Groundwork

What is bias in ML datasets?

Bias in the ML dataset indicates prejudices and unfair inaccuracies[1] in the data that is used to train ML models. In simple terms, certain elements of a dataset are overweighted or overrepresented. This bias can have a significant impact on the fairness of ML models[2].

Problems with biased datasets

Apart from the technical problems of having ML models having incorrect outcomes, it also creates a major issue from a systemic point of view where a certain age, race, culture, or sexual orientation faces discrimination.
A real-time example is flawed Amazon's hiring system. Upon research, it was found that Amazon's ML model to select the top 5 candidates was discriminating against Women by deteriorating resumes that included the word "Women" in the resume.[3]

Existing tools that are used to identify bias

1. **AuditAI**: AuditAI is an open-source tool that helps you assess dataset bias. It provides various metrics and visualizations to identify bias in your data based on different attributes. https://github.com/pymetrics/audit-ai
2. **IBM AI Fairness 360**: IBM's AI Fairness 360 is an open-source toolkit that offers a comprehensive set of metrics and algorithms to detect and mitigate bias in AI and machine learning models. It includes tools for bias detection, bias mitigation, and bias visualization. It uses the metric *Disparate impact ratio* which is defined as the ratio of the rate of favorable outcomes for the one group to the rate of favorable results for the other group, the two groups (unprivileged and privileged) predetermined by the evaluator or surfaced by some other method like the Multi-Dimensional Subset Scan(MDSS) https://github.com/Trusted-AI/AIF360
3. **Google's What-If Tool**: This tool allows you to analyze the fairness of machine learning models and datasets. It provides a user-friendly interface to explore different scenarios and understand how model predictions may be influenced by different groups within the data. It has an interactive dashboard and requires less coding. https://pair-code.github.io/what-if-tool/get-started/
4. **Aequitas**: Aequitas is an open-source bias audit toolkit specifically designed for the machine learning workflow. It helps identify bias in various stages, from data preparation to model evaluation. http://aequitas.dssg.io/
5. **Fairlearn**: Fairlearn is a Python library developed by Microsoft that focuses on fairness in machine learning. It provides metrics and algorithms to assess and mitigate bias in models. https://fairlearn.org/
6. **scikit-learn**: Scikit-learn, a popular machine learning library, includes tools for calculating fairness-related metrics. While it doesn't provide comprehensive bias detection, it's useful for getting started with fairness analysis
7. **FairTest**: FairTest is a Java-based library for detecting and quantifying discrimination in machine learning models. It's particularly useful for evaluating model fairness on structured datasets

8. **Amazon SageMaker Clarify**: Amazon SageMaker Clarify is a tool that can help identify bias in your data and models. It offers pre-built bias detection and explanation capabilities. It is a paid service from Amazon. https://aws.amazon.com/sagemaker/data-wrangler/
9. **Commercial Tools**: Some AI and machine learning platforms, such as DataRobot and H2O.ai, offer built-in bias detection features as part of their toolsets
10. **Custom Scripts and Data Analysis**: Depending on the complexity of the dataset and the specific bias developers are concerned about, ML experts develop custom scripts or conduct in-depth data analysis to identify subtle forms of bias

Important features a tool detecting biases in ML datasets should have

1. **Bias Metrics**: The tool should offer a range of bias metrics that can capture different aspects of bias, such as disparate impact, disparate treatment, and disparate mistreatment. These metrics help assess fairness from various angles. The data should be tested for equal parity, Proportional parity, False positive parity, and False Negative parity[4].
2. **Explainability**: It should provide explanations and insights as to why the data is biased, such as highlighting the specific features or data points contributing to bias
3. **Compatibility with Frameworks**: The tool should be compatible with common machine learning frameworks and libraries like TensorFlow, PyTorch, sci-kit-learn, etc
4. **Demographic selection**: The tool should allow users to specify and analyze bias with respect to different demographic groups, such as race, gender, age, and other protected attributes. It should also support multiple options to identify biases that affect multiple groups simultaneously
5. **Threshold Setting**: This is the most important feature of the tool. It should allow users to set thresholds for bias metrics to define what constitutes an unacceptable level of bias. This enables users to judge based on specific use cases and fairness requirements.

Nice to have features

1. **Visualization**: Visualizations can make it easier to understand bias results. It is especially beneficial for users who do not know coding. The main advantage of visualizing data is to quickly identify bias without even understanding the dataset
2. **Data Preprocessing**: Incorporating data preprocessing techniques such as reweighting, re-sampling, or data augmentation can be valuable. The tool could support or suggest preprocessing methods to mitigate bias
3. **Statistical hypothesis testing**: The tool should incorporate statistical tests to assess the significance of observed bias, helping users identify random fluctuations and systematic bias

Detailed look at considered tools with UI

1. IBM AI Fairness 360:
    - IBM AI Fairness 360 is an open-source toolkit that aims to help developers and data scientists detect and mitigate bias in machine learning models.

    - It provides a collection of algorithms and metrics for fairness assessment, including reweighing, adversarial debiasing, and disparate impact remover.
    - AI Fairness 360 is compatible with various machine learning frameworks, making it versatile and accessible.
2. Google's What-If Tool:
    - Google's What-If Tool is a visual tool designed to help users understand the behavior of machine learning models, including fairness-related aspects.
    - It allows users to explore model predictions across different groups and visualize the impact of various input features on predictions.
    - While it doesn't provide automated fairness mitigation techniques, it is a valuable tool for model interpretability and fairness exploration.
3. Aequitas:
    - Aequitas is an open-source bias and fairness auditing library specifically tailored for assessing and mitigating bias in machine learning models applied to various domains, including criminal justice and lending.
    - It offers features for bias measurement, visualization, and reporting, allowing users to identify and address bias in their models.

Usability experiment for the above-mentioned tools in this document

References

1. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. P. J. Kenfack, A. M. Khan, S. M. A. Kazmi, R. Hussain, A. Oracevic and A. M. Khattak, "Impact of Model Ensemble On the Fairness of Classifiers in Machine Learning," 2021 International Conference on Applied Artificial Intelligence (ICAPAI), Halden, Norway, 2021, pp. 1-6, doi: 10.1109/ICAPAI49758.2021.9462068.
3. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
4. http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/