**Literature Review Report:**

**Automated Data Cleaning Can Hurt Fairness in Machine Learning-based Decision Making**

*Group 19*

*Lisa He, Chris Chen, Rudra More*

**Introduction:**

Guha et al. investigates the relationship between data quality issues and the impact of automated cleaning techniques on sensitive demographic attributes in their publication *Automated Data Cleaning Can Hurt Fairness in Machine Learning-based Decision Making*. The paper begins by highlighting the ubiquity of ML systems in critical decision-making areas like loan approvals, hiring, and medical interventions. It emphasizes the potential for these systems to amplify existing biases if unchecked, especially since real-world data often contains errors that require automated cleaning​​.

When discussing data quality, missing values are often cited as a sign of poor data quality. It's common to remove tuples with missing values or impute them with parameters representing the entirety of the dataset. However, this approach can lead to removal of data from already underrepresented groups or a misrepresentation due to imputation with unfit data. Do these techniques take into account demographic characteristics or the potential loss of representation in already disadvantaged groups? Automation makes our lives easier, but is it fair that it may lead to even less representation in underrepresented groups? For instance, when a prediction is based on a poorly fitted model, can it lead to unfair hiring or loan approvals for these groups? These are the concerns that Guha et al. are interested in.

The focus of the paper delves on two main research questions:

1. Does the incidence of data errors track demographic group membership in ML fairness datasets?
2. Do common automated data cleaning techniques impact the fairness of ML models trained on the cleaned datasets?

Essentially, are we aware of the potential impact of any data that is cleaned (i.e. removed or imputed)?

**Methodology:**

The research utilized five datasets from domains like census, finance, and healthcare, each associated with a binary classification task. The study was designed to analyse the differences in error detection and data repair's impact on fairness metrics across datasets that are commonly used in decision-making.

*Sensitive Attributes:* The study involves partitioning data into privileged and disadvantaged groups based on sensitive attributes like sex, race, and age. These attributes were then flagged using the following statements:

These formulas give the ratio of erroneous privileged data to the entire dataset and the ratio of erroneous disadvantaged data to the entire dataset, respectively.

*Data Cleaning:* The process of detecting and correcting (or removing) corrupt or inaccurate records from a dataset.

*G2 Significance Test*: A statistical method for testing the independence of two variables in a contingency table

*Fairness in ML:* The study of biases in machine learning algorithms and models, particularly concerning how the output of these models affects different demographic groups.

*Fairness metrics*: Predictive parity is satisfied if the classifier has the same precision for the subjects in the privileged and disadvantaged groups. This is given by the formula

Equal opportunity is satisfied if the classifier has the same recall for the subjects in the privileged and disadvantaged groups. This is given by the formula

In their study on the impact of automated data cleaning on fairness in machine learning models as part of the first research question, Guha et al. utilized five specific datasets — *adult*, *folk*, *credit*, *German*, and *heart*. These datasets were chosen for their relevance in fairness research and the varied demographics they represent. Each dataset corresponds to the following domains: census (*adult* and *folk*), finance (*credit* and *German*), and healthcare (*heart*), making them suitable for a comprehensive analysis of automated data cleaning across different sectors.

Reasons for Dataset Selection

*Representation of Sensitive Attributes:* The datasets were chosen because they contain sensitive attributes like sex, race, and age, which are crucial for studying fairness. Each dataset allows for the partitioning of data into privileged and disadvantaged groups based on these attributes.

*Prevalence in Fairness Research:* These datasets are commonly used in research on responsible machine learning and data management. Their frequent use in previous studies provides a baseline for comparison and consistency in research methodologies.

*Binary Classification Tasks:* Each dataset is associated with a binary classification task, such as determining creditworthiness or prioritizing healthcare access. This consistency in task type ensures that the impact of data cleaning on decision-making outcomes can be compared across datasets.

Error Detection and Automated Repair Strategies

*Error Detection Strategies:* The study applied various error detection strategies to identify data quality issues like missing values, outliers, and label noise. The detection strategies were chosen based on their common usage in data cleaning processes and their ability to flag potentially erroneous tuples.

*Automated Data Repair Methods:* Once errors were detected, different data repair methods were employed to correct the flagged tuples. The study tailored the repair techniques to the type of error detected, considering the most effective methods for each error type. For instance, missing values were handled using different imputation techniques, while outliers were addressed using rules based on standard deviation or isolation forests.

*Analysis of Disparities:* A significant part of the study involved analysing whether these error detection and repair strategies led to disparities in treatment of different demographic groups. This analysis was crucial to understand if automated data cleaning inadvertently introduced or exacerbated fairness issues in ML models.

Based on the above strategies, it was suggested that the *adult* dataset be retired due to unclear origin and class label imbalance and replaced with the *folk, credit* and *German* datasets to provide more financial information, credit score detail and evaluate medical data to predict cardiovascular diseases. The choice of datasets provided a robust foundation for examining how data cleaning influences model fairness across various demographic groups and sectors, while the error detection and repair strategies offered insights into the practical aspects of implementing data cleaning in real-world ML applications.

**Results:**

After evaluating 26,400 models based on the datasets chosen that align with the first research question, the findings indicated that:

Automated cleaning of missing values usually does not degrade model performance and often improves accuracy. However, it can negatively impact fairness in some cases.

The impact of data cleaning varies based on the type of error, dataset, fairness metric, and the ML model used​​. Trends identified in some datasets showed an opposite trend in another. For example, *folk* and *heart* showed more errors overall in the disadvantaged group but showed little disparity between groups. On the other hand, credit and german showed large disparities but did not show more frequent errors in the disadvantaged group.

**Discussion and Implications:**

The paper suggests that automated data cleaning can distribute benefits unevenly across different demographic groups, potentially worsening fairness in decision-making processes. This finding is concerning for ML systems in production. However, the study also notes that there are configurations where cleaning does not negatively impact fairness, indicating the need for a principled approach to select cleaning procedures​​. Despite finding higher rates of missing values in disadvantaged groups, no significant observations regarding demographic dependency were made in data errors. However, the downstream effect of automated cleaning showed that it could distribute benefits unevenly, potentially worsening fairness in decision-making processes.

The study finds higher rates of missing values in data from historically disadvantaged groups but does not find that poor data quality generally tracks with demographic group membership. The hypothesis as implied, is presented as automated data cleaning trends to have an insignificant impact on both accuracy and fairness in most cases. However, it is slightly more likely to worsen fairness, especially when cleaning techniques are not carefully chosen​​.

**Critical Analysis:**

*Importance of Holistic Approach:* The paper effectively highlights the need for a holistic view of data quality and its impact on fairness in ML models. This comprehensive approach is crucial for developing fairness-aware data cleaning methods.

*Methodological Robustness:* The use of multiple datasets and the evaluation of a large number of models add robustness to the study. However, the reliance on only common and existing datasets and methods might limit the generalizability of the findings.

*Practical Implications*: The findings have significant practical implications, especially for the deployment of ML systems in critical areas. The need for careful selection of data cleaning techniques to avoid exacerbating fairness issues is a key takeaway.

*Limitations and Future Directions:* The paper acknowledges its limitations, such as focusing mainly on US-centric datasets. Future research could explore other areas of the world and develop universal fairness-aware data cleaning methods.

*Reproducibility and Openness:* Providing the code and results for reproducibility is commendable and aligns with the principles of open and responsible research.

**Conclusion:**

Overall, the study offers insightful information about the relationship between automated cleaning, fairness, and data quality in machine learning. It emphasizes the importance of giving fairness-aware data cleaning techniques due thought and development. The study provides a thorough and critical examination of how automated data cleaning affects machine learning models' fairness. The study's methodology is meticulous, making use of datasets from several domains critical towards certain socio-economic decisions. It offers a comprehensive view of how different demographic groups can be impacted by data cleaning. The insights on the connection between data quality problems, in particular missing values, and their occurrence in historically marginalized groups, as well as the subtle effects of automated data cleaning methods on model fairness and accuracy, are the paper's primary contributions.

# 

# References

[1] Automated Data Cleaning Can Hurt Fairness in Machine Learning-Based Decision Making, Shubha Guha, Falaah Arif Khan, Julia Stoyanovich, Sebastian Schelter, 2023 IEEE 39th International Conference on Data Engineering (ICDE), pp. 3747--3754, 2023.