# Investigating Bias in a Recruitment ML Model

Robinson Victor
School of Computing, Engineering and Digital Technology
Teesside University
United Kingdom
D3739635@live.tees.ac.uk

***Abstract*—This report explores the application of appropriate fairness criteria to investigate bias in a machine learning model. Bias can stem from various sources, like limited raw datasets, imputations, outliers, recording errors, etc. Investigating bias in ML models involves identifying the protected feature, splitting the outcomes into groups using the feature, computing performance metrics for the groups from each confusion matrix, and applying relevant fairness criteria such as equal accuracy, group fairness (demographic or statistical parity), and equality of opportunity.**

**Keywords—recruitment, bias, fairness criteria, decision tree, AI**

## I. Introduction

Recruitment involves attracting, screening, interviewing, and selecting candidates for roles within organizations, thereby connecting them with potential employees. AI-based recruitment employs AI/ML-powered technology to automate this process.

Reference [1] suggests that AI/ML-based recruitment offers numerous benefits for all stakeholders. These include mitigating implicit or unconscious human bias in recruitment by ensuring all applicants are assessed based on the same criteria and to the same standard, increasing the likelihood of hiring quality candidates, and improving efficiency by handling large-scale applications and conducting thorough assessments quickly. This reduces reliance on manual screening and the need for additional human recruiters, ultimately saving time and costs.

However, [17] and [3] highlight that despite these positive impacts, AI-based recruitment raises ethical and legal concerns. These include the potential for algorithmic bias, which can lead to discrimination in the hiring process based on various protected characteristics like gender stemming from limited raw data sets and biased algorithm designs. This can lead to a growing number of issues, as it can inadvertently filter out good candidates, and companies may be forced to pay huge sums in compensation and legal costs in cases of unintended regulatory breaches. The growing number of rejections can lead to a decline in the mental health of candidates.

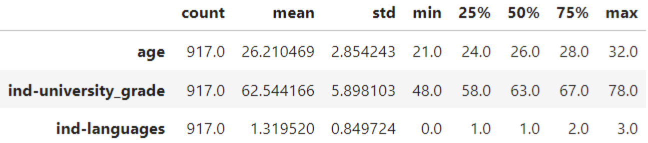
This report aims to investigate bias in a binary classification ML model that predicts the outcome of candidates' applications based on details from their CVs. The model, utilizing the decision tree algorithm, is trained on a dataset sourced from Kaggle. The dataset consists of several protected characteristics. Gender has been chosen as the protected feature for this investigation.

## II. Machine Learning Pipeline

This section highlights all the necessary steps taken to understand, clean, prepare and split the dataset, as well as feature scaling, developing the model using the decision tree algorithm, computing various performance metrics to evaluate the model's performance on different sets, splitting the model based on the protected feature, and applying fairness criteria.

### A. Structure Investigation

Exploring the general shape and structure of the dataset, as well as the data types of the features, reveals that the dataset contains a total of 917 samples and 14 features of which 3 are numerical features, and 11 are non-numerical features. The *describe* method was useful in displaying the statistical summaries of both categories of features in the dataset.



| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| age | 917.0 | 26.210469 | 2.854243 | 21.0 | 24.0 | 26.0 | 28.0 | 32.0 |
| ind-university_grade | 917.0 | 62.544166 | 5.898103 | 48.0 | 58.0 | 63.0 | 67.0 | 78.0 |
| ind-languages | 917.0 | 1.319520 | 0.849724 | 0.0 | 1.0 | 1.0 | 2.0 | 3.0 |

Fig. 1. Statistical summary of numerical features

### B. Quality Investigation and Data Preprocessing

By investigating the quality of the dataset to clean or remove data that may compromise subsequent analysis, the aim is to assess duplicate entries, missing values, unwanted data, and recording errors. The dataset initially contained 11 duplicate entries, which were removed, reducing the number of rows to 906. Further analysis using appropriate methods indicates the absence of missing values. Plots for numerical features indicate no unwanted entries, recording errors and outliers for all numerical features.

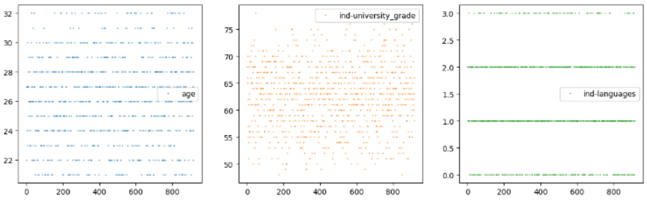


Fig. 2. Pandas plot showing numerical distribution

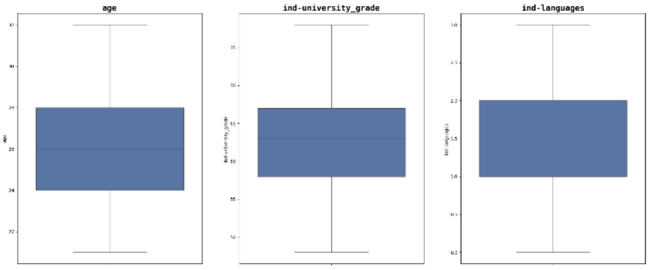


Fig. 3. Box plot showing age distribution

### C. Exploratory Data Analysis

Further analysis, employing relevant charts, aims to better understand the distribution of the protected feature against some features with categorical data, as well as the correlation between the target and predictor variables in the dataset.

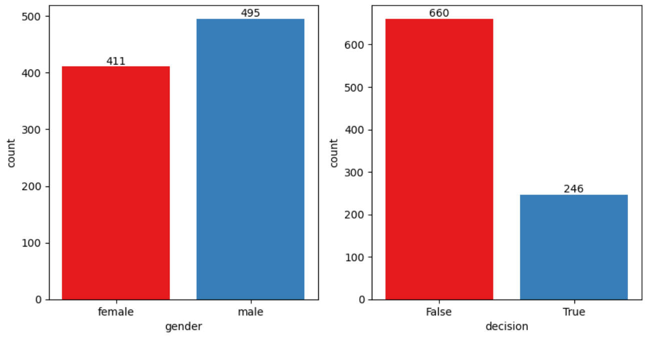


Fig. 4. Distribution of gender feature and target

Fig. 4 reveal a slight gender disparity and a relatively high number of rejections in the dataset. The gender distribution, depicted in the fig. 4, shows 495 (55%) entries for 'male' and 411 (45%) for females, automatically making 'female' the underrepresented group; a model trained with this dataset may thus be biased towards women.

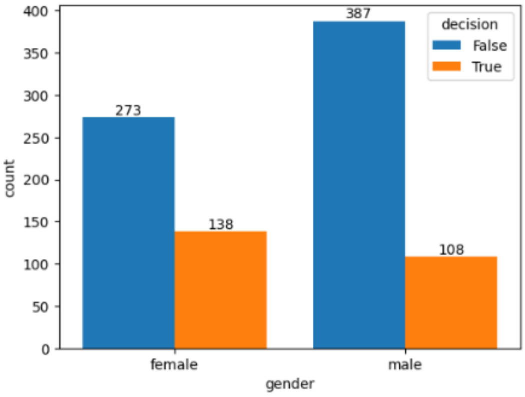


Fig. 5. Distribution of gender against the target

Fig. 5 shows the distribution of the gender-protected feature against the target. From this, male candidates, with a relatively higher number of entries, have an approval rate of 22% (108 out of 495), while female candidates, with a relatively lower number of entries, have an acceptance rate of 34% (138 out of 411).

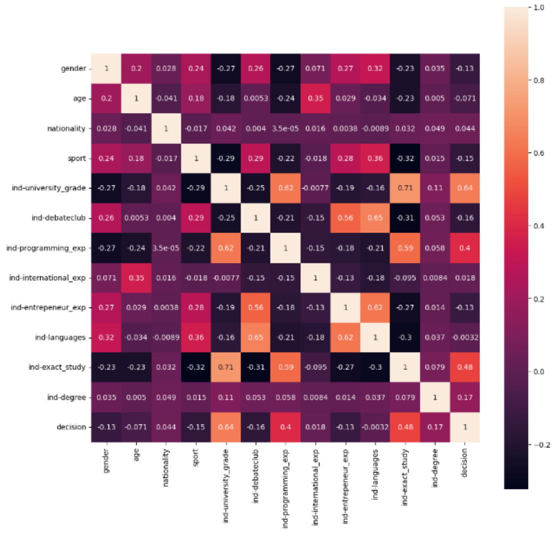


Fig. 6. Correlation matrix

Correlation matrix shows the strength of association between features. Fig. 6 shows low positive correlation.

## D. Feature Engineering

It is crucial to exclude unimportant attributes during the preparation of data. The 'Id' feature was dropped because it served no actual purpose in the dataset and was only used as a unique identifier for each entry. This reduced the number of features to 13.

*1) Data Encoding:* Categorical features were manually and correctly encoded, which is an essential part of every ML model development process.

*2) Feature Selection:* Involves selecting only the most important and relevant features [23]. With the help of fig. 6, no highly correlated features were found; hence, all features were used in the model development.

## E. Model Development

In this section, we fit and train the model using decision tree algorithm, which is a popular ML algorithm used for classification.

*1) Train-test split:* The label or target is separated from the predictors or independent variables. The dataset is then divided into two parts: one set for training the model and the other for testing its performance. It is imperative that 80% constitute the training set, with the remaining 20% allocated for testing. The *random_state* of 42 ensure consistency in the split whenever the program is executed.

*2) Feature Scaling:* Scaling is essential in model development, as it adjusts the values of features to have a standard scale of measure. MinMaxScaler was used.

*3) Fitting the Model:* Decision tree algorithm was used in the development of the model. It's a supervised learning method used for classification [19]. This model involves classifying samples into one of two classes, predicting the likelihood of an applicant getting selected for a role based on several pieces of information.

## F. Model Evaluation

Here, the model is evaluated based on performance metrics such as accuracy, precision, recall, and the F1 score computed from the 4 components of the confusion matrix (TP, TN, FP, and FN). The test is done using the test set that was originally split from the dataset. This step is simply to determine the effectiveness of the model in making accurate predictions.

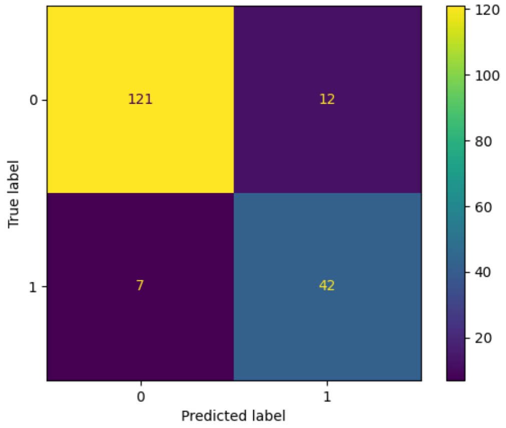


Fig. 7. Confusion matrix

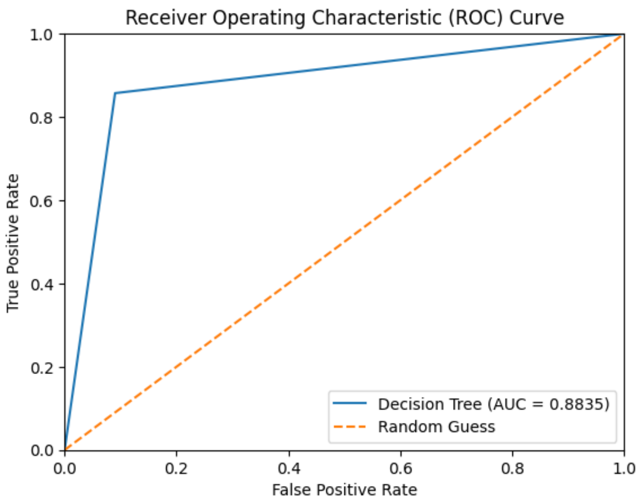


Fig. 8. ROC curve of the model

TABLE I. MODEL PERFORMANCE METRICS

| | *Accuracy* | *Recall* | *Precision* |
|---|---|---|---|
| **Train set** | 1.00 | 1.00 | 1.00 |
| **Test set** | 0.90 | 0.86 | 0.78 |

[a] Model performance metrics.

## G. Splitting Outcomes

Using the protected characteristic - gender, the outcome is split into female and male groups.

*1) Female:* Regarded as the underrepresented group in the original dataset, females totaled 411. In the test data, out of a total of 182, females make up 76.

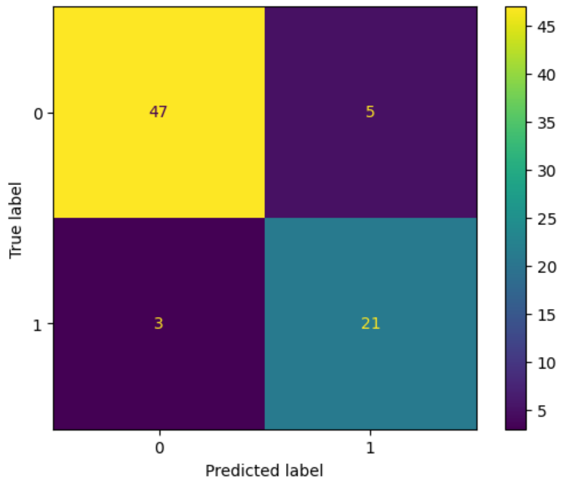


Fig. 9. Confusion matrix for 'female' group

TABLE II. PERFORMANCE METRICS FOR FEMALE GROUP

| Metrics | *Actual* | *Fraction* | *Percentage (%)* |
|---|---|---|---|
| **Accuracy** | 0.90 | $^{68}/_{76}$ | 90 |
| **Recall** | 0.88 | $^{21}/_{24}$ | 88 |
| **Precision** | 0.81 | $^{21}/_{26}$ | 81 |
| **Positive rate** | 0.34 | $^{26}/_{76}$ | 34 |

Performance metrics for female group

*2) Male:* Regarded as the majority in the original dataset, males make up 106 in the test data after the split.

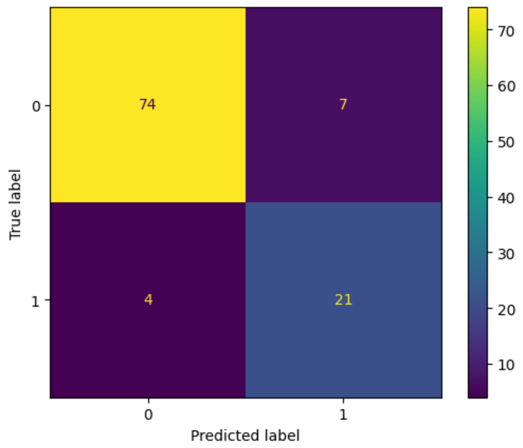


Fig. 10. Confusion matrix for 'male' group

TABLE III. PERFORMANCE METRICS FOR MALE GROUP

| Metrics | *Actual* | *Fraction* | *Percentage (%)* |
|---|---|---|---|
| **Accuracy** | 0.90 | $^{95}/_{106}$ | 90 |
| **Recall** | 0.84 | $^{21}/_{25}$ | 84 |
| **Precision** | 0.75 | $^{21}/_{28}$ | 75 |
| **Positive rate** | 0.26 | $^{28}/_{106}$ | 26 |

Performance metrics for male group

## H. Fairness Criteria

This step involves the use of confusion matrix to investigate bias in an ML model. The performance metrics computed from the various confusion matrices of each group will be evaluated to determine if the fairness criteria are met.

*1) Equal Accuracy:* Percentage of correct classifications. Table IV shows the same accuracy across groups.

TABLE IV. EQUAL ACCURACY

| **Metrics** | **Female** | | **Male** | |
|---|---|---|---|---|
| | *Fraction* | *Percentage (%)* | *Fraction* | *Percentage (%)* |
| **Accuracy** | $^{68}/_{76}$ | 90 | $^{95}/_{106}$ | 90 |

Equal Accuracy.

*2) Group Fairness:* This is evaluated using the positive rate metric. Table V clearly shows unequal positive rates, which signify demographic disparity.

TABLE V. GROUP FAIRNESS

| **Metrics** | **Female** | | **Male** | |
|---|---|---|---|---|
| | *Fraction* | *Percentage (%)* | *Fraction* | *Percentage (%)* |
| **Positive Rate** | $^{26}/_{76}$ | 34 | $^{28}/_{106}$ | 26 |

Demographic/statistical parity.

*3) Equality of Opportunity:* Percentage of actual positives predicted as positive. Table VI shows unequal true positive rate (recall) for both groups.

TABLE VI. EQUAL OPPORTUNITY

| **Metrics** | **Female** | | **Male** | |
|---|---|---|---|---|
| | *Fraction* | *Percentage (%)* | *Fraction* | *Percentage (%)* |
| **Recall** | $^{21}/_{24}$ | 88 | $^{21}/_{25}$ | 84 |

Equality of Opportunity.

### I. *Model Optimization*

Here, several steps were taken to improve the overall performance of the model and to compare the metrics of the distinct groups under various circumstances.

*1) SMOTE:* as part of the steps taken to improve the performance of the model on the test data and to mitigate bias, Synthetic Minority Over-sampling Technique [12] was utilized to balance the gender disparity and equal the number of the underrepresented and dominant groups in the train set.

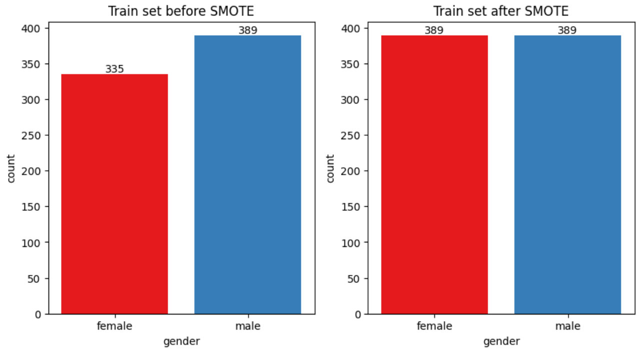


Fig. 11. Distribution of gender before and after SMOTE

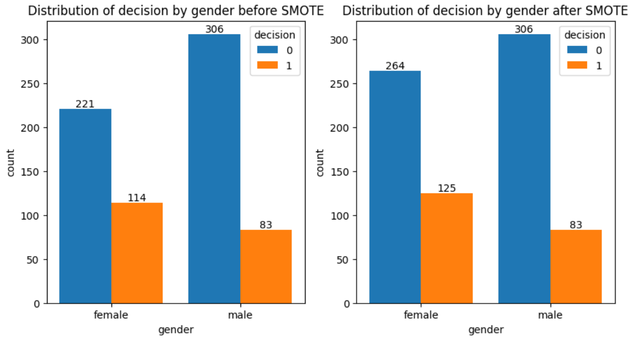


Fig. 12. Distribution of gender against target before and after SMOTE

*2) K-fold Cross-validation:* The purpose of this is to test the model with non-overlapping test sets [24]. As suggested in [23], 10-fold cross-validation was performed. This helped to reduce the variance, leading to an improvement in the overall accuracy score..

*3) Hyperparameter Tuning:* Hyperparameters are a set of values specified to influence the performance of a model. Gridsearch hyperparameter tuning technique was used to find the best hyperparameters for the model because it performs an exhaustive search and ensures that the best values within the defined search space are not missed. Parameters specified are *max_depth* of 3,5,7,9 and *min_samples_split* of 2, 4, 6. This subsequently improved the performance of the model.

TABLE VII. METRICS OF THE TUNED MODEL

| | *Accuracy* | *Recall* | *Precision* |
|---|---|---|---|
| **Train set** | 0.93 | 0.97 | 0.80 |
| **Test set** | 0.94 | 1.00 | 0.82 |

Tuned model performance metrics.

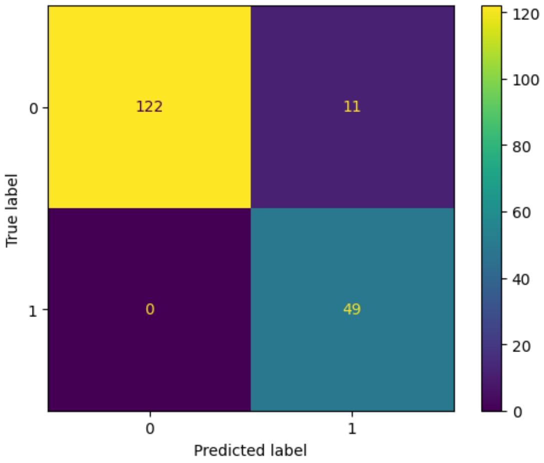


Fig. 13. Confusion matrix of the best model

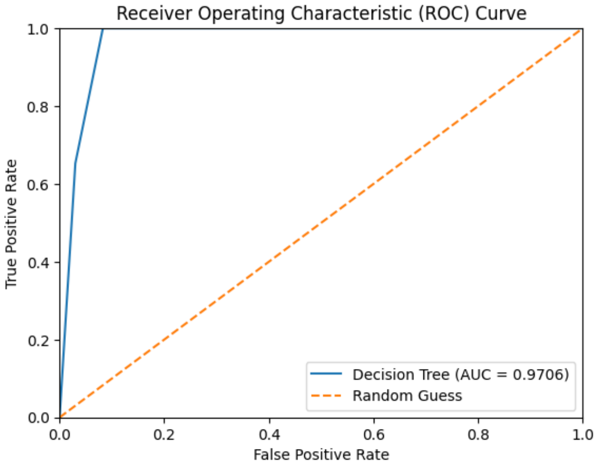


Fig. 14. ROC curve of the best model

*4) Splitting Outcomes:* After several techniques to optimize the performance of the model, the outcome of the best model is also split into groups using same gender-protected feature.

*a) Female:* After balancing the gender disparity, females are no longer underrepresented.

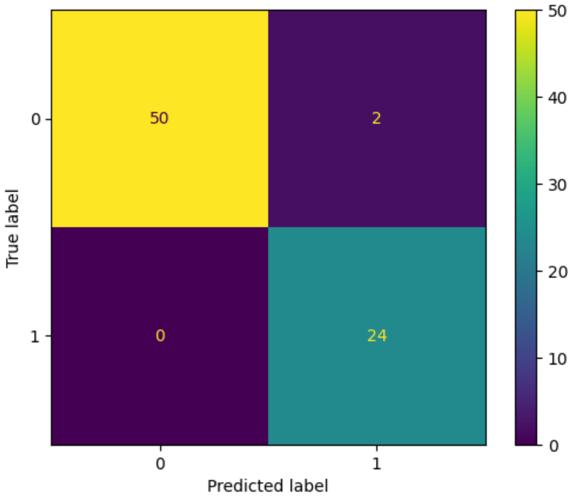


Fig. 15. Confusion matrix for 'female' group

TABLE VIII. BEST MODEL METRICS FOR FEMALE GROUP

| Metrics | *Actual* | *Fraction* | *Percentage (%)* |
|---|---|---|---|
| **Accuracy** | 0.97 | $^{74}/_{76}$ | 97 |
| **Recall** | 1.00 | $^{24}/_{24}$ | 100 |
| **Precision** | 0.92 | $^{24}/_{26}$ | 92 |
| **Positive rate** | 0.34 | $^{26}/_{76}$ | 34 |

Performance metrics for female group

*b) Male:* After balancing the gender disparity, males are no longer the majority group.

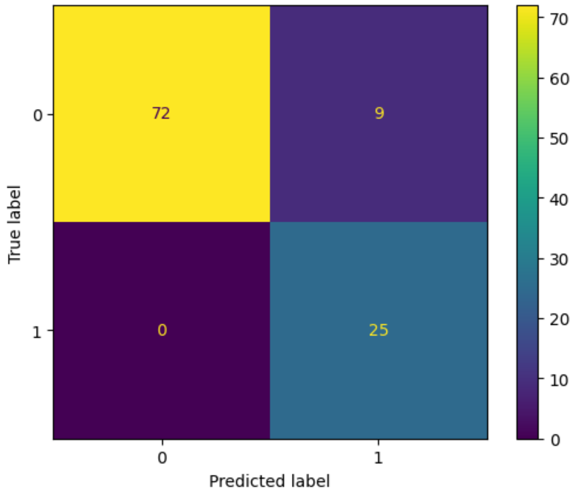


Fig. 16. Confusion matrix for ‘male’ group

TABLE IX. BEST MODEL METRICS FOR MALE GROUP

| Metrics | *Actual* | *Fraction* | *Percentage (%)* |
|---|---|---|---|
| **Accuracy** | 0.92 | $^{97}/_{106}$ | 92 |
| **Recall** | 1.00 | $^{25}/_{25}$ | 100 |
| **Precision** | 0.74 | $^{25}/_{34}$ | 74 |
| **Positive rate** | 0.32 | $^{34}/_{106}$ | 32 |

Performance metrics for male group

### 5) *Applying Fairness Criteria*

Checking whether fairness conditions are met for the best tuned model.

*a) Equal Accuracy:* Table X below shows unequal accuracy across both groups.

TABLE X. EQUAL ACCURACY

| Metrics | Female | | Male | |
|---|---|---|---|---|
| | *Fraction* | *Percentage (%)* | *Fraction* | *Percentage (%)* |
| **Accuracy** | $^{74}/_{76}$ | 97 | $^{97}/_{106}$ | 92 |

Equal Accuracy.

*b) Group Fairness:* Table XI shows unequal positive rates, but a slight increase in the positive rate for the male group.

TABLE XI. GROUP FAIRNESS

| Metrics | Female | | Male | |
|---|---|---|---|---|
| | *Fraction* | *Percentage (%)* | *Fraction* | *Percentage (%)* |
| **Positive Rate** | $^{26}/_{76}$ | 34 | $^{34}/_{106}$ | 32 |

Demographic/statistical parity.

*c) Equality of Opportunity:* Table XII shows equal true positive rate (recall) for both groups.

TABLE XII. EQUAL OPPORTUNITY

| Metrics | Female | | Male | |
|---|---|---|---|---|
| | *Fraction* | *Percentage (%)* | *Fraction* | *Percentage (%)* |
| **Recall** | $^{24}/_{24}$ | 1.00 | $^{25}/_{25}$ | 1.00 |

Equality of Opportunity.

## III. Findings

The analysis indicates the existence of bias in the recruitment model. Fig. 4 clearly illustrates a slight gender disparity, with females being underrepresented and males, the majority. However, despite this unequal distribution, Fig. 5 reveals that females have a higher success rate at being hired than males, with 34% of 411 females being successful compared to only 22% of 495 males. This suggests that the system favors women over men, even though they make up a smaller proportion. Further assessment through the application of fairness criteria confirms this bias. Comparing relevant performance metrics for the outcomes of both groups shows that not all fairness conditions are met, indicating the existence of bias.

From the model development, Table IV shows equal accuracy across groups, indicating the same proportion of correct classifications in both groups. While this condition is met, all others must also be fulfilled.

However, the situation appears different for group fairness, as Table V reveals demographic or statistical disparity between both groups, indicating that the distribution of acceptance or hiring (positive predictions) does not align with the representation of each group. The proportion of acceptance varies between the two groups.

The inequality of opportunity in Table VI further confirms bias in the model, as females tend to have a higher chance of being hired than males. This suggests that candidates with the same level of competence in both groups are not afforded the same opportunities.

## IV. Limitations and Conclusion

After taking several steps to optimize the model by addressing gender disparity and attempting to mitigate bias, the performance of the model improved. However, it still exhibited a slight bias toward males. Table X indicates unequal accuracy, suggesting improvement. Table XI demonstrates almost equal group fairness, with a slight increase in the acceptance or hiring of males. Lastly, Table XII reveals equality of opportunity, which means that candidates with the same qualification or skills are now afforded the same opportunity. This investigation yields promising outcomes; however, there are still areas for improvement, such as exploring alternative algorithms, applying sampling techniques to the target rather than just the protected feature, or balancing both. Also, blinding the protected feature from model development and computing results.

In conclusion, this report aimed to investigate bias based on protected characteristics in an ML model through the application of relevant fairness criteria. This approach can be beneficial for evaluating ML models to prevent potential harm after deployment.

# References

[1] Albassam, W.A. (2023) 'The Power of Artificial Intelligence in Recruitment: An Analytical Review of Current AI-Based Recruitment Strategies', International Journal of Professional Business Review, 8(6), pp. e02089. Available at: https://doi.org/10.26668/businessreview/2023.v8i6.2089

[2] Bhatt, P. (2023) 'AI adoption in the hiring process – important criteria and extent of AI adoption', Foresight (Cambridge), 25(1), pp. 144-163. Available at: https://doi.org/10.1108/FS-07-2021-0144

[3] Chen, Z. (2023) 'Ethics and discrimination in artificial intelligence-enabled recruitment practices', Humanities & Social Sciences Communications, 10(1), pp. 567-12. Available at: https://doi.org/10.1057/s41599-023-02079-x

[4] Dobson, S. (2018) 'IS AI BIASED IN RECRUITMENT?', Canadian HR Reporter, 31(12), pp. 8. Available at: https://tees.summon.serialssolutions.com/

[5] Eubanks, B. (2022) Artificial Intelligence for HR: Use AI to Support and Develop a Successful Workforce. 2; Second edn. London: Kogan Page.

[6] Figueroa-Armijos, M., Clark, B.B. and da Motta Veiga, S.P. (2023) 'Ethical Perceptions of AI in Hiring and Organizational Trust: The Role of Performance Expectancy and Social Influence', Journal of Business Ethics, 186(1), pp. 179-197. Available at: https://doi.org/10.1007/s10551-022-05166-2

[7] Hunkenschroer, A.L. and Luetge, C. (2022) 'Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda', Journal of Business Ethics, 178(4), pp. 977-1007. Available at: https://doi.org/10.1007/s10551-022-05049-6

[8] Jeske, D. and Shultz, K.S. (2016) 'Using social media content for screening in recruitment and selection: pros and cons', Work, Employment and Society, 30(3), pp. 535-546. Available at: https://doi.org/10.1177/0950017015613746

[9] Johnson, R.D., Stone, D.L. and Lukaszewski, K.M. (2021) 'The benefits of eHRM and AI for talent acquisition', Journal of Tourism Futures, 7(1), pp. 40-52. Available at: https://doi.org/10.1108/JTF-02-2020-0013

[10] Kammerer, B. (2022) 'Hired by a Robot: The Legal Implications of Artificial Intelligence Video Interviews and Advocating for Greater Protection of Job Applicants', Iowa Law Review, 107(2), pp. 817-849.

[11] Larsson, S., White, J. and Ingram Bogusz, C. (2024) 'The Artificial Recruiter: Risks of Discrimination in Employers' Use of AI and Automated Decision-Making', Social Inclusion, , pp. 1.

[12] Lema ^itre, G., Nogueira, F. and Aridas, C.K. (2017) 'Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning', Journal of Machine Learning Research, 18(17), pp. 1-5. Available at: http://jmlr.org/papers/v18/16-365.html

[13] Løkke, A., Villadsen, A.R. and Bach, A.S. (2023) 'Recruitment and Selection in the Public Sector: Do Rules Shape Managers' Practices?', Public Personnel Management, 52(2), pp. 218-239. Available at: https://doi.org/10.1177/00910260221146145

[14] Malik, A. et al. (2022) 'May the bots be with you! Delivering HR cost-effectiveness and individualised employee experiences in an MNE', International Journal of Human Resource Management, 33(6), pp. 1148-1178. Available at: https://doi.org/10.1080/09585192.2020.1859582

[15] Mastracci, S.H. (2009) 'Evaluating HR Management Strategies for Recruiting and Retaining IT Professionals in the U.S. Federal Government', Public Personnel Management, 38(2), pp. 19-34. Available at: https://doi.org/10.1177/009102600903800202

[16] Miasato, A. and Silva, F.R. (2019) 'Artificial Intelligence as an Instrument of Discrimination in Workforce Recruitment', Acta Universitatis Sapientiae, Legal Studies, 8(2), pp. 191-212.

[17] Mujtaba, D.F. and Mahapatra, N.R. (2019) 'Ethical Considerations in AI-Based Recruitment', IEEE Available at: 10.1109/ISTAS48451.2019.8937920.

[18] Ore, O. and Sposato, M. (2022) 'Opportunities and risks of artificial intelligence in recruitment and selection', International Journal of Organizational Analysis (2005), 30(6), pp. 1771-1782. Available at: https://doi.org/10.1108/IJOA-07-2020-2291

[19] Pedregosa, F. et al. (2011) 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, 12(85), pp. 2825-2830. Available at: http://jmlr.org/papers/v12/pedregosa11a.html

[20] Pinto, J., Borrego, M. and Cardoso, R. (2023) 'Artificial Intelligence as a booster of a Business Intelligence System to help the recruitment process : Business Intelligence, Human Resources and Talent', ITMA Available at: 10.23919/CISTI58278.2023.10211627.

[21] Tilmes, N. (2022) 'Disability, fairness, and algorithmic bias in AI recruitment', Ethics and Information Technology, 24(2) Available at: https://doi.org/10.1007/s10676-022-09633-2

[22] Yam, J. and Skorburg, J.A. (2021) 'From human resources to human rights: Impact assessments for hiring algorithms', Ethics and Information Technology, 23(4), pp. 611-623. Available at: https://doi.org/10.1007/s10676-021-09599-7

[23] Campesato, O. 2021. Chapter 7 Introduction to Machine Learning. Natural Language Processing and Machine Learning for Developers. Berlin, Boston: Mercury Learning and Information, pp. 283-328. https://doi.org/10.1515/9781683926177-008

[24] Campesato, O. 2020. 4. Intro to Machine Learning. Angular and Machine Learning Pocket Primer. Berlin, Boston: Mercury Learning and Information, pp. 135-172. https://doi.org/10.1515/9781683924685-005