Artificial Intelligence Life Cycle: The Detection and Mitigation of Bias

Ashionye Aninze and Jagdev Bhogal
Birmingham City University, UK

ashionye.aninze@mail.bcu.ac.uk
jagdev.bhogal@bcu.ac.uk

Abstract: The rapid expansion of Artificial Intelligence(AI) has outpaced the development of ethical guidelines and regulations, raising concerns about the potential for bias in AI systems. These biases in AI can manifest in real-world applications leading to unfair or discriminatory outcomes in areas like job hiring, loan approvals or criminal justice predictions. For example, a biased AI model used for loan prediction may deny loans to qualified applicants based on demographic factors such as race or gender. This paper investigates the presence and mitigation of bias in Machine Learning(ML) models trained on the Adult Census Income dataset, known to have limitations in gender and race. Through comprehensive data analysis, focusing on sensitive attributes like gender, race and relationship status, this research sheds light on complex relationships between societal biases and algorithmic outcomes and how societal biases can be rooted and amplified by ML algorithms. Utilising fairness metrics like demographic parity(DP) and equalised odds(EO), this paper quantifies the impact of bias on model predictions. The results demonstrated that biased datasets often lead to biased models even after applying pre-processing techniques. The effectiveness of mitigation techniques such as reweighting(Exponential Gradient(EG)) to reduce disparities was examined, resulting in a measurable reduction in bias disparities. However, these improvements came with trade-offs in accuracy and sometimes in other fairness metrics, identified the complex nature of bias mitigation and the need for precise consideration of ethical implications. The findings of this research highlight the critical importance of addressing bias at all stages of the AI life cycle, from data collection to model deployment. The limitation of this research, especially the use of EG, demonstrates the need for further development of bias mitigation techniques that can address complex relationships while maintaining accuracy. This paper concludes with recommendations for best practices in Artificial Intelligence development, emphasising the need for ongoing research and collaboration to mitigate bias by prioritising ethical considerations, transparency, explainability, and accountability to ensure fairness in AI systems.

Keywords: Artificial intelligence, Machine learning, Ethics in AI, Data bias, Machine learning fairness, Bias in AI

1. Introduction

This paper aims to raise awareness of the increasing importance of mitigating biases within Machine Learning(ML) algorithms, detecting them, and effectively reducing them. We interact with Artificial intelligence(AI) systems without being aware. It is in our home devices, banking systems, search engines, and any system that requires intelligence. As AI is increasingly becoming a part of our daily lives, these systems must be fair, reliable, and trustworthy. The issue regarding the fast expansion of AI is that there are not many rules or regulations regarding ethics considered or developed. While investment schemes run by(GOV.UK, 2023), have been set up and AI bills are in the process of being made(Parliament.uk, 2023), those that do exist are in their early days. There is a gap regarding ethics in AI and without a deeper understanding of how our own biases can be transferred into the algorithms we create; biased AI systems will continue to be developed. This could have many repercussions socially and legally such as providing harmful information or reinforcing biased opinions. AI can have a negative influence and there should be more guidance for developers to effectively address bias.

1.1 Bias in AI

According to Google Cloud(2024), AI is a broad field referring to technologies that build machines and computers that can replicate cognitive functions associated with human intelligence. This includes seeing, understanding, and responding to written or spoken languages. AI can analyse data, and make predictions, often thought of as a system - AI is a set of technologies implemented in systems to learn and solve complex problems.

ML is a subset of AI that enables a machine or system to learn and improve from experience. ML algorithms improve their performance over time as they are trained on more data. ML is the application of AI that allows machines and systems to extract knowledge from the data and learn by itself. Further stated by Google Cloud(2024), AI Is the idea that a machine can mimic human intelligence, but ML does not. ML aims to teach machines how to perform a task based on patterns provided by data.

Bias in AI refers to the systematic and structured errors in AI systems that result in unfair outcomes and inequalities. As ML is becoming widely used, biases within these decision-making machines must not replicate the biases that exist within training data. Park and Hu(2023) state that biases can arise at various stages of AI development, including data-generating processes and human decision-making. As ML algorithms learn from data if the data given is biased, the algorithm will learn from that and continue perpetuating the biases within that dataset.

1.2 Prevalence of Bias in AI and Potential Consequences

In recent years, the growing implications of AI have garnered significant attention within the scientific community and our society. As stated by Rishi Sunak during the AI safety summit in November 2023 the UK will be taking unique steps to establish the first AI safety Institute to 'Understand what each new model is capable of, exploring all the risk from social harms like bias and misinformation'(The Telegraph, 2023). Sunak's announcement highlights the growing prevalence of biases within AI systems, prompting nations to address these challenges and mitigate their impact collaboratively. This highlights that biased AI algorithms are becoming an international issue and therefore must be addressed. Bias is the act of supporting or opposing a particular thing or person unjustly because of personal opinions that influence judgment. In AI and ML, bias refers to the tendency of algorithms to produce results that favour certain outcomes or groups of people over others, this is due to the underlying human bias or training data(Chopra and Khurana, 2023). This can lead to unfair outcomes or inaccurate predictions, especially in sensitive industries like healthcare or justice. The outcomes often disadvantage specific groups of people, especially those from marginalised communities, reinforcing existing societal inequalities(Akselrod, 2021). There is ample evidence of discriminatory harm that AI tools can cause to already marginalised groups, Akselrod(2021) explains that AI is built by humans and deployed within systems that have been established with discrimination – from housing to the workplace. The bias preexisting in the data that are used to train the AI is often discriminatory or under-representative for people of colour, women, and any other marginalised group. Only those who understand can work to address the potential harm.

2. Literature Review

This literature review examines multiple topics and themes relevant to ethics within AI. In a comprehensive exploration of bias in AI algorithms, this report will address the ethical considerations within ML highlighting real-world consequences.

Bias mitigation between high-income and low-middle income is explored in a paper by Yang et al.(2024). Within their research, they revealed the preexisting biases occurring within their dataset that were further manifested with their Neural Network Model. Conclusions revealed OD disparities were reduced at the expense of model accuracy. In another paper by González-Sendino, Serrano and Bajo(2024), they explore several datasets. This includes the adult dataset. The paper goes beyond observing correlations and takes a causal model approach, which aims to represent a cause-and-effect relationship between the different variables.

2.1 Different Variations of Bias in AI

Within the context of AI, Ferrara(2023) states that bias can arise from various sources of the ML pipeline. ML models can learn and replicate patterns of bias present in the data used to train them, resulting in unfair or discriminatory outcomes. Algorithmic bias refers to the systematic and structured errors in an AI system that generates unfair results. It can emerge from intentional design choices or unexpected decisions associated with how data is gathered(Shin and Shin, 2023). Understanding and addressing algorithmic bias is fundamental for ensuring fairness and equality in decision-making processes, biased algorithms can lead to discriminatory outcomes, perpetuating existing inequalities. Data bias occurs when data used to train or test a model is not representative of the real-world population, or when certain groups are over or under-represented(Wamburu et al., 2022). This can happen when the data is collected from biased sources. Social bias manifests when individuals or groups tend to favour or discriminate against social characteristics, such as race, gender, or socioeconomic status. It involves the formation of stereotypes, prejudices, and discriminatory behaviours(Masha Fedzechkina, Lucy Hall Hartley, and Roberts, 2022.

The development of fairer and more ethical AI systems can be developed by addressing diverse types of biases within the AI lifecycle.

2.2 Examples of Bias in AI

The problem of bias is shown in a paper by Raji and Buolamwini(2019) which found statistics that facial recognition algorithms used by law enforcement are up to 100 times more likely to misidentify Black or Asian faces compared to White faces. Algorithms like this could lead to the misidentification of Black or Asian individuals, resulting in wrong arrests and allowing actual perpetrators to avoid justice.

Nazer et al.(2023) states, ‘Race, sex and disability status are inequities that may determine which health-related problems are prioritised and funded.’ This decides what research, including those related to AI, is produced. Biases would result in research questions and problems favouring a segment of the population regardless of the burden of the disease. This can lead to unbalanced research in certain demographics leading to skewed data which produces skewed algorithms. This draws attention to the importance of fair and balanced research questions, as the failure to address such biases could perpetuate health disparities as the focus may be disproportionally on diseases affecting certain ethnic groups.

3. Methodology

This section provides an overview of the steps to detect and mitigate biases within the ML model and dataset.

The data mining methodology follows the KKD process referred to in a paper by Zavaleta-Sánchez et al.(2024) excluding the closing section of deployment. The Adult dataset Becker & Kohavi(1996) was selected for its popularity in fairness and bias research due to its demographic attributes and imbalanced income target(reflecting real-world inequalities). Fairlearn, a Python toolkit, is used for building AI systems that are fairer. DP was chosen as a key fairness metric; it measures whether each group have similar proportions of positive outcomes. This promotes fairness by striving for all groups to be selected equally. EO was also measured, focusing on whether the error rates were equal for all groups. The model was mitigated using EG where training sample weights are adjusted based on DP. After applying mitigation techniques DP and EO were re-evaluated to assess the effectiveness of the mitigation attempt. The impact of the accuracy was analysed to understand any potential trade-offs. To conclude, ethical implications were addressed and recommendations for future works were suggested focusing on the importance of responsible AI development.

4. Data Collection and Analysis

The adult dataset will be explored, and an analysis will be presented.

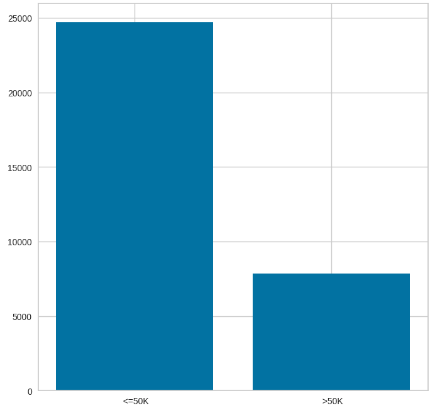


Figure 1: Distribution of Target

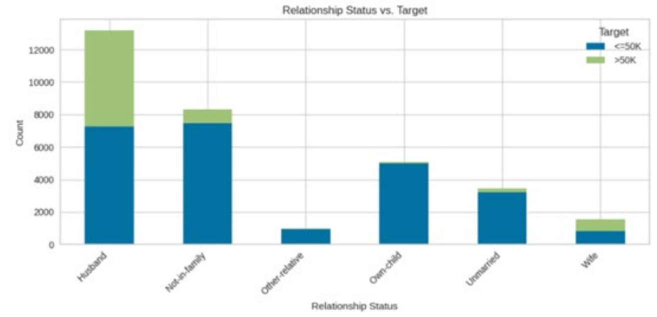


Figure 2: Relationship status vs target

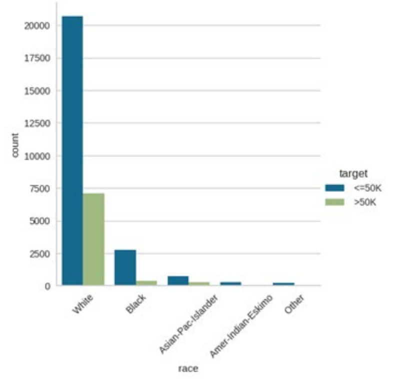


Figure 3: Distribution of race to target

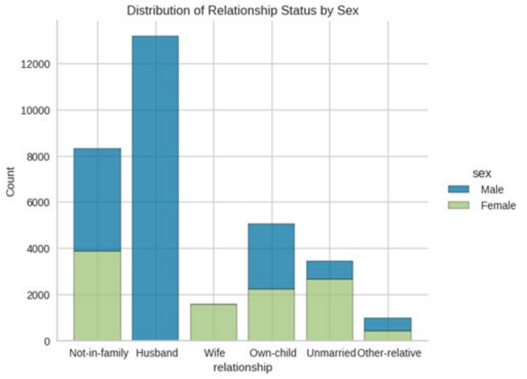


Figure 4: Distribution of relationship status by sex

4.1 Dataset Exploration

Explanatory Data Analysis(EDA) is performed on the adult dataset. EDA is the process of investigating the dataset to provide insights and better understanding. One of the goals of EDA is to uncover sensitive data attributes(sex, race, etc.), any correlations, and existing data quality issues in the dataset for potential bias. Python will be utilised to perform EDA on the adult dataset, including but not limited to pandas and others to visualise distributions and correlations. The goal is to use EDA to uncover potential biases such as unbalanced attributes and high correlations.

The dataset exhibits imbalances shown in Figures 3(racial) and 1(income). The over-representation of white individuals is an example of data bias. This imbalance can affect the model's ability to accurately predict income for those in the minority groups. Data bias is shown again in Figure 4, where Females are the underrepresented group. Fewer women make over 50k, potentially leading the model to favour males for higher income prediction.

Figure 4 reveals that 'Husband' perfectly correlates with 'Male' and vice versa making the relationship status attribute redundant. Husband is the most represented group, directly correlating with Figure 4 which shows men as the larger group. To address this, dropping the relationship attribute could be considered. Figure 5 highlights the unbalanced distributions within the relationship status attribute. The majority class is Husband again, while other categories lack high earners. Husband and wife classes have similar distributions, but the wife class is significantly underrepresented. This imbalance likely reflects historical bias and societal inequalities in earning potential between relationship groups(Norris et al., 2015).

Figure 5 highlights the correlation between all the features and the target variable. It highlights potential irrelevant features. Removing said features may be able to boost the accuracy score for the Gradient Boosting Classifier(GBC) Model. The features with the highest correlation are the education years, capital gain and the person's relationship status. The moderate positive correlation of 'education-num' and 'capital gain' suggests that those with higher education and capital gain are correlated with higher income.

The negative correlation for relationships needs more investigation. It is hard to tell which relationship status is negatively correlated as it has multiple classes. Figure 6 provides insights into the distribution in greater depth and Table 3 provides a label key for understanding. According to Figure 6, approximately 47% of wives are high earners, 45% of husbands are high earners and the other groups 7% are high earners. While wives are higher earners than husbands on average, the whole dataset will not reflect that as the class is outnumbered. As the husband group has the most individuals earning over 50k the ML model may inaccurately reflect that husbands make more than other relationship status.

Table 1: Relationship Status Value Label key

Relationship Status Value	Label
0	Husband
1	Not-in-family
2	Other-relative
3	Own-child
4	Unmarried
5	Wife

To summarise, this dataset is heavily unbalanced. If this dataset were used to train a model to detect whether an individual gets a loan by giving loans to a high earner it would show prejudice against those who are not a white married man.

5. Bias Mitigation Strategies

5.1 Bias Measurement

This section explores bias mitigation using Fairlearn, a Python toolkit created for building fairer AI systems. The metrics that will be used to measure the bias in the model are DP and EO. Unknown values were marked as '?' in the work class, occupation, and native country attributes were replaced with the mode to avoid data loss. Label encoding was used to relabel the categorical values to numerical ones for model interpretability. The data was split 80/20 for training and testing. Afterwards, SMOTE oversampling was performed on the training data to help mitigate the class imbalance of the target groups. GBCs are excellent models as they consistently rank high for classification tasks. Boosting combines the predictions of multiple weak learners to create stronger and more accurate learners.

Figure 10 provides different accuracy metrics on the model classification outcome; the combined accuracy score of the model is 87%.

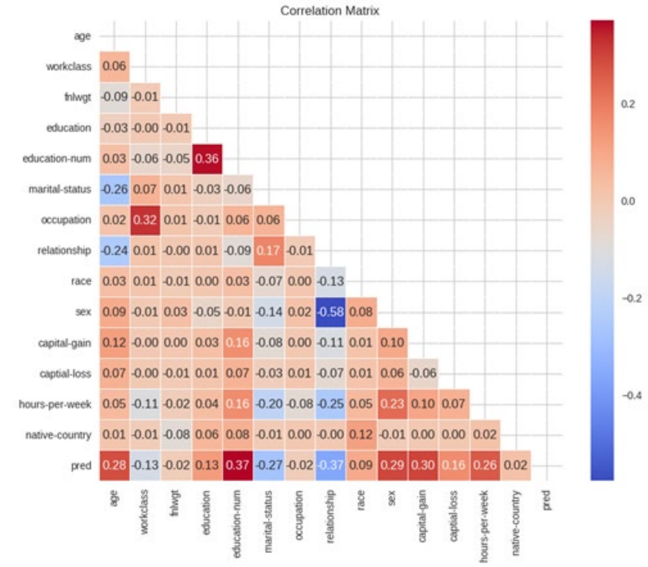


Figure 5: Correlation Matrix

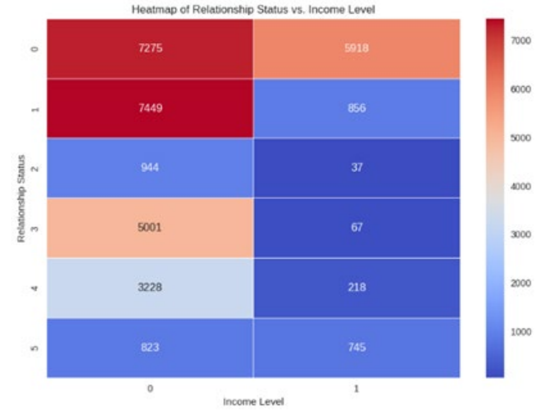


Figure 6: Heatmap of Relationship Status vs Income Level

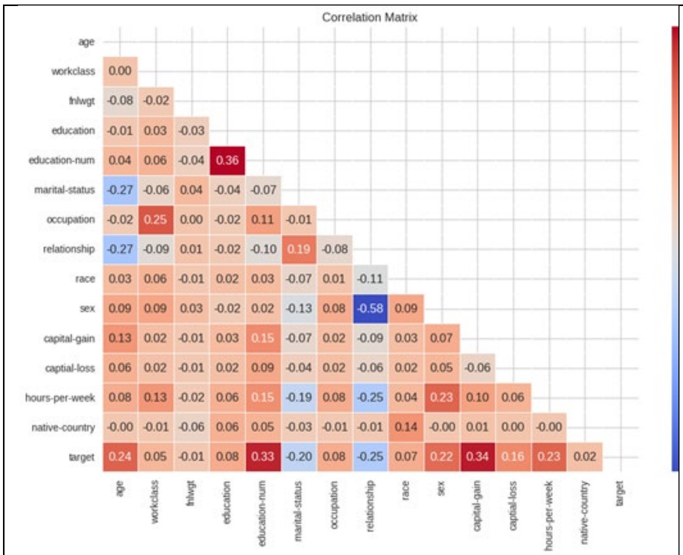


Figure 7: Correlation Matrix of Predicted Target

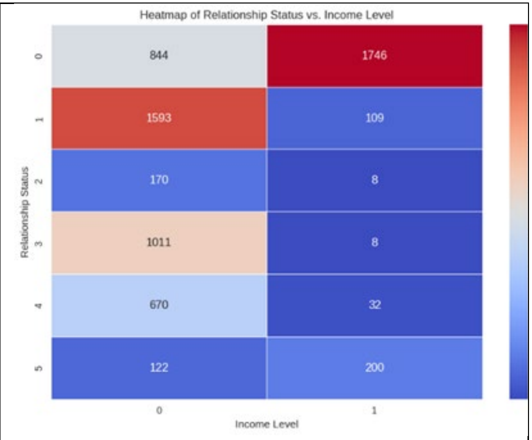


Figure 8: Heatmap of Relationship Status vs Predicted Income Level

Figure 7 is the correlation matrix of the test data and the targets the GBC model predicted. Compared with the correlation matrix on the data set before the ML model(Figure 5), the correlation matrix value for education-num, sex and relationship values have increased slightly; this is shown in Figure 7. This suggests that the algorithm slightly perpetuates the preexisting correlations. According to this diagram, education years positively correlate with higher earners while relationship negatively correlates. The distribution of high and low earners has significantly changed from Figure 8. In Figure 8 the percentage of women being labelled as higher earners is 62%, 67% for men and 6% for all other groups combined. While the original data showed that wives had a higher percentage of high earners, the husband class outnumbered the wife class. With this class imbalance, the ML model is disproportionally trained on husband data, which in turn, causes the ML model to favour that class. Even with class balancing, the model has amplified the existing bias within the dataset. This is an example of algorithmic bias, where the ML model has learned and reinforced the biased patterns in the data. Feature importance refers to techniques that calculate a score for all the input features Shin(2023), explains that a larger score suggests more importance. According to Figure 9, marital status is the biggest deciding factor. Marital status has an importance value of 50% while all other attributes have values under 10% - this suggests very high importance; therefore, this feature likely has a powerful influence.

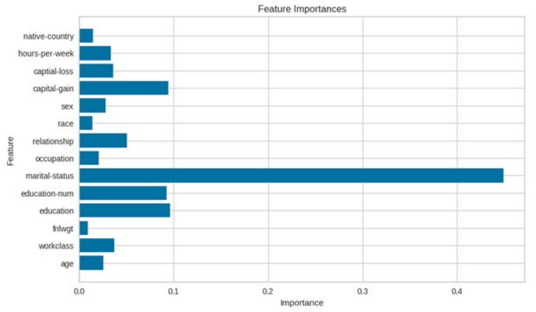
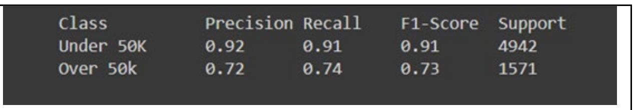


Figure 9: Feature Importances of the GBC Model



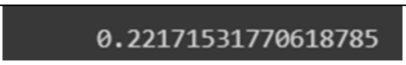
```
Class       Precision Recall    F1-Score  Support
Under 50K   0.92      0.91      0.91      4942
Over 50k    0.72      0.74      0.73      1571
```

Figure 10: Accuracy score for GBC model

This could indicate that the model depends solely on one factor. From the EDA performed, the insights derived from that was that most individuals are married with the biggest group being husbands.

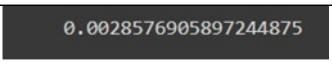
After identifying the biases, the next steps are to measure and quantify the biases to be able to compare values before and after mitigation processing. DP is a fairness metric that measures whether all groups have an equally positive outcome(Google for Developers, 2023). EO is another metric that will be utilised. Instead of measuring whether the model is predicting outcomes equally, EO measures the error rates for each class. The fairness metric calculates the true positive and negative rates, assessing whether the model is predicting outcomes equally well for all sensitive attributes(Google for Developers, 2024).

Figure 11 illustrates the DP difference between sexes, revealing a disparity of 0.22 in selection rates for high earners. This disparity is evidenced by Figure 14, which demonstrates that men(coded as 1) are more likely to be classified as high earners than women. It further supports how men are likelier to be falsely identified as a higher earner and women are more likely to falsely be identified as a lower earner through the error rates. This suggests that the ML model has learned the patterns that disproportionally benefit men and disadvantage women. Figure 12 provides the DP difference after mitigation using EG. The selection rate has been equalised between males and females resulting in the new DP score now being approximately 0.003. This shows that the disparities between the classes have been reduced, resulting in equal selection rates, and reducing the over-prediction of males.



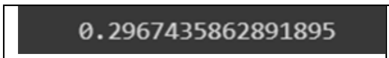
```
0.22171531770618785
```

Figure 11: DP Difference of Sex before Mitigation



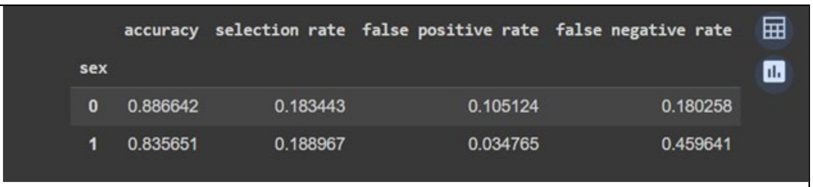
```
0.0028576905897244875
```

Figure 12: DP Difference of Sex after Mitigation



```
0.2967435862891895
```

Figure 13: Equalised Odds Difference on Sex attribute after mitigation



sex	accuracy	selection rate	false positive rate	false negative rate
0	0.886642	0.183443	0.105124	0.180258
1	0.835651	0.188967	0.034765	0.459641

Figure 14: Metrics for Sex attribute after Mitigation

5.2 Effectiveness of Mitigation Techniques

EG was a mitigation technique utilised to reduce the DP score. The weight change had a positive effect on the distribution of target classes. While the accuracy score was reduced(87% to 85%) the demographic parity scores were reduced(22% to 3%). This suggests that the classes now have an equal selection rate. Figure 14 reveals the complexities of bias mitigation. While a fairer DP has been achieved, the false negative rate of the male class has increased resulting in the EO gap widening. This suggests a trade-off between improving one fairness metric at the expense of another. Achieving parity often comes at the cost of higher error rates in specific groups. This highlights the challenge of different fairness considerations. Additionally, the accuracy score has been reduced when implementing mitigation methods. This also illustrates the trade-off between fairness and accuracy. The next section explores the ethical dilemmas of biased algorithms, their negative consequences, and the role of AI developers in mitigating these issues.

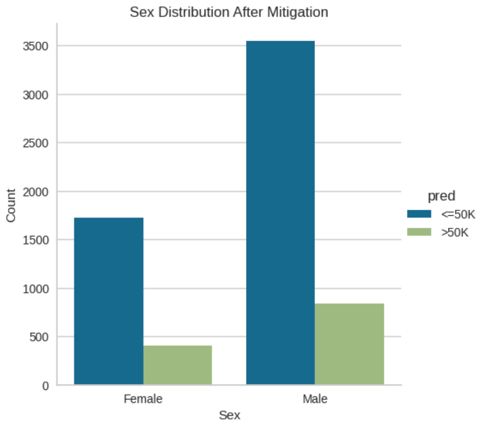


Figure 15: Sex distribution after mitigation

6. Ethical Considerations

This section will explore the ethical dilemmas of biases within AI and the repercussions., this will highlight the negative effects of biased AI systems to reiterate the importance of mitigation. After exploring the ethical dilemmas, implications, and social and legal consequences of those biases; strategies for responsible AI development will be discussed.

6.1 Ethical Dilemmas

Ethical dilemmas arise from biased AI systems, affecting various sectors like health, finance, and system development processes Ludovica Ilari et al.(2023). Khare(2023) explains that biased AI systems can have severe consequences for individuals and society. For example, if an AI scoring system is biased it can perpetuate financial inequalities; making it difficult for individual groups in marginalised communities to access credit. The adult dataset poses an ethical challenge, the data reflects historical and socioeconomic bias which raises questions about using datasets that perpetuate biases to train AI models. With efforts to mitigate biases through over-sampling and EG, disparities were still observed.

6.2 Ethical and Societal Implications of Biased AI

As AI becomes integrated into our lives, the potential for technologies to shape cultural norms and social structures becomes more significant(Ferrara, 2023). This could lead to new forms of discrimination based on skin colour or physical appearance. It is even possible for humans to inherit AI bias, according to a study conducted by Vicente and Matute(2023) where participants completed a medical-themed classification task with the help of a biased AI system. When participants were asked to complete the task again without assistance, they mimicked the same mistakes that biased AI systems had made during the previous phase. This suggests that biased AI could potentially shape future societal norms.

Additionally, bias amplification is another issue that causes societal and social implications. When initial models are biased, they produce biased outputs which are later used to train newer models resulting in the newer models learning more biases. To determine the effects, Bohdal et al.(2023) explained that if a job hiring decision AI is based on previous hiring decisions, bias in the subgroup can reinforce more bias in the future. This could lead to worse jobs and worse access to finance in that subgroup. Over time the subgroup would be socially and economically marginalised, widening existing inequalities.

6.3 Responsibilities of AI Developers

Ethical considerations, transparency and explainability in models must be prioritised. Developers should address biases by fostering diversity in their teams and conducting testing in diverse societal groups(Sanderson et al., 2023). Engaging with diverse stakeholders, including marginalised groups, throughout AI development can ensure these systems meet everyone's needs(Petar Radanliev et al., 2024). This ensures that the training data is diverse, while actively identifying and addressing any potential biases present in the data. Developers need to be able to apply mitigation techniques, to reduce any discriminatory outcomes. Doris and Potter(2024) suggest continuous monitoring, this allows developers to proactively identify any new biases. By following these principles developers can build more trust into these systems, ensuring that AI serves as a tool for good rather than harm.

Overall, this section explored the ethics surrounding biases within AI, highlighting the potential of societal harm due to unfair systems. The adult dataset served as a case study to showcase the presence of biases and demonstrate the challenges when mitigating biases and their trade-offs with accuracy scores and fairness metrics.

7. Conclusions and Future Work

This section summarises the key findings of this research and offers practical guidelines for future AI developers in promoting fair and ethical AI practices.

7.1 Summary of Findings

This paper investigated the potential biases in ML models trained on the Adult Census Income dataset, focusing on the relationship between income prediction and sensitive attributes such as marital status. Mitigation techniques were applied, and the disparities were reduced. While the dataset is valuable for income prediction, it contains sources of bias that could negatively impact model fairness. The bias in the ML model manifested from the over-representation of the male group resulting in the ML disproportionality predicting men as high earners than women. If this dataset were solely to be used for training a credit card application model; women would have a harder time getting credit, perpetuating financial inequalities. These biases were addressed through preprocessing, feature engineering, and fairness awareness metrics. This was to strive for a model that would produce equal outcomes for all individuals regardless of their demographic characteristics.

While a fairer model was achieved it had its trade-offs in accuracy and equal odds. The false negative rate for men increased suggesting that men were more frequently misclassified as low earners compared to women. While this was the case, the number of men being predicted as higher earners was still higher than women. This poses an ethical dilemma, even though the goal was to reduce bias, the mitigation strategy has introduced a new form of unfairness. This poses a question of whether DP should be prioritised at the expense of equalised odds. In this case for this dataset, DP should be prioritised as EO could not be achieved.

7.2 Implications of Findings in the AI Field

This section will provide a brief overview of how the findings in this dissertation contribute to the field of AI.

This paper has demonstrated an understanding of how bias manifests within ML models, often originating from the data itself. It contributes to data bias awareness, a critical issue as data bias is the biggest source of bias in ML models. The findings support existing theories on how biases are present in data and how they are further perpetuated, often through algorithmic amplification. This study's findings challenge the assumption that bias is solely created by the ML model and reiterate that bias can be introduced at several stages of an algorithm's development. This paper offers new insights into whether all mitigation steps should be taken, even if the trade-off is a worse-performing model, revealing an ethical dilemma on whether it is worth under-predicting certain groups in the name of fairness. Overall, by raising awareness of these issues and exploring potential solutions, this research contributes to the ongoing efforts to develop fairer algorithms.

7.3 Recommendations for Future Works

Future research should focus on developing stronger fairness metrics and exploring more effective mitigation techniques that address bias at every stage in the ML pipeline. Synthetic data generation should be considered to lessen the impact of overrepresented groups where more data collection is not possible. For future developers, they should focus on potentially developing new algorithms to specifically address fairness issues in models, users need to understand, trust, and manage AI results.

Encouraging collaboration between AI researchers and experts in social science, ethics, and law to ensure that AI development is guided by diverse perspectives with an understanding of societal impact is essential.

7.4 Guidelines

AI systems should be lawful, ethical and robust as discussed in the European Commission(2019).

References

Akselrod, O.(2021). How Artificial Intelligence Can Deepen Racial and Economic Inequities | ACLU. [online] American Civil Liberties Union. Available at: https://www.aclu.org/news/privacy-technology/how-artificial-intelligence-can-deepen-racial-and-economic-inequities .

Becker, B. and Kohavi, R.(1996). UCI Machine Learning Repository. [online] Uci.edu. Available at: https://archive.ics.uci.edu/dataset/2/adult.

Bohdal, O., Hospedales, T., Philip and Barez, F.(2023). Fairness in AI and Its Long-Term Implications on Society. [online] arXiv.org. Available at: https://arxiv.org/abs/2304.09826#:~:text=AI%20fairness%20focuses%20on%20mitigating,act%20as%20a%20social%20stressor.

Chopra, D. and Khurana, R.(2023). Bias and Fairness in Ml. BENTHAM SCIENCE PUBLISHERS eBooks, [online] pp.116–122. doi:https://doi.org/10.2174/9789815124422123010012.

Doris, L. and Potter, K.(2024). Continuous Monitoring and Improvement: Implement continuous monitoring of AI models to detect and correct... [online] ResearchGate. Available at: https://www.researchgate.net/publication/378342382_Continuous_Monitoring_and_Improvement_Implement_continuous_monitoring_of_AI_models_to_detect_and_correct_issues_in_real-time#:~:text=By%20continuously%20monitoring%20AI%20models,continued%20effectiveness%20of%20deployed%20models.

European Commission(2019). HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION ETHICS GUIDELINES FOR TRUSTWORTHY AI. [online] Available at: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

Ferrara, E.(2023). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. Sci, [online] 6(1), pp.3–3. doi:https://doi.org/10.3390/sci6010003.

González-Sendino, R., Serrano, E. and Bajo, J.(2024). Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making. Future Generation Computer Systems, [online] 155, pp.384–401. doi:https://doi.org/10.1016/j.future.2024.02.023.

Google Cloud.(2024). AI vs. Machine Learning: How Do They Differ? | Google Cloud. [online] Available at: https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning .

Google for Developers.(2023). Machine Learning Glossary: Fairness. [online] Available at: https://developers.google.com/machine-learning/glossary/fairness#:~:text=demographic%20parity&text=A%20fairness%20metric%20that%20is,on%20a%20given%20sensitive%20attribute.

OV.UK(2023). New innovation challenge launched to tackle bias in AI systems. [online] GOV.UK. Available at: https://www.gov.uk/government/news/new-innovation-challenge-launched-to-tackle-bias-in-ai-systems .

Ludovica Ilari, Giulia Rafaiani, Baldi, M. and Benedetta Giovanola(2023). Ethical Biases in Machine Learning-based Filtering of Internet Communications. [online] doi:https://doi.org/10.1109/ethics57328.2023.10154975.

Markou, G., Bakas, N. and Megan, A.(2023). Use of AI and ML Algorithms in Developing Closed-Form Formulae for Structural Engineering Design. Advances in civil and industrial engineering book series, [online] pp.73–105. doi:https://doi.org/10.4018/978-1-6684-5643-9.ch004.

Masha Fedzechkina, Lucy Hall Hartley and Roberts, G.(2022). Social biases can lead to less communicatively efficient languages. [online] 30(3-4), pp.230–255. doi:https://doi.org/10.1080/10489223.2022.2057229.

Nazer, L., Razan Zatarah, Waldrip, S., Janny, Moukheiber, M., Khanna, A.K., Hicklen, R.S., Lama Moukheiber, Moukheiber, D., Ma, H. and Mathur, P.(2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS digital health, [online] 2(6), pp.e0000278–e0000278. doi:https://doi.org/10.1371/journal.pdig.0000278.

Park, Y. and Hu, J.(2023). Bias in Artificial Intelligence. PubMed, [online] 18(3), pp.394–396. doi:https://doi.org/10.2215/cjn.0000000000000078.

Parliament.uk.(2023). Artificial Intelligence(Regulation) Bill [HL] - Parliamentary Bills - UK Parliament. [online] Available at: https://bills.parliament.uk/bills/3519.

Petar Radanliev, Santos, O., Alistair Brandon-Jones and Joinson, A.(2024). Ethics and responsible AI deployment. Frontiers in artificial intelligence, [online] 7. doi:https://doi.org/10.3389/frai.2024.1377011.

Raji, I.D. and Buolamwini, J.(2019). Actionable Auditing. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. doi:https://doi.org/10.1145/3306618.3314244.

Shin, D. and Shin, E.Y.(2023). Data's Impact on Algorithmic Bias. IEEE Computer, [online] 56(6), pp.90–94. doi:https://doi.org/10.1109/mc.2023.3262909.

Shin, T.(2023). Understanding Feature Importance in Machine Learning. [online] Built In. Available at: https://builtin.com/data-science/feature-importance .

The Telegraph(2023). Rishi Sunak Warns of Risks Posed by AI in Tech Summit. YouTube. Available at: https://www.youtube.com/watch?v=262YJQWE7Ec.

Vicente, L. and Matute, H.(2023). Humans inherit artificial intelligence biases. Scientific reports, [online] 13(1). doi:https://doi.org/10.1038/s41598-023-42384-8.

Wamburu, J., Girmaw Abebe Tadesse, Cintas, C., Adebayo Oshingbesan, Akumu, T. and Speakman, S.(2022). Systematic Discovery of Bias in Data. 2022 IEEE International Conference on Big Data(Big Data). [online] doi:https://doi.org/10.1109/bigdata55660.2022.10020781.

Yang, J., Clifton, L., Dung, N.T., Phong, N.T., Yen, L.M., Thy, D.B.X., Soltan, A.A.S., Thwaites, L. and Clifton, D.A.(2024). Mitigating machine learning bias between high income and low–middle income countries for enhanced model fairness and generalizability. Scientific Reports, [online] 14(1). doi:https://doi.org/10.1038/s41598-024-64210-5.

Zavaleta-Sánchez, E., Domínguez-Sánchez, G., Cecilia-Irene Loeza-Mejía and Sánchez-DelaCruz, E.(2024). Comparative Study of KDD and CRISP-DM Methodologies for Phishing Identification. Lecture notes in networks and systems, [online] pp.317–330. doi:https://doi.org/10.1007/978-981-97-3559-4_25.