

Ethics and Fairness in Machine Learning: Equity Guidelines for Brazilian Public Policies.

Vicente Julião MR Barros   [Federal University of Alagoas | vjmrb@ic.ufal.br]

George Brandão Souza   [Federal University of Alagoas | gsbs@ic.ufal.br]

Baldoino Fonseca   [Federal University of Alagoas | baldoino@ic.ufal.br]

Davy Baia   [Federal University of Alagoas | davy.baia@ceca.ufal.br]

Tiago Vieira   [Federal University of Alagoas | tvieira@ic.ufal.br]

 Institute of Computing, Federal University of Alagoas, Campus A.C.Simões, Avenida Lourival Melo Mota, s/n, Tabuleiro dos Martins, Maceió, AL, 57.072-970, Brazil.

Received: 23 September 2025 • **Accepted:** 27 July 2026 • **Published:** 22 September 2026

Abstract.

The application of machine learning technologies in public policy formulation represents a significant advance in contemporary government management, but raises critical issues related to equity, transparency, and algorithmic justice. The present study investigates the presence of systemic biases in data and predictive models applied in Brazilian public policy scenarios through a comprehensive audit using open government data. A pipeline-structured methodology was developed that covers ten social scenarios in critical areas such as education, health, work, and technology, using data from five Brazilian government agencies (IBGE, MS, MEC, INEP, CETIC.BR).

Eight supervised binary classification algorithms were evaluated. To rigorously address the severe class imbalance inherent in public policy microdata, the final predictive models were selected strictly based on the highest 5-fold Cross-Validated Area Under the ROC Curve (CV AUC-ROC) rather than global accuracy. This unified criterion resulted in the selection of Gradient Boosting for seven social scenarios and Random Forest for three scenarios (Food Insecurity, Digital Inclusion, and Tech Vulnerability) as the most robust models, which were subsequently subjected to systematic fairness auditing.

The results indicate the presence of consistent patterns of disparity among sensitive demographic groups, with frequent violations of established disparity limits, particularly in the Demographic Parity and Predictive Equality metrics. Because the reference group is the modal category of each sensitive attribute—Brown individuals in every racial audit—the reported ratios express deviation from the predictive pattern of that group, and such deviations affect all contrasted categories: minority groups (black, yellow, and indigenous people) record the most extreme cases, including predictive collapses, while the white group registers some of the largest deviations in positive prediction and false positive rates. These patterns show concrete risks of algorithmic discrimination and perpetuation of historical inequalities. The Equal Opportunity metric suggests better equity for the gender sensitive attribute in eight of the ten scenarios analyzed, the exceptions being Poverty Prediction (CENSO) and Live Births Vulnerability (SINASC), where the metric fails against the sex reference of the respective dataset.

This research contributes significantly to the field of ethical and responsible Artificial Intelligence in the Brazilian context, providing empirical evidence of the presence of systematic biases in government data and models and establishing a replicable protocol for fairness audit in automated decision support systems. The findings demonstrate the urgent need for the implementation of continuous audit practices in the life cycle of Machine Learning projects in the public sector. Furthermore, this study identifies a critical gap for future research and policy-making: the urgent necessity to formulate specific legal instruments and strengthen existing data protection frameworks (such as the LGPD in Brazil) to formally enforce algorithmic equity and regulate automated decision-making. By addressing this regulatory void, we propose guidelines for the development of more just and equitable Artificial Intelligence solutions that are rigorously aligned with democratic principles, legal accountability, and fundamental rights.

Keywords: Fairness, Algorithmic Bias, Public Policy, Government Open Data, AI Ethics, Algorithmic Justice.

1 Introduction

There has been extraordinary growth in the application of Artificial Intelligence (AI) technologies, more specifically Machine Learning (ML), in the formulation and decision-making of public policies. These technologies enhance operational efficiency and automate complex processes, enabling predictive analytics that previously required substantial human resources (Floridi *et al.* [2018]). Although AI presents numerous advantages and benefits in both the public and private sectors by

improving decision-making and automation capacity, it is not immune to errors, especially in Machine Learning contexts. The implementation of AI solutions in governmental settings raises critical concerns related to equity, transparency, and algorithmic justice, often presenting situations in which socially indifferent algorithms favor privileged groups (Cheng *et al.* [2021]). To make predictions, ML systems rely on historical data, inevitably carrying and amplifying the biases present in the datasets prepared for training (Mehrabi *et al.* [2021]). Consequently, ensuring that decisions are fair and free from

bias remains one of the main challenges of the field (Pagano *et al.* [2023]; Li *et al.* [2023]). Ultimately, machine learning algorithms have profound social, economic, and ethical implications (Ortega-Bolaños *et al.* [2024]).

When these predictions are applied in public policies, such biases manifest themselves as systematic discrimination against specific demographic groups, perpetuating and amplifying existing social inequalities. “As democracies suffer and structural biases are exacerbated, the free enjoyment of civil rights is no longer available to all. Thus, democracy and civil rights should be considered in the ethics of AI” (Huang *et al.* [2022]).

Emblematic cases are reported in the literature, such as the COMPAS system in the U.S. criminal justice system, which presents disproportionality of false positives for African-American defendants (Saleiro *et al.* [2018]). Minorities and disadvantaged groups—race (Black, Indigenous) and gender (women)—are the most affected by socially indifferent algorithms, both in frequency and severity (Cheng *et al.* [2021]).

In response to these limitations, the concept of Fairness in Machine Learning emerges, defined as the absence of prejudice or favoritism toward an individual or group based on intrinsic characteristics or those acquired in the context of decision-making (Cheng *et al.* [2021], Mehrabi *et al.* [2021], Liu *et al.* [2022]). This critical and complex research field seeks to ensure that the actions triggered by AI systems are fair, promote equity, and eliminate discrimination, while recognizing the multiplicity of definitions and the practical challenges involved in implementation (Mehrabi *et al.* [2021], Pagano *et al.* [2023]).

In the Brazilian context, public policies take on a particular dimension due to deep socioeconomic inequalities and the demographic diversity of the population. A critical aspect of this demographic diversity is the influence of smaller sample sizes on the predictive results. Minority groups are frequently underrepresented in public datasets, which degrades the ML model’s ability to identify accurate patterns for these populations. This data scarcity directly increases error rates and exacerbates algorithmic discrimination against these specific smaller samples. The use of open government data to power ML systems represents an opportunity to democratize access to public services, but simultaneously amplifies the risks of algorithmic discrimination. Morley *et al.* [2020] point out that “not all principles will have the same importance in all contexts,” and that AI ethics should be “culturally and contextually specific.” The authors state that the increased availability of open data, as well as the sophistication of ML techniques, “increased opportunities to improve public and private life,” allowing society to have “more control than ever over outcomes related to: (1) who people can become; (2) what people can do; (3) what people can achieve; and (4) how people can interact with the world.”

Thus, it is imperative to identify biases in sensitive areas of public policies (health, education, public security, social assistance), where automated decisions directly impact the well-being of vulnerable citizens Ortega-Bolaños *et al.* [2024]. The perpetuation of historical injustices, the exclusion of marginalized groups, and the inefficient allocation of public resources are primarily caused by the absence of adequate audit mechanisms (Liu *et al.* [2022]).

To detail how we address these challenges, this document is structured as follows: Section 2 establishes the Theoretical Framework and Related Work, mapping the foundational concepts of algorithmic fairness and responsible Artificial Intelligence. Section 3 details our Methodology, encompassing the public data acquisition, the Machine Learning pipeline design, and the fairness auditing processes using the AE-QUITAS toolkit (Saleiro *et al.*, 2018). Section 4 presents the empirical Results of the fairness metrics across the ten social scenarios. Section 5 provides the Discussion and Implications of these findings, addressing mathematical trade-offs, policy-level impacts, and study limitations. Finally, Section 6 concludes the study and outlines future works.

2 Theoretical Framework and Related Work

To properly contextualize the algorithmic fairness audit conducted in this study, it is essential to establish the theoretical framework that supports Responsible Artificial Intelligence (RAI) and to review the related works that guide current bias mitigation practices.

Specifically in Machine Learning, the discrepancy between high-level ethical principles and their practical implementation remains a critical challenge. As vehemently argued by Morley *et al.* [2020], while AI ethics has reached a broad consensus on the “what” (defining the principles), there is still a profound deficit of tools and methods addressing the “how” (operationalizing these principles in real-world workflows). Consequently, adopting a multidisciplinary approach to engineer practical, fully usable tools is indispensable for the realization of fair and responsible AI. In this context, Ortega-Bolaños *et al.* [2024] analyze several justice-focused frameworks developed within academic circles, highlighting both the potential of these contributions and the persistent need for greater operationalization of ethics across the software development lifecycle.

2.1 Theoretical Foundations of AI Ethics

The proposal of the AI4People framework, established an initial framework for the conceptualization of reliable AI, as one that produces “benefits and respect for people and the environment (beneficence), is robust and safe (non-maleficence), respectful of human values (autonomy), fair (justice), and explainable, accountable and understandable (explicability)” (Floridi *et al.* [2018]).

This concept was later refined and expanded by other authors. “Process guided by human values, considered principles, values such as Equity, Transparency, Accountability, Reliability and Safety, Privacy and Security, and Inclusivity, considering the development of socially responsible AI algorithms as a means, as well as with the approach of social expectations to generate shared value” is the main objective proposed by Cheng *et al.* [2021].

The conceptual evolution is evidenced by the proposal that “Reliable AI is a holistic and systematic approach to ensure that artificial intelligence systems are robust, fair, transparent, explainable, private and accountable, aiming not only for

performance but also for sustainability and long-term social benefit.” proposal by Li *et al.* [2023].

There is a significant contribution, the proposal of a three-dimensional categorization of reliable AI on three levels: individual, societal, and environmental. Straightforward categorization comprehensively addresses the ethical issues of AI. Trusting AI is a fundamental prerequisite for its effective adoption by individuals and societies. “Trust in AI can be achieved through impartiality, transparency, accountability, and control, which are key principles directly addressed by this three-dimensional categorization of ethical challenges,” (Huang *et al.* [2022]).

The analysis reveals a growing sophistication in understanding ethical trade-offs. An ideal AI system should simultaneously satisfy all six crucial dimensions: Safety & Robustness; Non-Discrimination & Equity, Explainability, Privacy, Accountability & Auditability, and Environmental Welfare, but in reality, conflicts exist where the satisfaction of one dimension can prevent the achievement of another (Liu *et al.* [2022]). Likewise, several authors recognize tensions between privacy and explainability, between equity and predictive performance, and between different definitions of fairness.

2.2 Related Work in Bias Auditing

Bridging the gap between abstract ethical principles and their technical operationalization requires robust methodological frameworks. To properly contextualize our approach, this subsection reviews the current landscape of algorithmic bias auditing by examining the evolution of practical mitigation tools, mapping the persistent sociotechnical challenges, exploring domain-specific applications in high-stakes environments, and ultimately situating our research within the applied auditing paradigm.

2.2.1 Practical tools and methods

The materialization of the development of specific tools and methodologies promoted the transition between proposals for principles and practice. The identification of 106 tools available for the implementation of ethics in AI, categorized by ethical principles and stages of the development life cycle, was analyzed by Morley *et al.* [2020].

Representing one of the most important and significant contributions to practical audit of bias and fairness, the AEQUITAS toolkit “allows users to test models for various metrics of fairness bias concerning multiple population subgroups”, (Saleiro *et al.* [2018]). Its interface serves both AI practitioners, developers, data scientists, and public policy makers. The evolution is in AEQUITAS FLOW, providing an experiment, “pipeline for training models with fairness, optimization of hyperparameters and evaluation,” (Jesus *et al.* [2024]).

We highlight the Systematic Literature Review that identified 352 tools and resources for the practical application of ethics in AI, conducted by Ortega-Bolaños *et al.* [2024]. These tools and resources were classified into three categories: (1) hybrid tools combining technical and non-technical approaches; (2) non-technical tools providing conceptual, ethi-

cal, and legal guidance; and (3) technical tools offering practical and applicable solutions, (Ortega-Bolaños *et al.* [2024]).

The analysis of fairness metrics reveals a proliferation of often incompatible definitions. “Different fairness metrics do not present uniform results for a given use case”, (Pagano *et al.* [2023]). Observation of paramount importance, noting the need for the selection of appropriate metrics to be applied to the context.

Considering “from data acquisition to model development, system development and deployment, finally to continuous monitoring and governance”, Li *et al.* [2023] propose a systematized framework organizing multidisciplinary approaches in a continuous workflow. The design and development of full lifecycle frameworks represents an extraordinary advance from principles to practice, that is, in operationalization.

2.2.2 Challenges and Limitations

The literature reveals multiple dimensions of challenges in the implementation of ethics in AI. The gap between principles and practice is an identified fundamental challenge. “Almost all the guidelines that have been produced to date suggest that technical solutions exist, but very few provide technical explanations” is what Morley *et al.* [2020] note. Resulting in frustration among developers about the limited support they receive from highly abstract principles when it comes to daily work.

Modern AI systems have an inherent complexity, presenting specific technical challenges. Cheng *et al.* [2021] demonstrate that “measurement errors” arise at the moment when “we assume that training and test samples are representative of the target population and their distributions are sufficiently similar”, but question “how often does this assumption hold in practice?”

Another fundamental challenge is the incompatibility between different fairness metrics; several proposed metrics often conflict with each other. This incompatibility requires that ML practitioners, together with public policy managers, make conscious choices about which aspect of justice to prioritize, considering the implications and application domain of the AI system (Mehrabi *et al.* [2021]).

Organizational and cultural challenges are also significant. At least three causes make high-level principles insufficient: (1) focus on “what” rather than “how”; (2) total absence of an enforcement mechanism; and (3) conflict of short-term commercial interests (Morley *et al.* [2020]).

2.2.3 Sector Applications

Analysis conducted by sector reveals distinct patterns of challenges and ethical solutions depending on the context of the application.

In **Criminal Justice**, several studies have analyzed the COMPAS system noting the risk in the assessment of recidivism, presenting a higher rate of False Positive for black defendants compared to white defendants, “African Americans have a False Omission Rate disparity of 1.6x, having

FOR of 5.6% compared to the reference group (White) having FOR of 3.5%”, (Saleiro *et al.* [2018]).

In the **health** area, AI solutions face challenges related to population representativeness and generalization. ”Machine learning system improves on expert heuristic by 20%-50%” but identified significant biases: ”white individuals have higher FOR (False Omission Rate) than African Americans”, Saleiro *et al.* [2018], demonstrating the complexity of Fairness in medical contexts.

In the **financial** sector, the challenges are related to historical discrimination and regulation. ”Systems based on prediction of algorithmic decisions are being widely adopted by governments and organizations in lending, contracting, and online advertising,” according to Pagano *et al.* [2023], making the consideration of biases in credit decisions indispensable.

Complex trade-offs between effectiveness and fairness are found in **public safety** applications. Case of the Early Police Intervention System, where ”Despite this improved effectiveness, the audit of the model selected and implemented with the Aequitas tool revealed that it did present biases. Biases were identified concerning attributes such as gender, race, and marital status of officers”, (Saleiro *et al.* [2018]).

Finally, the **technology** sector is demonstrating unique challenges related to large-scale algorithmic bias. We found that ”ad delivery algorithm that promoted jobs in science, technology, engineering and mathematics (STEM) fields” that ”delivered ads in a discriminatory manner, although its original and pure intention was to be gender-neutral” (Mehrabi *et al.* [2021]).

2.2.4 Auditing in Practice

This research is distinguished by the practical application of a structured pipeline for auditing algorithmic bias in **ten** public policy scenarios, an area of high social sensitivity. The methodological process begins with the collection of open public microdata from **five** Brazilian government agencies. It continues with the selection of the optimal predictive model among **eight** supervised binary classification algorithms, selected strictly based on the highest 5-fold Cross-Validated Area Under the ROC Curve (CV AUC-ROC) to rigorously account for class imbalance. Subsequently, the audit of biases in sensitive subgroups is performed through the toolkit of fair ML Aequitas, with the analysis of **three** justice metrics (fairness). Thus, this work fills the gap between the development of fairness techniques and their systematic application in public impact decisions. The processing pipeline was fully developed in Python and is openly documented in a public GitHub repository. By providing an automated execution architecture and Docker containerization, this study ensures the exact reproducibility of the model selection, cross-validation metrics, and fairness audits.

3 Methodology

To ensure algorithmic equity, it is crucial to prevent discrimination resulting from biases — in both predictive models and datasets — by incorporating responsible practices at each

stage of the AI lifecycle. Therefore, the exploratory objective of this research adopts a dual methodological approach: (i) a qualitative mapping of ten social scenarios within public policies to define context-specific fairness constraints and to identify the sensitive attributes subject to audit, and (ii) a quantitative approach to train predictive classification models and systematically measure empirical bias through mathematical fairness metrics.

To achieve this, we utilized open and disaggregated microdata sourced from five official Brazilian government agencies (IBGE, MS, MEC, INEP, and CETIC.BR), as illustrated in Figure 1. The study is operationalized through a structured Machine Learning pipeline and subjected to a rigorous fairness audit using the open-source AEQUITAS toolkit.

To ensure full reproducibility, the following subsections detail this methodology: Section 3.1 describes the public data acquisition and the ten scenario definitions; Section 3.2 delineates the Exploratory Data Analysis (EDA) and the training framework for the binary classification models; and Section 3.3 presents the formal fairness audit process and the applied metrics.

3.1 Government

Government data play a central role in this study, as they represent the primary source of information for building the social scenarios analyzed. Produced by official institutions, these datasets are essential for guiding investments, interventions, and decision-making in public policy. In this section, we describe the main government data sources, their connection with public policies, the social scenarios developed for analysis, and the construction of the datasets used in the experiments.

3.1.1 Public Data

For this study, we selected Brazilian government data generated by official surveys conducted by public agencies. As illustrated in Figure 1, our overall methodological structure is strictly divided into three interconnected phases: the extraction of Government Data from ten public policy scenarios, the execution of the Machine Learning Pipeline, and the subsequent Fairness Audit. The specific databases were sourced from the following agencies:

- IBGE - Brazilian Institute of Geography and Statistics¹;
- MS - Ministry of Health²;
- MEC - Ministry of Education³;
- INEP - National Institute of Educational Studies and Research Anísio Teixeira⁴;
- CETIC .BR - Regional Study Centre for the Development of the Information Society⁵.

The starting point and main reference of this research was IBGE, where most of the studies were concentrated. Other government agencies were also included, either through joint

¹<https://www.ibge.gov.br/>

²<https://www.gov.br/saude/pt-br>

³<https://www.gov.br/mec/pt-br>

⁴<https://www.gov.br/inep/pt-br>

⁵<https://cetic.br/>

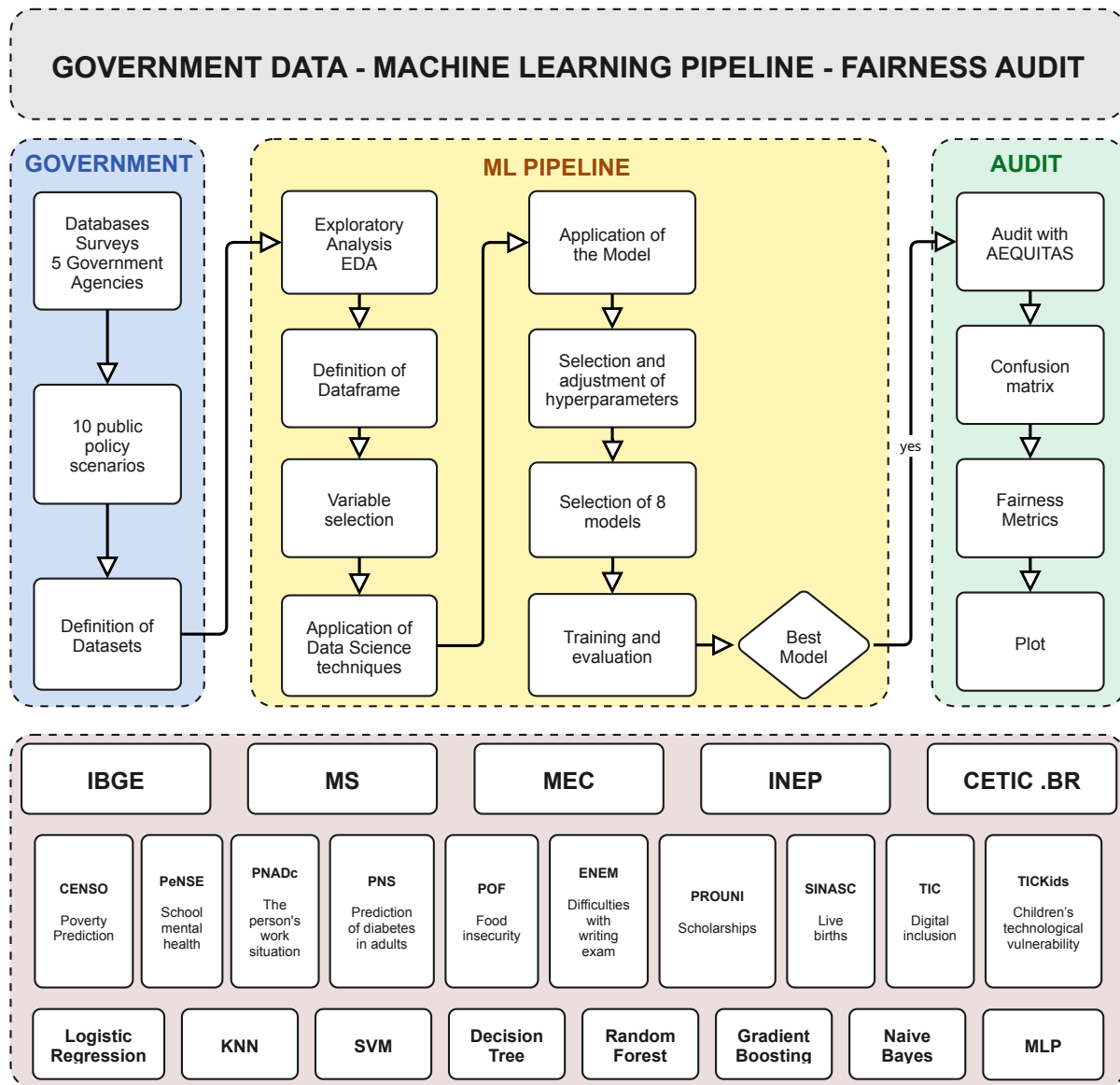


Figure 1. Methodological structure of the reference research: a study with the Bases of the Brazilian government. The *Best Model* decision node denotes the selection of the algorithm with the highest 5-fold Cross-Validated AUC-ROC (CV AUC-ROC) among the eight candidates, as detailed in Section 3.2.5.

surveys with IBGE or independent research. In selecting each dataset, we prioritized: 1. The availability of open and accessible data; 2. The granularity of information (preferring microdata at the individual level, such as student, father, or mother, instead of aggregated statistics like household, municipality, or state); 3. The presence of sensitive demographic variables, particularly race and gender.

3.1.2 Public Policy

The datasets produced by IBGE and other government agencies provide a relevant basis for investments, interventions, and decisions in the formulation of public policies. For this reason, public policies were adopted as the starting point of our research.

Public policy is a multidisciplinary field with significant conceptual diversity. Souza [2006], in a comprehensive literature review, emphasizes that there is no single, definitive definition of public policy. Mead [1995] defines it as a field

of policy studies that analyzes government in relation to major public issues. Lynn and Gould [1980] describe it as a set of government actions that produce specific effects. Peters *et al.* [1986] argue that public policy is the sum of governmental activities, whether carried out directly or through delegation, that influence the lives of citizens. One of the most influential classical definitions, offered by Dye [1972], states that public policy is “what government chooses to do or not to do.”

Both IBGE and other Brazilian government agencies produce a considerable volume of official statistical research that is fundamental to society. These surveys provide the government, the economy, and the public with essential data on Brazil’s economic, demographic, social, and environmental conditions.

3.1.3 Social Scenarios

In this study, we designed an experiment based on ten social scenarios, understood as contexts within public policies

that support the analysis and understanding of how these policies are formulated, implemented, and evaluated. These scenarios incorporate social, economic, political, institutional, and technological factors that influence the identification of public problems and the definition of possible solutions and strategies for action. Each scenario is applied to research developed by government agencies and is suitable for predictive modeling in public policy, where we carefully consider the following:

- the social relevance and potential impact on vulnerable populations;
- the availability of disaggregated demographic data of sufficient quality;
- the clear identification of sensitive variables for fairness analysis;
- the practical applicability to support public policy formulation.

The scenarios developed were as follows:

1. **Poverty Prediction**

Using the database of the Brazilian Demographic Census (CENSO)⁶, the primary reference for understanding the living conditions of the population in all municipalities of the country and their territorial subdivisions, with the household resident as the unit of collection on the reference date across the National Territory, we created a scenario to predict high poverty rates in *grotas*⁷. The purpose is to support decision-making for implementing targeted credit policies.

2. **School mental health**

Based on data from PeNSE⁸, the National School Health Survey, conducted by IBGE in partnership with the Ministry of Health⁹ and with the support of the Ministry of Education¹⁰, and referencing the registry of public and private schools from the National Institute of Educational Studies and Research Anísio Teixeira (INEP)¹¹, we created a scenario to predict which schools are at higher risk of mental health issues among students. The purpose is to support decision-making actions such as allocating psychologists and implementing mental health prevention programs.

3. **The occupational situation of the individual**

The PNADc¹², Continuous National Household Sample Survey, developed by IBGE, aims to monitor quarterly fluctuations, as well as the short-, medium-, and long-term evolution of the labor force, along with other information necessary for studying the country's socioeconomic development. In this scenario, we predict the

probability of unemployment based on variables such as sector of activity, family income, and educational level. The purpose is to provide support for unemployment programs.

4. **Prediction of diabetes in adults**

The PNS¹³, National Health Survey, the result of an agreement between IBGE and the Ministry of Health¹⁴, was first conducted in 2013 to expand the thematic scope of the Health Supplements of the National Household Sample Survey (PNAD), which had been investigated by IBGE until 2008. In this scenario, we propose to predict the prevalence of diabetes in adults aged 35 years or older, with the aim of supporting decisions to optimize supplementary health and quality-of-life programs, promote prevention campaigns, and expand access to medicines.

5. **Food insecurity**

The POF¹⁵, Household Budget Survey, conducted by IBGE, evaluates consumption structures, expenses, income, and part of the wealth variation of families, providing a profile of the population's living conditions through the analysis of household budgets. In this scenario, we predict food insecurity risks based on food expenditures, offering decision support for optimizing access to programs such as the Family Grant and the National Food and Nutrition Plan.

6. **Difficulties with writing exam**

Research conducted by the National Institute of Educational Studies and Research Anísio Teixeira (INEP) through the National High School Examination ENEM¹⁶, created in 1998 to evaluate the performance of students completing high school, is used in this scenario to identify those most likely to obtain critical grades (below 600 points) in the writing exam. This scenario supports the investment of public policies aimed at strengthening critical and argumentative writing skills, particularly in regions with historical educational inequalities.

7. **Scholarships**

From the PROUNI¹⁷ database, a program that grants full and 50% partial scholarships in private higher education institutions for undergraduate and sequential specific training courses to Brazilian students without a higher education diploma, and based on research conducted by the Ministry of Education, we built a scenario to predict the type of scholarship (full or partial) awarded to a PROUNI candidate. The purpose is to contribute to the continuous improvement of the program and ensure greater transparency, equity, and efficiency in the allocation of public resources.

8. **Live births**

The SINASC¹⁸ survey provides data from the Information System on Live Births, developed by the Informatics Department of SUS (DATASUS)¹⁹ of the Ministry of

⁶Available through the Brazilian Institute of Geography and Statistics (IBGE).

⁷The term *grotas* is a regional Portuguese designation used in Maceió, Alagoas, to describe highly precarious, informal urban areas established along deep valleys, and steep slopes. Morphologically similar to favelas or slums, these areas present extreme socioeconomic vulnerability and severe deficits in basic infrastructure, sanitation, and safety.

⁸<https://www.ibge.gov.br/>

⁹<https://www.gov.br/saude/pt-br>

¹⁰<https://www.gov.br/mec/pt-br>

¹¹<https://www.gov.br/inep/pt-br>

¹²<https://www.ibge.gov.br/>

¹³<https://www.ibge.gov.br/>

¹⁴<https://www.gov.br/saude/pt-br>

¹⁵<https://www.ibge.gov.br/>

¹⁶<https://www.gov.br/inep/pt-br>

¹⁷<https://www.gov.br/mec/pt-br>

¹⁸<https://www.gov.br/saude/pt-br>

¹⁹<https://datasus.saude.gov.br/>

Health. SINASC is managed by the Health Situation Analysis Department of the Health Surveillance Secretariat, together with the State and Municipal Health Secretariats. In this scenario, we defined a model to identify vulnerable groups, aiming at early interventions that support better allocation of neonatal ICUs, the Maternity Benefit Program, and the training of community health agents.

9. Digital inclusion

The ICT²⁰ Households survey has been conducted annually since 2005 to map access to ICT infrastructure in urban and rural households across the country, as well as the ways in which these technologies are used by individuals aged 10 years or older. In 2013, ICT Households also incorporated ICT Children, which investigates ICT use among individuals aged 5 to 9 years and has been conducted separately since 2009. The survey is carried out by the Regional Study Center for the Development of the Information Society (Cetic.br), a department of the Brazilian Network Information Center (NIC.br).

10. Children's technological vulnerability

With the central objective of mapping possible online risks and opportunities, the first wave of the ICT Kids²¹ Online Brazil survey was conducted in 2012. The study generates indicators of internet use by children and adolescents aged 9 to 17. It seeks to understand young people's perceptions regarding online safety and to outline the mediation practices of parents and guardians related to internet use. In this scenario, we created a model to identify factors associated with negative online experiences among children and adolescents. These predictions allow policymakers to promote the development of preventive interventions and evidence-based public policies.

3.1.4 Datasets

In this step, we identify, in each research developed by public bodies, the respective available database, safeguarding ethical procedures to preserve identity and minimize risks. The preparation of the datasets followed the procedures below:

- Extraction of data relevant to the chosen scenario, standardized in CSV format (comma-separated values), where each row represents a data entry;
- Anonymized open public data to ensure privacy;
- Detailed documentation of the data source (Data Dictionary);
- Clear definition of time and geographic boundaries;
- Stratified sampling to ensure representativeness of minority groups;
- Documentation of inclusion and exclusion criteria (Document for each developed scenario);
- Preparation of multiple dataset versions:
 1. Raw – the original version of the data, collected directly from the source;
 2. Preprocessed – after collection, including cleaning and transformation steps such as noise removal,

normalization, format standardization, missing value treatment, and relevant feature extraction;

3. Modeling-ready – the final version, fully prepared for direct use in Machine Learning algorithms. At this stage, the data is organized, stratified with respect to the target and the sensitive attributes, partitioned into training and test sets under the 80/20 split, and with the relevant variables selected. No class re-sampling was applied: the original class imbalance of each scenario was preserved, which is precisely why model selection relies on CV AUC-ROC rather than on global accuracy (Section 3.2.5).

To ensure full transparency, scientific rigor, and absolute reproducibility of our predictive pipeline, Table 1 summarizes the quantitative characteristics of the microdata used across all ten public policy scenarios. We clarify that the values reported in the “Total Records (Samples)” column correspond to the final preprocessed, modeling-ready datasets prior to the partition of the data. For all evaluated scenarios, these complete samples were uniformly subjected to a standardized 80/20 train-test split. The 20% partition reserved for testing corresponds precisely to the exact volume of audited records analyzed during the subsequent AEQUITAS fairness evaluation phase in Section 4.

3.2 Machine Learning Pipeline

Through a structured process composed of interconnected steps that guide the transformation of raw data into useful predictions, we applied a supervised learning pipeline. We trained the model with labeled data using a binary classification technique designed to predict the category to which each instance belongs.

Supervised learning algorithms have been consolidated as decision-support tools in public policies, especially in contexts where binary classification is essential for allocating resources, identifying risks, and defining priorities in government interventions. Binary classification systems, which categorize data into two distinct classes, are particularly valuable for public managers, who often must decide between yes/no, eligible/ineligible, or high risk/low risk Floridi *et al.* [2018] Li *et al.* [2023].

3.2.1 Exploratory Analysis

We conducted exploratory data analysis (EDA), identifying potential biases and disparities among demographic groups in the collected data. Without this initial investigation, we would risk building models based on erroneous, incomplete, or distorted data, which could lead to wrong conclusions and poor decisions—especially in government public policy projects, where the impact can be disastrous.

We started the process by examining the variables of the ten datasets, using the available CSV files, the data dictionary, and the scenario definition: (1) we checked the names and descriptions of the columns; (2) their respective data types; and (3) we searched for missing or incomplete data.

²⁰<https://cetic.br/>

²¹<https://cetic.br/>

Table 1. Quantitative summary of the datasets across the ten social scenarios, detailing the source, timeframe, and total records of the final modeling-ready samples prior to the 80/20 train-test split.

Public Policy Scenario	Government Agency	Exact Timeframe	Total Records (Samples)
1. Poverty Prediction (CENSO)	IBGE	2010	93,245
2. School Mental Health (PeNSE)	IBGE / MS / MEC	2019	162,010
3. Occupational Situation (PNADc)	IBGE	4th Quarter 2024	470,225
4. Diabetes in Adults (PNS)	IBGE / MS	2019	279,360
5. Food Insecurity (POF)	IBGE	2017–2018	50,000
6. Exam Writing Difficulties (ENEM)	INEP	2023	50,000
7. Scholarships Allocation (PROUNI)	MEC	2020	166,760
8. Live Births Vulnerability (SINASC)	MS (DATASUS)	1979–2023	23,140
9. Digital Inclusion (ICT)	CETIC.BR	2024	20,487
10. Tech Vulnerability (TICKids)	CETIC.BR	2024	2,610

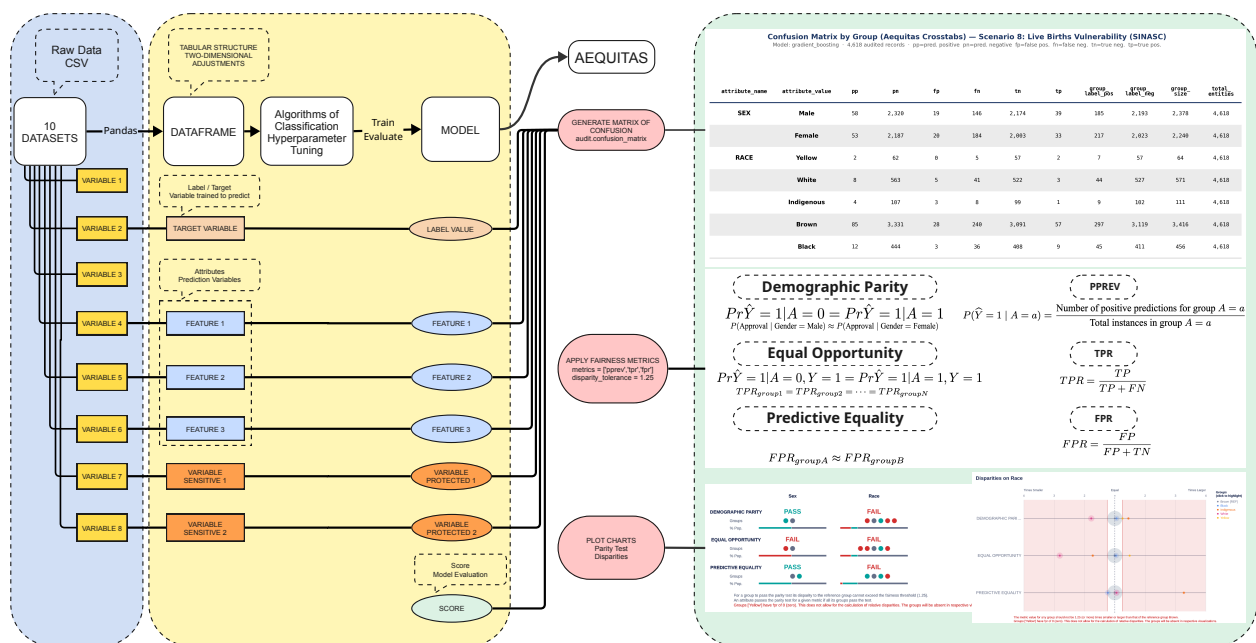


Figure 2. Detail of the structure of the Machine Learning Pipeline and Fairness Audit. The *Score / Model Evaluation* node corresponds to the CV AUC-ROC computation used as the sole selection criterion; the audit block reports the parameters effectively applied to the AEQUITAS Audit call (`metrics = ['pprev', 'tpr', 'fpr'], disparity_tolerance = 1.25`, majority-group reference selection).

We then verified values outside the expected range, detected outliers and inconsistent records, and promoted investigation and adjustment by omitting or correcting such cases. Finally, we evaluated the patterns and relationships between variables and identified inconsistent or missing data, correcting them by either removing or imputing values.

Figure 2 presents the detailed architecture of our Machine Learning pipeline and Fairness Audit, mapping the tabular structure adjustments—from raw variables to defined target and sensitive features—down to the application of the AE-QUITAS toolkit and its generated fairness evaluation matrices.

As detailed in Figure 2, the overarching architecture of our Machine Learning pipeline and Fairness Audit is thoroughly mapped. The flowchart is structured into three continuous

phases: the initial block (Government) outlines the microdata extraction and public policy scenario definition; the central block (ML Pipeline) encapsulates the EDA, variable selection, and hyperparameter tuning processes; and the final block (Audit) illustrates the integration of the AEQUITAS toolkit to generate confusion matrices and fairness evaluations.

3.2.2 Dataframe

Using the open-source Pandas library for data manipulation and analysis, we organized the two-dimensional tabular structures, or dataframes. We adhered to the recommended practices for transparency and reproducibility:

1. Hierarchically structuring the dataframes (raw, clean, feature, modeling):

2. Organizing detailed metadata documentation (variable types, units, definitions);
3. Implementing data integrity checks;
4. Creating repeatable transformation pipelines;
5. Versioning the dataframes at each significant step;
6. Implementing validation tests to ensure consistency.

3.2.3 Variables

A fundamental point in the structuring of the predictive model is the selection of target variables (label values) and predictor variables (features), in addition to sensitive variables, which we selected by considering both technical and ethical aspects:

a) **Definition of the target variable (label value):**

The target variable is the one on which the model is trained to make predictions in supervised learning; it represents the correct answer associated with each training example. Therefore, we considered:

1. the relevance of the variable for the selected public policy scenario;
2. its potential impact in terms of equity and social justice;
3. the quality and reliability of its measurement;
4. possible biases in the definition of the variable itself; and
5. the distribution of the variable across different demographic groups.

b) **Definition of predictor variables (Features):**

Predictor variables describe the data and serve as input for the model, representing the information both quantitatively and qualitatively. The model uses these variables to learn and make predictions.

We selected the features by choosing those that added value and discarding those of little use, which reduced the model's performance. For feature engineering, we transformed some variables to facilitate the model's interpretation, creating new features from existing ones and modifying others.

c) **Definition of sensitive variables:**

Sensitive variables in a data and machine learning context are those considered critical from ethical, legal, or privacy perspectives. Attributes that can identify, discriminate, or expose personal information are protected by legislation, such as the General Data Protection Law (LGPD) ²² in Brazil and the General Data Protection Regulation (GDPR) ²³ in Europe. Both set strict rules regarding the collection, storage, and use of personal data. In addition to legal requirements, special attention is given to fairness in machine learning, aiming to identify algorithmic biases in models. When decisions are based on sensitive variables (e.g., gender, ethnicity, disability, socioeconomic status), the results can be considered unfair. In the context of this research, we

defined gender and ethnicity as sensitive variables in all established public policy scenarios.

3.2.4 Data Science

After completing the data exploration, we applied data science techniques to organize the dataframe. We used interpolation methods and other approaches to handle missing and inconsistent data.

In the preprocessing step, the data were prepared for training by correcting issues identified during the exploratory phase, such as incomplete or incorrect records, and ensuring proper organization and transformation. We used the interpolation technique (Little and Rubin [2019]) to fill existing gaps in the dataframe. This method estimates values that were not recorded, allowing for a more complete and accurate analysis. Interpolation is particularly useful in areas where data integrity is fundamental for informed decision-making, such as public policy scenarios. The technique calculates missing values between two points by assuming default values among known points, thus avoiding the problem of Machine Learning models becoming inaccurate due to missing information.

3.2.5 Models

Models represent the specific materialization of an algorithm after being trained with particular data, i.e., a computer program capable of extracting patterns and making autonomous decisions as a result of applying algorithms to a dataset.

a) **Model selection**

The eight algorithms chosen are suitable for the context of supervised learning with binary classification. In public policy contexts, developing predictive solutions with Machine Learning prioritizes models that reconcile good predictive performance, interpretability, robustness, and ease of implementation. Recent literature and experiences point to the following models as the most frequently used:

1. **Logistic Regression**

Despite the term "regression," it is a classification algorithm. It models the probability of an instance belonging to a specific class using a logistic function to map the output of a linear equation between 0 and 1. If the calculated probability exceeds a set threshold, the instance is classified as positive; otherwise, it is classified as negative. "*Logistic Regression Analysis of Graduate Student Retention*" (Pyke and Sheridan [1993]). Although focused on student retention, this study uses Logistic Regression to identify variables that predict course completion, which is analogous to predicting the success or failure of a policy based on specific indicators.

2. **KNN (k-Nearest Neighbors)**

This algorithm is often called "lazy" because it does not learn a discriminative model explicitly but memorizes the entire training set. To classify a new point, it looks at the nearest k training points (its "neighbors") and assigns the most common class among them. "*Introduction to Machine Learning: k-nearest neighbors*" (Zhang [2016]). Presents KNN and its use in classification, for

²²LGPD - Lei nº 13.709/2018. Available at: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm

²³GDPR - Regulation (EU) 2016/679. Available at: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

example, in medical contexts. Although medical, the principle of classifying new instances based on similarity with known cases can be applied to public policies — for example, assessing whether a new policy proposal has a chance of success based on similar policies implemented in the past; *"High-Level K-Nearest Neighbors (HLKNN): A Supervised Machine Learning Model for Classification Analysis"* (Ozturk Kiyak et al. [2023]). Discusses the broad applicability of KNN in several fields, including GIS (Geographic Information Systems), which may be relevant to spatially dependent policies.

3. SVM (Support Vector Machine)

SVM seeks to find the "hyperplane" that best separates two classes in the feature space. It maximizes the margin between the hyperplane and the nearest data points of each class (the "support vectors"). Kernels can map data to higher dimensions, allowing non-linear separations. In applied contexts, *"Application of SVM to Speed Up and Accurate Nursing Decisions for Mentally Disordered Patients"* (Pohan et al. [2024]) demonstrates the use of SVM to support clinical decisions, functioning as an effective decision-support system. Building upon such methodological success, *"Support Vector Machines: A Literature Review on Their Application in Analyzing Mass Data for Public Health"* (Khyathi et al. [2025]) reviews the broader use of SVMs in analyzing public health data for disease classification and health trend forecasting, highly supporting decision-making in public health policies.

4. Decision Tree

This algorithm builds a classification or regression model in the form of a tree structure. It divides the dataset into smaller, more homogeneous subsets based on feature values, asking sequential questions as in a flowchart until a decision is reached at a "leaf" node. Early applications, such as *"Evaluation of Student's Performance Using Decision Tree Model"* (Hasan et al. [2018]), and subsequent studies like *"Decision Tree Algorithm Use in Predicting Students' Academic Performance in Advanced Programming Course"* (MURAINA et al. [2022]), successfully applied decision trees to predict student performance—a type of modeling that can be easily adapted to predict the "performance" or "impact" of a public policy. More recently, *"Reframing Educational Inquiry: Uncovering the Value of Decision Tree Analysis — A Practical Guide for Educational Researchers"* (Du [2025]) highly highlights the transparency of decision trees, a characteristic that is crucial in public administration to ensure that automated decisions remain understandable and justifiable.

5. Random Forest

An ensemble learning method that builds multiple decision trees during training and produces a prediction by taking the mode of classes (classification) or the mean (regression) across individual trees. Each tree is trained on a random sample of the data and a random subset of features, reducing overfitting and increasing robustness. *"Predicting University Students' Academic Success and Major using Random Forests"* (Beaulac and Rosenthal [2019]). Uses Random Forests to predict academic suc-

cess and course choice. The method's ability to identify the importance of variables is very useful for understanding which factors are most determinant for policy success.

6. Gradient Boosting

Another ensemble method that builds models sequentially, where each new model corrects the mistakes of the previous one. It "boosts" learning by focusing on misclassified instances. Gradient boosting is powerful for capturing complex relationships in data and often achieves high discriminative performance in predicting policy outcomes. *"An Explained Extreme Gradient Boosting Approach for Identifying the Time-Varying Determinants of Sovereign Risk"* (Giraldo et al. [2023]). Applies XGBoost to identify sovereign risk determinants, relevant for macroeconomic analysis and public investment decisions; *"Analysis of Gradient Boosting, XG-Boost, and CatBoost on Mobile Phone Classification"* (Ptr et al. [2024]). Although focused on mobile phone ratings, it demonstrates the high performance of boosting algorithms in complex classification tasks.

7. Naive Bayes

A probabilistic classifier based on Bayes' theorem with a "naive" assumption of conditional independence among features. That is, it assumes that the presence of a particular feature is unrelated to the presence of others. Despite its simplicity, it often performs surprisingly well. Foundational literature, such as *"Naive Bayes and Text Classification I - Introduction and Theory"* (Raschka [2014]), along with contemporary works like *"An Introduction to the Naive Bayes Classifier and How to Use It"* (Cadili [2024]), robustly demonstrate its theoretical bases and modern utility. While not exclusively applied to public policies in these specific works, the method is widely used in text classification. In governmental contexts, it could be optimally applied to classify proposal documents, analyze public feedback, or identify emerging issues in social discourse.

8. MLP (Multilayer Perceptron)

A type of feedforward artificial neural network composed of at least three layers of nodes: an input layer, one or more hidden layers, and an output layer. Each node in the hidden and output layers uses a nonlinear activation function. MLP learns complex and nonlinear mappings between inputs and outputs through back-propagation, adjusting connection weights to minimize prediction error. In the socio-medical domain, *"Multilayer Perceptron Modeling for Social Dysfunction Prediction Based on General Health Factors in an Iranian Women Sample"* (Bagheri et al. [2023]) applies MLP to predict social dysfunction, highlighting its enormous potential in social science research to accurately identify important predictors. Extending its capacity, *"Predicting Student Academic Success Using Deep Learning: A Multi-Factor Approach to Performance Prediction"* (Al Husaini et al. [2025]) develops an MLP model to predict student academic success, effectively proving that MLPs can outperform traditional predictive models in complex forecasting tasks involving multiple socio-demographic variables.

b) Hyperparameters

After separating the data into independent predictor variables (features) and the dependent target variable, we defined the external settings of the model that control the training process and directly affect the performance and efficiency of the algorithms: the selection and adjustment of hyperparameters. Unlike the internal parameters learned during training, hyperparameters are set before the process begins and require manual or automated adjustment for optimization.

After defining the search space of the most common hyperparameters for each algorithm, we applied the Grid Search optimization technique. This method systematically explores and exhaustively scans the specified combinations of hyperparameters, selecting the optimal configuration that maximizes the 5-fold Cross-Validated Area Under the ROC Curve (CV AUC-ROC). This optimization setup guarantees that the hyperparameter tuning process is robust against severe class imbalance, prioritizing class discriminability and baseline predictive quality over global accuracy.

c) Training and evaluation

After configuring the hyperparameter grid, we proceeded to the training and evaluation phase for each binary classification model. A critical methodological consideration in public policy datasets is the severe class imbalance, which renders global accuracy a highly misleading metric. Optimizing solely for accuracy often results in models that superficially favor the majority class while systematically failing to predict minority occurrences. To ensure algorithmic robustness and absolute transparency in our model selection prior to the fairness audit, 5-fold Cross-Validated AUC-ROC (CV AUC-ROC) was utilized as the sole decisive criterion to determine the optimal algorithm for each scenario. AUC-ROC is insensitive to classification thresholds and is highly recommended for imbalanced domains. Other metrics (Accuracy and F1-Macro) were calculated and documented strictly as secondary, multi-dimensional evaluation parameters, with no influence on the final model choice. Based on this unified CV AUC-ROC selection criterion, Gradient Boosting was selected in seven scenarios, while Random Forest attained the highest CV AUC-ROC in the remaining three (Food Insecurity, Digital Inclusion, and Children's Tech Vulnerability). Table 2 presents a comparative analysis of the top-performing algorithms across all ten social scenarios, providing empirical justification for our baseline models before the subsequent AEQUITAS fairness evaluation phase.

3.3 Audit

Throughout all stages of the Machine Learning pipeline for developing predictive models in the ten public policy decision-support scenarios, we considered issues related to fairness. However, such care alone is insufficient, as it lacks a mechanism to visually demonstrate potential biases in protected groups. Therefore, we proceeded with the indispensable step

of auditing, using a fair Machine Learning tool. Highly sensitive topics require increased attention in the search for evidence of biases that may infiltrate through data, algorithms, and user interactions.

3.3.1 Aequitas toolkit

Among the numerous tools developed to support fairness audits in Machine Learning, the most cited in academic publications are Aequitas (Saleiro *et al.* [2018]), AI Fairness 360 (AIF360) (Bellamy *et al.* [2019]), and Fairlearn (Bird *et al.* [2020]). We opted for Aequitas, a fairness-oriented Machine Learning toolkit, because it:

- is open source;
- provides technical interaction for data scientists and developers within the Machine Learning pipeline, rather than for public policy managers, allowing external audit of models before deployment;
- offers multiple interfaces to check disparities through reports and graphs.

With the Aequitas toolkit, we audited each of the ten public policy scenarios defined with their respective pre-trained models:

- we identified and standardized the sensitive/protected attributes for the ten scenarios as Gender and Ethnicity, and adopted the majority-group reference selection provided by the Audit API of Aequitas 1.1.0, with a disparity tolerance of 1.25. This selection yielded the Brown group as the racial reference in all ten scenarios and, for the sex attribute, Female in the CENSO, PeNSE, PNADc, PNS, POF, ENEM, PROUNI and ICT scenarios and Male in SINASC and ICT Kids. All disparity ratios reported in Section 4 and in Appendix A must be read against these references;
- we prepared a specific dataframe for intersectional analysis;
- we structured the data to facilitate the application of fairness metrics.

3.3.2 Confusion matrix

We created a new dataframe and prepared it with the protected groups. We then generated the confusion matrix to support the analysis. Since the fairness metrics audited in this study (PPREV, TPR and FPR) are derived from hard class assignments, the predicted probabilities were binarised using a fixed classification threshold of 0.50, applied uniformly across the ten scenarios and to both selected architectures. We stress that this threshold governs the audit stage only: model selection relies on CV AUC-ROC, which integrates over all thresholds and is therefore unaffected by this choice.

Designed to visualize and evaluate the performance of a binary classification algorithm, the confusion matrix presents a detailed summary of how the model classifies instances by comparing actual values with predicted values. Understanding both the overall discriminative capacity and the specific errors the model makes is particularly useful. In this way:

Table 2. Comparative performance analysis of the top-performing Machine Learning models across the ten public policy scenarios. Selection was strictly determined by the highest Cross-Validated AUC-ROC (CV AUC-ROC) to account for dataset imbalance. Hold-out AUC-ROC, F1-Macro and Accuracy are reported as secondary parameters and had no influence on the selection.

Public Policy Scenario	Algorithm	CV AUC-ROC (selection)	AUC-ROC (hold-out)	F1-Macro (secondary)	Accuracy (secondary)	Status
1. Poverty Prediction (CENSO)	Gradient Boosting	73.30%	74.15%	61.25%	87.66%	Selected
	Random Forest	72.99%	73.60%	62.39%	75.98%	
	Decision Tree	72.96%	73.72%	62.30%	75.90%	
2. School Mental Health (PeNSE)	Gradient Boosting	68.03%	68.39%	60.63%	64.81%	Selected
	Random Forest	67.96%	68.24%	62.12%	62.40%	
	MLP	66.16%	67.38%	61.56%	63.48%	
3. Occupational Situation (PNADc)	Gradient Boosting	91.78%	91.81%	80.91%	81.89%	Selected
	Random Forest	91.73%	91.77%	79.80%	79.82%	
	Decision Tree	91.55%	91.63%	79.73%	79.74%	
4. Diabetes in Adults (PNS)	Gradient Boosting	99.81%	99.80%	98.64%	98.86%	Selected
	Random Forest	99.79%	99.79%	98.28%	98.54%	
	Decision Tree	99.77%	99.76%	98.34%	98.60%	
5. Food Insecurity (POF)	Random Forest	91.03%	93.66%	83.12%	87.82%	Selected
	Gradient Boosting	84.33%	84.06%	71.04%	79.52%	
	KNN	78.36%	84.40%	74.92%	82.52%	
6. Exam Writing Difficulties (ENEM)	Gradient Boosting	87.66%	87.66%	76.68%	78.70%	Selected
	Random Forest	87.38%	87.45%	76.61%	77.34%	
	Decision Tree	86.20%	86.36%	75.69%	76.63%	
7. Scholarships Allocation (PROUNI)	Gradient Boosting	92.99%	93.09%	77.78%	86.45%	Selected
	Random Forest	91.20%	92.04%	80.22%	86.01%	
	Decision Tree	86.90%	88.45%	71.91%	75.47%	
8. Live Births Vulnerability (SINASC)	Gradient Boosting	79.01%	79.22%	61.92%	92.03%	Selected
	Random Forest	78.84%	79.25%	72.25%	89.76%	
	Logistic Regression	78.69%	79.11%	70.79%	88.61%	
9. Digital Inclusion (ICT)	Random Forest	98.37%	98.18%	92.59%	95.39%	Selected
	Gradient Boosting	98.34%	98.20%	92.77%	95.58%	
	MLP	98.14%	98.20%	92.33%	95.36%	
10. Tech Vulnerability (ICT Kids)	Random Forest	78.09%	77.04%	72.05%	86.02%	Selected
	Gradient Boosting	76.65%	76.76%	73.23%	86.97%	
	SVM	73.89%	77.43%	73.05%	86.59%	

- it organizes predictions in a table with respect to real classes, separating positive and negative classes;
- it correctly predicts: (1) the positive class, generating a true positive (e.g., eligible beneficiaries identified as such); and (2) the negative class, generating a true negative (e.g., non-eligible beneficiaries identified as such);
- it incorrectly predicts: (1) the positive class, generating a false positive (e.g., non-eligible people classified as eligible), which may indicate bias in resource allocation; and (2) the negative class, generating a false negative (e.g., real beneficiaries excluded), signaling under-representation bias for vulnerable groups;
- rows represent the instances of the actual class, while columns represent the instances of the predicted class (or vice versa);
- correct forecasts appear on the diagonal of the matrix,

making errors easier to inspect since they appear outside the diagonal.

In fairness audits, the confusion matrix is analyzed separately by demographic groups. Critical results include significant differences in false positive and false negative rates between protected and unprotected groups, systematic error patterns that reinforce historical inequalities, and trade-offs between baseline predictive performance (such as CV AUC-ROC) and algorithmic equity metrics.

3.3.3 Metrics

For performance metrics, Aequitas evaluates models by group, checking whether they perform equally well across all groups. For example, in granting a specific benefit directed by public policy, a low false positive rate helps avoid costly errors.

Fairness metrics, in turn, verify whether the model treats all groups equitably, analyzing disparities between them. For instance, in the demographic parity fairness metric, the tool checks whether the approval rate is equal for all groups.

The application of these metrics generates graphical views of potential algorithmic discrimination and historical bias in pre-trained models, supporting the design of mitigation strategies. Performance metrics should always be analyzed alongside fairness metrics to provide a more comprehensive assessment of potential discriminatory impacts.

In each of the ten public policy scenarios, with the optimal baseline models selected strictly according to the highest 5-fold Cross-Validated AUC-ROC (Gradient Boosting for seven scenarios and Random Forest for three), we conducted audits after generating the confusion matrix to verify disparities that indicate possible bias. In each scenario, we applied the following three fairness metrics:

a) Demographic parity

A fundamental metric for evaluating algorithmic fairness, especially in public policies where equity in allocating resources or opportunities is critical. It requires that a model's favorable prediction rate (PPREV) be statistically independent of sensitive attributes. Its relevance lies in detecting systemic disparities that perpetuate historical inequalities.

The mathematical expression is strictly defined as statistical independence between the model's prediction and the sensitive attribute. It requires the marginal probability of a positive prediction to be equal across all groups, formally expressed as:

$$Pr\hat{Y} = 1|A = 0 = Pr\hat{Y} = 1|A = 1 \quad (1)$$

Where:

$\hat{Y}$: the model's prediction (1 = approved, 0 = disapproved); A : the sensitive demographic attribute (e.g., gender, race); 0, 1: the distinct group values evaluated within the sensitive attribute, where one of them is the reference group—defined in this study as the modal category of the attribute—and the other is any of the groups contrasted against it. The reference is therefore not assumed to be a non-protected or privileged group.

In practice, the rate of positive predictions (PPREV) should be similar across different groups, ensuring that no group systematically receives a higher or lower proportion of positive outcomes.

The calculation of disparity in the demographic parity metric involves quantifying the difference or ratio between the positive prediction rates of groups defined by sensitive attributes. It consists of: (1) collecting the positive prediction rate for each group individually, given by:

Equation 2 defines the probability of a positive prediction for a given group:

$$P(\hat{Y} = 1 | A = a) = \frac{N_{\text{pos}}(a)}{N_{\text{total}}(a)} \quad (2)$$

where $N_{\text{pos}}(a)$ represents the number of positive predictions for group $A = a$, and $N_{\text{total}}(a)$ denotes the total number of instances within the same group.

(2) identifying the reference group, which the toolkit sets by default as the demographic group with the largest sample representativeness within each sensitive attribute rather than as a function of the observed metric values—a behaviour that keeps the reference constant across all three metrics and allows disparity ratios to assume values both above and below 1.0; (3) assessing disparity either by absolute difference, which measures the percentage gap between each group's positive prediction rate and that of the reference group, or by relative ratio, which measures the proportion between the positive prediction rate of each group and that of the reference group.

For example, in a credit-granting model where gender is the sensitive variable, the demographic parity metric verifies whether the proportion of credit approvals—i.e., positive predictions—between men and women is similar:

$$P(\hat{Y} = 1 | A = m) \approx P(\hat{Y} = 1 | A = f) \quad (3)$$

where $\hat{Y} = 1$ denotes the credit approval (positive prediction), and $A \in \{m, f\}$ represents the sensitive attribute for gender (male and female, respectively). Thus, if the approval rate is 70% for men and 40% for women, demographic parity is violated, indicating a possible disparity between the two groups.

b) Equal opportunity

This is a fundamental metric for evaluating fairness in Machine Learning systems, especially in contexts where algorithmic decisions affect public policies and the equitable distribution of resources. It evaluates whether a classification model provides equal opportunities for qualified individuals across different protected demographic groups.

We achieve equal opportunity when a classifier has the same true positive rate (TPR) across demographic groups. In other words, for individuals who deserve the positive outcome, the probability of receiving a positive prediction should be the same regardless of the group to which they belong. Formally, a binary classifier $\hat{Y}$ satisfies equal opportunity with respect to the sensitive attribute A (such as race or gender) and the true label Y when individuals strictly belonging to the positive class ($Y = 1$) have equal probabilities of receiving a correct positive prediction. This ensures equal True Positive Rates (TPR):

$$Pr\hat{Y} = 1|A = 0, Y = 1 = Pr\hat{Y} = 1|A = 1, Y = 1 \quad (4)$$

This mathematical definition guarantees that qualified individuals ($Y = 1$) have the exact same likelihood of receiving an accurate positive prediction ($\hat{Y} = 1$), regardless of their demographic group ($A=0$ or $A=1$). In this way:

$$TPR_{\text{group}_1} = TPR_{\text{group}_2} = \dots = TPR_{\text{group}_N} \quad (5)$$

$$\text{where: } TPR = \frac{TP}{TP + FN} \quad (6)$$

TP (True Positives): the number of positive examples correctly classified as positive;
 FN (False Negatives): the number of positive examples incorrectly classified as negative;

To measure equal opportunity, we proceed as follows: (1) divide the test data by demographic groups of interest, such as protected attributes (e.g., race, gender); (2) calculate the TPR for each group by counting the number of true positives and false negatives and applying the TPR formula; (3) compare the TPRs across groups to evaluate whether the disparity between them is acceptable. For example, consider a pre-trained model for school drop-out classification with the following metrics: TPR (male) = 0.85 and TPR (female) = 0.65. In this case, equal opportunity is not satisfied because there is a significant difference between the TPRs, indicating preference for group A (male) over group B (female). This demonstrates bias and highlights the need for model adjustments or data balancing.

c) Predictive Equality

The predictive equality metric is particularly relevant for analyzing biases in systems used for the development and implementation of public policies. It is based on the premise that different demographic groups should experience equivalent false positive rates (FPR), making it an essential criterion to ensure fair and equal treatment across populations.

We define predictive equality as a mathematical condition in which the probability that an individual belonging to the actual negative class receives a positive prediction must be identical across distinct groups.

Formally, for a classifier operating on protected and non-protected groups, the metric requires that:

$$P(\hat{Y} = 1 \mid Y = 0, G = w) = P(\hat{Y} = 1 \mid Y = 0, G = b) \quad (7)$$

where $G \in \{w, b\}$ represents the race attribute in a generic illustrative formulation, with w standing for White and b for Black. In the audits reported in this study, the reference group is the Brown group, and the remaining categories (White, Black, Yellow and Indigenous) are each contrasted against it.

Also known as the False Positive Rate (FPR) balance or delta false positive rate, this metric refers to the condition in which both protected and privileged groups have equal FPR. The FPR represents the probability that an instance belonging to the negative class is incorrectly predicted as positive. The mathematical definition is as follows:

$$FPR = \frac{FP}{FP + TN} \quad (8)$$

FP (False Positives): cases where the model predicts a positive outcome, but the true label is negative;
 TN (True Negatives): cases where the model correctly predicts a negative outcome.

When the predictive equality metric is satisfied, we have:

$$FPR_{g_1} \approx FPR_{g_2} \quad (9)$$

where g_1 and g_2 represent the distinct demographic groups under evaluation. In the audits reported here, g_1 is always the reference group (Brown for the race attribute) and g_2 is each of the remaining categories in turn, so that every ratio reported in Section 4 and in Appendix A has the reference group in the denominator. In a fair system, from the perspective of the Predictive Equality metric, the probability that a person is mistakenly classified as positive should be the same regardless of whether they belong to a protected group, such as women or Black individuals. This type of error is particularly relevant in applications where false positives can generate serious consequences, such as: (1) being mistakenly denied a loan because the applicant is considered high risk; and (2) being unfairly punished in predictive justice systems.

While Demographic Parity (PPREV) analyzes individuals in general, the metrics Equal Opportunity (TPR) and Predictive Equality (FPR) focus on qualified individuals. PPREV verifies the overall probability of positive outcomes; TPR ensures that correct predictions occur in the same proportion across groups; and FPR ensures that incorrect positive predictions occur in the same proportion.

When associating each of the three metrics with the ten public policy scenarios, we note:

1. in contexts that seek diversity, demographic parity is applicable by promoting proper equity verification;
2. the Equal Opportunity metric is more appropriate in critical contexts where correct predictions are essential, such as in health and justice;
3. Predictive Equality is useful for verifying fairness in contexts where false positives may generate disproportionate impacts, such as credit granting.

The practitioner defines the disparity tolerance limit applied to the metrics in each of the ten scenarios. The reference group, in turn, was not specified by the authors: it was obtained through the default reference selection of the Audit class in Aequitas 1.1.0, which falls back on the most frequent group within each sensitive attribute whenever the `reference_groups` argument is not supplied. This procedure yielded the Brown group as the racial reference across all ten scenarios, and sex references that vary according to the composition of each dataset. This majority-group default has two operational consequences for the present audit:

1. the reference is, by construction, the modal category of each sensitive attribute, and therefore the group with the largest sample representativeness and the largest proportion of the training data, which stabilises the denominator of the disparity ratios;
2. the reference is fixed by frequency alone and carries no normative claim: it is not selected as a privileged, non-discriminated or socially neutral group. Consequently, the ratios reported in Section 4 and in Appendix A measure deviation from the predictive pattern of the modal group, not from an ideal of equity.

Regarding the tolerance limit for disparity, the range [0.8 – 1.25], widely used in fairness studies, is based on guidelines established by the Equal Employment Opportunity Commission (EEOC). Fair machine learning tools adopt this convention as a practical threshold to detect disparities without overestimating minor statistical variances. However, a critical theoretical consideration in algorithmic fairness is the inherent mathematical incompatibility among different fairness metrics. As rigorously established by Chouldechova [2017] and Kleinberg *et al.* [2016], it is mathematically impossible to simultaneously satisfy Demographic Parity, Equal Opportunity, and Predictive Equality when the base rates (prevalence of the target variable) differ across protected demographic groups. Consequently, mitigating algorithmic bias requires navigating an unavoidable fairness trade-off. This impossibility theorem dictates that policymakers and machine learning practitioners cannot simply aim for “universal fairness” across all indicators. Instead, they must deliberately prioritize specific metrics based on the domain context, the punitive or assistive nature of the policy, and the socio-technical goals of the public administration. For instance, in health-related scenarios (such as predicting diabetes in adults or mapping neonatal vulnerabilities in SINASC), the cost of failing to identify a patient in need is critically high. Therefore, Equal Opportunity (focusing on True Positive Rates) often takes priority, ensuring that vulnerable individuals requiring medical assistance are not disproportionately overlooked. Conversely, in labor and socio-economic contexts (such as the occupational situation in PNADc or the allocation of PROUNI scholarships), historical and structural exclusions are deeply entrenched. In these cases, Demographic Parity might take precedence to guarantee a structurally fair distribution of resources and opportunities across demographic groups, actively combating historical marginalization even if it mathematically compromises perfect Equal Opportunity or overall predictive performance. By applying constraints to make the model fairer, predictive performance as measured by CV AUC-ROC may decrease, as the model is no longer optimized exclusively for discriminative power but also incorporates these targeted socio-technical fairness criteria. Our audit framework explicitly recognizes these trade-offs, guiding the choice of the primary fairness metric aligned with the specific objectives of each evaluated public policy.

3.3.4 Plot

The AEQUITAS toolkit provides Machine Learning practitioners, developers, data scientists, and public policy managers with an audit environment that generates visual resources and interactive reports for analyzing prediction scenarios in Machine Learning systems applied to public policies.

For each of the ten public policy scenarios built from open public data produced by five government agencies, we developed ten pre-trained models. For each scenario, we verified the optimal predictive model selected via 5-fold CV AUC-ROC after testing eight Machine Learning classification algorithms, and we conducted a fairness audit to detect possible biases using the three most traditional fairness metrics. To ensure full scientific reproducibility and methodological transparency, we documented and executed all pipeline processes

across the three levels (Government, ML Pipeline, and Audit) using an automated Python framework. The complete source code, execution guidelines, datasets, and containerized environment settings (Docker) are openly accessible in our public GitHub repository ²⁴.

These results allow public policy managers to interact with and monitor the possibility of bias after model implementation through web-based reports. The strength of the tool lies in its ability to plot graphs that test parity between sensitive demographic groups and show levels of disparity within each group. We conducted audits of the ten scenarios, generating two graphs per scenario: the Parity Test Dashboard and the Fairness Disparity Plot.

a) Parity Test Dashboard

Also known as the parity evaluation graph, this visual feature integrates bias audits into AI solutions that generate Machine Learning predictions. It is widely used in AEQUITAS and other fairness toolkits.

The fairness evaluation chart shows the results of the parity test for three fundamental metrics of algorithmic fairness: Demographic Parity, Equal Opportunity, and Predictive Equality. Its organizational structure is arranged as follows: on the top axis appear the sensitive attributes analyzed (sex and race); on the side axis appear the three fairness metrics; on the right, a legend displays the possible states of the parity tests.

To illustrate the visual components of the AEQUITAS toolkit, we present the parity evaluation graphs generated for the Food Insecurity (POF) risk prediction scenario (Figure 3). The visual elements are interpreted as follows: red indicates a failure in the parity test (FAIL), meaning statistically significant bias; cyan indicates approval (PASS), revealing acceptable equity; and gray represents the reference group. Colored circles identify the demographic groups within each sensitive attribute. In this didactic example, the graphical interface provides a clear visual dichotomy. For the gender-sensitive attribute, all three metrics receive a PASS evaluation (cyan), demonstrating that the Random Forest model achieved optimal equity between men and women in this specific prediction. Conversely, for the race-sensitive attribute, while Equal Opportunity receives a PASS, both Demographic Parity and Predictive Equality receive a FAIL evaluation (red). This visually alerts the practitioner to the presence of systemic disparities, indicating that positive prediction rates and false alarms are disproportionately distributed across racial groups.

Colored circles identify demographic groups, with several circles representing different racial groups and fewer circles representing sex groups. To visually demonstrate this mechanism, Figure 3 exhibits the Parity Test Dashboard corresponding specifically to the Food Insecurity scenario (developed from the POF research), clearly indicating which fairness metrics passed or failed the established threshold.

This type of dashboard plays a crucial role in fairness audits by quickly identifying metrics that fail parity tests, comparing multiple sensitive attributes simultane-

²⁴<https://github.com/George-Souza/Fairness>

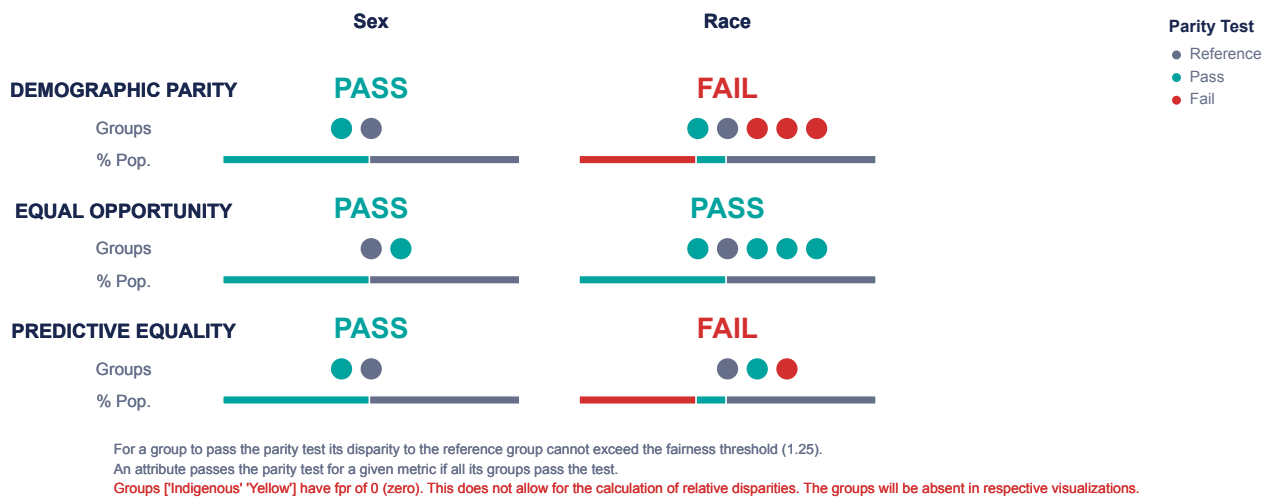


Figure 3. Parity Test Dashboard illustrating the visual elements (PASS/FAIL) in the Food Insecurity risk prediction scenario (POF).

ously, and visualizing compliance with equity standards, thereby supporting bias mitigation decisions.

The combination of approvals and disapprovals suggests that the audited model contains bias: different aspects of fairness are affected in distinct ways, requiring specific mitigation strategies for each metric and sensitive attribute.

b) Fairness Disparity Plot

AEQUITAS uses this visualization in Machine Learning audits to identify possible inequalities or biases in metrics across different sensitive demographic groups. The x-axis represents the ratio of sensitive groups (Yellow, White, Indigenous, and Black) to the reference group (Brown). The center marked "Equal" indicates that the group performs identically to the reference group. Values to the left ("times smaller") indicate lower performance, while values to the right ("times larger") indicate higher performance. The scale spans up to three times higher or lower, making discrepancies easier to identify. When the evaluated indicator yields positive results (e.g., access to opportunities, approvals, or benefits), values "times larger" represent overrepresentation, showing that the group receives more benefits than the reference group. When the indicator yields negative results (e.g., rejections, punishments), values "times larger" indicate greater exposure to adverse outcomes compared to the reference group, characterizing bias.

Each horizontal line corresponds to a traditional fairness metric in Machine Learning models: Demographic Parity (positive outcomes), Equal Opportunity (true positives), and Predictive Equality (false positives). Demographic Parity ensures that the probability of a positive outcome is independent of the sensitive attribute. Equal Opportunity requires equal true positive rates between groups. Predictive Equality requires equal false positive rates between groups.

The legend on the right side of the plot identifies the reference group in gray — in all racial audits reported in this study, the Brown group — and the other evaluated

groups (Yellow, White, Indigenous, and Black), each represented by circles with distinct colors to facilitate visual differentiation. The diameter of the circles can be configured to reflect the sample size of the respective demographic group.

To rigorously understand the magnitude of the bias, Figure 4 details the Fairness Disparity Plot for the race-sensitive attribute from the aforementioned Food Insecurity (POF) scenario. The shaded red areas represent the unacceptable disparity zones, mathematically defined as values exceeding the fairness threshold (times smaller than 0.80 or times larger than 1.25). The central dotted line marks perfect algorithmic equality (1.0). Circles falling within the shaded zones visually confirm severe algorithmic bias. For instance, in the Predictive Equality metric (bottom line), the orange circle representing the White group is positioned deeply into the right shaded area, indicating that this group suffers from a false positive rate more than twice as large as the reference group. Ideally, for a model to be considered strictly fair, all colored circles should align within the central unshaded band.

At the bottom of the chart, a message reminds the user of the fairness criterion: 'The metric value for any group should not be 1.25 (or more) times smaller or larger than that of the reference group (Brown).' Values above this threshold constitute possible evidence of unjustifiable bias. As explicitly detailed in Figure 4, the Fairness Disparity Plot for the Food Insecurity scenario allows for a granular visualization of these precise disparities across different demographic segments.

Disparities on Race

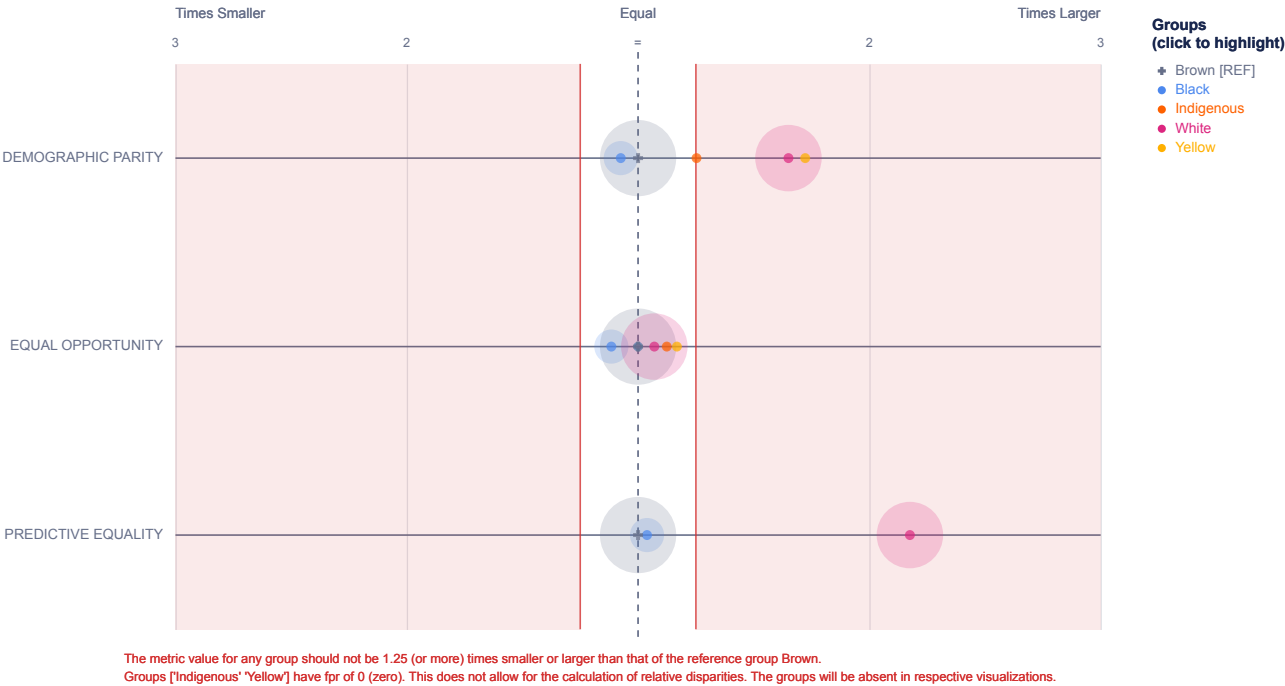


Figure 4. Fairness Disparity Plot (Race) for the Food Insecurity scenario (POF), highlighting the shaded red zones of unacceptable disparity and the precise distribution of demographic groups.

4 Results

To ensure quantitative transparency and to separate data-intrinsic bias from inductive model artifacts, we conducted our fairness audit on the classification model selected, in each scenario, as the best performing under the CV AUC-ROC criterion, a metric recommended for imbalanced domains. Following our CV AUC-ROC evaluation (detailed in Section 3.2.5), we applied the AEQUITAS toolkit to the final predictions generated by Gradient Boosting (in 7 scenarios) and Random Forest (in 3 scenarios). We note that in four scenarios the margin between the selected model and the runner-up is below 0.1 percentage points of CV AUC-ROC (Table 2); since no dispersion across folds is reported, the criterion should be read as a deterministic and uniformly applied selection rule rather than as evidence of statistically significant superiority.

Due to manuscript length constraints, the complete set of Parity Test Dashboards, Fairness Disparity Plots, and detailed analytical matrices containing the exact PPREV, TPR, and FPR values for all ten scenarios are provided in the Supplementary Appendix A. However, to illustrate the algorithmic behavior, Table 3 provides a consolidated synthesis of the fairness audit across all evaluated public policies.

Because the reference group is determined by sample frequency within each attribute, the racial reference is invariably the Brown group across all ten scenarios, whereas the sex reference varies with the composition of each dataset (Female in the CENSO, PeNSE, PNADc, PNS, POF, ENEM, PROUNI and ICT scenarios; Male in SINASC and ICT KIDS). All gender disparity ratios reported in Section 4 and in Appendix A must therefore be read against the reference group indicated in the respective scenario.

Table 3. Consolidated Synthesis: Main findings of the algorithmic fairness audit across the 10 social scenarios after robust model selection (CV AUC-ROC).

Analysis Dimension	Main Finding
Frequency of equity failures	Even with robust AUC-ROC model selection, 9 out of 10 scenarios presented severe equity violations, proving the bias is deeply data-intrinsic.
Most frequently failing metric	Predictive Equality (FPR) consistently failed, particularly for the race attribute across both Gradient Boosting and Random Forest predictions.
Most critical attribute	Race: systematic disparities across all non-modal groups relative to the Brown reference, with extreme deviations observed in Indigenous and Yellow groups (including predictive collapses / FPR=0) and pronounced deviations in the White group.
Attribute with better performance	Gender: Achieved Equal Opportunity (TPR) parity in eight of the ten scenarios; the exceptions are Poverty Prediction (CENSO) and Live Births Vulnerability (SINASC).
Reference scenario (Parity)	School Mental Health Risk (PeNSE): optimal equity in both sensitive attributes using Gradient Boosting.
Critical scenario (Disparity)	Live Births Vulnerability (SINASC): extreme systemic disparities (FPR 3.28 times larger for Indigenous infants) using Gradient Boosting.

The quantitative analysis directly exposes that optimizing models with threshold-independent metrics (AUC-ROC) does not cure underlying historical disparities. Demographic parity (PPREV) and predictive equality (FPR) systematically failed the parity tests for the groups contrasted against the modal reference. To objectively demonstrate these extremes, we present the graphical and tabular results of two representative scenarios in Sections 4.1 and 4.2.

4.1 Representative Case I: Algorithmic Parity (School Mental Health Risk)

Based on the 2019 PeNSE survey, the model predicting schools at higher risk of mental health issues achieved remarkable equity. As shown in the AEQUITAS evaluation, the model received parity approval (PASS) across all three fairness metrics for both gender and race attributes.

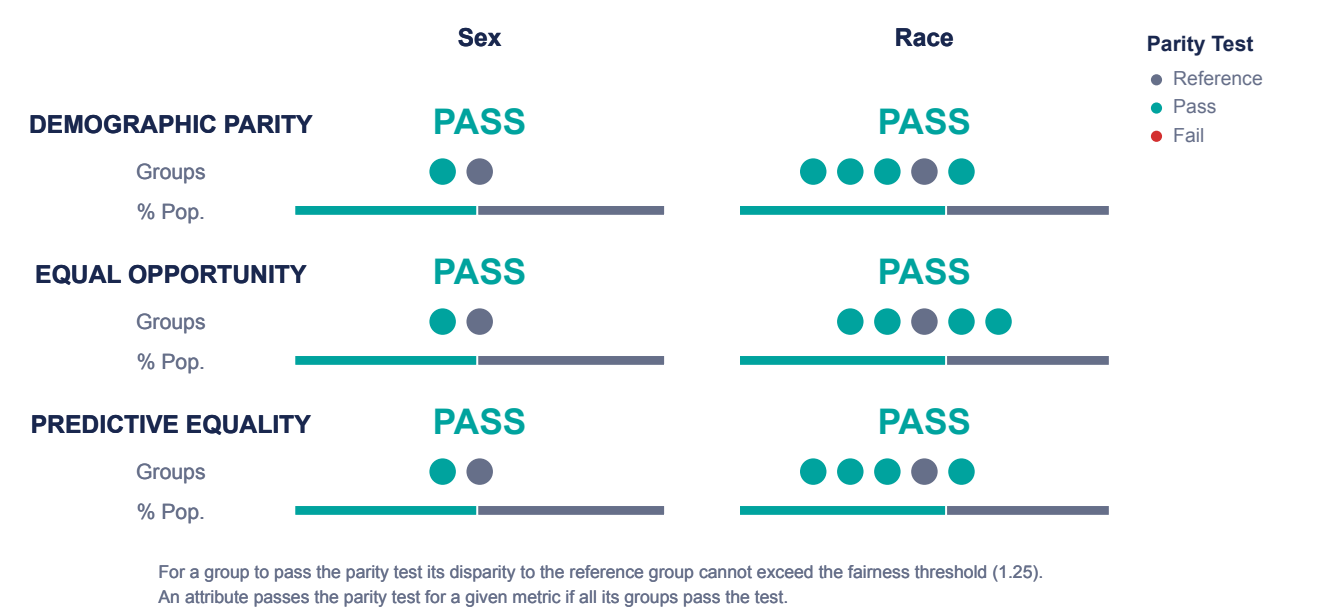


Figure 5. Chart DASHBOARD PARITY TEST: Presents the results of the audit of disparities between the sensitive attributes Gender and Race of the school mental health risk scenario.

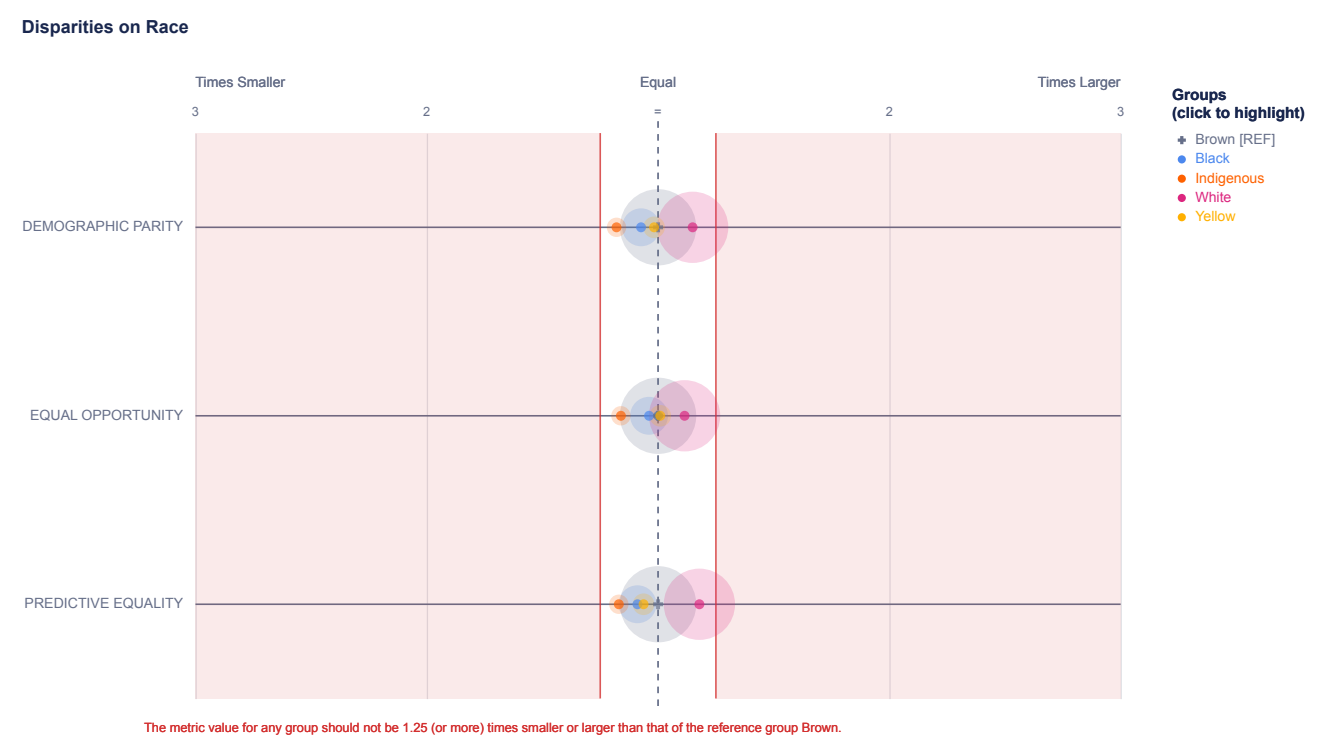


Figure 6. Graph DISPARITIES ON RACE: Result of the race-sensitive attribute audit demonstrating the disparities among their demographic groups in a school mental health risk scenario.

Table 4. Detailed fairness audit results for the race-sensitive attribute demonstrating optimal algorithmic parity among demographic groups in the School Mental Health scenario (Model: Gradient Boosting, selected by CV AUC-ROC = 0.6803; hold-out AUC-ROC = 0.6839).

Fairness Metric	Reference	Group	Result of the Disparity (Value & Evaluation)
Demographic Parity	Brown (Gray)	Reference	gray circle - Reference group, Axis Equal (1.0000).
		Yellow	yellow circle - 0.9830 disparity. Central band indicates parity (PASS).
		White	pink circle - 1.1494 disparity. Central band indicates parity (PASS).
		Indigenous	orange circle - 0.8477 disparity. Central band indicates parity (PASS).
		Black	blue circle - 0.9315 disparity. Central band indicates parity (PASS).
Equal Opportunity	Brown (Gray)	Reference	gray circle - Reference group, Axis Equal (1.0000).
		Yellow	yellow circle - 1.0095 disparity. Central band indicates parity (PASS).
		White	pink circle - 1.1146 disparity. Central band indicates parity (PASS).
		Indigenous	orange circle - 0.8616 disparity. Central band indicates parity (PASS).
		Black	blue circle - 0.9624 disparity. Central band indicates parity (PASS).
Predictive Equality	Brown (Gray)	Reference	gray circle - Reference group, Axis Equal (1.0000).
		Yellow	yellow circle - 0.9410 disparity. Central band indicates parity (PASS).
		White	pink circle - 1.1786 disparity. Central band indicates parity (PASS).
		Indigenous	orange circle - 0.8549 disparity. Central band indicates parity (PASS).
		Black	blue circle - 0.9181 disparity. Central band indicates parity (PASS).

4.2 Representative Case II: Systemic Bias and Disparity (Vulnerable Groups in Live Births - SINASC)

The analysis of the race attribute, using the Brown group as a reference, revealed severe, statistically significant failures (FAIL) across Equal Opportunity (TPR) and Predictive Equality (FPR) metrics. Given the critical nature of this neonatal health scenario, the Equal Opportunity metric (TPR) ensures that early interventions are distributed fairly to all truly vulnerable infants. However, the Gradient Boosting model (hold-out AUC-ROC: 0.7922) exhibited severe algorithmic bias. The White group suffered from a massive under-detection, registering a TPR disparity of 0.3553 (2.81 times smaller than the Brown reference group). Most alarmingly, the Predictive Equality metric (FPR) revealed that the Indigenous group has a disparity ratio of 3.2763, meaning non-vulnerable Indigenous infants are erroneously classified as vulnerable at a rate 3.28 times higher than the reference group, leading to highly noisy classifications and inefficient resource allocation.

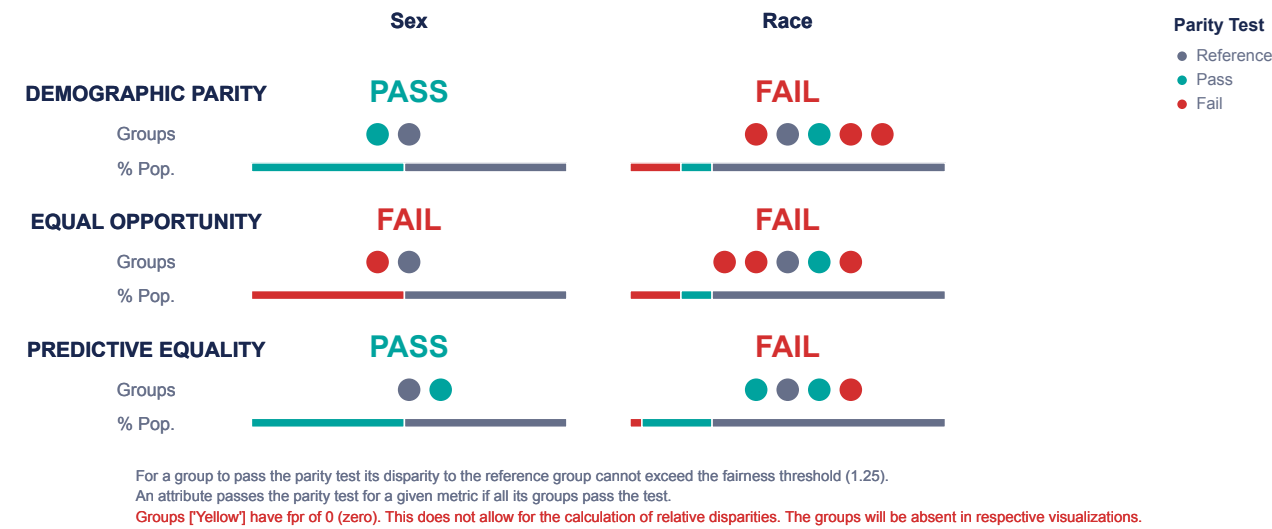


Figure 7. Chart DASHBOARD PARITY TEST: Presents the results of the audit of disparities between the sensitive attributes Gender and Race of Identification of vulnerable groups in live births scenario.

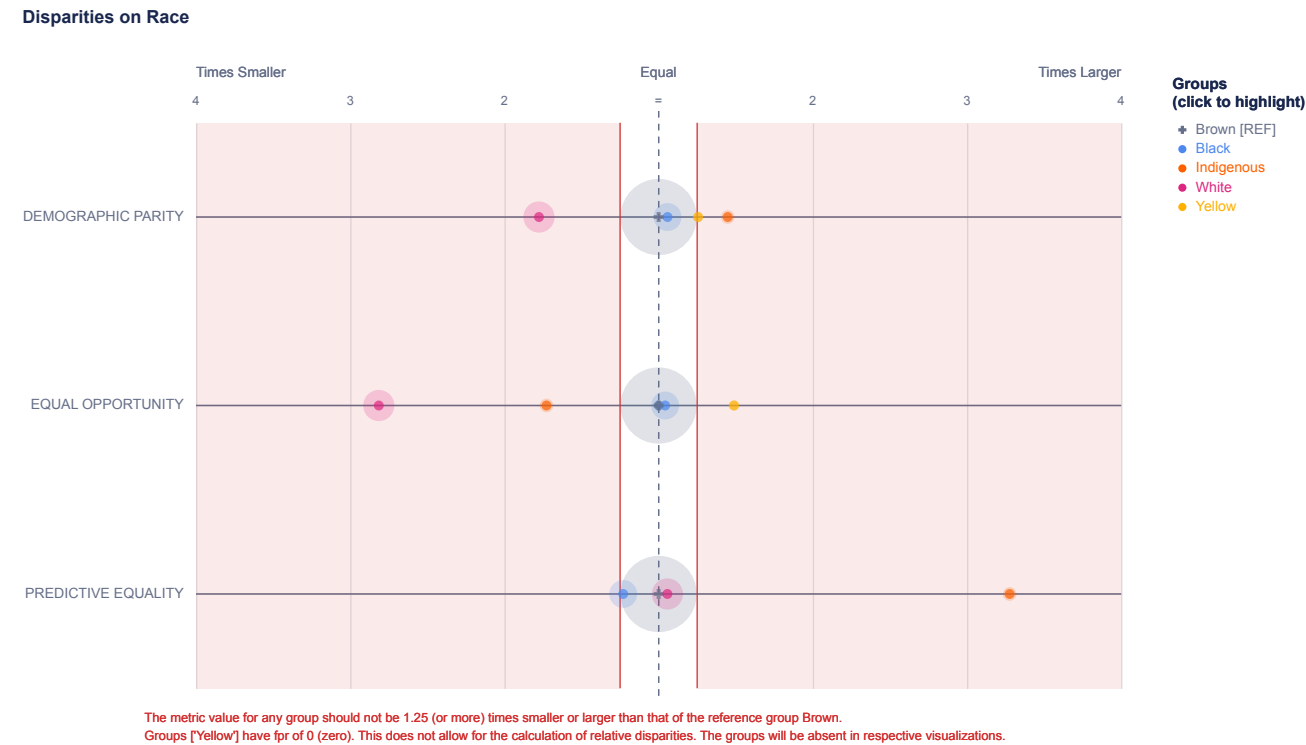


Figure 8. Graph DISPARITIES ON RACE: Result of the race-sensitive attribute audit demonstrating the disparities among their demographic groups in a Identification of vulnerable groups in live births scenario.

Table 5. Detailed fairness audit results for the race-sensitive attribute demonstrating extreme systemic disparities among demographic groups in the Live Births Vulnerability (SINASC) scenario (Model: Gradient Boosting, selected by CV AUC-ROC = 0.7901; hold-out AUC-ROC = 0.7922).

Fairness Metric	Reference	Group	Result of the Disparity (Value & Evaluation)
Demographic Parity	Brown (Gray)	Reference	gray circle - Reference group, Axis Equal (1.0000).
		Yellow	yellow circle - 1.2559 disparity. Shaded zone (FAIL/Unacceptable).
		White	pink circle - 0.5631 disparity (1.78x smaller). Shaded zone (FAIL/Unacceptable).
		Indigenous	orange circle - 1.4482 disparity. Shaded zone (FAIL/Unacceptable).
		Black	blue circle - 1.0576 disparity. Central band indicates parity (PASS).
Equal Opportunity	Brown (Gray)	Reference	gray circle - Reference group, Axis Equal (1.0000).
		Yellow	yellow circle - 1.4887 disparity. Shaded zone (FAIL/Unacceptable).
		White	pink circle - 0.3553 disparity (2.81x smaller). Shaded zone (FAIL/Unacceptable).
		Indigenous	orange circle - 0.5789 disparity (1.73x smaller). Shaded zone (FAIL/Unacceptable).
		Black	blue circle - 1.0421 disparity. Central band indicates parity (PASS).
Predictive Equality	Brown (Gray)	Reference	gray circle - Reference group, Axis Equal (1.0000).
		Yellow	yellow circle - 0.0000 disparity. FPR = 0 (zero) in the audited test partition. The disparity ratio relative to the Brown reference is therefore not computable, and the group is absent from Figure 8, as reported in the AEQUITAS note. This extreme predictive collapse is not evaluated by the parity test.
		White	pink circle - 1.0569 disparity. Central band indicates parity (PASS).
		Indigenous	orange circle - 3.2763 disparity (3.28x larger). Shaded zone (FAIL/Unacceptable).
		Black	blue circle - 0.8131 disparity. Central band indicates parity (PASS).

Note: Groups whose raw metric equals zero yield an undefined disparity ratio and are omitted from the corresponding visualization by AEQUITAS. Such cases are reported as “not computable” and are not assigned a PASS/FAIL verdict; see Section 5.3 on the limitations imposed by minority under-representation.

5 Discussion and Implications

The empirical evidence extracted from the two representative cases reveals a profound sociotechnical dichotomy in the application of Machine Learning to Brazilian public policies. The transition from the quantitative metrics presented in Section 4 to a qualitative analysis demonstrates that achieving universal algorithmic fairness is not merely a technical challenge, but a highly contextual endeavor intricately tied to the structural realities of the data and the specific domain of the public policy.

In the first representative case (School Mental Health Risk - PeNSE), the achievement of perfect algorithmic parity demonstrates that datasets with adequate demographic representativeness across all audited groups can yield models that distribute predictions equitably, even when the target variable itself remains imbalanced. The absence of unacceptable disparities in this scenario proves that it is methodologically possible to develop fair predictive solutions within the national context, establishing a baseline for equitable machine learning in public health screening. Conversely, the second representative case (Live Births Vulnerability - SINASC) exposes how historical inequalities are seamlessly encoded into predictive models. The alarming under-detection of White infants (TPR disparity of 2.81 times smaller), juxtaposed with the severe over-representation of false positives for Indigenous infants (FPR disparity 3.28 times larger), highlights a critical algorithmic failure.

This stark contrast between the two representative scenarios (PeNSE and SINASC) empirically instantiates the structural boundaries of algorithmic equity. Under the mathematical constraints formalized by Chouldechova [2017] and Kleinberg *et al.* [2016], mutual optimization of Demographic Parity, Equal Opportunity, and Predictive Equality remains strictly incompatible when target base rates diverge across protected groups. Consequently, rather than a failing of the predictive pipeline, the observed disparities reflect an unavoidable mathematical trade-off. Our findings fundamentally shift the decision-making paradigm: achieving algorithmic fairness is not a universal optimization problem with an absolute mathematical resolution, but a contextual, socio-technical choice that must be explicitly guided by the policy goals of the public administration.

5.1 Practical Impacts on Public Policies and Society

The decision-making processes and the management of public resources are directly impacted by the existence of the systemic biases evidenced in our results. The proof of algorithmic disparities imposes new responsibilities on public managers:

- **Risk of perpetuating structural inequalities:** Adopting algorithmic solutions based on unmitigated data unintentionally expands the exclusion patterns of historically vulnerable groups (such as Black, Yellow, and Indigenous profiles). As seen in the SINASC scenario, this compromises the ultimate mission of public policies, which is the promotion of equity and protection.

- **Ethical and legal challenges in implementation:** Deploying models that present severe disparities denies constitutional principles and human rights standards, increasing the risk of judicialization and delegitimization of public initiatives based on Artificial Intelligence.
- **Articulation of mitigation strategies:** It is incumbent upon policymakers to adopt strategies to mitigate these inequalities. This requires mechanisms of transparency, accountability, and the active participation of civil society to debate equity criteria and the real-world impact of automated decisions.

5.2 Technical Implications for AI Systems

For Machine Learning practitioners, developers, and data scientists utilizing open government databases, our findings demonstrate that algorithmic fairness must transcend theoretical discourse and become an operational requirement:

- **Continuous Fairness Audits:** In highly sensitive domains, systematic audits (such as the ones conducted using AEQUITAS) must be integrated as an indispensable part of the CI/CD pipeline, rather than an accessory step at the end of the project.
- **Adaptation to the Brazilian Social Context:** Pre-trained models invariably reflect the distortions present in historical data. Therefore, the fine-tuning of hyperparameters, feature selection, and the definition of reference groups must be strictly tailored to the sociocultural specificities of the country, avoiding the blind importation of foreign fairness thresholds. In the present audit, the reference groups were not manually tailored: they were derived from the majority-group default of the auditing toolkit, and the deliberate specification of reference groups on socio-technical grounds remains a prospective recommendation rather than a procedure adopted here.
- **Rigorous Pre-processing Approaches:** There is an urgent need to intensify Exploratory Data Analysis (EDA) specifically targeted at detecting and mitigating biases before modeling, implementing balancing techniques aligned with strict fairness standards.

5.3 Limitations and Future Paths

While this study provides a comprehensive fairness audit framework, it presents certain limitations. A critical limitation identified across the evaluated scenarios was the frequent occurrence of null performance indicators (e.g., zero false positive rates) for extremely under-represented minority groups (such as specific Indigenous populations). This data scarcity prevents the mathematical calculation of relative disparities and masks the true extent of algorithmic discrimination.

It should also be noted that disparity ratios are sensitive to the choice of reference group: adopting the majority group (Brown) as the denominator means that the reported disparities measure deviation from the predictive pattern of the modal group, not from a normative ideal of equity. Robustness analyses using alternative reference groups constitute a pertinent extension of this work.

To overcome these limitations and advance the field of algorithmic fairness in the Brazilian public sector, specific future paths must be paved:

- **Methodological expansion:** Future work should explore the integration of advanced bias mitigation techniques—including pre-processing (e.g., synthetic data generation), in-processing, and post-processing methods—evaluating their effectiveness in reducing identified disparities without significantly compromising predictive performance (CV AUC-ROC). Furthermore, exploring Explainable AI (XAI) alongside fairness audits will provide deeper insights into the causal mechanisms underlying the observed biases.
- **Diversification of scenarios and longitudinal data:** Extending this research to other high-stakes public policy domains, such as environmental licensing and public security, will provide a more holistic view of fairness challenges. Incorporating longitudinal temporal data will also allow researchers to evaluate the evolution of biases over time and measure the effectiveness of applied interventions.
- **Development of customized tools and frameworks:** Establishing an integrated fairness audit platform explicitly adapted to the Brazilian legal and socioeconomic context is paramount for establishing standardized governance practices in government AI projects.

Therefore, metric prioritization must be domain-dependent. In high-stakes healthcare scenarios like SINASC, Equal Opportunity (TPR) must take absolute priority to ensure that life-saving interventions are not disproportionately withheld from vulnerable minorities due to algorithmic misclassification. Conversely, in socio-economic resource allocation scenarios, such as the distribution of PROUNI scholarships, Demographic Parity (PPREV) might take precedence to actively counteract historical marginalization and ensure a structurally fair distribution of opportunities. Future paths in this research field must prioritize advanced preprocessing strategies—such as synthetic data generation or targeted oversampling—to ensure that minority groups are adequately represented before fairness audits are even conducted.

6 Conclusion

This study conducted a systematic algorithmic fairness audit across ten Brazilian public policy scenarios using open government microdata. By evaluating multiple classification algorithms and strictly selecting the most robust models (Gradient Boosting and Random Forest) based on Cross-Validated AUC-ROC to account for severe data imbalance, the results unequivocally confirm the presence of systemic biases in Machine Learning predictions, with nine out of ten scenarios exhibiting statistically significant equity violations.

Fairness metrics such as Predictive Equality and Demographic Parity frequently failed for the racial groups contrasted against the Brown reference, affecting both minority groups (Black, Indigenous, and Yellow)—exposed to disproportionate risks of exclusion and misclassification—and the

White group, which deviates markedly from the modal predictive pattern in several scenarios. This was most critically evidenced by the severe disparity in false positive rates for Indigenous infants in the neonatal health scenario (SINASC). Conversely, the achievement of parity in the School Mental Health scenario (PeNSE) demonstrates that equitable algorithmic decision-making is attainable with representative data and rigorous auditing. Ultimately, this research empirically proves that algorithmic fairness is not a universal mathematical absolute, but a contextual imperative that requires prioritizing specific metrics aligned with the socio-technical goals of the public policy, preventing the perpetuation of historical inequalities.

To advance this critical field, future work must focus on developing and integrating advanced bias mitigation techniques tailored to the Brazilian socio-demographic context. Subsequent research should expand this auditing framework to other high-stakes domains—such as public security and resource allocation—integrating explainable AI (XAI) and targeted data augmentation to overcome the limitations of minority under-representation and ensure the deployment of responsible, transparent, and equitable Artificial Intelligence in the public sector.

Declarations

Authors' Contribution

Under the supervision of Baldoino Fonseca and Tiago Vieira, this research had the contribution of Davy Baia, Tiago Vieira, and Vicente Julião in the design of the study. Vicente Julião and George Brandão were responsible for conducting the experiments and executing the machine learning pipeline, while the methodological development and algorithmic fairness auditing framework were elaborated by Vicente Julião, Baldoino Fonseca, Davy Baia, and George Brandão. The manuscript was written and edited by Vicente Julião and Davy Baia. Baldoino Fonseca, Davy Baia, Vicente Julião, and Tiago Vieira acted as the main contributors and editors of this work. All authors have read, revised, and approved the final version of the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Availability of Data and Materials

All government datasets used in this study (CENSO, PeNSE, PNADc, PNS, POF, ENEM, PROUNI, SINASC, ICT, and TICKids) are open, fully anonymized, and in the public domain. The complete python-based execution framework, hyperparameter grids, evaluation logs, and systematic AEQUITAS auditing scripts are openly available in a public GitHub repository. To guarantee absolute reproducibility and seamless execution, the repository includes a standardized Docker containerization environment, as described in Section 3.2.4. The permanent access link is provided in the manuscript.

Use of AI-assisted Technologies

During the preparation of this manuscript, the authors utilized AI-assisted technologies exclusively to enhance linguistic fluency, perform grammatical proofreading, and optimize academic formatting. Following the use of these services, the authors thoroughly reviewed, verified, and edited the content as necessary. The authors assume full and sole responsibility for the scientific integrity, conceptual accuracy, and originality of the final published work.

References

- Al Husaini, Y., Alkishri, W., Abdulla, M., Mahmood, A.-B., and Abrar, M. (2025). Predicting student academic success using deep learning: A multi-factor approach to performance prediction. *Journal of Logistics Informatics and Service Science*, 12:263–283. DOI: 10.33168/JLISS.2025.0114.
- Bagheri, S., Taridashti, S., Farahani, H., Watson, P., and Rezvani, E. (2023). Multilayer perceptron modeling for social dysfunction prediction based on general health factors in an iranian women sample. *Frontiers in Psychiatry*, 14:1283095. DOI: 10.3389/fpsy.2023.1283095.
- Beaulac, C. and Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, 60(7):1048–1064. DOI: 10.1007/s11162-019-09546-y.
- Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al. (2019). Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1. DOI: 10.1147/jrd.2019.2942287.
- Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in ai. Available at: <https://www.jmlr.org/papers/v24/23-0389.html>.
- Cadili, R. (2024). An introduction to the naive bayes classifier and how to use it. *Medium.com*. Available at: <https://medium.com/low-code-for-advanced-data-science/an-introduction-to-the-naive-bayes-classifier-and-how-to-use-it-bba78826ef8f>.
- Cheng, L., Varshney, K. R., and Liu, H. (2021). Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71:1137–1181. DOI: 10.1613/jair.1.12814.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163. DOI: 10.1089/big.2016.0047.
- Du, Y. (2025). Reframing educational inquiry: Uncovering the value of decision tree analysis—a practical guide for educational researchers. DOI: 10.31219/osf.io/pcvbe_v1.
- Dye, T. R. (1972). *Understanding Public Policy*. Prentice-Hall, Englewood Cliffs, NJ. DOI: 10.2307/975069.
- Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and machines*, 28(4):689–707. DOI: 10.1007/s11023-018-9482-5.
- Giraldo, C., Giraldo, I., Gomez-Gonzalez, J. E., and Uribe, J. M. (2023). An explained extreme gradient boosting approach for identifying the time-varying determinants of sovereign risk. *Finance Research Letters*, 57:104273. DOI: 10.1016/j.frl.2023.104273.
- Hasan, R., Palaniappan, S., Raziff, A. R. A., Mahmood, S., and Sarker, K. U. (2018). Student academic performance prediction by using decision tree algorithm. In *2018 4th international conference on computer and information sciences (ICCOINS)*, pages 1–5. IEEE. DOI: 10.1109/iccoins.2018.8510600.
- Huang, C., Zhang, Z., Mao, B., and Yao, X. (2022). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, 4(4):799–819. DOI: 10.1109/tai.2022.3194503.
- Jesus, S., Saleiro, P., e Silva, I. O., Jorge, B. M., Ribeiro, R. P., Gama, J., Bizarro, P., and Ghani, R. (2024). Aequitas flow: Streamlining fair ml experimentation. *Journal of Machine Learning Research*, 25(354):1–7. DOI: 10.48550/arxiv.2405.05809.
- Khyathi, G., Indumathi, K., Jumana Hasin, A., Lisa Flavin Jency, M., Krishnaprakash, G., LISA, F. J. M., et al. (2025). Support vector machines: A literature review on their application in analyzing mass data for public health. *Cureus*, 17(1). DOI: 10.7759/cureus.77169.
- Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*. DOI: 10.4230/lipics.its.2017.43.
- Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., and Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46. DOI: 10.1145/3555803.
- Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley & Sons. DOI: 10.2307/1533221.
- Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., and Tang, J. (2022). Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–59. DOI: 10.1145/3546872.
- Lynn, L. E. and Gould, S. G. (1980). Designing public policy: A casebook on the role of policy analysis: instructor's manual. (*No Title*). Book.
- Mead, L. (1995). Public policy: vision, potential, limits. *Policy Currents (Newsletter of the Public Policy Section, APSA)*, 68(3). Book.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35. DOI: 10.1145/3457607.
- Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. (2020). From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4):2141–2168. DOI: 10.2139/ssrn.3830348.
- MURAINA, I. O., Aiyegbusi, E., and Abam, S. (2022). Decision tree algorithm use in predicting students' academic performance in advanced programming course. *International*

- Journal of Higher Education Pedagogies*, 3(4):13–23. DOI: 10.33422/ijhep.v3i4.274.
- Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M., Galeano Sarmiento, J., Ruz, G. A., and Tabares-Soto, R. (2024). Applying the ethics of ai: a systematic review of tools for developing and assessing ai-based systems. *Artificial Intelligence Review*, 57(5):110. DOI: 10.1007/s10462-024-10740-3.
- Ozturk Kiyak, E., Ghasemkhani, B., and Birant, D. (2023). High-level k-nearest neighbors (hlknn): A supervised machine learning model for classification analysis. *Electronics*, 12(18):3828. DOI: 10.3390/electronics12183828.
- Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., Araujo, M. M., Santos, L. L., Cruz, M. A., Oliveira, E. L., et al. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing*, 7(1):15. DOI: 10.3390/bdcc7010015.
- Peters, B. G. et al. (1986). American public policy. *Chatham, NJ: Chatham House*, pages 11–28. DOI: 10.1007/978-1-349-18388-3.
- Pohan, S., Faisal, R. A., Nasution, F., Ramadani, P., and Hasibuan, A. I. Y. (2024). Application of svm to speed up and accurate nursing decisions for mentally disordered patients. *International Journal of Health Engineering and Technology (IJHET)*, 3(4). DOI: 10.55227/ijhet.v4i1.277.
- Ptr, A. F. L., Siregar, M. M., and Daniel, I. (2024). Analysis of gradient boosting, xgboost, and catboost on mobile phone classification. *J. Comput. Networks, Archit. High Perform. Comput*, 6(2):661–670. DOI: 10.47709/cnaphc.v6i2.3790.
- Pyke, S. W. and Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention. *Canadian Journal of Higher Education*, 23(2):44–64. DOI: 10.47678/cjhe.v23i2.183161.
- Raschka, S. (2014). Naive bayes and text classification i-introduction and theory. *arXiv preprint arXiv:1410.5329*. DOI: 10.13140/2.1.2018.3049.
- Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., and Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*. DOI: 10.48550/arxiv.1811.05577.
- Souza, C. (2006). Políticas públicas: uma revisão da literatura. *Sociologias*, pages 20–45. DOI: 10.1590/s1517-45222006000200003.
- Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of translational medicine*, 4(11):218. DOI: 10.21037/atm.2016.03.37.

A Supplementary Material: Comprehensive Fairness Audit Results

This Appendix presents the complete AEQUITAS algorithmic fairness audit results for the remaining eight public policy scenarios not detailed in the main text. To fully satisfy the quantitative transparency requirements requested during the peer-review process, the following subsections provide the exact disparity ratios for Demographic Parity (PPREV), Equal Opportunity (TPR), and Predictive Equality (FPR). These results were generated using the models selected via CV AUC-ROC (Gradient Boosting or Random Forest) in order to assess whether data-intrinsic bias persists across different modeling architectures.

A.1 Scenario I: Poverty Prediction (CENSO) - Model: Gradient Boosting

The Gradient Boosting model (hold-out AUC-ROC: 0.7415) exhibited severe disparities. For the gender attribute (reference: Female), the Male group violated all metrics, notably showing an Equal Opportunity (TPR) disparity of 1.5660 and a Predictive Equality (FPR) disparity of 1.2547. For the race attribute (reference: Brown), the White group presented extreme over-representation (PPREV disparity: 2.1189; FPR disparity: 1.9314). Conversely, the Indigenous group suffered from a complete predictive collapse, with PPREV, TPR and FPR all equal to zero (absolute under detection); no disparity ratio could be computed for this group, which is consequently absent from Figure 10.

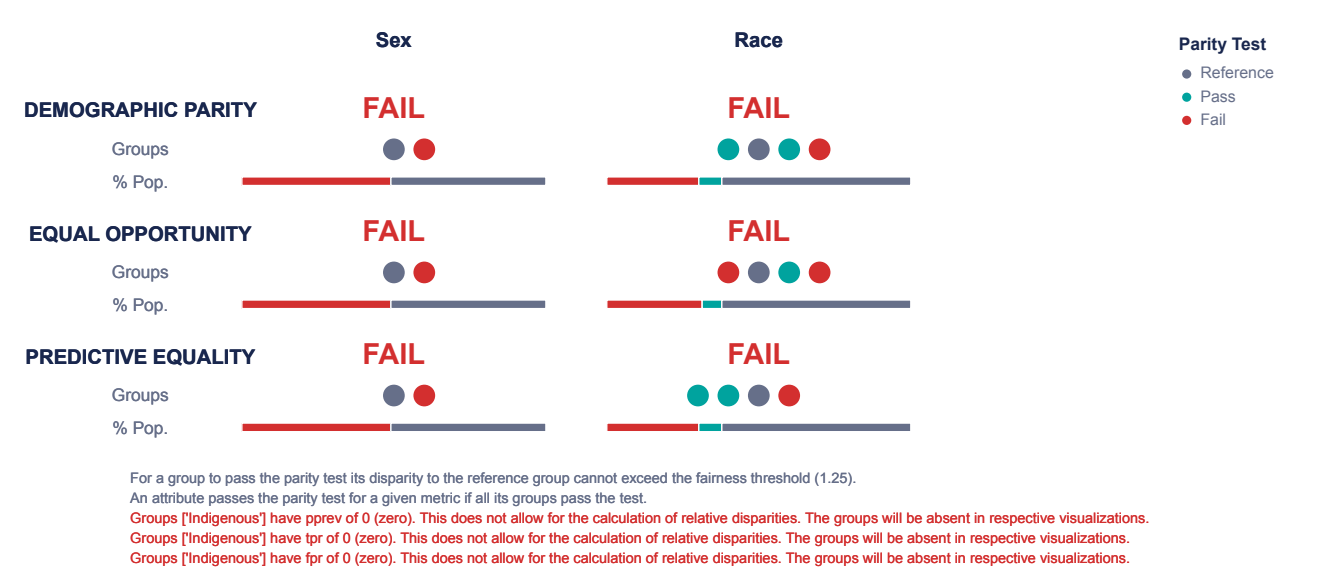


Figure 9. Parity Test Dashboard: Poverty Prediction (CENSO).

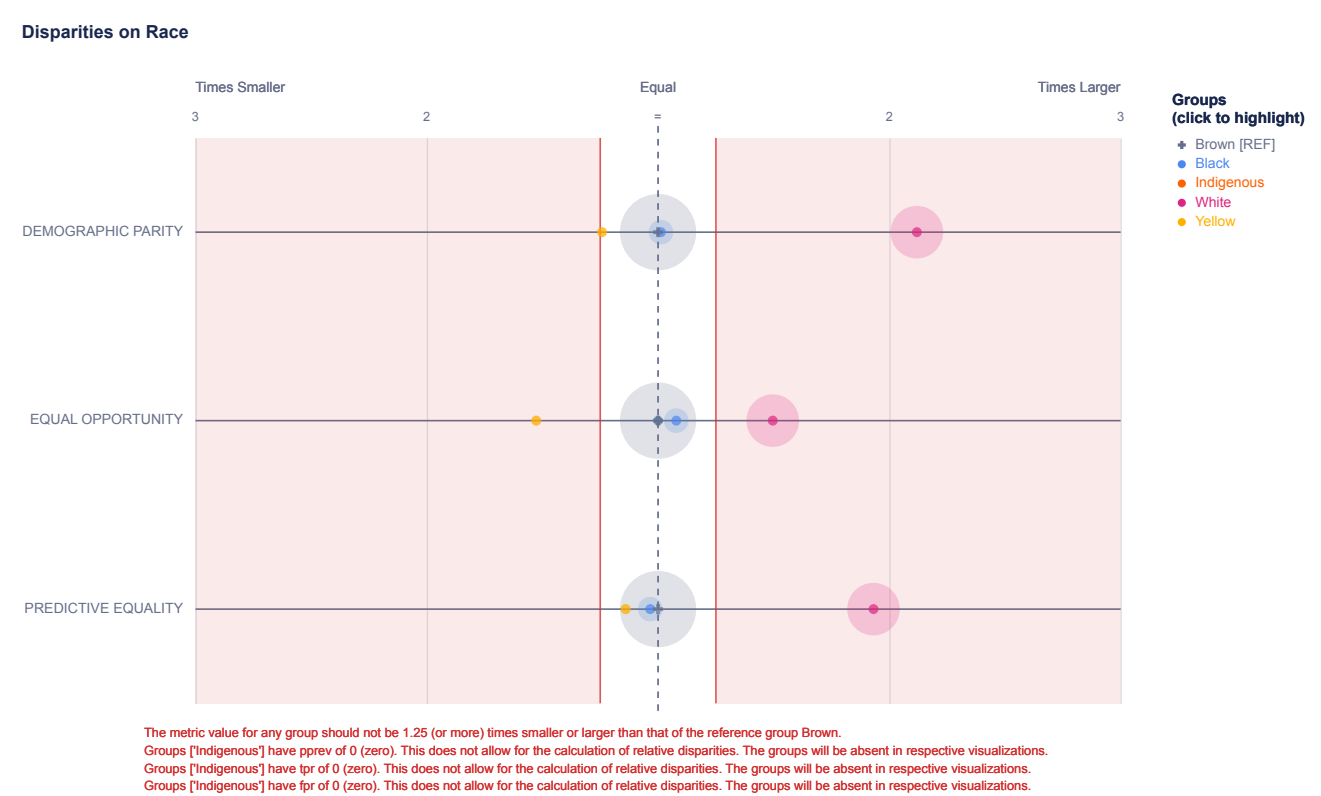


Figure 10. Fairness Disparity Plot (Race) for the Poverty Prediction scenario (CENSO).

A.2 Scenario III: Occupational Situation (PNADc) - Model: Gradient Boosting

Auditing the occupational probability scenario (hold-out AUC-ROC: 0.9181), the FPR metric was systematically violated for every racial group contrasted against the Brown reference. The Indigenous group presented the largest deviation, failing all three metrics, with an FPR disparity of 0.4343 (2.30 times smaller than the reference); the White and Yellow groups recorded FPR disparities of 0.5139 and 0.4929, respectively. Since all three ratios fall below 1.0, the direction of the violation is the reverse of the usual pattern: false positives are concentrated in the reference group itself, so that Brown individuals are disproportionately exposed to false alarms while the remaining racial groups are under-signalled.

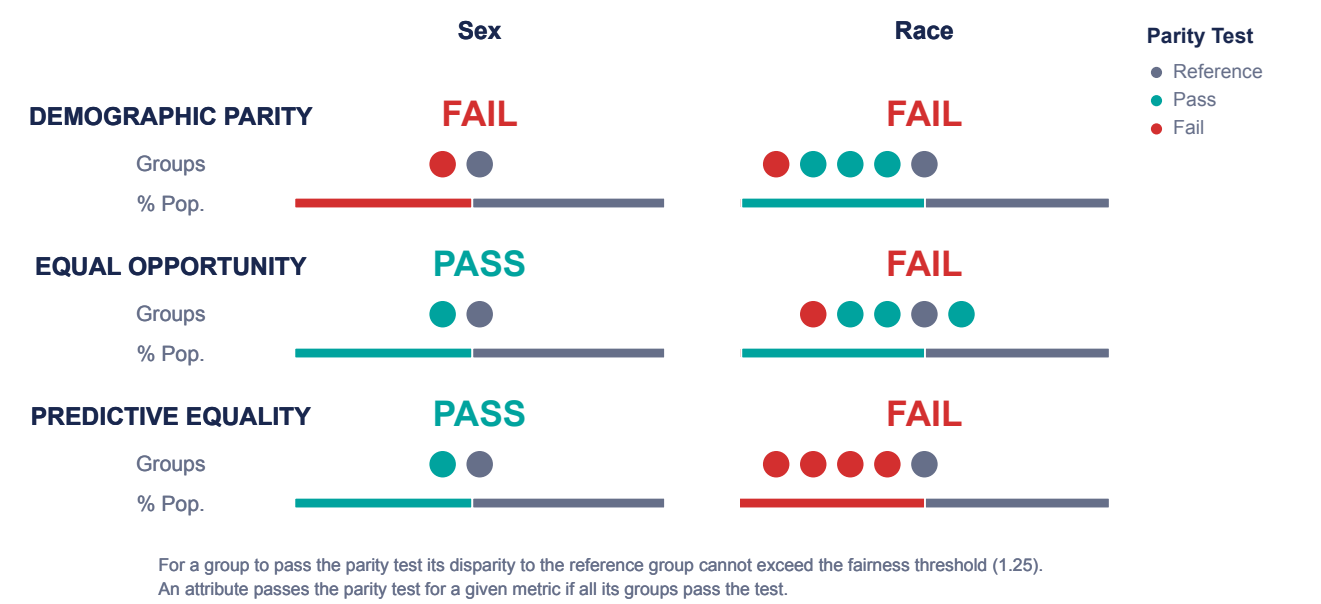


Figure 11. Parity Test Dashboard: Occupational Situation (PNADc).

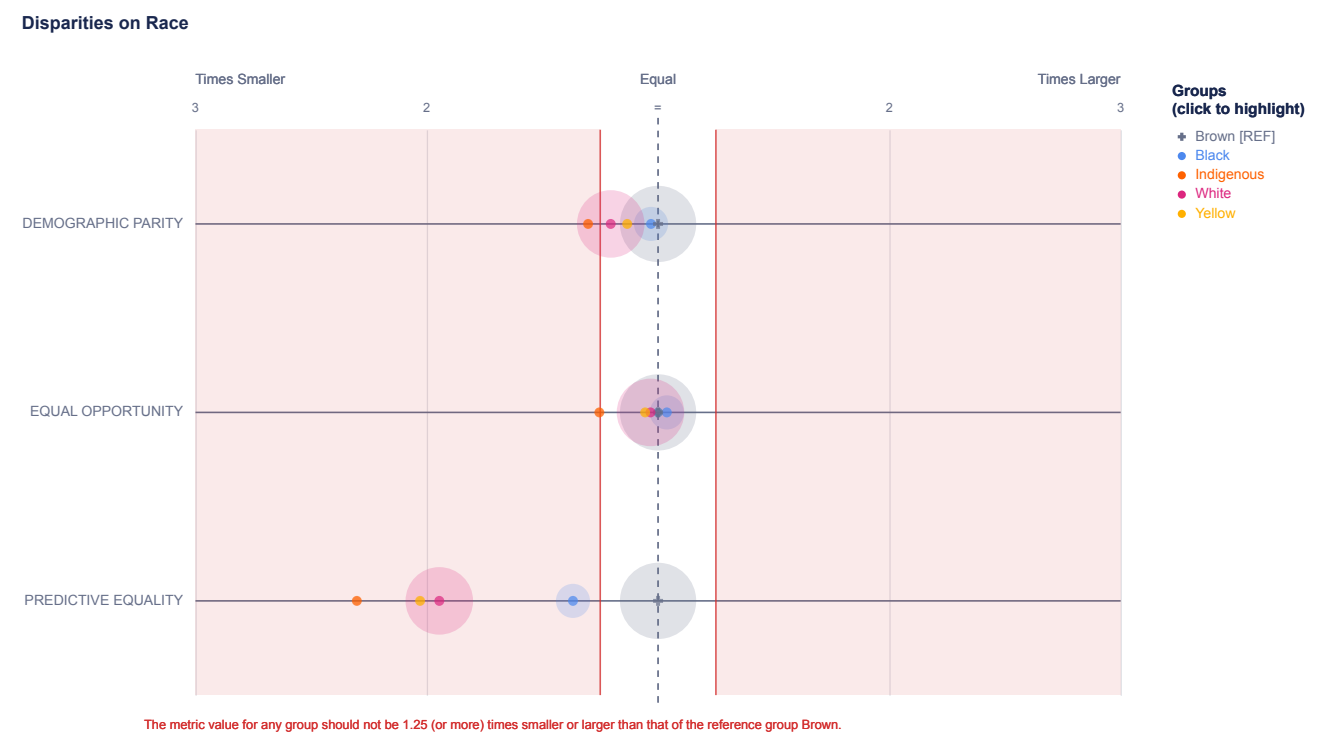


Figure 12. Fairness Disparity Plot (Race) for the Occupational Situation (PNADc).

A.3 Scenario IV: Diabetes in Adults (PNS) - Model: Gradient Boosting

While this model achieved excellent predictive performance (hold-out AUC-ROC: 0.9980) and maintained parity in TPR, it failed critically in Predictive Equality (FPR). Men registered an FPR disparity of 2.0594 (more than twice the false alarms compared to women). Racially, White individuals recorded an FPR disparity of 0.4690, meaning they receive less than half the false positive predictions compared to the Brown reference group.

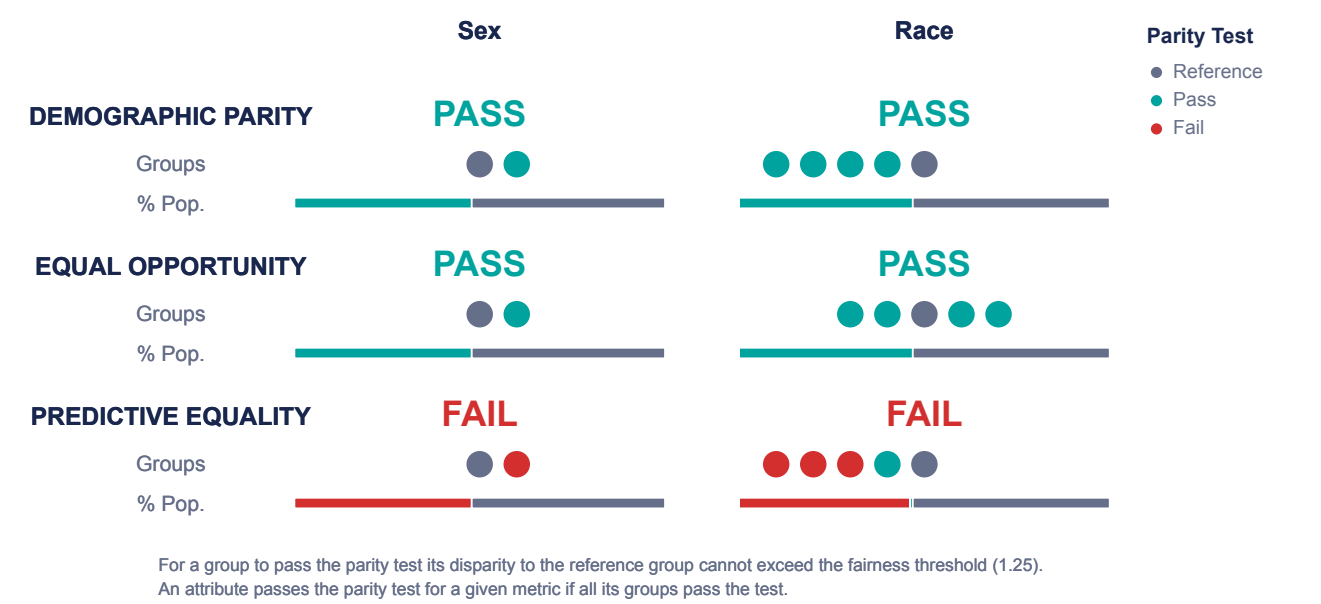


Figure 13. Parity Test Dashboard: Diabetes in Adults (PNS).

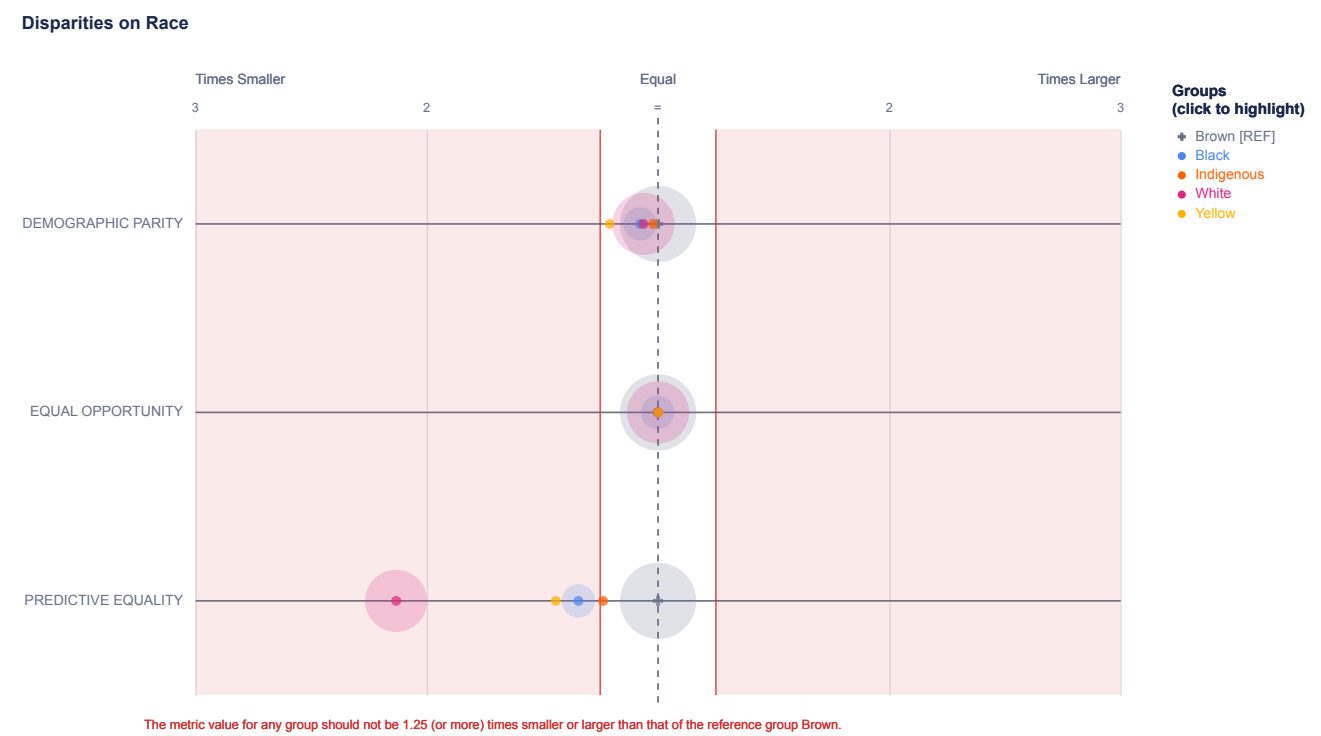


Figure 14. Fairness Disparity Plot (Race) for the Diabetes in Adults (PNS).

A.4 Scenario V: Food Insecurity (POF) - Model: Random Forest

In this scenario (hold-out AUC-ROC: 0.9366), the Random Forest model showed perfect equity for gender (reference: Female). However, racial disparities relative to the Brown reference group were profound. The White group recorded a PPREV disparity of 1.6501 and an FPR disparity of 2.1752, suffering more than double the false alarms of the reference group. The Yellow group registered the highest prevalence disparity (1.7226) but an FPR of zero, so that no false positive rate disparity could be computed for this group, revealing highly unbalanced error distributions across demographic segments.

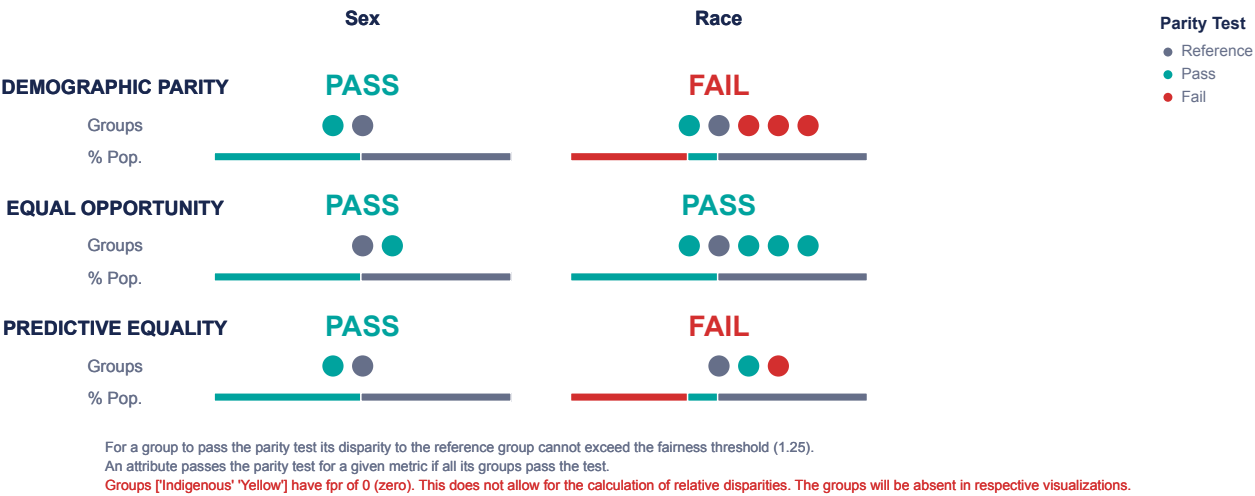


Figure 15. Parity Test Dashboard: Food Insecurity (POF).

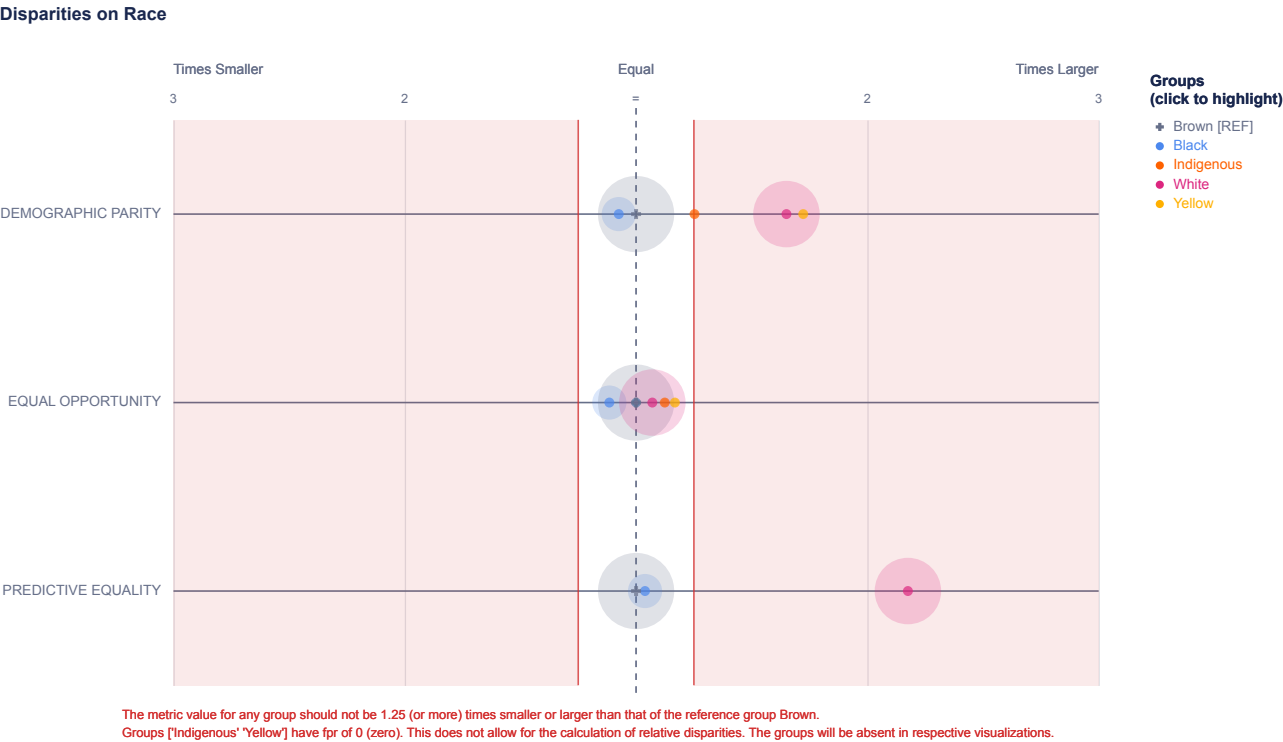


Figure 16. Fairness Disparity Plot (Race) for the Food Insecurity (POF).

A.5 Scenario VI: Exam Writing Difficulties (ENEM) - Model: Gradient Boosting

Predicting critical difficulties in the ENEM exam (hold-out AUC-ROC: 0.8766), the model proved equitable in True Positive Rates (TPR) but failed in FPR. Men showed an FPR disparity of 0.6690 compared to women. Racially, the Indigenous group recorded a severe FPR disparity of 0.3692 (2.71 times smaller than the Brown reference group), indicating that Indigenous students are highly under-signaled for false alarms.

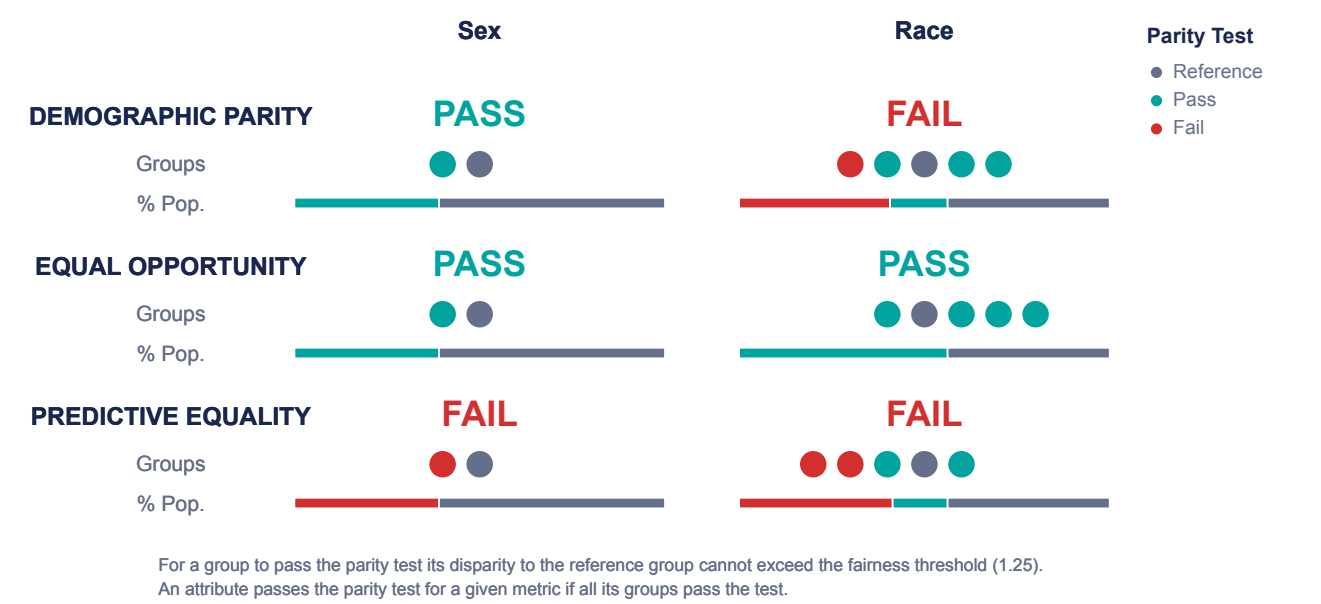


Figure 17. Parity Test Dashboard: Exam Writing Difficulties (ENEM).

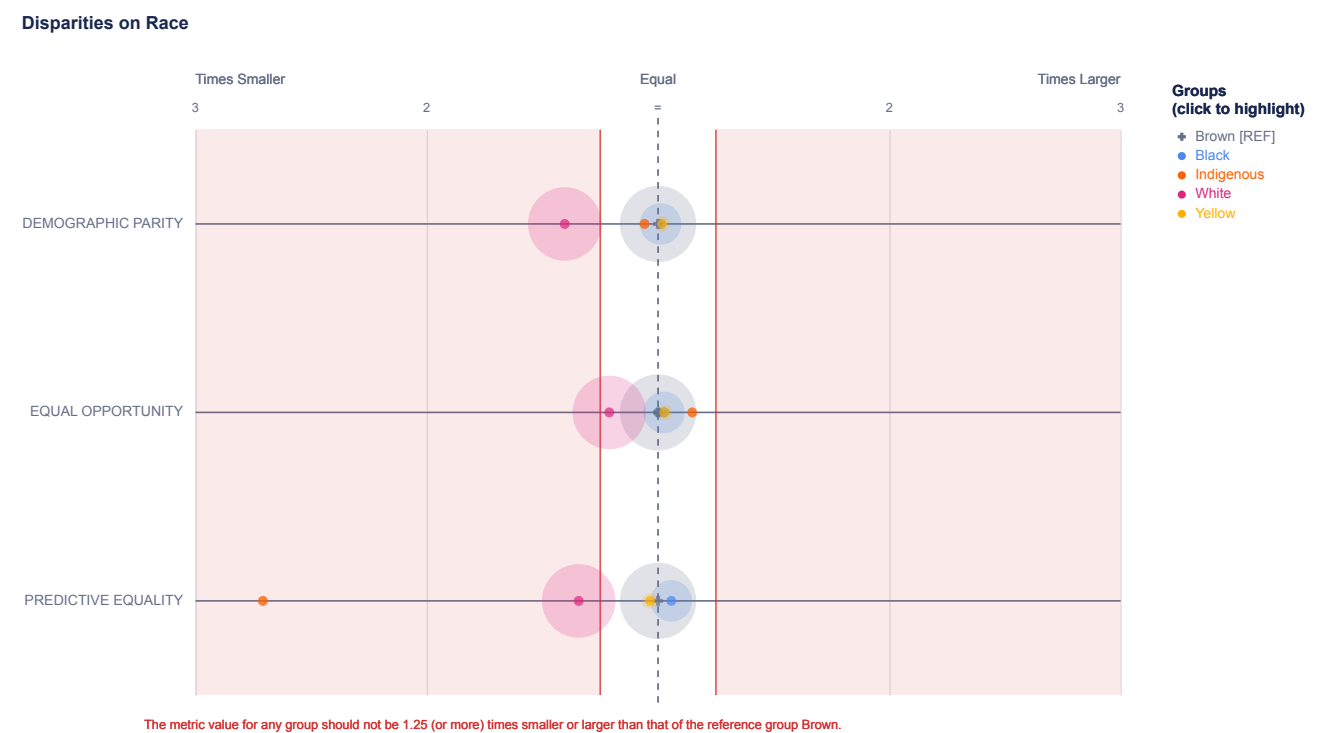


Figure 18. Fairness Disparity Plot (Race) for the Exam Writing Difficulties (ENEM).

A.6 Scenario VII: Scholarships Allocation (PROUNI) - Model: Gradient Boosting

For scholarship allocation (hold-out AUC-ROC: 0.9309), the model demonstrated strong parity across gender. The only critical fairness violation occurred in the race attribute for the Indigenous group, which recorded a Predictive Equality (FPR) disparity of 2.4325. This indicates that non-eligible Indigenous candidates are erroneously classified as eligible at a rate 2.43 times higher than the Brown reference group.

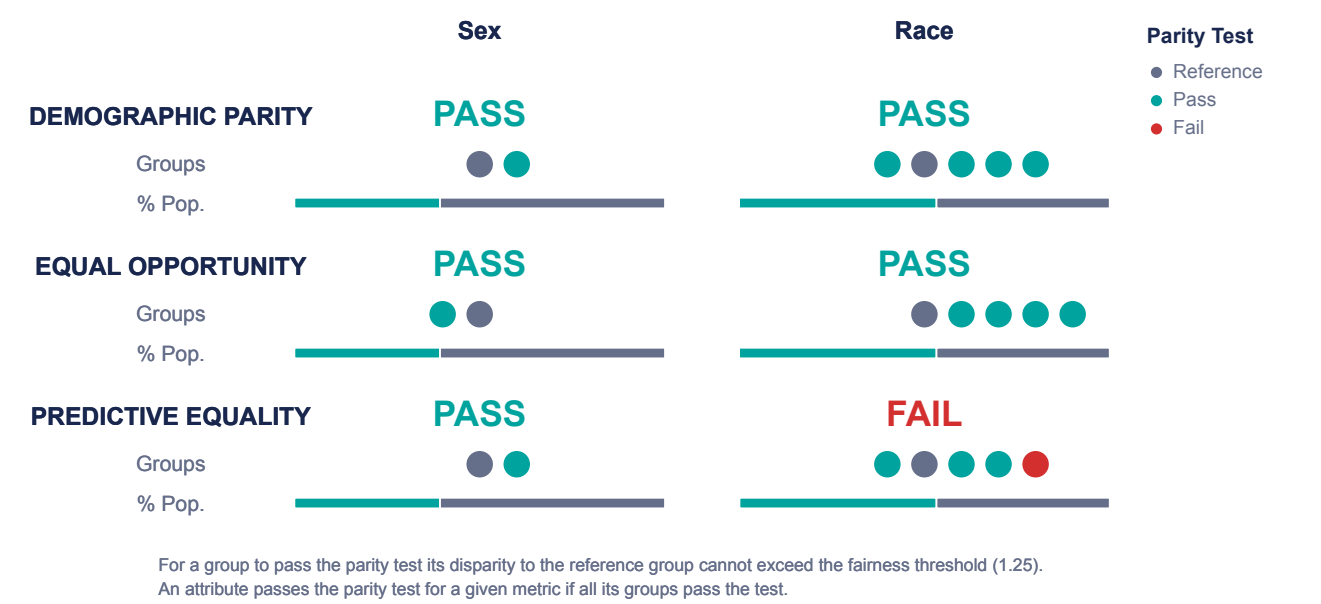


Figure 19. Parity Test Dashboard: Scholarships Allocation (PROUNI).

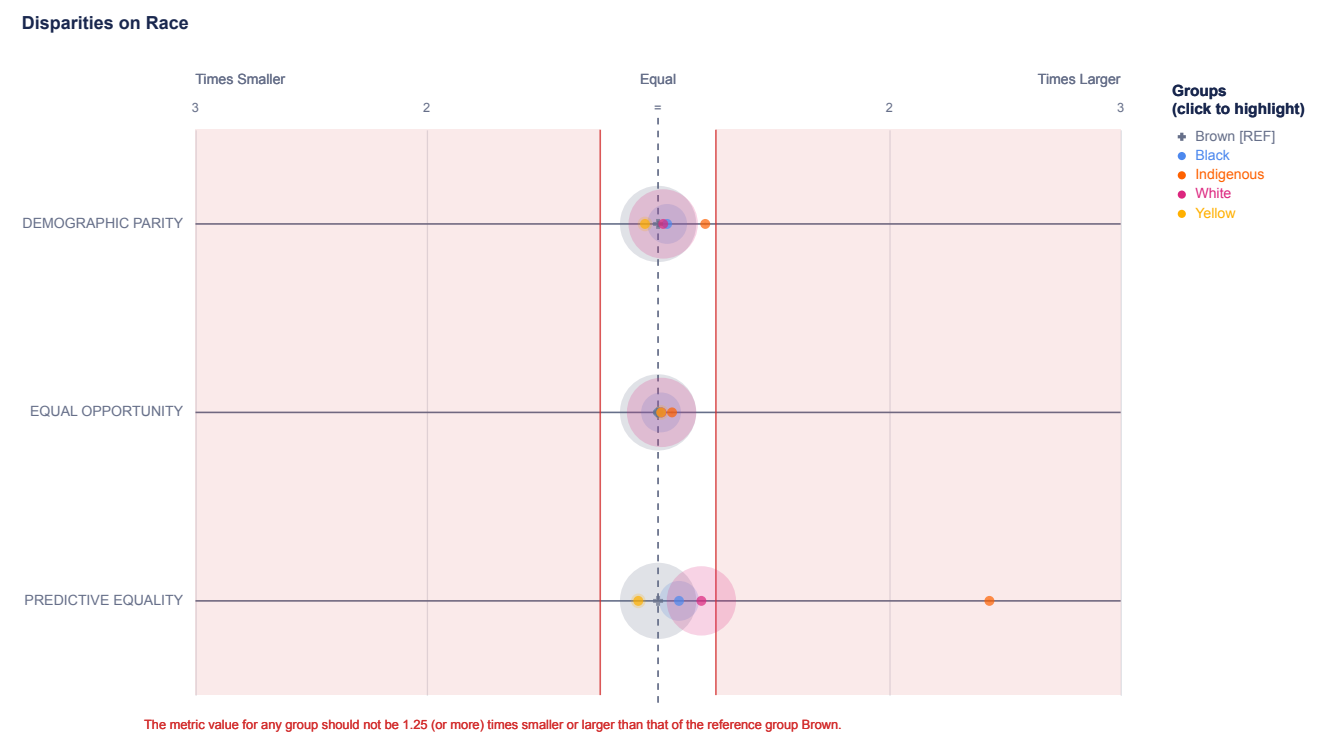


Figure 20. Fairness Disparity Plot (Race) for the Scholarships Allocation (PROUNI).

A.7 Scenario IX: Digital Inclusion (ICT) - Model: Random Forest

The Random Forest model (hold-out AUC-ROC: 0.9818) maintained Equal Opportunity parity for all groups but failed in FPR. Men had an FPR disparity of 0.7349. The Indigenous group was the most penalized, being significantly under-represented in the positive prediction (PPREV: 0.7994) while simultaneously suffering almost double the false positive rate (FPR: 1.9261) compared to the Brown group.

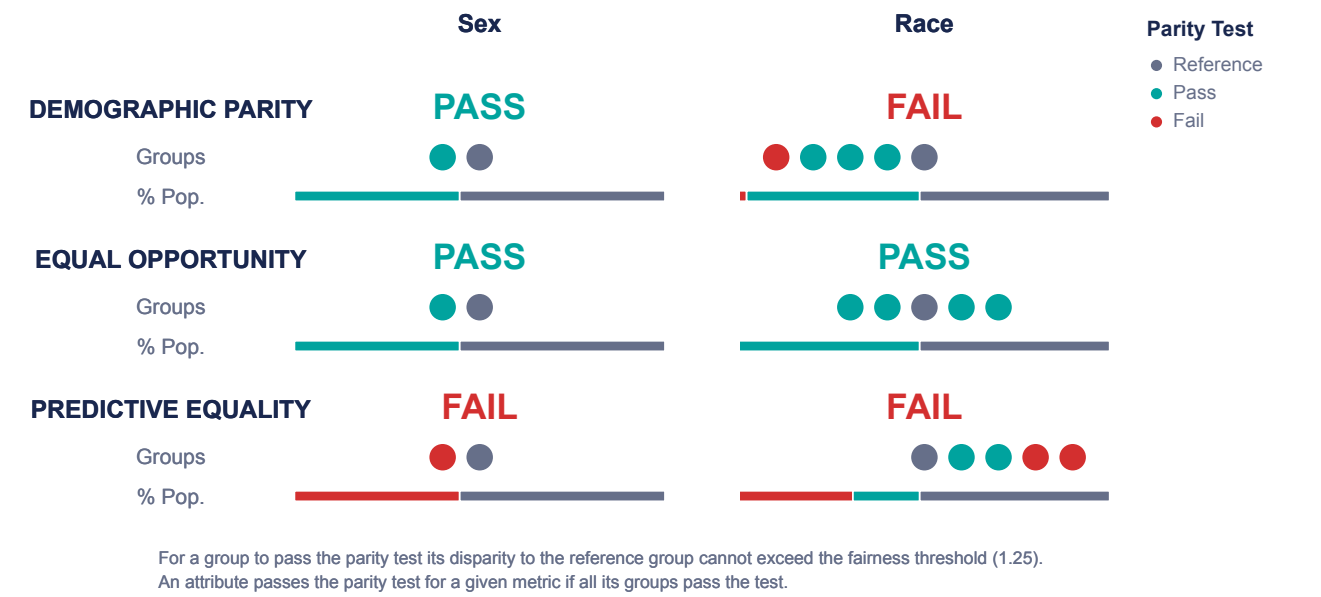


Figure 21. Parity Test Dashboard: Digital Inclusion (ICT).

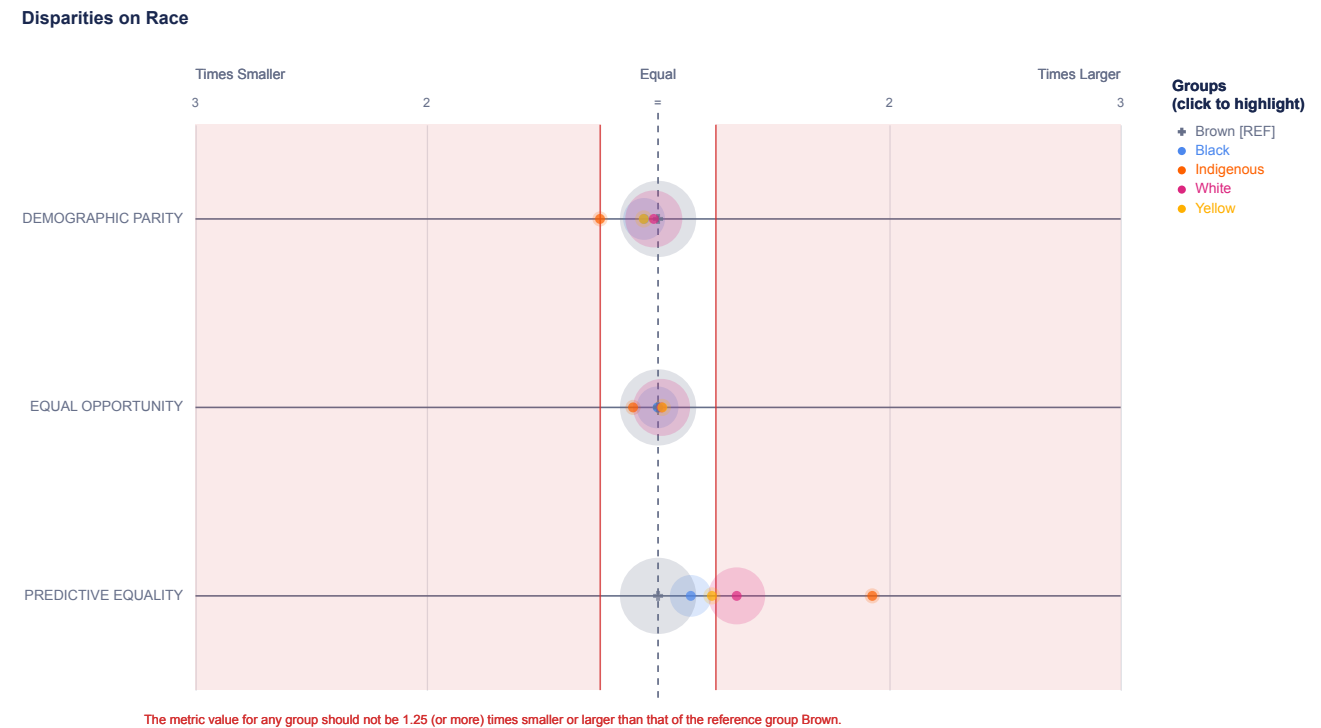


Figure 22. Fairness Disparity Plot (Race) for the Digital Inclusion (ICT).

A.8 Scenario X: Children’s Tech Vulnerability (ICT Kids) - Model: Random Forest

In this critical scenario (hold-out AUC-ROC: 0.7704), the Random Forest model exhibited extreme algorithmic over-representation for the Indigenous group relative to the Brown reference group (PPREV disparity: 2.8558; TPR disparity: 1.8012). Conversely, the Yellow group suffered a total predictive collapse, with all three raw metrics equal to zero and, consequently, no computable disparity ratio. Furthermore, girls registered an FPR disparity of 0.5427 against the Male reference of this dataset, demonstrating a highly noisy and disparate classification environment.

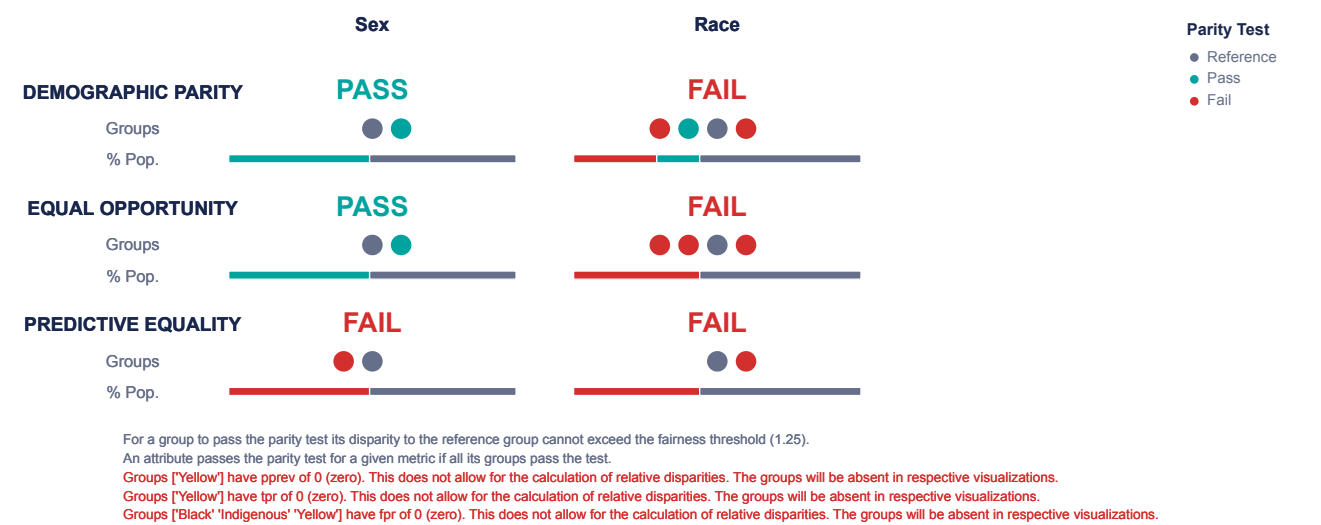


Figure 23. Parity Test Dashboard: Children’s Tech Vulnerability (ICT Kids).

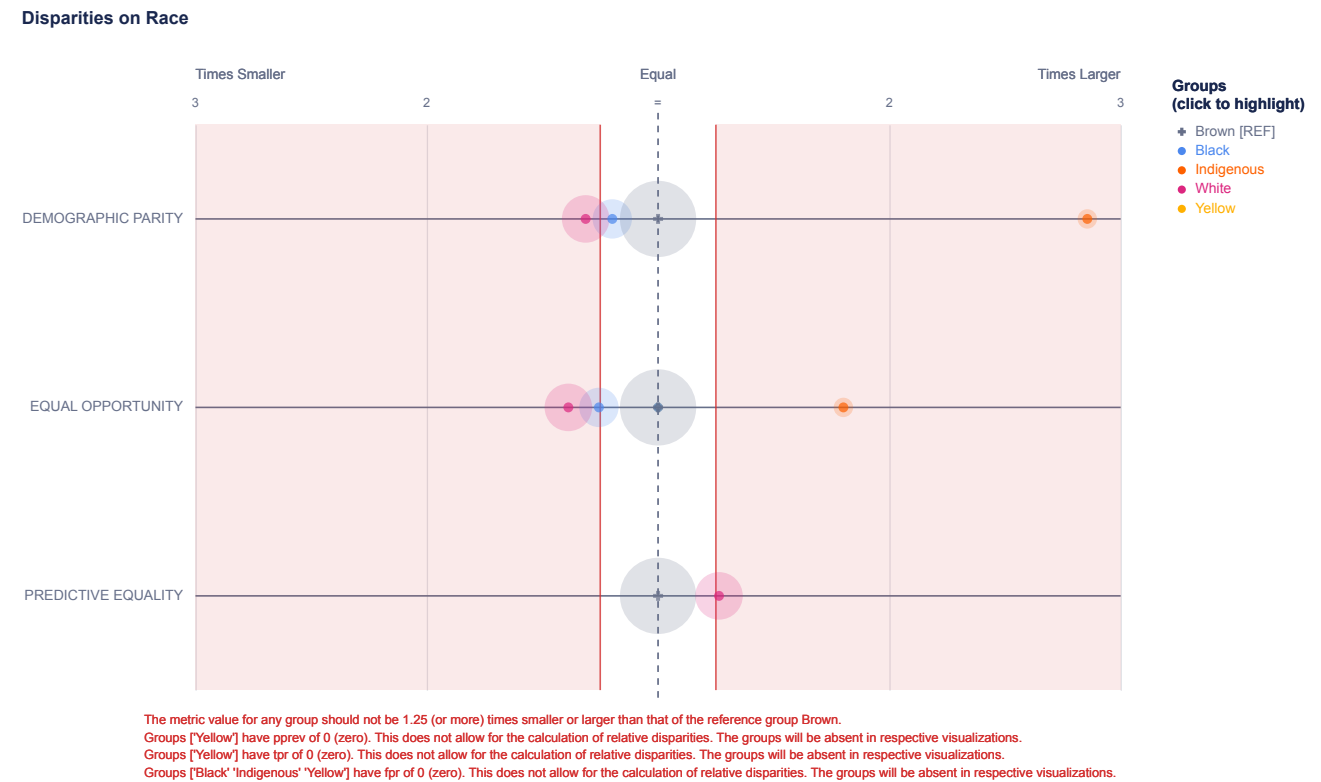


Figure 24. Fairness Disparity Plot (Race) for the Children’s Tech Vulnerability (ICT Kids).