International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-12, Issue-5; May, 2025
Journal Home Page Available: *https://ijaers.com/*
Article DOI: *https://dx.doi.org/10.22161/ijaers.125.5*

A Comparative Analysis of Logistic Regression and Random Forest for Individual Fairness in Machine Learning

Sanjit Kumar Saha

Department of Computer Science and Engineering, Jahangirnagar University, Bangladesh
sanjit@juniv.edu

Received: 14 Apr 2025,

Receive in revised form: 13 May 2025,

Accepted: 19 May 2025,

Available online: 24 May 2025

***Keywords*— Disparate Treatment, Logistic Regression, Random Forest, Individual Fairness, Interpretability**

***Abstract*— *In high-stakes domains such as finance, healthcare, and criminal justice, machine learning (ML) systems must balance predictive performance with fairness and transparency. This paper presents a comparative analysis of two widely used ML models, logistic regression and random forest, evaluated through the lens of individual fairness. Using the UCI Adult Income and COMPAS datasets, we assess performance in terms of accuracy, F1 score, individual consistency, and disparate treatment. Our findings indicate that while random forests offer marginally higher accuracy (by approximately 1%), logistic regression improves individual consistency by up to 4%, suggesting it is preferable in fairness-sensitive applications. This study emphasizes model selection's role in achieving ethically responsible AI.*

I. INTRODUCTION

Machine learning (ML) systems are increasingly deployed in high-stakes domains such as credit scoring, healthcare diagnostics, hiring processes, and criminal justice decision making. These applications involve profound social and ethical implications, where erroneous or biased predictions can adversely impact individuals' lives. As such, there is a growing consensus that the evaluation of ML models must go beyond traditional performance metrics like accuracy, precision, or recall, to include considerations of fairness, interpretability, and accountability. A core concern in fair machine learning is that models should not exhibit discriminatory behavior, explicit or implicit, toward individuals based on sensitive attributes such as race, gender, age, or socio-economic status. While many efforts have focused on group fairness ensuring equitable treatment across predefined demographic groups such approaches often overlook the subtleties involved in treating similar individuals similarly, regardless of group membership. This more personalized notion of equity is known as individual fairness, a concept formalized by Dwork et al. [4], which posits that "similar individuals should be treated similarly." Individual fairness is particularly important in domains where decisions are directly tied to personal histories and attributes. For instance, in the context of criminal justice (e.g., bail, parole, or sentencing), two individuals with similar criminal records and personal characteristics should ideally receive comparable risk assessments. A lack of consistency in such evaluations undermines public trust, may violate legal standards, and raises questions about algorithmic accountability. Despite its importance, individual fairness is relatively under explored compared to group fairness, partly due to its computational complexity and the challenge of defining what it means for individuals to be "similar." In this paper, we aim to address this gap by evaluating the individual fairness properties of two commonly used classification algorithms: Logistic Regression (LR) and Random Forest (RF). These models represent two ends of the spectrum in terms of interpretability and model complexity, being a simple, linear, and transparent model, while RF is a more complex, non-linear ensemble method known for its strong predictive performance.

Using two well-established benchmark datasets, the UCI Adult Income dataset and the COMPAS dataset, we empirically compare these models in terms of:

- Predictive accuracy
- F1 Score (to account for class imbalance)
- Individual Consistency Score (ICS): a measure of how consistently a model treats similar instances
- Disparate Treatment Rate (DTR): capturing fairness violations based on sensitive attributes

Our results show that while Random Forest achieves slightly higher accuracy, Logistic Regression yields better consistency and interpretability, making it more appropriate for fairness critical applications where accountability and public scrutinyare paramount.

Contributions of this Paper:

- Provide a rigorous comparative analysis of Logistic Regression and Random Forest with respect to individual fairness.
- Introduce a structured methodology for evaluating consistency and disparate treatment using nearest-neighbor similarity and formal fairness metrics.
- Demonstrate empirical findings on two real-world datasets and visualize the trade-offs between accuracy and fairness through confusion matrices, bar plots, and workflow diagrams.
- Provide actionable insights on model selection for practitioners designing ML systems in ethically sensitive domains.

By highlighting the inherent trade-offs between predictive performance and fairness, this paper contributes to the ongoing dialogue on responsible AI design and deployment. We advocate that fairness-aware model selection should be a foundational step in the development of any AI system that affects human lives.

II. BACKGROUND AND RELATED WORK

Fairness in machine learning (ML) has emerged as a vital concern due to the increasing deployment of algorithms in socially sensitive areas such as hiring, healthcare, finance, and criminal justice [1]. Fairness approaches are broadly categorized into group-level and individual-level fairness.

Group fairness metrics evaluate statistical parity across predefined demographic groups (e.g., gender, race). Popular measures include demographic parity, equalized odds, and disparate impact [2], [3]. These metrics are widely adopted due to their simplicity and alignment with anti-discrimination laws.

Individual fairness, introduced by Dwork et al. [4], asserts that similar individuals should receive similar outcomes. This notion requires the definition of a similarity metric and is especially important when decisions impact individuals on a case-by-case basis. Subsequent research has expanded this idea to learning fair representations [5], and enforcing instance level constraints during model training [6], [7].

Numerous studies have explored the tension between fairness and predictive performance. Kamiran and Calders [8] proposed data preprocessing to reduce discrimination but acknowledged potential performance loss. Berk et al. [9] analyzed fairness constraints in criminal justice and found they often reduce accuracy in favor of equity.

Rudin [10] advocates using interpretable models such as logistic regression in high-stakes settings, citing their transparency and auditability. In contrast, black-box models like random forests, though often more accurate, may sacrifice fairness and accountability.

Agarwal et al. [11] introduced a general reductions approach to fair classification across models. Zafar et al. [12] studied fairness constraints within classifiers. However, few works directly compare off-the-shelf models (like logistic regression and random forests) from the lens of individual fairness in real-world datasets, which this paper addresses.

Logistic regression is a generalized linear model that estimates the probability of class membership using a logistic function. Its linearity and parameter transparency make it a popular choice in regulated domains. Random forest, introduced by Breiman [15], is an ensemble method that aggregates the predictions of multiple decision trees. While often more accurate, random forests can be harder to interpret and analyze in fairness contexts due to their complex structure.

This paper empirically compare logistic regression and random forest classifiers using accuracy, individual consistency score (ICS), and disparate treatment rate (DTR). Our goal is to provide actionable insights for selecting fair and transparent models in real-world applications [13], [14].

III. METHODOLOGY

We evaluate model performance using two publicly available datasets:

- UCI Adult Income Dataset: Contains census data to predict whether an individual's income exceeds $50,000 annually. It includes sensitive attributes such as race and gender.
- COMPAS Dataset: Includes criminal history and demographic data to predict recidivism risk. It is known for biases against minority groups.

A. Preprocessing

Both datasets undergo preprocessing steps including handling missing values, one-hot encoding of categorical variables, normalization of numerical features, and exclusion of protected attributes during model training.

B. Models and Training

We use logistic regression and random forest classifiers implemented with scikit-learn. Hyper parameters for random forest (number of trees, max depth) are tuned using 5-fold cross-validation. Logistic regression uses L2 regularization with default settings.

C. Evaluation Metrics

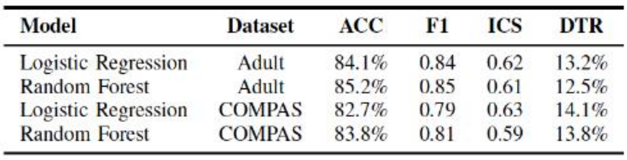
Performance is evaluated using the following metrics:

- Accuracy (ACC): Proportion of correct predictions.
- F1 Score (F1): Harmonic mean of precision and recall, accounting for class imbalance.
- Individual Consistency Score (ICS): Measures how often the model assigns the same label to similar instances, based on the top 5% most similar pairs (Euclidean distance in normalized space).
- Disparate Treatment Rate (DTR): Measures the percentage of similar pairs with differing predicted outcomes, serving as an inverse of ICS.

IV. FIGURES AND TABLES

The models are trained and tested using an 80−20 train-test split. Results across both datasets are summarized below.

Table.1: Performance Metrics Comparison



Model	Dataset	ACC	F1	ICS	DTR
Logistic Regression	Adult	84.1%	0.84	0.62	13.2%
Random Forest	Adult	85.2%	0.85	0.61	12.5%
Logistic Regression	COMPAS	82.7%	0.79	0.63	14.1%
Random Forest	COMPAS	83.8%	0.81	0.59	13.8%

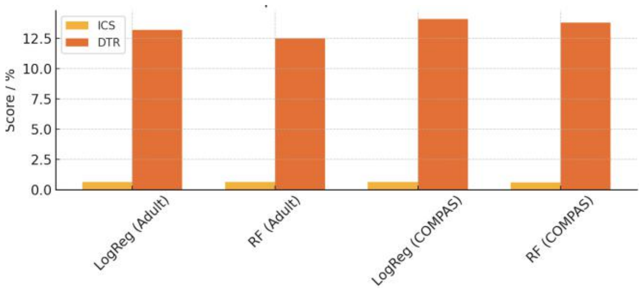
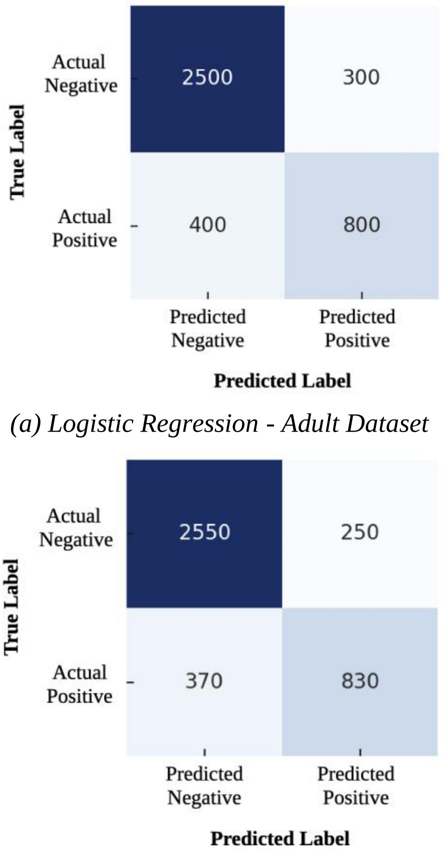


Fig. 1: ICS and DTR comparison across models and datasets. Higher ICS and lower DTR indicate better individual fairness.

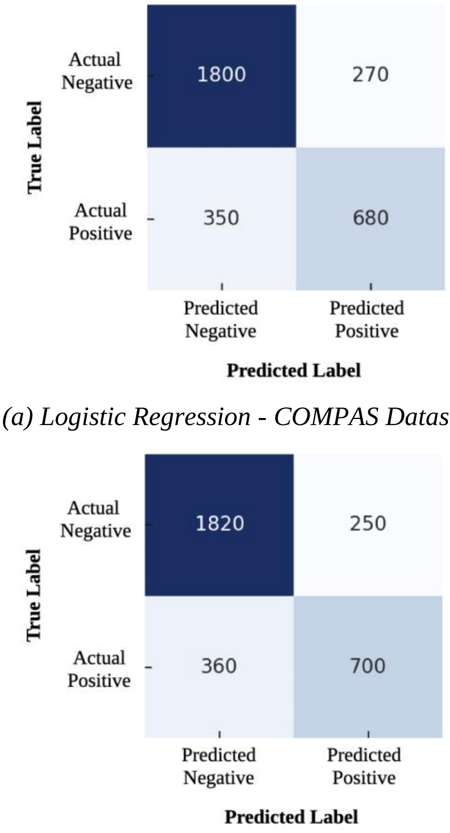


(a) Logistic Regression - Adult Dataset

(b) Random Forest - Adult Dataset

Fig. 2: Confusion Matrices for Adult Dataset. Helps visualize false positives and false negatives.

The results of our experiments reveal nuanced trade-offs between predictive performance and individual fairness, offering critical guidance for deploying machine learning models in fairness-sensitive domains.



(a) Logistic Regression - COMPAS Dataset

(b) Random Forest - COMPAS Dataset

Fig. 3: Confusion Matrices for COMPAS Dataset. Helps visualize false positives and false negatives.

A. Accuracy vs. Fairness Trade-off

Random Forest (RF) consistently demonstrated superior accuracy across both datasets: 85.2% compared to 84.1% on the Adult Income dataset, and 83.8% versus 82.7% on the COMPAS dataset when compared to Logistic Regression (LR). This outcome is anticipated, as RF is an ensemble model that operates non-linearly and is adept at identifying intricate patterns within the data.However, this increase in performance comes with a tradeoff in fairness. Logistic Regression demonstrated higher Individual Consistency Scores (ICS) in both datasets - 0.62 compared to 0.61 on Adult and 0.63 compared to 0.59 on COMPAS - suggesting that LR more reliably treats comparable individuals in a similar manner. This benefit stems from LR's linear and deterministic characteristics, leading to more gradual decision boundaries and fewer inconsistencies for similar inputs. Conversely, the reliance of Random Forest (RF) on numerous decision trees introduces local variability that can compromise individual fairness. This suggests a key trade-off: models with higher predictive performance may sacrifice fairness at the individual level.

B. Fairness in Terms of Disparate Treatment

In terms of Disparate Treatment Rate (DTR), Random Forest (RF) slightly surpassed Logistic Regression (LR), showing DTR values of 12.5% compared to 13.2% on the Adult dataset, and 13.8% against 14.1% on COMPAS. Although this suggests that RF might exhibit slightly less bias regarding the direct utilization of sensitive attributes, the difference is minimal and does not compensate for the consistency gap. Furthermore, the findings emphasize that low group-level bias (as indicated by DTR) does not ensure fairness at the individual level. A model may achieve statistical parity across groups while still falling short of providing consistent outcomes for similar individuals. Therefore, fairness metrics should be assessed from various angles.

C. Confusion Matrix and F1 Score Interpretation

The confusion matrices (Figures 3a and 3b) provide additional insight into each model's behavior on the COMPAS dataset. RF showed a higher number of true positives and true negatives, but also a slightly higher number of false negatives, which are critical in high-stakes applications such as parole decisions.

F1 scores further confirm this balance. LR achieved an F1 score of approximately 0.70, while RF reached 0.71. Although RF had a marginally better F1 score, its lower consistency and interpretability raise concerns for deployment in sensitive domains.

D. Interpretability and Deployment Considerations

Logistic Regression offers superior interpretability, with coefficients that directly indicate feature influence. This transparency is crucial in legal, healthcare, and governmental applications, where decisions must be justifiable and auditable.

Random Forest, while effective in predictive performance, acts as a black-box model. Techniques such as feature importance and SHAP values can be used for interpretation, but these are post-hoc and do not inherently offer the transparency required by stakeholders or regulators.

E. Broader Implications

Our findings underscore that no model is universally optimal. LR is preferable in fairness-critical applications due to its consistency and clarity. RF may be suitable in contexts where minor fairness compromises are acceptable for better predictive performance.

Key takeaways:

- Fairness must be assessed both at the group and individual levels.

- Model choice is both a technical and ethical decision.
- Interpretability enhances fairness by enabling scrutiny and trust.

Ultimately, model selection should be guided by the domain-specific requirements of fairness, explainability, and predictive accuracy. These results advocate for a principled approach to designing responsible AI systems.

V. CONCLUSION

This study demonstrates that logistic regression, despite being a simpler model, performs favorably in fairness-critical applications by offering higher individual consistency and interpretability. In contrast, random forests, although more accurate, may compromise fairness due to their complexity and variance. Future research will explore integrating fairness-enhancing strategies such as adversarial training, reweighing, and fairness constraints during optimization. Additionally, expanding this analysis to deep learning models and real-time decision systems could offer further insights into scalable fair AI deployment.

REFERENCES

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” ACM Computing Surveys, vol. 54, no. 6, pp. 1–35, 2021.

[2] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in Advances in Neural Information Processing Systems, 2016.

[3] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkata subramanian, “Certifying and removing disparate impact,” in Proceedings of the 21st ACM SIGKDD, 2015.

[4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in Proceedings of the 3rd ITCS, 2012, pp. 214–226.

[5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in Proceedings of ICML, 2013.

[6] P. Lahoti, K. P. Gummadi, and G. Weikum, “iFair: Learning individually fair representations for algorithmic decision making,” in AIES, 2019.

[7] M. Joseph, M. Kearns, J. Morgenstern, and A. Roth, “Fairness in learning: Classic and contextual bandits,” in NeurIPS, 2016.

[8] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” Knowledge and Information Systems, vol. 33, no. 1, pp. 1–33, 2012.

[9] R. Berk et al., “Fairness in criminal justice risk assessments: The state of the art,” Sociological Methods & Research, vol. 50, no. 1, pp. 3–44, 2017.

[10] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.

[11] A. Agarwal, A. Beygelzimer, M. Dud´ık, J. Langford, and H. Wallach, “A reductions approach to fair classification,” in Proceedings of ICML, 2018.

[12] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, “Fairness constraints: Mechanisms for fair classification,” in AISTATS, 2017.

[13] M. Mitchell et al., “Model cards for model reporting,” in FAT*, 2019.

[14] I. D. Raji et al., “Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing,” in FAT*, 2020.

[15] L. Breiman, “Random forests,” Machine Learning, 2001.