BiasNet: A Framework for Unsupervised Discovery and Quantification of Bias in Datasets

Anmol Devansh[1], Sudhanshu Ranshevare[2]

[1]*Independent Researcher, Hyderabad, India*
[2]*Independent Researcher, Nashik, India*

---***---

Abstract - This paper introduces BiasNet, a comprehensive, end-to-end data analysis tool designed to uncover hidden biases and inequities in datasets without requiring pre-existing labels. BiasNet serves as an exploratory engine, leveraging a suite of unsupervised machine learning algorithms to segment a population into naturally occurring groups. By analyzing the demographic composition and characteristics of these discovered clusters, the tool quantifies potential disparities across sensitive attributes. The methodology encompasses versatile data ingestion, advanced preprocessing for structured and text data, a comprehensive suite of clustering algorithms, and the calculation of unsupervised fairness metrics. The entire workflow is encapsulated in an interactive web interface, culminating in a detailed, AI-generated PDF report with rich visualizations, making data auditing accessible to data scientists, analysts, and decision-makers.

Key Words: unsupervised learning, bias detection, algorithmic fairness, data auditing, clustering, explainable AI, responsible AI

1.INTRODUCTION

The proliferation of data-driven decision-making systems across various domains—from credit scoring and hiring to medical diagnoses and criminal justice—has highlighted the critical problem of inherent bias within datasets. Models trained on biased data can perpetuate and even amplify existing societal inequities, leading to unfair or discriminatory outcomes. Identifying such biases is often challenging, especially in the absence of explicit labels that would allow for supervised fairness assessments. This paper presents BiasNet, an unsupervised bias discovery engine designed to address this challenge directly.

BiasNet provides a systematic and accessible framework for auditing datasets for potential disparities across sensitive demographic attributes such as race, gender, or age. It empowers users to explore their data, identify naturally occurring clusters of individuals with shared characteristics, and critically assess whether these clusters are disproportionately composed of certain demographic subgroups. The primary contribution of this work is a novel, end-to-end tool that operationalizes the process of unsupervised bias discovery. By combining advanced clustering techniques with robust fairness metrics and an intuitive user interface, BiasNet democratizes the ability to perform preliminary fairness audits, promoting greater transparency and accountability in the development of artificial intelligence systems. The objective of this paper is to detail the architecture, methodology, and technical implementation of BiasNet, demonstrating its potential as a crucial tool for promoting fairness and transparency in data analysis.

2. Related Work

The field of algorithmic fairness has produced a wide range of tools and techniques. Many existing solutions, such as IBM's AI Fairness 360 and Google's What-If Tool, are powerful but primarily focus on *supervised* scenarios where ground-truth labels and model predictions are available. They excel at measuring fairness metrics like equal opportunity or equalized odds post-training.

BiasNet differentiates itself by operating in a purely *unsupervised* context. It is designed for the exploratory phase of data analysis, *before* a model is even built. Its approach is more akin to exploratory data analysis tools that aim to uncover latent structures in data. While tools like fairlearn (which BiasNet incorporates) provide the metrics, they are libraries, not end-to-end applications. BiasNet integrates these components into a seamless workflow, from data ingestion to automated reporting. It fills a crucial gap by providing a user-friendly engine to probe for potential biases at the data level itself, which is the root cause of many downstream fairness issues.

3. THE BIASNET METHODOLOGY

The core of BiasNet is an end-to-end pipeline that transforms raw data from various sources into an actionable, comprehensive bias report. The pipeline consists of four main stages: Data Ingestion and Preprocessing, Unsupervised Clustering, Analysis and Fairness Quantification, and finally, Reporting and Visualization.

3.1. Data Ingestion and Preprocessing

The process begins with flexible data ingestion, accepting user-uploaded files in common formats like CSV, Excel, and even PDF. For PDF documents, BiasNet

DOI: 10.55041/IJSREM52445

automatically extracts and structures the text content for analysis. A key feature is its ability to automatically detect whether the data is structured (tabular) or unstructured (text-based) and apply the appropriate preprocessing pipeline.

- **For Structured Data:** It identifies categorical and numerical features. Categorical data is transformed using LabelEncoding, while numerical features are scaled using a user-selectable method—either StandardScaler (for algorithms sensitive to feature variance) or MinMaxScaler (for algorithms requiring features within a specific range).
- **For Unstructured Text Data:** The system utilizes a powerful Sentence-BERT model (all-MiniLM-L6-v2) to convert text content into high-dimensional numerical embeddings. This process captures the semantic meaning of the text, allowing clustering algorithms to group documents based on their content and context.

3.2. Unsupervised Clustering

The preprocessed data is then fed into a comprehensive suite of clustering algorithms. This variety allows users to select the most appropriate algorithm for their data's structure, size, and complexity. The implemented algorithms include:

- **Centroid-based (K-Means):** Efficient for identifying simple, spherical clusters.
- **Hierarchical (Agglomerative Clustering):** Useful for understanding nested cluster structures.
- **Distribution-based (Gaussian Mixture Models):** Assumes clusters are Gaussian distributions, allowing for more flexible cluster shapes.
- **Density-based (DBSCAN, HDBSCAN):** Excels at finding arbitrarily shaped clusters and identifying noise points.
- **Graph-based (Spectral Clustering):** Effective for complex, non-convex cluster shapes.
- **Deep Learning-based (Autoencoder with K-Means):** Reduces data dimensionality with a neural network before clustering, which can uncover more intricate patterns.

3.3. Analysis and Fairness Quantification

Once cluster assignments are made, BiasNet generates detailed profiles for each cluster by calculating statistical summaries (mean, median, mode) of its features. This helps in understanding the characteristics of each discovered group. More importantly, it quantifies bias using established unsupervised fairness metrics against user-selected sensitive attributes:

- **Demographic Parity Difference:** This metric measures whether any subgroup is over- or under-represented within a cluster. It calculates the difference in the proportion of a subgroup in a cluster compared to its proportion in the overall dataset. A large difference suggests that the clustering algorithm is disproportionately grouping individuals based on the sensitive attribute.
- **Chi-Squared Test of Independence:** This statistical test assesses whether there is a significant association between the cluster assignments and the sensitive attribute. A low p-value (e.g., < 0.05) indicates that the observed distribution is unlikely to be due to random chance, suggesting a systematic relationship between an individual's demographic group and the cluster they are assigned to.

3.4. Reporting and Visualization

The final stage involves compiling all results into an accessible and interpretable format. The system constructs a detailed prompt containing the quantitative results and queries the Google Gemini API to generate a concise, three-paragraph executive summary. This summary, along with a rich set of interactive visualizations, is assembled into a professional, multi-page PDF report using the reportlab library.

Visualizations are a core component of the output and include:

- **UMAP Cluster Visualization:** A 2D representation of the high-dimensional data, allowing for visual inspection of cluster separation.
- **Disparity Distribution Plot:** A bar chart showing the distribution of sensitive attribute subgroups within each discovered cluster.
- **Cluster Profile Heatmap:** A normalized heatmap of numeric features for each cluster, allowing for easy comparison of cluster characteristics.
- **Silhouette Plot:** A diagnostic tool to help validate the quality and cohesion of the clustering result.

4. CASE STUDY: ANALYSIS OF THE ADULT CENSUS DATASET

To demonstrate the practical application of BiasNet, we performed an analysis of the "Adult" dataset from the UCI Machine Learning Repository. This dataset contains 14 attributes extracted from the 1994 US Census database and is a common benchmark for fairness research. The prediction task is to determine whether a person makes over $50K a year. We used BiasNet to explore potential biases related to the 'race' and 'sex' attributes without using the income label.

- **Setup:** We uploaded the adult.csv file, selected 'race' as the primary sensitive attribute and 'sex' as the secondary one. We chose the K-Means clustering algorithm, with the number of clusters

(k) set to 6 based on the elbow method and silhouette analysis.

- **Results:** The K-Means algorithm partitioned the dataset into six distinct clusters. The cluster profiles revealed that the groups were largely defined by combinations of age, education level, and hours worked per week. For instance, one cluster might represent young, part-time workers with lower education, while another might represent older, highly-educated professionals.
- **Fairness Analysis Findings:** The fairness evaluation of the race_sex attribute uncovered substantial demographic disparities across the six k-means clusters. A pronounced demographic parity difference of 0.4353, coupled with a statistically significant Chi-Squared test result (p-value = 0.0000), confirmed a strong association between demographic characteristics and cluster membership. Clusters characterized by higher levels of education and income, such as Cluster 2 and Cluster 5, were disproportionately composed of White males (78% and 87%, respectively), indicating a concentration of privilege within these groups. Conversely, Cluster 4 demonstrated a comparatively balanced gender distribution but was predominantly associated with lower income levels, with more than 99% of individuals earning ≤50K and exhibiting reduced educational attainment. Across clusters, Black individuals and women were systematically underrepresented in higher-income segments, highlighting persistent inequities in the distribution of socioeconomic resources and opportunities.

Metric	Value	Interpretation
Demographic Parity	0.4353	Closer to 0 is fairer.
Chi-Squared p-value	0.0000	< 0.05 suggests bias.

Table -1: Fairness Metrics for Attribute: 'Race_Sex'

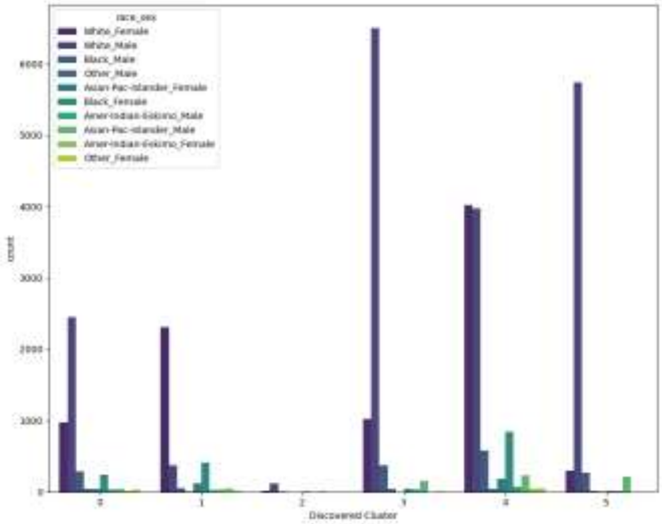

Fig -1: Disparity Plot for 'Race_Sex' Attribute

5. CONCLUSIONS AND FUTURE WORK

BiasNet provides a robust, user-friendly, and accessible solution for the critical task of unsupervised bias discovery in datasets. By integrating a flexible data processing pipeline, a diverse suite of clustering algorithms, and automated fairness metrics, it lowers the barrier for data practitioners to conduct thorough and responsible data audits. The generation of an AI-powered, comprehensive PDF report ensures that the findings are interpretable and actionable for a broad audience, fostering better communication between technical and non-technical stakeholders.

This tool represents a significant step towards enabling more equitable and responsible data science practices. It allows organizations to proactively identify and understand potential biases in their data before these biases are codified into automated systems.

Future work will focus on several key areas. First, we plan to expand the library of fairness metrics to include more nuanced measures. Second, we aim to incorporate bias mitigation algorithms that can suggest or even automatically apply corrections to the dataset. Finally, we will continue to enhance the visualization capabilities, providing even more interactive ways for users to explore and understand the complex relationships within their data.

5. ACKNOWLEDGEMENT

The authors acknowledge the use of the publicly available UCI Adult Census dataset, which provided the empirical basis for this study. This research was supported by open-source software ecosystems, including Python and its associated libraries (scikit-learn, pandas,

NumPy, and Matplotlib), which facilitated data processing, statistical analysis, and visualization. The utilization of machine learning frameworks and deployment platforms, particularly Hugging Face Spaces, further enabled the implementation and dissemination of the proposed system. The accessibility of these resources was essential in ensuring the rigor, reproducibility, and transparency of the findings presented in this work.

5. REFERENCES

1. Caton, S., Haas, C.: Fairness in Machine Learning: A Survey. ACM Comput. Surv. 56(7) (2024) 1–38
2. Baraldi, A., Brucato, M., Dudík, M., Guerra, F., Interlandi, M.: FairnessEval: A Framework for Evaluating Fairness of Machine Learning Models. In: Proc. EDBT 2025. OpenProceedings (2025)
3. Hasanzadeh, F., Josephson, C.B., Waters, G., de Adedinsewo, D., Azizi, Z., White, J.A.: Bias Recognition and Mitigation Strategies in Artificial Intelligence Healthcare Applications. npj Digit. Med. (2025)
4. Mackin, S., et al.: Identifying and Mitigating Algorithmic Bias in the Safety Net. npj Digit. Med. (2025)
5. Small, E.A., Shao, W., Zhang, Z., Liu, P., Chan, J., Sokol, K., Salim, F.D.: How Robust Is Your Fair Model? Exploring the Robustness of Prominent Fairness Strategies. Data Min. Knowl. Discov. (2025)
6. Ma, S.-C., Ermakova, T., Fabian, B.: FairGridSearch: A Framework to Compare Fairness-Enhancing Models. arXiv:2401.02183 (2024)
7. Parziale, A., Voria, G., Giordano, G., Catolino, G., Robles, G., Palomba, F.: Contextual Fairness-Aware Practices in ML: A Cost-Effective Empirical Evaluation. arXiv:2503.15622 (2025)
8. Cohen-Inger, N., Cohen, S., Rabaev, N., Rokach, L., Shapira, B.: BiasGuard: Guardrailing Fairness in Machine Learning Production Systems. arXiv:2501.04142 (2025)
9. Jung, D., Park, J., Kim, G., Park, K., Kim, J.: FLEX: A Benchmark for Evaluating Robustness of Fairness in LLMs under Extreme Scenarios. In: Findings of NAACL 2025 (2025)
10. Laakom, F., et al.: Fairness Overfitting in Machine Learning. arXiv:2506.07861 (2025)
11. Qin, Q., Wang, J., Li, Q., Li, J., Wang, S.: Representation-Based Fairness Evaluation and Bias-Correction Robustness via Computational Profile Distance. Knowl.-Based Syst. (2025)
12. Uddin, S., et al.: A Novel Approach for Assessing Fairness in Deployed ML Algorithms Using k-Fold Cross-Validation and t-Tests. Sci. Rep. (2024)
13. Wang, Y., Singh, R.: Impact on Bias Mitigation Algorithms of Variations in Inferred Sensitive-Attribute Accuracy. Front. Artif. Intell. (2025)
14. Liu, M., et al.: A Scoping Review and Evidence Gap Analysis of Clinical AI Fairness. Digit. Health (PLOS) (2025)
15. Sasseville, M., et al.: Bias Mitigation in Primary Health Care Artificial Intelligence: A Scoping Review. J. Med. Internet Res. 27(1) (2025) e60269
16. Huang, Y., et al.: Fair ML Techniques in Healthcare Data Applications: A Scoping Review. Future Gener. Comput. Syst. (2024)
17. Ni, H., Han, L., Chen, T., Sadiq, S., Demartini, G.: Fairness Without Sensitive Attributes via Knowledge Sharing. arXiv:2409.18470 (2024)
18. Duong, M.K., Conrad, S.: Measuring and Mitigating Bias for Tabular Datasets with Multiple Protected Attributes. arXiv:2405.19300 (2024)
19. Caton, S.: Fairness in Machine Learning – Keynote Overview. NCI Research Day (Slides) (2025)
20. Fabris, A., et al.: Fairness and Bias in Algorithmic Hiring. ACM (2025)
21. Nepomuceno, K., et al.: The AI Fairness Myth: A Position Paper on Context-Aware Bias. arXiv:2505.00965 (2025)
22. Voria, G., Fazakis, N., Davrazos, G., Kotsiantis, S.: A Comprehensive Review and Benchmarking of Fairness-Aware Variants of Machine Learning Models. Algorithms 18(7) (2025) 435
23. Hoche, M., et al.: What Makes Clinical Machine Learning Fair? A Practical Framework. npj Digit. Med. (2025)
24. Ramineni, V., et al.: Dataset Construction Beyond Internal Data for Fairness Testing. arXiv:2507.18561 (2025)
25. Cohen, S., et al.: BiasGuard: Ensuring Fairness in Production ML Pipelines. arXiv:2501.04142 (2025)