ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING MINI PROJECT REPORT

DEPARTMENT OF INFORMATION SCIENCE AND

TECHNOLOGY

COLLEGE OF ENGINEERING,

GUINDY ANNA UNIVERSITY,

CHENNAI

SUBMITTED BY:

BRITTNEY LAUREN (2022179020)

KAMALESHWARI (2022179035)

PRATHIYUSHA (2022179037)

SUBMITTED TO:

MR. M. DEIVAMANI

(ASSISTANT PROFESSOR

DIST)

TABLE OF CONTENTS

Introduction

In the realm of artificial intelligence, the effectiveness and fairness of AI models depend significantly on the quality of the data and the preprocessing steps applied to it. Data preprocessing, a pivotal stage in AI development, involves refining and transforming raw data for model training. However, lurking within this process are potential biases that, if unaddressed, can lead to skewed outcomes, posing ethical and practical challenges. This project focuses on the vital task of "Detecting and Tackling Biases in Data Preprocessing for Responsible AI." Leveraging the concepts of labelling bias and selection bias, we aim to mitigate these biases, ensuring not only improved model performance but also ethical AI development.

This introduction underscores the gravity of biases in data preprocessing and their implications for fairness and ethical AI development. To confront this challenge, our project employs a multifaceted approach, drawing on the concepts of labelling bias and selection bias. By identifying and addressing these biases, we aim not only to enhance the robustness and fairness of AI models but also to contribute to the broader discourse on responsible AI development.

The proposal of utilizing labelling bias and selection bias as key tools in our endeavor emphasizes a commitment to a nuanced and comprehensive strategy for bias detection and mitigation. As we delve into the intricacies of these methodologies, we seek not only to improve the technical aspects of AI development but also to reinforce the ethical foundations that underpin responsible AI. Through this project, we aspire to pave the way for AI systems that not only excel in performance but also stand as beacons of fairness, transparency, and ethical responsibility in the ever-evolving realm of artificial intelligence.

Background

In recent years, the widespread integration of artificial intelligence (AI) systems into various facets of our lives has underscored the critical importance of addressing biases inherent in the development and deployment of these technologies. As AI models increasingly influence decision-making processes, the potential for biased outcomes has become a focal point of concern. One pivotal stage in the AI development pipeline that significantly influences model performance and fairness is data preprocessing.

Data preprocessing involves the cleaning, transformation, and preparation of raw data before it is fed into machine learning models. Despite its foundational role, this stage is susceptible to biases that may stem from historical data collection practices, human annotation, or inherent societal prejudices. Biases in data preprocessing can lead to skewed model outcomes, perpetuating or exacerbating existing disparities and posing ethical and practical challenges.

Recognizing the imperative to ensure responsible and fair AI, this research project delves into the nuanced complexities of bias detection and mitigation during the data preprocessing phase. The project acknowledges that a thorough understanding of biases at various stages of AI development is indispensable, and it places particular emphasis on the need for expertise in the specific domain under consideration.

The proposed research aims to address two primary forms of bias: selection bias and labeling bias. Selection bias, which arises when training data fails to represent the entire population, can significantly impact the generalizability and fairness of AI models. On the other hand, labeling bias, stemming from incorrect or subjective labels, can introduce unfairness into model predictions.

In response to these challenges, the research project outlines a comprehensive framework that encompasses methods for identifying and analyzing biases, proposing effective mitigation techniques, and ensuring ethical and legal compliance during data preprocessing. By exploring automation for bias detection and correction, as well as benchmarking against industry standards, the project seeks to contribute to the ongoing discourse on responsible AI development.

Through an in-depth examination of diverse datasets and the application of sophisticated metrics and algorithms, this research project aims to provide practical insights into the complexities of bias in data preprocessing. The ultimate goal is to contribute valuable knowledge and methodologies that will empower developers, data scientists, and policymakers in their pursuit of building fair, transparent, and ethically sound AI systems.

Data

Data Preparation

Selection Bias

1. Adult
2. BodyPerformance
3. Credit Score
4. Crime
5. Heart Failure
6. Insurance
7. LawSchool
8. Spotify
9. Student
10. Titanic

Labelling Bias

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases

1. Diabetes Dataset

Approach Used

Selection Bias

Bias Measurement methods:

Statistical parity: Computes the difference in success rates between the protected groups. Values below 0 are considered unfair towards group_a while values above 0 are considered unfair towards group_b, the range (-0.1, 0.1) is considered acceptable.

Disparate Impact: Shows the ratio of success rates between the protected groups for a certain quantile. Values below 1 are unfair towards group_a. Values above 1 are unfair towards group_b. The range (0.8, 1.2) is considered acceptable.

Four Fifths: Computes the ratio of success rates between the protected groups. Values below 1 are considered unfair while a range between (0.8, 1) is considered acceptable.

Cohen D: Computes the normalised statistical parity between the protected groups. Values below 0 are considered unfair towards group_a while values above 0 are considered unfair towards group_b.

Equality of opportunity difference: Computes the difference in true positive rates between the protected groups. Values below 0 are considered unfair towards group_a while values above 0 are considered unfair towards group_b.

False positive rate difference: Computes the difference in false positive rates between the protected groups, negative values indicating bias against group_a while positive values indicating bias against group_b.

Average Odds Difference: Computes the difference in average odds between the protected groups, negative values indicating bias against group_a while positive values indicating bias against group_b, a range between (-0.1, 0.1) is considered acceptable.

Accuracy Difference: Computes the difference in accuracy of predictions for the protected groups, positive values show bias against group_b while negative values show bias against group_a.

classification_bias_metrics: function allows us to select which metrics we want to calculate, if equal_outcome, equal_opportunity or both, where equal_outcome shows how disadvantaged groups are treated by the model and equal_opportunity shows if all the groups have the same opportunities.

multiclass_bias_metrics: function allows us to select which metrics we want to calculate, if equal_outcome, equal_opportunity or both, where equal_outcome shows how disadvantaged groups are treated by the model and equal_opportunity shows if all the groups have the same opportunities.

Multiclass Statistical parity: Computes the statistical parity between multiple classes and a protected attribute with multiple groups. For each group it computes the vector of success rates for entering each class, finally uses the mean or max strategies to aggregate them. Same as the 1d case, values in the range (-0.1, 0.1) are considered acceptable.

Multiclass Equality of Opportunity: Computes the matrix of error rates for each group, then computes all distances (mean absolute deviation) between such matrices, finally uses the mean or max strategies to aggregate them. Same as the 1d case, values in the range (-0.1, 0.1) are considered acceptable.

Multiclass Average Odds: Computes the matrix of error rates for each group, then averages these matrices over rows, and computes all pairwise distances between the resulting vectors, finally uses the mean or max strategies to aggregate them. Same as the 1d case, values in the range (-0.1, 0.1) are considered acceptable.

Multiclass True Positive Difference: Computes the matrix of error rates for each group, then computes all pairwise distances between the diagonal of such matrices, finally uses the mean or max strategies to aggregate them. Same as the 1d case, values in the range (-0.1, 0.1) are considered acceptable.

Average score difference: Computes the difference in average scores between the protected groups. Negative values indicate that group_a has lower average score, so bias against group_a, while positive values indicate group_b has lower average score, so bias against group_b.

Z score difference: Computes the spread in Z Scores between the protected groups, the Z Score is a normalised version of Disparate Impact.

Max Statistical Parity: Computes the maximum over all thresholds of the absolute statistical parity between the protected groups, values below 0.1 in absolute value are considered acceptable.

RMSE ratio: Computes the RMSE for the protected groups, lower values show bias against group_a while higher values show bias against group_b.

MAE ratio: Similar to the previous metric, computes the MAE for the protected groups, lower values show bias against group_a while higher values show bias against group_b.

Correlation difference: Computes the difference in correlation between predictions and targets for the protected groups, positive values show bias against group_a while negative values show bias against group_b.

clustering_bias_metrics: function allows us to select which metrics we want to calculate, if equal_outcome, equal_opportunity or both, where equal_outcome shows how disadvantaged groups are treated by the model and equal_opportunity shows if all the groups have the same opportunities.

Cluster Balance: Given a clustering and protected attribute. The cluster balance is the minimum over all groups and clusters of the ratio of the representation of members of that group in that cluster to the representation overall. A value of 1 is desired. That is when all clusters have the exact same representation as the data. Lower values imply the existence of clusters where either group_a or group_b is underrepresented.

Minimum Cluster Ratio: Given a clustering and protected attributes. The min cluster ratio is the minimum over all clusters of the ratio of number of group_a members to the number of group_b members. A value of 1 is desired. That is when all clusters are perfectly balanced. Low values imply the existence of clusters where group_a has fewer members than group_b.

Cluster Distribution Total Variation: This function computes the distribution of group_a and group_b across clusters. It then outputs the total variation distance between these distributions. A value of 0 is desired. That indicates that both groups are distributed similarly amongst the clusters. The metric ranges between 0 and 1, with higher values indicating the groups are distributed in very different ways.

Cluster Distribution KL Div: This function computes the distribution of group_a and group_b membership across the clusters. It then returns the KL distance from the distribution of group_a to the distribution of group_b. A value of 0 is desired. That indicates that both groups are distributed similarly amongst the clusters. Higher values indicate the distributions of both groups amongst the clusters differ more.

Social Fairness Ratio: Given a centroid based clustering, this function computes the average distance to the nearest centroid for both groups. The metric is the ratio of the resulting distance for group_a to group_b. A value of 1 is desired. Lower values indicate the group_a is on average closer to the respective centroids. Higher values indicate that group_a is on average further from the respective centroids.

Silhouette Difference: We compute the difference of the mean silhouette score for both groups. The silhouette difference ranges from -1 to 1, with lower values indicating bias towards group_a and larger values indicating bias against group_b.

Exposure ratio: Calculates the relation between the exposure of non-protected and protected elements from the dataset. For a fairer model we seek to have this value lower, indicating that the protected examples are gaining more exposure.

Exposure difference: Calculates the difference of exposure between the two groups this value will be zero when the protected group achieves more exposure than the non-protected group.

Bias Mitigation methods:

Correlation remover: is a pre-processing technique that applies a linear transformation to the non-sensitive features of the dataset to remove the correlation with respect to the sensitive columns. This process is done aiming to maintain as much as possible to prevent lost information

- **alpha** - parameter to control how much to filter, for alpha=1.0 we filter out all information while for alpha=0.0 we don't apply any.

Disparate impact remover:is a pre-processing technique that uses perturbation to modify the values of the features such that the distributions of privileged and unprivileged groups are close in order to increase fairness.

- **repair_level value**- parameter to control the repair amount, where 0 means no repair while 1 is full repair.

Reweighing: is an pre-processing technique that adapts the impact of the training instances by reweighing their importance according to its label and the protected attributes to ensure fairness before classification.

Disparate impact remover RS: is a preprocessing algorithm that edits feature values to increase group fairness while preserving rank-ordering within groups.

Labelling Bias

Supervised Classifier

In supervised learning, the model is trained on a labelled dataset, where each data point is associated with a correct label. Human annotators provide the labels for the training data, and the model learns to make predictions based on this labelled information. Labelling bias in supervised learning can occur if the training data is biased or if the labelling process introduces bias.

Semi-Supervised Classifier

Semi-supervised learning involves a combination of labelled and unlabelled data for training. A smaller portion of the dataset is labelled, and the model learns from both the labelled and unlabelled examples. In the context of labelling bias, semi-supervised learning can help mitigate bias by leveraging a larger pool of unlabelled data, which may be less biased compared to the labelled subset.

Unsupervised Classifier

Unsupervised learning is used when the training data is not labelled, and the algorithm must find patterns or structure in the data without explicit guidance. Since there are no labelled examples, biases introduced through the labelling process are not a concern in unsupervised learning. However, biases may still exist in the data itself, and unsupervised learning methods can inadvertently learn and propagate these biases.

Results

Selection Bias

As we can see from the CorrelationRemover method, we are able to get closer to the reference when we increase the alpha parameter, but we need to keep in mind that, as the alpha parameter value increases, the information filtered will be higher, but accuracy will also be diminished. This also applies to the other methods. In the case of a proportional decrease between accuracy and fairness, the dataset is already almost fair.

The Choice of model parameters depends on our main objective, whether fairness or accuracy is our goal.

Labelling Bias

The dataset initially exhibits class imbalance, and efforts have been made to address this issue by adjusting class weights. Three different classifiers (supervised, semi-supervised, and unsupervised) have been employed, each with its own set of class weights, indicating an exploration of different strategies to handle imbalanced data. Clustering has also been performed, leading to the identification of four unique clusters.

Discussion and Challenges

Data preprocessing has a big impact on model performance and fairness in the field of artificial intelligence. The project aims to tackle biases during this crucial stage by identifying and mitigating biases, ensuring fairness and equity in resulting AI models. The proposal encompasses detailed methods, including bias detection, mitigation techniques, data imputation strategies, and ethical considerations. It also explores automated bias detection, performance evaluation, and benchmarking against industry standards.

Selection bias is addressed by testing model predictive performance across different subsets of data. Various datasets, including Adult, Heart Failure, and Titanic, are employed for bias measurement using statistical parity, disparate impact, and other metrics. Bias mitigation methods involve correlation remover, reweighing, and disparate impact remover. The project uses graphs and metrics to visualize the impact of methods like CorrelationRemover on binary classification, multi-classification, and regression.

Label bias is addressed in the context of supervised, semi-supervised, and unsupervised classifiers using the Diabetes dataset. Supervised learning involves training on labeled data, while semi-supervised learning mitigates bias by incorporating a larger pool of unlabeled data. Unsupervised learning, free from labeling bias, discovers patterns without explicit guidance.

The project highlights the effectiveness of methods like CorrelationRemover in achieving fairness, while emphasizing the trade-off between accuracy and fairness. The choice of model parameters depends on the primary objective—whether it is maximizing accuracy or ensuring fairness. In the context of labeling bias, efforts to address class imbalance in the Diabetes dataset involve adjusting class weights across various classifiers and exploring clustering strategies. The overall project underscores the importance of comprehensive approaches to detect and mitigate biases in data preprocessing for responsible AI. Below are some of the difficulties we encountered.

- Despite initial challenges in sourcing a suitable real-world dataset, the project proceeded with the creation of a synthetic dataset to ensure controlled data characteristics and meet specific project requirements.
- Due to the unavailability of an appropriate real dataset, a synthetic dataset was generated, allowing for precise customization of data features and overcoming limitations in existing data sources.
- Measurement and mitigation techniques were not effective for non-biased datasets. The outcome was unfavorable.

Conclusion

In conclusion, this project successfully fulfills its objective of providing a comprehensive set of validated techniques and tools for detecting and mitigating biases, with a specific focus on selection and labeling biases during data preprocessing. By addressing biases at their root, this initiative promotes responsible AI development, empowering practitioners to create more inclusive and equitable AI models. The outcomes of this endeavor mark a significant step forward in the pursuit of ethically sound artificial intelligence, as the tools and techniques provided serve as a foundation for fostering fairness and transparency in the development process. As a result, this project contributes substantially to the ongoing efforts to establish responsible AI practices and ensures that the impact of biases is proactively minimized, paving the way for a more socially responsible and unbiased artificial intelligence landscape.

References

- https://blogs.oracle.com/ai-and-datascience/post/4-approaches-to-overcoming-label-bias-in-positive-and-unlabeled-learning
- https://holistic-ai.readthedocs.io/en/latest/
- https://www.kaggle.com/code/liananapalkova/getting-started-with-responsible-ai/notebook
- https://ml.auckland.ac.nz/identification-and-mitigation-of-selection-bias/
- https://leena.ai/blog/mitigating-bias-in-ai/#:~:text=What%20does%20it%20mean%20to,%2C%20or%20 decision%2d Making%20 processes.