Fairness in Data

Ananya Balagonda, Shreshika Bommana, Alex Cai, Jiya Chachan, Jason Choe

1. Methodology

1.1 Evaluation Section B111 - Equity in Data

“Equity in Data (B111) evaluates the equality of data collection practices to confirm that no group is unfairly favored or disadvantaged” [1]. It further assesses whether the dataset fairly represents diverse groups across key factors such as gender, age, race, income, education, location, and disability. The goal is to prevent systematic bias and ensure that AI systems are trained on inclusive and representative data.

Criteria	**Description**	**Scoring Rubric**	**Dataset Score**
Gender	Indicates if the dataset is diverse/representative for different gender identities	**+1**: Equal gender representation, **0**: Slightly skewed to one gender, **-1**: Highly skewed to one gender	**0**
Age	Indicates diversity of age group representation in data	**+1**: Equal age representation, **0**: Slightly skewed to 1 age group, **-1**: Highly skewed to 1 age group	**0**
Race	Indicates whether the dataset has a diverse set of races represented	**+1**: Equal representation of race, **0**: Slightly skewed representation of race to 1 group, **-1**: Highly skewed representation of race to 1 group	**-1**
Income	Indicates whether the dataset has a diverse range of income levels	**+1**: Equal representation across income levels, **0**: Slightly skewed representation of income levels, **-1**: Highly skewed representation of income levels	**-1**
Education	Indicates whether the dataset has a diverse range of education levels	**+1**: Equal representation across education levels, **0**: Slightly skewed representation of education levels, **-1**: Highly skewed representation of education levels	**-1**
Location	Indicates if the dataset has a diverse range of locations it has been sampled from.	**+1**: Equal representation across different locations, **0**: Slightly skewed representation of different locations, **-1**: Highly skewed representation of different locations.	**-1**
Disability	Indicates if the dataset has disability as a diverse feature.	**+1**: Includes data from people with disabilities, **-1**: Doesn’t include data from people with disabilities	**-1**

1.2 Evaluation Section B112 - Bias Detection, Data

”Bias Detection in Data (B112) measures potential biases within the dataset, allowing for early identification and correction” [1]. This section includes a variety of different bias detection methods, starting with the existence of documentation detailing how data was collected. It checks for diverse representation and outliers, while also considering biases that don't fit under other evaluation sections.

Criteria	Description	Scoring Rubric	Dataset Score
Sourcing Metadata	There exists documentation regarding how the data was collected from fair and representative places.	**+1**: Metadata exists, and data was collected from a representative area, **0**: Metadata is not present, **-1**: Metadata exists, and data was collected from a non-representative area	**0**
Diversity Detection	Of the 7 diverse layers in Section B111, all of them should be representative.	**+1**: 6 or 7 of the requirements are fulfilled., **0**: 4 or 5 of the requirements are fulfilled., **-1**: 3 or less of the requirements are fulfilled.	**-1**
Examination of Outliers	Outliers in the data are examined and have reasonable explanations.	**+1**: Outliers have reasonable explanations., **0**: There are no outliers in the dataset., **-1**: Outliers do not have reasonable explanations.	**1**
Assignment Bias	If data is collected using an experiment, groups are assigned at random.	**+1**: Groups are assigned at random, **0**: Data is not collected via experimental format, **-1**: Groups are not assigned at random	**0**
Self Serving Bias	If data is collected with individuals reporting facts about themselves, the fact that they could be over-reporting their abilities is taken into consideration.	**+1**: Self-serving bias is noted as being taken into consideration., **0**: Data is not collected by surveying individuals about themselves., **-1**: Self-serving bias is not clearly noted as taken into consideration.	**-1**

1.3 Evaluation Section B113 - Sampling Integrity, Data

"Sampling Integrity (B113) inspect the methods used to gather data, ensuring they represent the target population accurately and fairly" [1]. This section considers the various biases that can come into effect when data is being collected and includes the use of randomization to generate a representative sample. Samples should be representative of the greater population of interest that takeaways will be generalized to.

Criteria	Description	Scoring Rubric	Dataset Score
Population of Interest (PoI)	The dataset should have a clear population of interest that takeaways from the data can be generalized to.	**+1**: PoI is stated and the data is representative of it., **0**: The PoI is stated but it is unclear if the data is representative., **-1**: The PoI is stated and the data is not representative OR no PoI given.	**1**
Random Sampling Methods	If sampling is used, a random method is used to determine sampling.	**+1**: Random sampling is used to collect data, **0**: Sampling is not used to collect data, **-1**: Random sampling is not used to collect data.	**0**

Sampling Bias	The data is sampled without having certain characteristics more likely to be sampled than others.	**+1**: Sampling bias is considered, and sampling is done fairly. **0**: There are no signs of sampling bias being considered. **-1**: Sampling bias causes the random sampling to be done unfairly.	**1**
Response Bias	The data is sampled without any pressure to answer a certain way (ex: no leading questions).	**+1**: Response bias is considered, and sampling is done fairly. **0**: There are no signs of response bias being considered. **-1**: Response bias causes the random sampling to be done unfairly.	**0**
Measurement Bias	If tools used to collect quantitative data (ex: thermometer) are used, tools measure properly.	**+1**: Measurement tools are accurate and functional. **0**: No tools are used to collect measurement data, **-1**: Measurement tools are inaccurate when collecting data.	**0**

1.4 Evaluation Section B114 - Fairness Metrics, Data

“Fairness metrics for data (B114) evaluates the fairness levels within the data to ensure they meet predefined fairness standards” [1]. It assesses the fairness levels within a dataset to ensure that no group is unfairly favored or disadvantaged. This means having fairness checks across gender, age, race, income, and other attributes and checking equal opportunity. The purpose is to ensure that AI systems do not perpetuate existing inequalities through biased data.

Criteria	**Description**	**Scoring Rubric**	**Dataset Score**
Demographic Equality	Equal representation for all demographic groups (e.g., gender/race).	**+1**: Equal representation. **0**: Slight skew. **-1**: Significant skew towards one group.	1
Equal Opportunity	Ensures each group has equal access to favorable outcomes (e.g., income, job opportunities).	**+1**: Equal access for all. **0**: Slight inequality. **-1**: Significant inequality.	-1
Statistical Equality	Measures if the probability of being included is the same for all groups.	**+1**: Equal probability. **0**: Slight inequality. **-1**: Significant underrepresentation.	1
Unequal Impact	Checks if any group faces disproportionate negative outcomes.	**+1**: No harm. **0**: Minor harm. **-1**: Significant harm.	-1
Fairness in Outcome Distribution	Ensure outcomes (e.g., income) are fairly distributed across groups.	**+1**: Equal outcome distribution. **0**: Slight skew. **-1**: Highly skewed.	-1

1.5 Evaluation Section B211 - Fairness, Data

"Fairness in Data (B211) evaluates the data for fairness, ensuring no discriminatory biases are present" [1]. This evaluation area focuses on identifying whether the dataset may result in unequal outcomes for different groups based on how the data is structured, labeled, or distributed. It considers if the data reflects inherent or historical biases and whether these biases could lead to unjust or skewed model behavior. The goal is to ensure the AI system is trained on data that treats all demographic groups equitably, minimizing the risk of perpetuating discrimination through automated decisions.

Criteria	Description	Scoring Rubric	Dataset Score
Representation Check	Checks for balanced demographic representation across key groups	**+1:** Data includes full demographic diversity and is statistically balanced. **0:** Some diverse representations, but incomplete or skewed. **-1:** Data excludes key demographic groups entirely.	**0**
Source Transparency	Evaluates whether the origin of the data is disclosed and includes fairness-related context	**+1:** Data source and demographic context are fully documented. **0:** Source is known but lacks fairness-related context. **-1:** Source is undocumented or lacks demographic transparency.	**1**
Preprocessing Bias Review	Looks at whether data preprocessing steps were evaluated for bias introduction	**+1:** Bias checks and documentation are included in preprocessing. **0:** Some mitigation attempted but lacks clarity. **-1:** No fairness considerations during preprocessing.	**-1**
Data Collection Equity	Assesses whether data was collected equally across populations and settings	**+1:** Collection ensured equal opportunity across demographics. **0:** Efforts were made, but the result was uneven. **-1:** Collection favors specific groups, introducing sampling bias.	**0**
Temporal Fairness	Checks if dataset remains fair over time, accounting for potential concept drift or temporal bias	**+1:** Dataset was reviewed/updated to ensure fairness over time. **0:** Temporal issues acknowledged but not mitigated. **-1:** Dataset is outdated or ignores temporal bias.	**-1**

1.6 Evaluation Section B212 - Impact Assessment, Data

"Impact Assessment, Data (B212) evaluates the potential effects of data handling and processing decisions on different demographic groups" [1]. This evaluation area focuses on whether the dataset's use has been critically examined for downstream social consequences, especially in high-stakes or sensitive contexts. It assesses whether risks are identified, documented, and addressed, and whether there are mechanisms in place to gather feedback post-deployment. The goal is to ensure that the dataset does not unintentionally reinforce harm or inequity and that its impact on real-world users, particularly marginalized communities—is actively considered and mitigated.

Criteria	Description	Scoring Rubric	Dataset Score
Demographic Impact Analysis	Assesses whether the dataset's use has been analyzed for potential impacts on different demographic groups.	**+1:** A thorough impact analysis was conducted across demographic lines. **0:** Some analysis exists but lacks depth or coverage. **-1:** No demographic impact analysis provided.	**-1**
Use Case Sensitivity	Checks if the dataset's social implications were considered relative to its intended application.	**+1:** Use case analysis identifies sensitive uses and risks. **0:** Basic consideration given to application context. **-1:** No attention to social sensitivity or application consequences.	**-1**
Documentation of Risks	Checks whether known risks associated with the dataset are openly documented.	**+1:** Potential harms are clearly documented and discussed. **0:** Risks are mentioned but not thoroughly detailed. **-1:** No documentation of potential risks.	**-1**
Contextual Relevance	Check whether data collection and attributes are appropriate to the social context in which the AI system operates.	**+1:** Contextual relevance is explicitly assessed and justified. **0:** Context is briefly mentioned or assumed. **-1:** Data context is unclear or mismatched.	**0**
Feedback Mechanism	Looks at whether there's a process to collect feedback on data impacts post-deployment.	**+1:** A formal process exists to monitor and address impact feedback. **0:** Feedback is informally or occasionally considered. **-1:** No mechanism for post-deployment feedback.	**-1**

1.7 Evaluation Section B311 - Bias Mitigation, Data

"Bias Mitigation for Data (B311) inspect using bias mitigation techniques such as data resampling or reweighing to mitigate bias in the data [1]." This section evaluates the techniques used for mitigating bias in the dataset. Common approaches include resampling or reweighting data to correct for imbalances, which ensures that the dataset better represents the diverse groups it is intended to model.

Criteria	Description	Scoring Rubric	Dataset Score
Use of Resampling Techniques	Evaluates whether the dataset was resampled to correct imbalances across different groups	**+1**: Resampling techniques (e.g., SMOTE) were applied successfully and balanced the dataset. **0**: Some resampling applied, but it didn't sufficiently balance the dataset. **-1**: No resampling was applied or attempted.	**0**

Use of Reweighting Techniques	Checks if the dataset was reweighed to give more importance to underrepresented groups or categories	**+1**: Reweighting techniques were applied successfully, ensuring fairness. **0**: Some reweighting applied, but it was not significant or effective. **-1**: No reweighting was applied to address imbalance.	**0**
Bias Identification and Documentation	Assesses whether the dataset includes clear documentation identifying biases and their potential impacts	**+1**: Biases in the dataset were clearly identified and documented with potential impacts discussed. **0**: Some biases identified, but documentation is limited or unclear. **-1**: No bias identification or documentation was present.	**0**
Evaluation of Bias Impact	Reviews of how the bias in the data was evaluated for potential harmful effects, particularly for marginalized groups	**+1**: Bias impact on marginalized groups was evaluated comprehensively and documented. **0**: Some evaluation but lacked depth or coverage. **-1**: No evaluation of bias impact was conducted.	**0**
Post Mitigation Monitoring	Checks whether there is an ongoing process to monitor and adjust for bias after the mitigation techniques have been applied	**+1**: There is a process in place to monitor and adjust for bias after mitigation. **0**: Some monitoring of bias is done, but it's informal or limited. **-1**: No post-mitigation monitoring is in place.	**0**

1.8 Evaluation Section B312 - Rebalancing Techniques

"(B312) evaluates the weight adjustment or presence of certain data points to substantiate a balanced and representative dataset. Bias Mitigation for Models [1]." This section assesses whether data rebalancing techniques have been used to adjust the weight or presence of certain data points, ensuring a more balanced and representative dataset.

Criteria	**Description**	**Scoring Rubric**	**Dataset Score**
Rebalancing of Sample Distribution	Evaluates whether the dataset's sample distribution across various categories (gender, age, race, etc.) was rebalanced to ensure equal representation	**+1**: The sample distribution was rebalanced effectively across all categories, ensuring equal representation. **0**: The sample distribution was slightly rebalanced but is still imbalanced in some areas. **-1**: No rebalancing was applied, and the distribution remains highly imbalanced.	**0**
Adjustments for Imbalance in Key Attributes	Examines if adjustments were made to the dataset to correct imbalances in important attributes, such as income, education, or geographic location	**+1**: Adjustments were made to ensure balanced representation of key attributes (e.g., income, education). **0**: Some adjustments were made but had limited impact on key attributes. **-1**: No adjustments were made to address imbalance in key attributes.	**0**

Impact of Rebalancing on Fairness	Assesses how the rebalancing techniques impacted the fairness of the dataset across different demographic groups	**+1**: Rebalancing improved fairness significantly across different demographic groups. **0**: Rebalancing had minimal or no significant impact on fairness. **-1**: Rebalancing worsened fairness or had no effect on fairness.	**0**
Transparency in Rebalancing Procedures	Reviews whether the dataset includes clear documentation of the rebalancing methods used and their rationale	**+1**: Rebalancing methods used were well-documented and clearly explained. **0**: The rebalancing methods were somewhat documented, but not in detail. **-1**: Rebalancing methods were not documented or explained.	**-1**
Evaluation of Rebalancing Effectiveness	Assesses whether the effectiveness of rebalancing techniques has been evaluated post implementation to ensure balanced representation	**+1**: The effectiveness of rebalancing was evaluated thoroughly, showing measurable improvements in fairness. **0**: The effectiveness was evaluated to some extent but lacked comprehensive analysis. **-1**: No evaluation of the effectiveness of rebalancing was done.	**0**

1.9 Evaluation Section B411 - Third-Party Assessment, Data

Third Party Assessments for Data (B411) involves evaluating the external audits to verify the fairness of data handling, providing independent verification of fairness practices [1]. It focuses on the quality and credibility of the audit process, including whether a recognized methodology was followed, if the audit covered key aspects like bias and representation, and whether the auditor was independent and qualified. It also examines how often these audits occur and whether their findings have led to meaningful improvements in data practices.

Criteria	Description	Scoring Rubric	Dataset Score
Existence of Third-Party Audit	Indicates whether any external audit of the dataset was performed.	**+1**: Audit was conducted by a third party **0**: Unclear or unverifiable **-1**: No audit conducted	**-1**
Auditor Independence	Indicates if the auditor is external and unbiased	**+1**: Fully independent. **0**: Affiliated but uninvolved. **-1**: Internal or no information	**-1**
Audit Methodology	Checks if formal, recognized framework was used	**+1**: Recognized standard. **0**: Internal/unverified method. **-1**: No information or unclear process	**-1**
Fairness Scope	Indicates if the audit covers key fairness aspects in data	**+1**: Covers key areas. **0**: Partial coverage **-1**: No info or not addressed	**-1**
Audit Frequency	Checks how often third-party audits are conducted	**+1**: Regular (e.g., annually). **0**: Once only **-1**: Never or no information available	**-1**

Corrective Actions	Indicates if the audit led to meaningful improvements	**+1**: Clear actions taken. **0**: Findings noted, no action. **-1**: No action or no information available	**-1**
Auditor Qualifications	Indicates if the auditor has relevant expertise	**+1:** Proven qualifications in data/fairness **0:** General background **-1:** No information available or unclear	**-1**

1.10 Evaluation Section B412 - Transparency Reports, Data

Transparency Reports for Data (B412) evaluates the mechanism for disclosing fairness practices and outcomes to the public, ensuring accountability and transparency [1]. This area assesses if the fairness information is easily accessible online, presented in a structured and user-friendly format, and written in plain language suitable for non-experts. It also considers whether updates to the dataset or disclosures are documented, whether fairness assessment methods are explained, and whether authors or maintainers are identifiable and reachable.

Criteria	Description	Scoring Rubric	Dataset Score
Public Accessibility	Checks if the fairness-related information is shared publicly	**+1**: Freely available. **0**: Limited information available. **-1**: Not publicly accessible	**0**
Format Clarity	Indicates if the fairness-related information is presented in a user-friendly, structured format	**+1**: Clear structure **0**: Some structure, but hard to follow **-1**: Confusing or not available	**0**
Language Simplicity	Indicates if the content is readable for non-experts	**1**: Written in plain, jargon-free language **0**: Some jargon but generally understandable **-1**: Highly technical language or no information	**0**
Versioning & Updates	Indicates if the metadata includes a version of history or documentation of changes	**+1:** Versioned with update history **0:** Updates mentioned vaguely **-1:** No update/version info	**0**
Transparency About Methods	Checks of whether the methods used to assess fairness (e.g., audits, metrics) are clearly explained	**1**: Methods detailed and reproducible **0**: Somewhat described **-1**: Not explained	**-1**
Contact or Attribution	Indicates if there is an author, maintainer, or point of contact listed for follow-up?	**+1**: Author's information is available **0**: Partial information available **-1**: No contact or attribution	**1**

2. Results

Development: In the development area, the Adult Income dataset [2] scored a total of -5 points, indicating that the dataset did not consider many fairness qualities in its collection. The only section it scored positively in was B113, Sampling Integrity, as the data was collected from the US Census, garnering points in how representative it was for the population of interest. However, in terms of considering fairness and the potential impacts this data could have, the data could have adverse impacts on different groups as it is highly skewed on all levels.

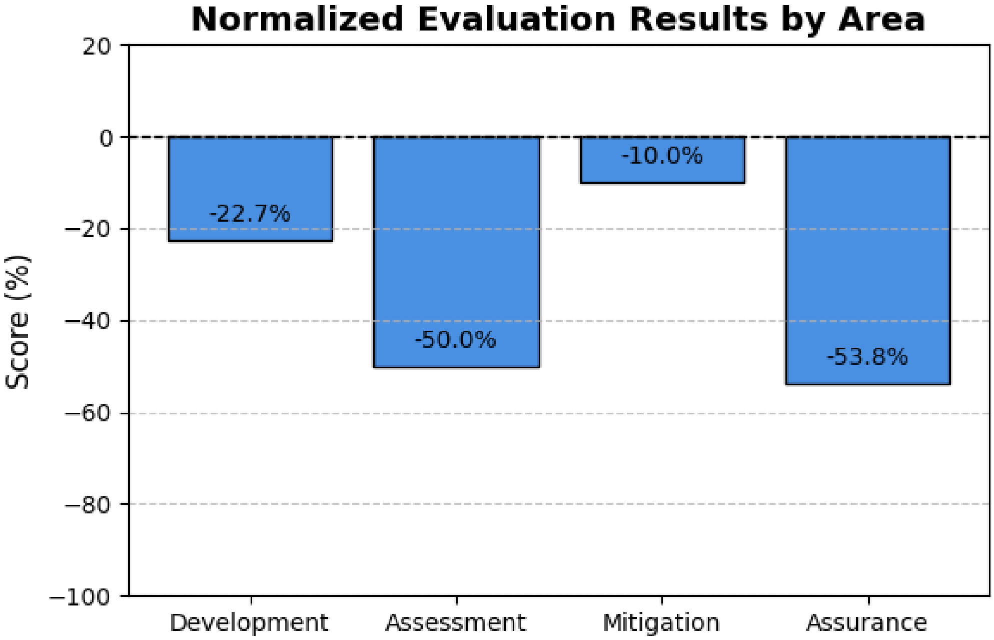
Assessment: In the assessment area, the Adult Income dataset [2] scored a total of -6 out of a total of 10 possible points, reflecting deeper structural issues tied to its legacy design and lack of responsible AI foresight. While the dataset includes diverse demographic attributes, the absence of fairness audits, bias mitigation steps, and documented impact assessments suggests it was never critically evaluated for how its structure might influence real-world outcomes. Its continued use in fairness research highlights a disconnect: the dataset serves as a benchmark for identifying bias, yet it remains misaligned with the very standards it’s used to test. This contradiction underscores the need to reassess foundational datasets not just for technical usability, but for their ethical and social implications in modern AI systems.

Mitigation: In the mitigation area, the Adult Income dataset [2] scored -1 points. For most categories, the dataset did not have any mitigation techniques applied. As this dataset was collected from census data, it makes sense that the data did not feature any mitigation in the form of rebalancing or reweighting. In the trade-off of fairness and accuracy, the census leans much more towards accuracy without considering the potential harmful and unfair impacts.

Assurance: In the assurance area, the Adult Income dataset [2] scored -7 out of +13, highlighting major gaps in both external auditing and transparency. Under B411, it received -7 out of +7 due to the absence of a third-party audit, no documented fairness verification, and no corrective actions. Under B412, the dataset scored 0 out of +6, with only minimal fairness-related disclosures such as author contact information and basic filtering criteria, but no structured reporting, clear update history, or fairness assessment methods. Overall, the dataset does not meet assurance standards for independent audit or public transparency.

Development		Assessment		Mitigation		Assurance	
Area	Score	Area	Score	Area	Score	Area	Score
B111	-5/7	B211	-1/5	B311	0/5	B411	-7/7
B112	-1/5	B212	-4/5	B312	-1/5	B412	0/6
B113	2/5						
B114	-1/5						

Total:	-5/22	Total:	-5/10	Total:	-1/10	Total:	-7/13
Overall Score: -18/55							



3. Discussion

3.1 Key Findings

The evaluation of the Adult Income dataset [2] revealed several important observations across development, assessment, mitigation, and assurance areas. In the Development area, the dataset scored -5, with only Sampling Integrity (B113) receiving a positive score, reflecting that the dataset was representative of the U.S. population but not equitable across all demographic categories. In the Assessment area, the dataset scored -5 out of 10, showing that bias detection and impact assessments were largely absent. In the Mitigation area, the dataset scored -1, as no bias mitigation techniques such as rebalancing or reweighting were applied. In the Assurance area, the dataset scored -7 out of +13, reflecting a complete absence of third-party audits (B411) and only minimal public fairness-related disclosures (B412).

While the Adult Income dataset [2] showed some positive aspects, including basic sampling integrity and the availability of author attribution, it demonstrated significant weaknesses overall. It lacked fair demographic representation, missed opportunities for bias detection and correction, had no mitigation strategies, and offered minimal transparency regarding fairness practices. Across all areas, the dataset consistently fell short of responsible

AI standards for fairness, bias mitigation, and accountability, highlighting the need for critical reassessment before continued use in fairness-related research or decision-making systems.

3.2 Interpretation of Results

The evaluation of the Adult Income dataset highlights serious concerns across the development, assessment, mitigation, and assurance areas. In development, although the dataset achieved sampling integrity by broadly representing the U.S. population, it lacked equitable demographic coverage, particularly across race and gender, as revealed in our exploratory data analysis. This finding directly reflects concerns raised in prior studies about how biased data can reproduce societal inequalities if demographic representation is not carefully managed. The significant gaps in assessment, with little evidence of bias detection or impact analysis, align with the gaps discussed in Ferrara's work [4], which emphasizes the dangers of training AI on unexamined or unbalanced datasets.

Similarly, the absence of any bias mitigation strategies, such as reweighting or causal adjustments, mirrors the challenges identified by González-Sendino et al. [3], who advocate for structured bias mitigation through methods like causal modeling. Without these measures, any structural imbalances present at the data layer persist throughout the modeling pipeline. Lastly, the dataset's poor performance in assurance—marked by no third-party audits or transparency practices, the need for robust documentation and accountability measures emphasized in fairness research across healthcare, finance, and NLP domains [5][6][7]. Our results thus show that the Adult Income dataset, despite its popularity, falls short of responsible AI standards. Without significant reassessment, continued use risks reinforcing systemic bias, undermining fairness objectives emphasized across recent literature.

3.3 Recommendations for Improvement

Recommendation 1: Apply resampling or reweighting techniques for bias mitigation. Use class weights in models (e.g., class_weight='balanced' in RandomForest or LogisticRegression) to ensure that the model does not favor the majority class (e.g., those earning <=50K). The Adult Income dataset [2] has shown significant class imbalances, particularly in income distribution and gender representation. To improve fairness and reduce bias in predictions, it is essential to apply resampling techniques like SMOTE (Synthetic Minority Oversampling Technique) or reweighting techniques during model training. These actions will help create a more balanced dataset and train the model to be fairer, reducing the likelihood of biased predictions for marginalized groups.

Recommendation 2: Introduce fairness metrics evaluation for improved monitoring. Implement fairness checks using libraries like fairlearn to evaluate how well the model performs across different demographic groups. The dataset lacks comprehensive evaluation of fairness metrics. Currently, there are no fairness evaluations applied to assess if the model disproportionately affects certain demographic groups, such as gender, race, or income level. By including fairness metrics such as Equal Opportunity, you can ensure that the model's decisions are not biased against certain groups. This recommendation will ensure that your model performs equitably across demographic groups, improving the transparency and accountability of the AI system.

Recommendation 3: Document and implement post-mitigation bias monitoring. Set up a formal process for post-deployment bias checks to regularly evaluate the impact of the dataset on different groups. Also, implement a feedback loop to update mitigation strategies based on performance across demographic groups. The report indicates no post-mitigation monitoring after applying bias mitigation strategies. Bias mitigation is not a one-time task; it requires ongoing monitoring to assess if the dataset and model maintain fairness over time, especially as new data is collected or used. This will help maintain long-term fairness in the model's performance, ensuring that new data or unforeseen trends do not cause disproportionate harm to any group

3.4 Limitations of Your Evaluation

One limitation of the evaluation is the biases in the Adult Income dataset. Despite being based on the U.S. Census, it shows skewed representation in gender, race, and income, with underrepresentation of certain groups, such as racial minorities and lower-income individuals. Future datasets should aim for a more balanced representation, especially regarding sensitive attributes like race and income, to enhance fairness.

A second limitation is the lack of bias mitigation techniques. As the dataset was sourced from census records, no modern bias correction methods, like reweighting or resampling, were applied. This gap means the dataset's fairness evaluation may not fully capture real-world harm, suggesting the need for future datasets to document and apply bias mitigation strategies.

Finally, the absence of transparency limits the reliability of the dataset's fairness assessment. Without external validation or fairness checks, it's difficult to confirm that the dataset aligns with ethical standards. Future evaluations should include fairness audits to ensure transparency and accountability.

4. Conclusion

To sum everything up, we began by cleaning the Adult Income dataset [2]. We then evaluated it using the different layers of the SystemCard+ framework with development, assessment, mitigation and assurance. In every single one of these sections, the dataset failed to achieve a positive score, indicating that the dataset did not highly take fairness into consideration. Given that the dataset was collected from US census data, it makes sense that many of the different sections involved in modifying the dataset have it be a fair and equal representation of many different diverse groups. The Adult Income dataset's [2] failures highlight a significant deficit in many datasets: lacking consideration for the disparate impact that data can have on different stakeholders. "Objective" data like the Adult Income dataset could be used to draw harmful conclusions about different groups, and potential like that should be considered and limited to prevent harm.

For the future, those interested in researching further should investigate the specific adverse impacts that a lack of fairness consideration in data could cause. They should implement the different techniques mentioned throughout the methodology to clean and include fairness in the data such as rebalancing or reweighting techniques as well as evaluating the effectiveness of including third party evaluations like audits. By understanding the correlation between these feature’s implementations and the level of impact on different groups, a set of best practices can be developed for all those who collect data to follow to ensure harm is mitigated.

Reference

[1] H. Tibebu, ‘System Card+: Responsible AI Framework for Decision Support Systems’, Jan. 2025. doi: 10.5281/zenodo.14736359.

[2] B. Becker and R. Kohavi. "Adult," UCI Machine Learning Repository, 1996. [Online]. Available: https://doi.org/10.24432/C5XW20.

[3] Rubén González-Sendino, E. Serrano, and J. Bajo, “Mitigating bias in artificial intelligence: Fair data generation via causal models for transparent and explainable decision-making,” *Future Generation Computer Systems*, vol. 155, Feb. 2024, doi: https://doi.org/10.1016/j.future.2024.02.023.

[4] E. Ferrara, “FAIRNESS AND BIAS IN ARTIFICIAL INTELLIGENCE: A BRIEF SURVEY OF SOURCES, IMPACTS, AND MITIGATION STRATEGIES,” *FAIRNESS AND BIAS IN ARTIFICIAL INTELLIGENCE: A BRIEF SURVEY OF SOURCES, IMPACTS, AND MITIGATION STRATEGIES*, no. 2, 2023, Available: https://arxiv.org/pdf/2304.07683

[5] S. K. B, A. Chandrabose, and B. R. Chakravarthi, “An Overview of Fairness in Data – Illuminating the Bias in Data Pipeline,” *ACLWeb*, Apr. 01, 2021. https://aclanthology.org/2021.ltedi-1.5/

[6] C. N. Nwafor, O. Nwafor, and Sanjukta Brahma, “Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach,” *Scientific Reports*, vol. 14, no. 1, Oct. 2024, doi: https://doi.org/10.1038/s41598-024-75026-8.

[7] L. Hu *et al.*, “Enhancing fairness in AI-enabled medical systems with the attribute neutral framework,” *Nature Communications*, vol. 15, no. 1, Oct. 2024, doi: https://doi.org/10.1038/s41467-024-52930-1.