Project Proposal: Ethical AI Dataset Auditing Tool for checking Representativeness, Diversity, and Data Governance in AI Systems

By Team Armah, AI Saturdays Lagos Cohort 9 - Flipped

Problem Statement

AI systems inherit strengths and weaknesses from the data they are trained on. Biases in data representation and diversity often lead to inaccurate results, particularly in sensitive industries such as healthcare. For instance, the varied human reactions to COVID-19 highlighted how underrepresented demographic groups in datasets can lead to harmful generalizations and inequitable health outcomes.

While AI is increasingly being adopted to solve societal problems, it also presents risks if left unchecked. As the world develops policies and frameworks, such as the EU AI Act, the NIST AI Risk Management Framework, China's AI regulations, Japan's G7 Code of Conduct, the UK's Pro-Innovation Framework, and Canada's AI Laws, there is a growing need for practical tools to operationalize these standards. One solution is to develop AI systems that can assess other AI systems for ethical compliance, safety, and security before deployment.

This project proposes the development of a lightweight AI dataset auditing tool that evaluates representativeness, diversity, and governance factors, ensuring AI systems are tested for ethical concerns before deployment.

Existing Solutions

- **Aequitas** – Open-source fairness audit toolkit (Aequitas).
- **AI Fairness 360** – IBM's extensible bias detection toolkit.
- **Eticas Audit Library** – Python-based fairness auditing tool (PyPI).
- **OECD Bias Detection Tool** – Unsupervised bias detection for NLP tasks (OECD).

Gap: Existing tools are often too technical, complex, or enterprise-oriented, limiting accessibility for researchers, students, and smaller organizations.

Objectives

1. To design a lightweight AI dataset auditing tool focused on accessibility and simplicity.
2. To compute representativeness and diversity metrics (class balance, demographic skewness, and missing values).
3. To incorporate data governance factors in the auditing process, aligning with emerging AI policies and frameworks.
4. To visualize dataset imbalance through clear, interpretable charts.
5. To generate human-readable audit reports that can guide corrective action before AI system deployment.

Proposed Dataset

- **Adult Census Income Dataset** (UCI/Kaggle) – includes demographic features (age, gender, race, income).
- **Additional public datasets** (Titanic dataset, survey data, healthcare-related datasets) for testing.

Proposed Methodology

1. **Data Input** – Users upload CSV/tabular datasets.
2. **Preprocessing** – Identify categorical and demographic features.
3. **Analysis** – Calculate class imbalance, diversity ratios, missingness, and governance-related metrics.
4. **Visualization** – Generate histograms, pie charts, and bar charts for interpretability.
5. **Modeling** – Baseline ML models (Naive Bayes, Logistic Regression) and neural network classifier, evaluate performance, and perform hyperparameter tuning to optimize results.
6. **Reporting** – Produce textual audit reports summarizing risks, gaps, and recommendations.
7. **Deployment** – Deploy the model on a Streamlit web app for users to audit AI datasets dynamically.

Modeling Plan

(Not modeling in the ML sense, but system auditing.)

We plan to experiment with multiple approaches to build a lightweight AI dataset auditing tool:

- Baseline checks for metadata completeness, demographic coverage, and governance compliance.
- Compute descriptive statistics and fairness metrics.
- Add governance compliance checks against selected frameworks.
- Provide simple exportable reports for practitioners.

Model Development Plan

We plan to test multiple models:

- **Logistic Regression** (regularized): interpretable baseline to detect subgroup imbalance.
- **Decision Trees**: Human-readable splits that highlight where imbalance occurs.
- **Random Forests**: Detect complex feature interactions leading to bias.
- **Gradient Boosted Trees** (XGBoost / LightGBM / CatBoost): high accuracy for non-linear imbalances, still explainable via SHAP/feature importance.

Deployment Plan

- Prototype as a **Streamlit or Jupyter Notebook tool**.
- Public deployment on **Streamlit Cloud** for accessibility.
- **Future scalability**: expansion into a full web app with modular evaluation stages.

Expected Outcomes

- A scalable, user-friendly dataset auditing tool.
- Clear, interpretable audit reports combining diversity and governance considerations.
- A system that balances simplicity, user experience, and responsible AI awareness.

Community Impact

- Promotes responsible AI adoption by embedding ethics at the dataset stage.
- Minimizes AI risks such as biased predictions, inaccuracy, and inequity.
- Supports policymakers, researchers, and practitioners in aligning with evolving AI governance frameworks.

Team Members

- Latifat Olabisi Idris (Team Lead)
- Abdusshakur Olarewaju Olabisi

Acknowledgement

I acknowledge that the inspiration for this work came from the rapid pace of AI innovation, which has led to premature deployments resulting in inequity, untrustworthy outcomes, inaccurate predictions, and social inequalities. This project is motivated by the urgent need to build safeguards that promote fairness, trust, and accountability in AI systems.

References

1. Adult Census Dataset – UCI/Kaggle
2. Aequitas Audit Toolkit
3. Bellamy, R. et al. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv:1810.01943*.
4. Eticas Audit Library – PyPI
5. Saleiro, P. et al. (2018). Aequitas: A Bias and Fairness Audit Toolkit. *arXiv:1811.05577*.