Deus Ex Machina: Power, Perception, and Prejudice

Bernardo Salgado, Daniel Dias, Lucas Santiago, Miguel Lopes, Rafael Conceição
(Dated: December 13, 2024)

This project aims to explore the societal implications of Artificial Intelligence in real-world case scenarios. We proposed a significant case study where AI has impacted society and thoroughly analyze the case to identify and explain its main challenges, outcomes, and ethical dilemmas. A technical experimentation was conducted to deepen the understanding of the project's concepts.

I. CASE STUDY DETAILS

Unmasking Dropout: NYC's School Algorithms Cement Segregation

This case study explores how the use of algorithms in NYC's school admission system has perpetuated racial and economic segregation. The analysis includes the impact of these algorithms on school assignments and student opportunities, particularly in underrepresented communities. The article from The Markup provides data-driven insights into the issue, highlighting the unintended consequences of algorithmic decision-making in public education.

We decided to break the project into phases for systematic progress and clarity. Each phase focuses on a specific aspect of the project to address specific issues:

1. **Case Study Analysis:** Analyze the case study, extract key data points, and identify primary issues and ethical dilemmas such as bias, fairness, and transparency.

2. **Processing Data:** Source or generate a dataset that reflects the under representation of a certain class.

3. **Technical Experimentation:** Hypothesize that underrepresented classes are negatively affected in the predictions of ML models. Run a basic classification model in the chosen dataset to verify bias.

4. **Bias Mitigation Techniques:** Use bias mitigation techniques. Evaluate the new model's performance on fairness metrics compared to the original. Analyze how reducing bias impacts overall accuracy and other metrics like Demographic Parity (DP) difference. The DP difference measures the disparity in selection rates (e.g., positive predictions) between different groups within a sensitive feature. A smaller DP difference indicates a more fair or unbiased model.

5. **Conclusions:** Prepare visualizations, a technical tutorial, and a final report to illustrate the results for easier comprehension. Discuss ethical aspects of the current world situation and explore how schools could use these methods to improve transparency and fairness.

II. CASE STUDY ANALYSIS

In this section, we discuss the selected case more thoroughly and bear upon the consideration of context and societal consequences of AI.

We identify who the major participants are, how the computational system of AI works, and the consequences that were observed when using AI. Further, it looks at the acceptability issues of the AI application, the public and media response and the legal or political ramifications.

This analysis aims to uncover critical factors of AI integration in real-world applications, shedding light on its practical benefits as well as the challenges and limitations it may present.

A. AI's Role and Societal Impact

Topic: Describe the context in which the AI was used and identify the issue that emerged, reflecting on AI's role and societal impact. Specify involved parties (institutions, stakeholders, companies, etc.).

This case study concerns more than 100 high schools in New York County, with the two main examples being Millennium Brooklyn High School and Park Slope Collegiate and the students trying to enter those schools where the applications of black and mexican students are disproportionately declined.

In the main example of the article it was shown that the use of AI and classification algorithms represent an extreme bias towards white and asian applicants and decline the application of black and mexican applicants in a disproportionate way, making it next to impossible even for outstanding students of those minories to be accepted into good schools.

Other than the apparent racism present in the AI screening another of the problems is the lack of transparency that the schools have in their selection process, independent of the use of AI or manual review, the screening process is hidden behind obtuse rules and bureaucratic obfuscation that make it very hard to understand the why of the acceptance/decline of the application **FIG. 1**.

One of the worst consequences of this problem is the perpetuation of the "segregation" present in New York where the opportunities of black and mexican children are put outside their reach and condemn them to social and economical hardships in the future.

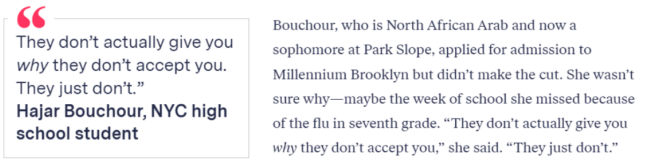

FIG. 1: NYC high school student quote

"I saw that they had a lot of AP classes, and they had photography class, and all this stuff. I was like, Whoa, whoa, whoa! That would be so cool." [...] "And then I was like, Oh, wait, but we don't have that."

B. AI Development

Topic: Discuss the technicalities of how the AI was developed and/or deployed, including relevant factors whenever you can.

Algorithmic choices: New York's school system employs both automated screening algorithms and conventional techniques to rank students applying to high schools. The automated systems use parameters such as test scores, attendance, and grades to produce scores for each student, but schools can format these parameters in different ways, for example, some of them prioritizing attendance over test scores. This leads to variability in how scores are calculated, resulting in unfair outcomes for certain groups. Since each school's algorithm is a black box, some families and students find it difficult to understand how decisions were made, which allows concerns about transparency and fairness to be raised.

Dataset quality: The datasets used by these algorithms have academic records of students, but the data doesn't say everything about them. For example, a student with a lower socioeconomic background may have less access to tutoring and other ways to improve their grades, and this student has a higher probability of accumulating absences due to the gap between access to private health. Furthermore, data from private school applicants needs to be fully included, and privacy and protection policies lead to some fractions of the dataset being redacted, increasing the difficulty of fully analyzing the data and addressing biases.

Outcomes: The statistics provided by "The Markup" illustrate a clear pattern of disparity, as we can see in **FIG. 2**.

These outcomes reveal how the AI systems reinforce already existing inequalities, favoring white and asian students, who tend to have better scores and attendance, but probably not due to how dedicated the students are, but due to various socioeconomic factors that have great influence on a young person's education. Even with efforts to make the system completely merit-based, these algorithms unintentionally perpetuate segregation.

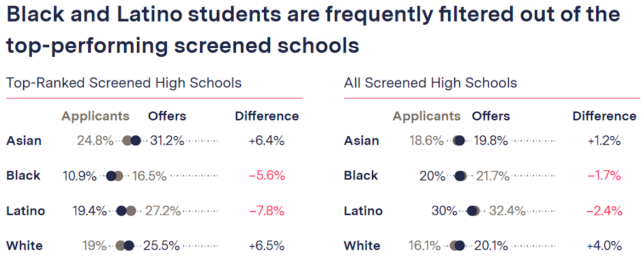

FIG. 2

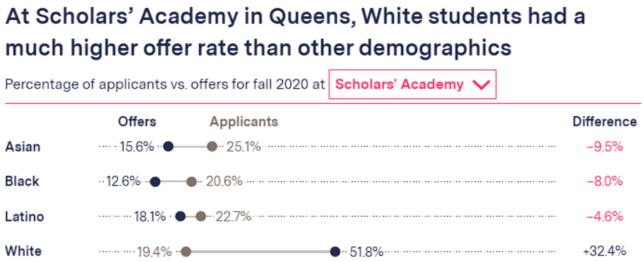

FIG. 3

Families with more resources can afford private tutoring, and have better access to private health, having effects in the students' test scores as well as in their attendance records. This results in another level of bias as most algorithms only look at the raw data, ignoring socioeconomic disparities of the students they are scoring.

C. AI System Outcomes and Overview

Topic: Focus on the outcomes of the AI application (positive or negative). Did the AI system improve efficiency or decision-making? Did it discriminate against certain groups? Was the logic behind the AI's decisions clear to its users? Was anyone accountable when (and if) the AI failed?

1. Did the AI system improve efficiency or decision-making? Indeed, the adoption of the AI system in the NYC high school increased efficiency through the automation of the processing of more than 80,000 applications annually. They were able to match students with the schools of their choice and the criteria that the school possessed, making it faster. However, this efficiency was achieved at the cost of the basic principle of fairness was not high on its agenda.

This approach made the system more efficient because it focused on such criterion as test scores and attendance data, but it was not fair in terms of decision-making.

"The algorithm used criteria such as test scores and attendance records..."

2. Did it discriminate against certain groups? Yes, what the algorithm did was to select more students that came from wealthy families thereby perpetuating segregation in such schools. The articles show that high schools which provide the best education for NYC students discriminated in admissions, and white and Asian students were accepted while Black and Hispanics students were not accepted.

This can be attributed to the elements that the algorithm employed including, performance on standard achievement tests that are closely linked to the SES making the problem worse by increasing separation in the school system.

"Selective high schools... overwhelmingly admitted white and Asian students..."

In addition, it was observed that the algorithm, which was not developed with bias, had inherent bias that privileged students with resources. For example, the schools that had competitive admission criteria applied geographic location, previous school performance and attendance, factors that worked against the students from low income areas. Source:

"Schools with competitive admissions used factors such as geography and previous performance..."

3. Was the logic behind the AI's decisions clear to its users? Unfortunately, few parents and students seem to understand the reasoning behind AI's decisions. The system was opaque, so many users failed to comprehend why the students were placed in certain schools. It was also evidenced that parents of the disadvantaged students were completely disoriented of the system and had no clue on how they could legally fight against the algorithm that gave priority to certain students than the others.

"Parents of marginalized students found the system opaque and confusing..."

Users never knew why some schools rejected their children since the algorithm depended on factors not easily understandable by users. This lack of transparency was felt to make the system untrustworthy for many parents from those groups.

"The opacity of the algorithm contributed to feelings of mistrust..."

4. Was anyone accountable when (and if) the AI failed? No, when the AI failed there was virtually no one to blame. The system was said to be operating with virtually no accountability, and there was no obvious means for families to find someone to blame for unfair placements or poor results. The schools, districts and the Department of Education did not accept much accountability for the way the algorithm worked, thus enabling the system to perpetuate unfairness without any interference.

"There was no clear way for families to hold anyone accountable for unfair placements..."

Further, it was unclear who the system's designers were to blame for the ways the AI reinforced racism and economic classism in the education system. Such lack of accountability entailed that the bias in the system was not corrected, and thus the inequalities in the distribution of education were maintained.

"The system's creators were not held accountable for perpetuating segregation..."

D. Societal, Political and Media Responses

Topic: Comment on the societal response (if there was any) or provide your own conclusions on the topic. How did society react? Were there legal consequences, public outrage, coverage on the news, political changes? What did the media report on the issue, how did stakeholders (government, public institutions, companies) react?

Public responded with a high level of concern, especially from parents, communities of color, and civil rights organizations. Education activist organizations such as the New York Civil Liberties Union compelled for reforms to abolish middle school screens and specialized high school admissions tests that disadvantaged Black and Latino students. The increasing coverage from reputable papers such as The New York Times and The Markup, shed light to the imperfections of the admission process.

"Public outrage has grown, with many calling for reforms, particularly focusing on dismantling policies that disproportionately exclude Black and Latino students."

Although social pressure triggered policy debates such as the suspension of middle school screening during the pandemic by Mayor Bill de Blasio, there were no major legal consequences. It might be possible to evaluate the effects of societal pressure by comparing the data on graduation rates and other metrics before and after specific policy. The absence of legal consequences indicates that changes were largely political rather than judicial.

"Under pressure, Mayor Bill de Blasio temporarily suspended middle school screens during the pandemic... [but] the pause did not immediately lead to a lawsuit or major legal consequences."

Community involvement was especially evident in ethnic diverse neighborhoods and the process was sharply criticized by advocates and parents. The media played a crucial role in raising awareness, with investigations by The Markup and coverage by The New York Times shedding light on race and economic disparity in enrollment of schools. The great media coverage was useful for disclosing the inadequacies of the systems and applying pressure for reforms.

"The media, including outlets like The Markup and The New York Times, provided extensive coverage of the segregation-driven policies embedded in these algorithms."

In response to public pressure, Mayor Bill de Blasio introduced measures like, suspension of screens for admission to middle school and elimination of geographic preferences as a way of diversifying selective schools. However, some affluent families resisted, fearing a reduction in academic standards, which caused more substantial reforms to remain at a proposal stage. The dataset contains some metrics that can be useful for evaluating the effectiveness of these political changes. For example, by comparing performance metrics before and after the suspension of middle school screening, we might get evidence on whether these reforms improved school diversity or student performance.

"Political changes came slowly, but Mayor de Blasio's reforms in the admissions process, such as pausing middle school screens, represented a step toward addressing these disparities."

III. TECHNICAL ANALYSIS

This section focuses on the technical steps of the project, including data processing, experimentation, and bias mitigation. We used a dataset involving underrepresented groups to determine the effects of skewed and biased data distributions.

To verify the hypothesis that these 'underprivileged' groups are discriminated by machine learning algorithms, we perform a simple classification check. This step involves training of a basic model in the data with a view of determining cases of unfair treatment as well as the performance of the groups in question. The use of measurements such as confusion matrices and error rates allows the identification of these subpopulations that have been biased.

Finally, we apply methods to reduce bias and compare the results, looking at how these changes affect fairness and performance metrics like accuracy and DP difference. This helps us understand the challenges of building more fair AI systems. Moreover, we discuss limitations of the fairness approach, and its trade-offs with interpretability and computational cost, so as to provide a realistic perspective on the potential of these solutions. Two distinct approaches are explored: one aimed at improving fairness while maintaining accuracy, and another focused on maximizing fairness at the expense of accuracy, providing a comparative view of their impacts and feasibility.

A. Data and Fairness Considerations

Since we did not have access to the dataset containing the individual applications of the students, we decided to use the Adult Census Income dataset, where class imbalance is evident, and the presence of morally sensitive features such as race and gender, can lead to biased model decisions against minorities.

The dataset used in this analysis was acquired from the 1994 U.S. Census Bureau database by Ronny Kohavi and Barry Becker and is intended to determine whether the respondent earns more than $50K per year. It comprises records which are filtered by age greater than 16, income greater than a certain threshold, and number of working hours.

The sensitive features include race, sex **FIG. 4**, income **FIG. 5** and native country. Each one of these categories also presents imbalance, with the most common type being a white male from the U.S. The potential for bias is further compounded by the skewed representation of these groups, making this dataset an excellent candidate for evaluating fairness-aware machine learning methods. To address these issues, we employ pre-processing techniques like oversampling and undersampling (in different stages), as well as post-processing strategies to adjust predictions. Additionally, we use fairness metrics such as disparate impact, demographic parity, and equalized odds to quantify bias and guide the provided model improvements.

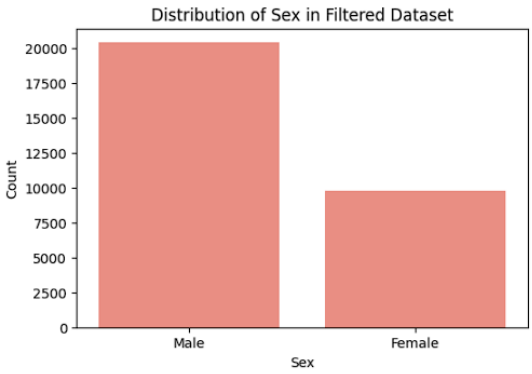

FIG. 4: Sex Distribution

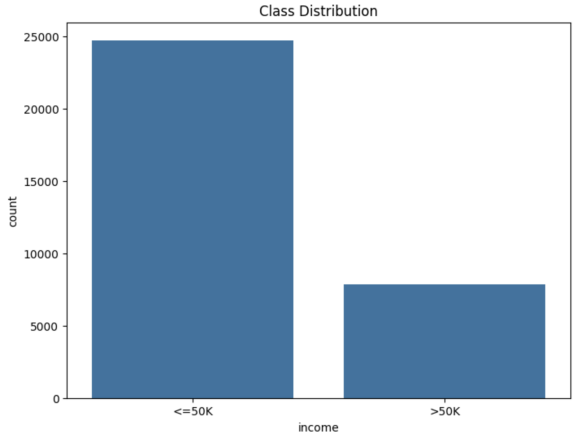

FIG. 5: Income Distribution

IV. FIRST APPROACH

A. EDA and data preparation

The analysis used the Adult Income dataset to perform the experiment.

Before starting the data cleaning process, YData Profiling was used to perform an in-depth analysis of the dataset, providing insights into its structure, missing values, correlations, and distributions. This step helped identify potential issues and patterns in the data that required attention before proceeding further. **FIG. 6**

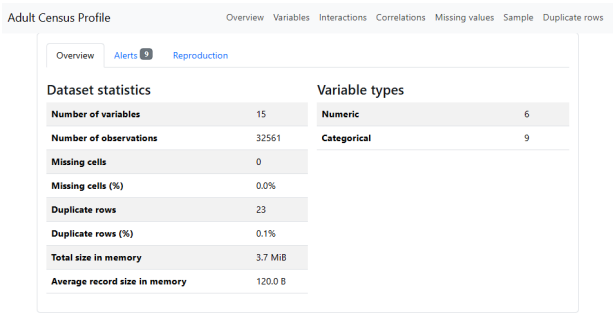

FIG. 6: YData Profiling Overview

The second step was to deal with missing values by deleting rows. Subsequently, the dataset went through data cleaning in order to eliminate any inconsistencies in the data collected. Unique values in each column were acknowledged to check if they could be grouped. Knowing that, the countries in the dataset were grouped into larger regions based on mapping of the countries. These regions were classified into Europe, Asia, United States, Latin America and Other. This regional mapping made easier the analysis based on geographical characteristics. **FIG. 7**

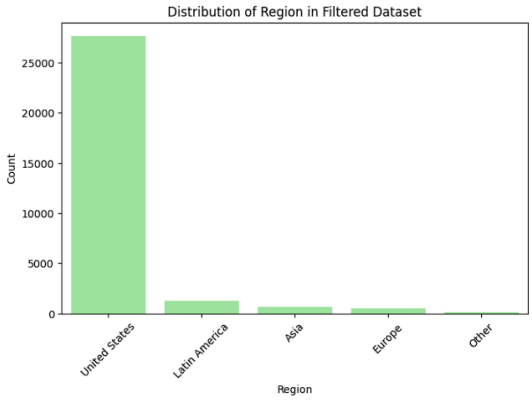

FIG. 7: Regions after Grouping

B. Technical Experimentation

The next step was to create a machine learning model, using the Fairlearn library. Fairlearn is a Python library which, in this case, allowed for the identification of fairness problems in machine learning models and addressing them. This particular project was centered on factors including accuracy and selection rate. The selection rate in this case is the ratio of correct outcomes (i.e., people with an income of $50,000 or more, being this the target class) to the various sensitive features. Sex, region, and race were the sensitive features in the study. Race was mapped into four categories: Asian-Pac-Islander, White, Black, and Other. This mapping was useful in simplifying the dataset while at the same time preserving components that would be useful in the evaluation of fairness.**FIG. 8**

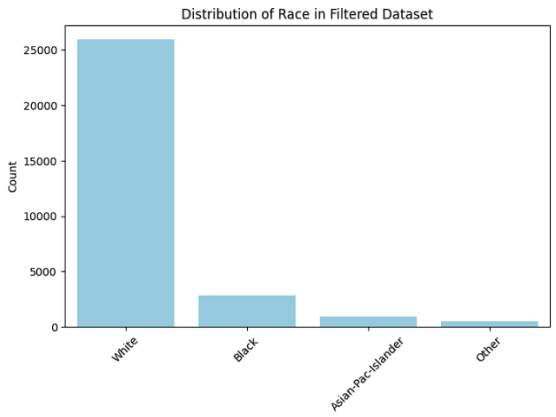

FIG. 8: Race after Grouping

C. Bias Mitigation Techniques

In this section, two distinct procedures were implemented.

In the first one, conditional oversampling of the dataset was performed using a tool known as Synthetic Data Vault (SDV), which learns from the original data and generates new data that aligns with its underlying patterns and trends. This technique made it possible to balance instances of male and female participants in the sample. Moreover, oversampling helped to solve the problem of the uneven distribution of the regions in the sample. At the start, the United States constituted 90% of the dataset, and in the final, it was cut to about 30%. The other regions were scaled up to approximately 20% each to attain a reasonable distribution of the data across the regions. However, this process led to high imbalance in the income class. The percentage of users with income greater than $50k decreased dramatically as the dataset increased. To address this, undersampling was used on the majority class (people with income less than or equal to $50k) to make the datasets balanced. The final dataset contained about 14,000 records for both groups (less than or equal to $50k and greater than $50k). The Fairlearn metrics were re-applied to evaluate the model's fairness and outcomes on the balanced dataset. The findings echoed those of the same sort in the study of New York school admissions discrimination, where a white American bias was evident. Interestingly, no significant differences were noted between the selection rates for men and women.

In the second procedure, rather than performing sex and region oversampling, a lighter oversampling was performed, with a focus on the cases with income greater than $50k. In addition, 15k cases from the USA with income less than or equal to $50k were excluded to keep the proportion of income levels. The final dataset ensured the same income distribution across all regions.

D. Conclusion

The fairness interventions had mixed results across sensitive features. For sex, the DP difference improved significantly, reducing disparity, but accuracy dropped as a trade-off, indicating fairness at the cost of performance. For region, DP difference worsened, and selection rates became imbalanced, suggesting the intervention failed to ensure fairness across regions despite maintaining accuracy. For race, the DP difference increased, highlighting worsened disparities, though selection rates became more inclusive overall. While the dataset became more inclusive in some cases, these interventions often compromised accuracy and fairness inconsistently across groups, underscoring the need for better-balanced fairness strategies. LIME (Local Interpretable Model-agnostic Explanations) is a technique used to explain machine learning model predictions by approximating them with interpretable models locally around the predicted instance, and it was used to get these results. **FIG. 9**

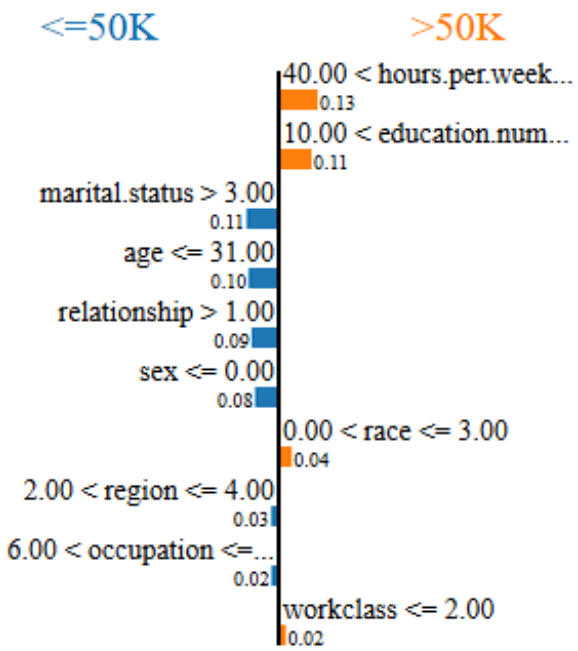

FIG. 9: Post-Processing of Data Explained Using LIME

V. SECOND APPROACH

A. EDA and data preparation

For data preparation, the dataset underwent specific transformations, most of them targeting the fight against imbalance. This process included three main steps: Data Simplification, Feature Engineering and Scaling and Encoding. **Simplification of Features**: Key features with multiple categories were mapped into binary or simplified representations. For instance, the race attribute was categorized into White and Non-White, and the native country attribute into United States and Foreign. **Feature Engineering**: Simplified features ensured better compatibility for one-hot encoding and fairness evaluation. For example, binary features retained a single column (e.g., White with values 0 or 1), eliminating redundant columns (e.g., Non-White). Data **Scaling and Encoding**: Numeric features were scaled using the standard scaler and categorical features were then encoded using one-hot encoding.

Note: Many parts of this approach are similar to the steps taken in the previous one.

B. Technical Experimentation

Using this processed data, a baseline XGBoost model was developed, and the obtained accuracy as well as fairness metrics, including selection rates, and fairness indicators such as Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD).

- **Accuracy**: 86.9
- **Disparate Impact**: Race: 0.5869096595502414. Sex: 0.3651704864740436. Region: 0.7602075054674209.
- **Average TPR (True Positive Rate**: Average Privileged groups. TPR: 0.65870.6587 Average Unprivileged groups TPR: 0.61670.6167

C. Bias Mitigation Techniques

To address the challenge of class imbalance and improve fairness, the ADASYN (Adaptive Synthetic Sampling) technique was applied as a targeted oversampling method. ADASYN works by generating synthetic samples for underrepresented classes, particularly focusing on groups with sensitive attributes such as Race, Sex, and Region, while also accounting for the positive income class label. This process ensures that the training data achieves a more balanced representation across these key features, which is essential for reducing bias in model predictions.

The ADASYN algorithm was configured to focus on the joint sensitive class distributions, meaning that oversampling prioritized groups that were simultaneously underrepresented across multiple sensitive attributes and income levels. This approach aimed not only to equalize the representation of sensitive groups but also to enhance the model's ability to learn from these balanced samples. The resulting rebalanced dataset was carefully analyzed to confirm that the distributions across sensitive attributes and labels were appropriately adjusted.

Following this preprocessing step, a new XGBoost model was trained on the oversampled dataset. The model was then evaluated across key performance metrics, including accuracy, F1-score, and recall, to ensure that predictive performance was not compromised. Fairness metrics such as Demographic Parity Difference (DPD), Equalized Odds Difference (EOD), and selection rates were recalculated to assess the impact of ADASYN on mitigating bias. The results showed a clear improvement in fairness metrics, with reduced disparities across sensitive groups, while maintaining high levels of accuracy and other performance metrics. This demonstrates the effectiveness of ADASYN as a tool for promoting equity in model outcomes without sacrificing predictive quality.

D. Conclusion

The application of ADASYN to oversample the dataset yielded insightful results, with notable improvements in fairness metrics while maintaining accuracy. The fairness interventions led to significant changes in metrics such as Demographic Parity Difference (DPD) and Equalized Odds Difference (EOD), particularly across sensitive features like Race, Sex, and Region. These metrics demonstrated the efficacy of the oversampling approach in achieving a balance between fairness and performance.

Here are some visualizations highlighting the balance/imbalance after the ADASYN application, improvements in fairness and maintenance of accuracy through the experiments:

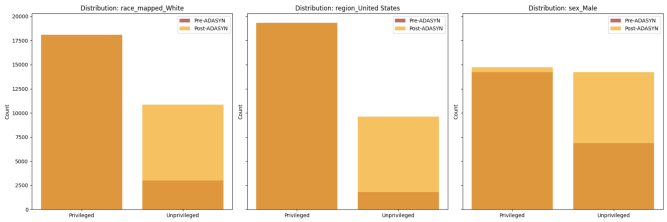

FIG. 10: Class Distribution Comparison: Pre-Oversampling vs. Post-Oversampling.

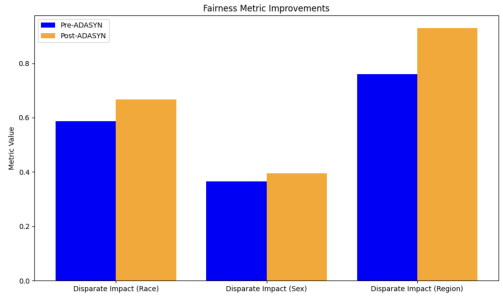

FIG. 11: Disparate Impact Evaluation Across Sensitive Features. The metric was plotted pre- and post-ADASYN, showcasing the reduction in bias while maintaining accuracy.

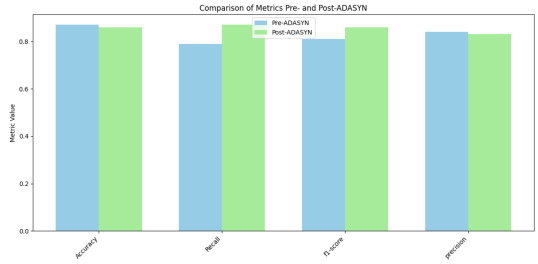

FIG. 12: Comparison of metrics pre- and post-ADASYN, showing consistent performance across accuracy, recall, F1-score, and precision.

VI. REFERENCES

- *The Markup - NYC's School Algorithms Cement Segregation* https://themarkup.

org/machine-learning/2021/05/26/nycs-school-algorithms-cement-segregation-this-data-shows-how

- *Algorithmic Fairness in Education:* https://arxiv.org/abs/2007.05443
- *The Markup - How We Investigated NYC High School Admissions:* https://themarkup.org/show-your-work/2021/05/26/how-we-investigated-nyc-high-school-admissions
- *Adult Census Income Dataset:* https://www.kaggle.com/datasets/uciml/adult-census-income
- *Fairlearn:* fairlearn.org
- *XGBoost:* xgboost.readthedocs.io/en/stable/
- *LIME:* https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanation
- *ADASYN:* https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.ADASYN.html
- *Aequitas Flow:* https://github.com/dssg/aequitas