longsheng Zhou. Fairness assessment for artificial intelligence in financial industry. *arXiv preprint arXiv:1912.07211*, 2019.