Policy Research Working Paper 11046

Gender Bias, Citizen Participation, and AI

Jose Antonio Cuesta Leiva
Natalia Gisel Pecorari

WORLD BANK GROUP
Social Development Global Department
January 2025

Policy Research Working Paper 11046

Abstract

This paper investigates the role of gender bias in artificial intelligence–driven analyses of citizen participation, using data from the 2023 Latinobarómetro Survey. The paper proposes that gender bias—whether societal, data driven, or algorithmic—significantly affects civic engagement. Using machine learning, particularly decision trees, the analysis explores how self-reported societal bias (machismo norms) interacts with personal characteristics and circumstances to shape civic participation. The findings show that individuals with reportedly low levels of gender bias, who express political interest, have high levels of education, and align with left-wing views, are more likely to participate. The paper also explores different strategies to mitigate gender bias in both the data and the algorithms, demonstrating that gender bias remains a persistent factor even after applying corrective measures. Notably, lower machismo thresholds are required for participation in more egalitarian societies, with men needing to exhibit especially low machismo levels. Ultimately, the findings emphasize the importance of integrated strategies to tackle gender bias and increase participation, offering a framework for future studies to expand on nonlinear and complex social dynamics.

This paper is a product of the Social Development Global Department. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at jcuesta@worldbank.org. A verified reproducibility package for this paper is available at http://reproducibility.worldbank.org, click **here** for direct access.

Gender Bias, Citizen Participation, and AI

Jose Antonio Cuesta Leiva[1] and Natalia Gisel Pecorari[1]

(1) Social Development Global Department

Keywords: Citizen participation; Gender bias; Machine learning; Latin America and the Caribbean

JEL Codes: D70, J16, C63

Acknowledgments: The authors thank Martin Carlos De Meio Reggiani for his invaluable comments, which certainly improved this paper. Any remaining errors are the entire responsibility of the authors

1. Introduction

Artificial intelligence (AI) enables computers and machines to mimic human thinking, problem-solving, and social skills (Haenlein and Kaplan, 2019). We encounter AI in our daily lives through digital assistants, GPS navigation, self-driving cars, and generative tools like OpenAI's ChatGPT. Many see AI as a catalyst for economic growth, impacting various sectors, from consumer goods to health care (Shrestha and Das, 2022). However, this promise also raises significant concerns, particularly regarding how AI can perpetuate societal inequalities and injustices linked to factors such as gender, race, ethnicity, or age. This paper specifically examines how bias in AI, rooted in social inequality, can hinder our understanding of citizen participation, a vital element of democracy.

While previous research has explored citizen participation, few studies have employed AI methods. The available evidence indicates that participation is a complex and nonlinear behavior that traditional linear statistical methods often oversimplify. For example, Pecorari and Cuesta (2024) highlight potential misinterpretations of participation patterns in Latin America. One of those is that women are generally less likely to engage in various forms of civic participation—such as signing petitions, or voting (which is not corroborated by data from Latinobarometro 2024).

Although the case for applying AI methods to study participation is compelling, little attention has been paid to how biases related to gender, gender identity, and sexual orientation impact research findings and the policy making that follows. We propose a two-part working hypothesis: first, that gender bias significantly affects our ability to predict citizen participation; and second, that all forms of gender bias—societal, data, and algorithmic—are crucial in shaping these predictions.

This paper contributes to the literature by identifying, assessing, and comparing different gender biases in the context of AI's analysis of social issues, rather than considering them separately as previous studies discussing AI biases have done. It is important to note that AI is closely linked to machine learning and deep learning: machine learning focuses on creating algorithms that improve predictions by learning from structured, labeled data, while deep learning uses unsupervised techniques to extract insights from large, unlabeled datasets (Gupta, Parra, and Dennehy, 2022). Although AI encompasses more than these two approaches, much of the literature emphasizes machine learning and its implications for fairness in research. Our analysis shares this emphasis.

The structure of this paper is as follows: Section 2 reviews the definitions and classifications of gender bias in AI, discussing their causes and potential mitigation strategies. Section 3 explains why decision trees are chosen as the preferred AI method for analyzing citizen participation and details how to operationalize societal, data, and algorithmic gender-related biases for comparison. Section 4 presents key predictions regarding citizen participation, focusing on how machismo societal norms influence engagement. Finally, Section 5 discusses the findings and concludes the paper.

2. A literature review of AI, gender bias, and citizen participation

2.1. What is gender bias in AI?

Bias occurs when outcomes are consistently less favorable for certain groups without valid justification, often due to socio-political power imbalances related to gender, race, or other factors (Beauvoir, 1949; Bem, 1993). Gender biases are pervasive across multiple domains—political, economic, and civic—limiting women's opportunities and experiences. In the political realm, women often face barriers to leadership due to societal perceptions of leadership as a male trait, reinforcing gender inequalities in power and decision-making (Eagly and Karau, 2002; Ridgeway and Correll, 2004; Kabeer, 2005). In the economic sphere, gender biases are evident in labor markets where women, particularly mothers, encounter what is known as the "motherhood penalty," with assumptions about their competence and commitment leading to lower hiring rates and slower career advancement compared to their male counterparts (Correll, Benard, and Paik, 2007; Kleven et al, 2019). Additionally, biases affect civic engagement and participation, as women's voices are often marginalized in community decision-making processes, despite their active contributions in various areas (Cornwall, 2003). These biases reinforce unequal power dynamics and hinder the broader goal of gender equality in all aspects of social life.

Friedman and Nissenbaum (1996) linked bias in AI to discriminatory outcomes, defining it as systems that unfairly and systematically disadvantage certain groups. Gender bias in AI arises when data and outputs systematically harm individuals based on gender, encompassing biological sex, gender identity, and sexual orientation, such as AI diagnostic tools misdiagnosing women due to male-centric data (UN Women, 2024). Non-binary individuals are also underrepresented in automated systems, for example, in the allocation of social programs, reinforcing gender bias (Hicks, 2019). Facebook's job advertisement algorithm exemplified this by targeting jobs based on gender, suggesting stereotypically feminine jobs to women (e.g., nurses) and masculine jobs to men (e.g., taxi drivers), and perpetuating stereotypes (Ali et al. 2019).

Empirical evidence of gender bias in AI is widespread. Vlasceanu and Amoro (2022) found that higher national gender inequality (measured by the global gender gap index) in 52 countries correlated with greater disparities in Google image search results for the term "person." Smith and Rustagi (2021) identified 133 biased AI systems across industries, with 45 percent displaying gender bias and 26 percent showing both gender and racial biases. In Ireland, Donnelly and Stapleton (2021) documented the marginalization of transgender and non-binary individuals in online services due to mandatory gender data requirements, leading to the rejection of applications and feelings of exclusion.

2.2.Causes of gender bias in AI

To understand the origins of gender bias in AI, we first consider general biases in AI. Garcia-Gathright, Springer, and Cramer (2018) categorize biases into input data, algorithmic decisions, and outcome biases affecting user groups. Friedman and Nissenbaum (1996) and Bender and Friedman (2018) similarly discuss pre-existing, technical, and emergent biases (from the misapplication of AI systems). Glymour and Herington (2019) focus on procedural, outcome,

and behavioral biases, different stages of the generation of AI-based predictions. Baeza-Yates (2018) introduces a taxonomy of algorithmic, user interaction, and data-related biases. Other studies highlight issues like unrepresentative training data and feedback loops (Cowgill and Tucker, 2020). Common in all these explanations is the distinction between data-related and algorithmic causes of gender bias in AI.

Data biases arise when training data reflects societal inequalities or is unrepresentative, such as AI models trained mainly on male data, which perform poorly for women, or facial recognition systems struggling with non-Caucasian faces. AI perpetuates societal biases through mechanisms like "word embedding," which encodes words linking skills or interests to gender (UN Women, 2024). Other Natural Language Processing (NLP) methods also exhibit biases, impacting applications like dialogue generation, translation, text parsing, hate speech detection, and sentiment analysis (Blodgett et al., 2020). Limited or poor-quality data further exacerbates these issues. Pedestrian detection algorithms often fail to detect children and women, while facial recognition systems misidentify darker-skinned faces due to their underrepresentation in training data (Lee, Resnick, and Barton, 2019).

Biases in AI can also emerge from stages beyond training data, including filtering, coverage, ranking, and presentation, which can amplify social biases entrenched in data. Algorithmic biases arise from design processes that embed societal biases, such as job recommendation algorithms showing fewer STEM ads to women because of fewer female STEM graduates. Confirmation bias is amplified through information filtering, with personalization algorithms creating "filter bubbles" that reinforce users' existing beliefs (Pariser, 2011). Search engines and social media platforms further influence user encounters, increasing the risk of bias (Van Couvering, 2007). Ranking bias, driven by factors like popularity and novelty, affects search results, making less popular content harder to find (Nissenbaum and Introna, 2000). For example, Twitter's trending topics highlight new content over persistent issues, potentially making a derogatory meme about women go viral faster than a slowly emerging gender equality movement (Gillespie, 2012).

The lack of diversity among AI developers also contributes to these biases, as their perspectives shape AI systems, potentially neglecting underrepresented groups. Only 22 percent of AI professionals are women, often in lower-status roles (Young, Wajcman, and Sprejer, 2021). Gebru (2020) questions if automatic gender recognition tools, harmful to transgender communities, would exist without tech industry dominance by cisgender men. Feminist approaches to AI go further. Walcott (2019) argues that unconscious bias narratives often overlook institutional discrimination, obscuring accountability for the racism and sexism of programmers—influenced by patriarchal, racial, and colonial hierarchies in AI development.

2.3. Consequences of gender bias in AI

Gender-biased AI systems have considerable consequences. Gender bias in AI leads to poor service, misrepresentation, and discrimination. Biased AI systems can deliver inferior services to women and non-binary individuals, as seen in voice recognition in the automotive sector (Smith and Rustagi, 2021). In health care, biased AI poses risks to underrepresented groups. For example, AI in skin cancer detection struggles with melanoma detection for Black people, endangering Black women who are already underserved by health care (Smith and Rustagi, 2021). Gender-biased systems also impact physical and mental well-being, including those of

women and non-binary individuals (Shrestha and Das, 2022). In criminal justice, biased AI affects female prisoners unfairly, with biased algorithms skewing recidivism predictions (Karimi-Haghighi and Castillo, 2021).

More relevant to our analysis are the implications of gender bias in AI to democracy, for which the verdict is still out. Advancements in AI can enhance the government's understanding of public needs and citizen participation in policy making. AI can analyze large datasets to detect corruption, boosting accountability and trust. Automated tools for summarizing public feedback can improve officials' ability to process inputs, while grievance systems enhance transparency and provide data for AI models to predict policy impacts (Rahim, Mahony, and Bandyopadhyay, 2024).

AI enhances government communication, particularly in multilingual settings. In India, the Jugalbandi platform uses AI chatbots to deliver services in multiple languages across 171 government programs, while the Bangsamoro region in the Philippines employs AI to analyze social media for development insights (Rahim, Mahony, and Bandyopadhyay, 2024). AI platforms also improve marginalized communities' access to essential services, promoting inclusivity. For example, Chile's national human rights action plan incorporated diverse perspectives from indigenous peoples and incarcerated individuals, using large language models to analyze their input and ensure fair representation (Fajardo-Hayward and Cuesta, 2024).

However, the benefits of AI in citizen engagement depend on the accuracy and fairness of the systems to avoid reinforcing sociohistorical inequalities and marginalizing specific communities. Ensuring AI systems provide accurate, accessible information requires rigorous testing, high-quality data, and human oversight. For example, AI bots may respond differently based on a user's location or struggle with various accents, disadvantaging marginalized communities (Panditharatne, Weiner, and Kriner, 2023). Such biases have led to issues like increased surveillance of marginalized groups for immigration decisions (Eubanks, 2018) and manipulation of public comments on net neutrality rules (Panditharatne, Weiner, and Kriner, 2023). Studies have shown similar rates of bias detection in AI-generated content compared to human responses, risking neglect of genuine needs (Kreps and Kriner, 2023).

2.4. ***Detection and mitigation of gender bias***

Detecting biases in AI requires systematic objective methods. Biases are typically identified by comparing AI outcomes with human-coded benchmark datasets. For instance, Booth et al. (2021) use psychometrics to assess gender bias in video recruitment. Census data is often used to evaluate income prediction biases (Feldman and Peake, 2021). More advanced methods include Winograd schemas, which involve ambiguous sentence pairs to test AI's language comprehension and contextual understanding, a method typically used in sentiment analysis (Sarraf et al., 2021). Beyond those methods, fairness metrics are also used for detecting AI bias. Fairness through unawareness ensures models ignore attributes like gender in predictions, while counterfactual fairness checks if altering someone’s gender affects outcomes. Equal opportunity fairness compares true positive rates across groups for equity, and average odds fairness assesses equal rates of true and false positives. Disparate impact and statistical parity fairness focus on

consistent outcome probabilities for different genders. Feldman and Peake (2021) detail these metrics, which have been applied in credit risk predictions and criminal re-offending assessments.

Mitigating AI bias, contrary to detection, involves both technical and policy aspects, requiring collaboration between technical experts, policy makers, and other stakeholders. Once biases are identified, they can be addressed through multiple ways, which we grouped, consistent with the causes of bias, around data, algorithms, and public policy solutions. Technical approaches to correcting data biases include formal evaluations, balanced datasets, and counterfactual data augmentation. Bender and Friedman (2018) highlight the importance of data statements, which provide essential details about datasets, helping identify and address biases. Cramer et al. (2019) advocate using checklists and understanding the model's application context to mitigate bias. Baeza-Yates (2018) stresses anticipating diverse usage contexts to avoid bias from mismatched assumptions. On balanced datasets, Wang et al. (2021) recommend datasets that reflect all demographic groups accurately, either by increasing data from underrepresented groups or adjusting proportions. Wu et al. (2020) propose creating datasets that are racially and gender-inclusive, including non-binary identities. Counterfactual data augmentation techniques address gender biases in NLP by adding or modifying data to balance gender representation. For instance, it involves creating gender-equivalent statements to ensure balanced representation, reducing biases in NLP models (Maudslay et al., 2019; Shrestha and Das, 2022). This can be done by randomly incorporating new data or modifying parts of the existing data.[1]

Technical approaches to correcting algorithmic biases include adversarial debiasing, greedy algorithms, and regularization approaches. Adversarial debiasing enhances prediction accuracy while minimizing the ability to reveal protected attributes like gender. An "adversary" challenges the model to uncover biases, aiming for the algorithm to assess attributes without considering protected information. Morales et al. (2020) and others apply this to facial analysis, visual recognition, and dialogue systems, ensuring fairness and privacy (Hong et al., 2021; Wang et al., 2020; Liu et al., 2020; Dhar et al., 2020). Greedy algorithms optimize fairness metrics during model training. These algorithms aim to maximize fairness based on criteria set by the model creators. For example, Barnabò et al. (2019) use greedy algorithms to balance team diversity—ensuring representation across gender, race, and other factors—while meeting task requirements and minimizing labor costs. Regularization approaches include techniques like Equalized Odds, which ensure similar prediction rates across groups (Singh and Hofenbitzer, 2019). Post-process regularization corrects biases after training. Multi-task Convolutional Neural Networks (MTCNN) improve performance by learning multiple tasks simultaneously (Das, Dantcheva, and Bremond, 2018). Lagrangian relaxation optimizes decision-making by breaking problems into simpler parts (Zhao et al., 2017).

Technical data and algorithm-related mitigation strategies should be primarily handled by programmers, but policy interventions are also needed. The first key recommendation includes raising algorithm literacy. Smith and Rustagi (2021) emphasize the need for public understanding of algorithms, akin to computer literacy. Training should cover responsible use,

[1] For example, if a dataset mostly includes phrases linking the male gender with the profession of a doctor, such as “He is a doctor” or “The doctor provided his expert advice,” Canonical Discriminant Analysis would create and include equivalent statements with female pronouns, like “She is a doctor” or “The doctor provided her expert advice.” See Shrestha and Das (2022).

ethics, and gender equity, with increased funding for lower digital literacy groups. A second recommendation refers to supporting research on inclusive AI models. The focus of funding should be the development of equitable AI models and datasets, addressing data gaps through community-driven efforts and partnerships, and promoting a wider programmer diversity. Initiatives like the Data Empowerment Fund and USAID's Equitable AI Community of Practice support such efforts (Smith and Rustagi, 2021).

Despite increasing evidence of gender biases in AI, key gaps persist. Current research, often based on small-scale studies and inconsistent metrics, fails to fully capture the scope and impact of these biases. We need comparative analyses of bias mitigation techniques to better understand how gender biases in AI influence assessments of citizen participation and the potential for improving its quality and quantity. The next section outlines an analytical strategy to align various data and algorithmic gender biases, comparing them to societal gender bias—measured by reported levels of machismo from individual survey responses.

3. Methodology and data: Analyzing gender bias in ML models of participation

3.1. Decision trees

Decision trees, a core method in supervised machine learning, effectively capture non-linear patterns in complex datasets, including those reflecting intricate social dynamics like citizen participation. Compared to methods such as Lasso, Ridge, and Elastic Net, they not only excel in modeling non-linearities, but also offer superior interpretability, as demonstrated by Hastie, Tibshirani and Friedman (2009); James, Witten, Hastie and Tibshirani (2013); Rudin (2019); and Pecorari and Cuesta (2024).

The use of decision trees allows for clear visualization of the pathways leading to citizen participation, offering intuitive representations of how various factors influence the outcome. While advanced methods like Random Forests and XGBoost deliver stronger predictive performance through ensemble and boosting techniques, they lack the interpretability that decision trees provide, especially in visualizing decision paths. In our study, these pathways will help uncover whether and how different levels of gender bias affect civic participation. Moreover, decision trees enable us to explore the interplay between gender bias and other variables, shedding light on how interactions between social norms, personal circumstances, and identity shape participation decisions.

The algorithm operates by recursively partitioning the dataset into increasingly distinct subsets based on specific features, continuing until each subset becomes homogenous, representing a unique label or value. Homogeneity is achieved by selecting features at each split that either minimize variance or maximize purity within the resulting groups. Ideally, the final subsets (leaf nodes) contain data points that share the same label or exhibit highly similar values, ensuring uniformity with respect to the target variable. The process begins by selecting an optimal predictor from a set {x1, x2, ..., xk} for a given dependent variable y, then partitioning the data at an optimal point. This optimization seeks to minimize an impurity measure for classification tasks or variance for regression. For classification, the objective can be formalized as:

$$Split\ criterion\ =\ arg\min_{f,t}(\frac{n_{left}}{n}.I_{left}+\frac{n_{right}}{n}.I_{right})$$

where f is the feature chosen for splitting, t is the threshold or split point for the feature, n_{left} and n_{right} are the number of samples in the left and right child nodes after the split, n is the total number of samples at the current node, and I_{left} and I_{right} are the impurity measures (such as Gini index, entropy, or misclassification error) for the left and right child nodes. By iteratively selecting f and t that achieve this minimization, the algorithm builds a branching structure, recursively splitting the data until the terminal nodes (leaf nodes) are reached. These nodes effectively illustrate the relationships between predictors and the target variable, offering a clear, interpretable map of the decision-making process.

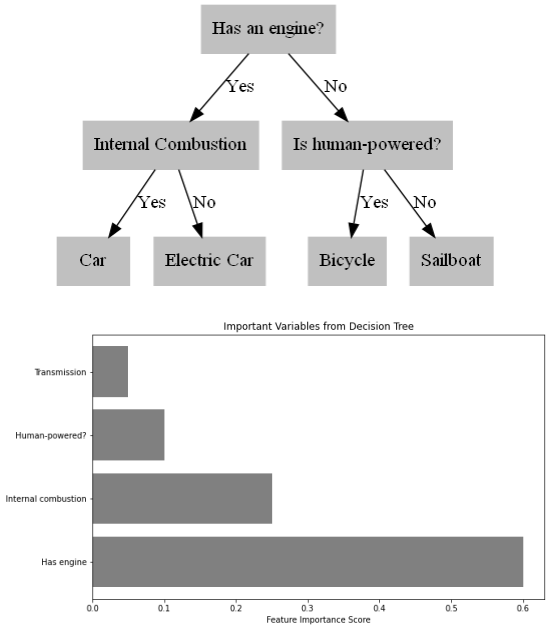
In the context of citizen participation, the dependent variable can capture activities like voting, signing petitions, participating in demonstrations, or engaging in community problem-solving. Independent variables often include demographic factors such as age, gender, employment status, and education level. Of particular relevance to our study are variables related to political and gender-related beliefs, which have been explored in prior research on participation, including Verba and Nie (1987) and Lee and Schachter (2019).

A major advantage of decision trees is their capacity to highlight feature importance, offering clear insights into the predictors that most significantly affect the outcome. They are also highly interpretable, with their rules and splits being intuitive and easy to visualize. To mitigate overfitting, pruning techniques are commonly employed to streamline the tree by removing branches that add minimal value to its accuracy (James, Witten, Hastie and Tibshirani, 2013).

Decision trees are essential for exploring complex nonlinear relationships among variables, enabling us to observe how independent variables influence the outcome at different levels, particularly in interaction with one another. In our context, understanding how the gender bias score—a measure of machismo—interacts non-linearly with education levels to shape citizen participation is especially pertinent. These intricate interactions will be further analyzed in the findings section, using interaction plots to illustrate their dynamics.

Despite these advantages, decision trees can be unstable, with small changes in the data potentially causing significant shifts in the tree's structure (James et al., 2013). To mitigate this, we complement our decision tree analysis with alternative models, including counterfactual scenarios, to validate our results. A key focus of our analysis will be to assess how different model choices influence the thresholds at which gender bias becomes a significant factor in predicting civic participation.

Figure 1. Illustration of a decision tree and feature importance



Source: authors

3.2 Using decision trees to understand the impacts of gender bias in AI

Having established that decision trees are the most suitable AI tools for predicting participation in civic activities, we now outline how they will be used to address gender biases. This involves three distinct approaches: first, correcting for the gender bias reported by individuals in surveys, which we term society's revealed gender bias; second, controlling for gender bias in the data by altering its distribution through scaling and shifting scenarios, as discussed in the previous section; and third, examining the mitigation of algorithmic biases by blinding the data to gender variables and switching gender among respondents. We also remove proxies closely correlated with gender in a deproxing scenario. Finally, we examine algorithmic biases by analyzing participation in male- and female-only samples.

Baseline Model: Decision Tree on Citizen Participation after Correcting for Society's Gender Bias

In the baseline model, we incorporate a gender bias score that reflects societal perceptions of women's roles, specifically measuring machismo levels. This allows us to directly address gender bias. The model will evaluate how varying degrees of machismo influence an individual's decision to engage in civic participation.

Model 1: Decision Tree on Citizen Participation after Correcting for Bias in the Algorithm: No Gender Variable (Gender Blinding)

In Model 1, we train the decision tree without allowing the algorithm to access the individual's gender. This approach helps correct for potential algorithm biases associated with the detection of gender in the data, as well as the absence of nonbinary options to capture gender identity.

Model 2: Decision Tree on Citizen Participation after Correcting for Bias in the Algorithm: No Gender Variable Nor Gender Proxies (Gender Deproxing)

In Model 2, we apply a deproxing strategy, which involves removing not only the gender variable but also any covariates that serve as proxies—those highly correlated with gender and thus likely to carry gender-related bias. In our analysis, employment status was identified as the most strongly correlated proxy for gender. As a result, we remove both the gender variable and employment status from the list of covariates.

Model 3: Decision Tree on Citizen Participation after Correcting for Bias in the Algorithm: Women and Men-only

In this model, we analyze how algorithms predict participation separately for female and male samples, excluding any information about the opposite gender.

Model 4: Decision Tree on Citizen Participation after Correcting for Bias in the Algorithm: Gender Switching

In Model 4, we simulate a gender-blind society by switching the gender of each individual—assigning males as females and vice versa.

Model 5: Decision Tree on Citizen Participation after Simulating a Less/More Machista Society: Scaling Machismo

In Model 5, we use counterfactuals to explore how the success path to citizen participation would differ in alternative scenarios, representing both a less machista and a more machista Latin American society, as reflected in Latinobarómetro data. We construct two main counterfactuals: first, by scaling down individual observations of the composite gender score while preserving the original distribution, simulating a society with reduced levels of machismo.[2] Our primary counterfactual envisions a Latin American society with half the current machismo level, but we also examine scenarios with machismo levels ranging from 0.01 to 1, where 1 represents the current distribution of the score. Additionally, we consider a scenario where machismo is doubled. Since the composite gender score is normalized between 0 and 1, doubling the score truncates the distribution at 1, altering the original variable distribution.

Model 6: Decision Tree on Citizen Participation after Simulating a Less Machista Society: Shifting Machismo Distribution

In Model 6, we apply a second counterfactual approach by shifting the gender bias score for each individual observation, deducting a fixed value of 0.25 from each score.[3] This method alters the distribution of the data, simulating a less machista society. The goal is to explore different ways

[2] Maintaining the original distribution ensures the analysis remains realistic and reflective of the true essence of Latin American society, as represented by the current distribution of machismo. By shifting this distribution to the left, we aim to provide a more authentic depiction of how reduced levels of machismo could influence pathways to citizen participation.

[3] The choice of a shift amount of 0.25 is informed by the observation that beyond this threshold, the gender score ceases to appear in the success path (refer to Figure 4, Panel B).

of reducing machismo and assess how these changes impact the success path to citizen participation. This approach also allows us to evaluate the robustness of the results regarding the influence of reduced machismo on participation.

These methods are informed by the review of technical strategies to mitigate gender biases in AI presented earlier. The gender blinding, deproxing, and shifting/scaling strategies align with data and algorithmic bias mitigation techniques in the literature, such as adversarial debiasing and counterfactual data augmentation, aimed at reducing gender bias in AI models. The WOMEN-only and MEN-only models further explore the literature's emphasis on examining gender-specific bias and ensuring fairness across demographic groups, consistent with recommendations for balanced datasets and inclusive AI. The focus on societal gender bias as an overarching factor aligns with calls for considering broader social contexts in AI assessments, as highlighted by Baeza-Yates (2018) and Smith and Rustagi (2021), who stress the need for algorithms to account for societal biases.

3.3 Data and variables

This study uses data from the 2023 Latinobarómetro Survey (Latinobarómetro Corporation 2023), an annual survey designed to assess individual perceptions of socioeconomic and political issues. Covering 17 countries in Latin America and the Caribbean, it is the largest regional database on citizen attitudes toward democracy. The survey employs a stratified random sampling method, weighted to accurately represent each country's population. The latest data available during the preparation of this study, released in December 2023, includes 19,205 observations, representing over 600 million people in the region. However, it does not include data from Nicaragua, Haiti, Trinidad and Tobago, or other small Caribbean nations. Data collection was conducted face-to-face from February 20 to April 18, 2023.

The dependent variable in this study is citizen participation, specifically in the context of working to address community problems. The primary independent variable is a composite gender bias score, derived as the first component of a principal component analysis (PCA) based on four indicators that capture the prevalence of gender-biased social norms in the country of analysis: (i) women should focus on domestic roles while men should work; (ii) men are better political leaders than women; (iii) if a woman earns more than a man, she is likely to face difficulties; and (iv) in times of scarce jobs, men should have a greater right to employment than women.

Each of these four gender bias indicators is coded as 1 if the respondent agreed or strongly agreed, and 0 otherwise. The composite gender bias score is then normalized on a scale from 0 (low machismo) to 1 (high machismo). By using the first component of principal component analysis (PCA), we derive a single variable that captures the combined variability of the four gender bias indicators, simplifying the analysis of how gender bias influences individual decisions to engage in solving community problems.

The decision to use the first principal component from the PCA is based on two widely used, independent methods. First, the Kaiser-Guttman criterion, which recommends retaining components with an eigenvalue greater than one, supports using the first component (Kaiser, 1960; Guttman, 1954). Second, the elbow method, a visual approach, indicates that the first

component explains most of the variability, while the second component adds only marginally to the total variance in our set of machismo variables. Additionally, using the first component to construct the composite gender bias score is preferred over a simple weighted average, as it better captures the underlying structure, variability, and interactions among the indicators, providing a more nuanced and comprehensive measure of gender bias. In contrast, a weighted average may oversimplify these relationships and overlook key interactions (Jolliffe, 2002).

The analysis includes a set of control variables identified in previous research as predictors of individual participation (Verba and Nie, 1987; Lee and Schachter, 2019). These variables are categorized as follows: *Demographic and socioeconomic variables*: age, gender, education level, subjective social class, receipt of subsidies, employment status, ability to save money, home ownership, internet access, and sewage access. *Beliefs variables:* interest in politics, political orientation (left-right), perceived freedom of expression, perceived freedom to join organizations without fear, life satisfaction, experience as a victim of crime, food insecurity status, interpersonal trust, and trust in governmental institutions (government, police, parliament, courts, and elections).

4. Results

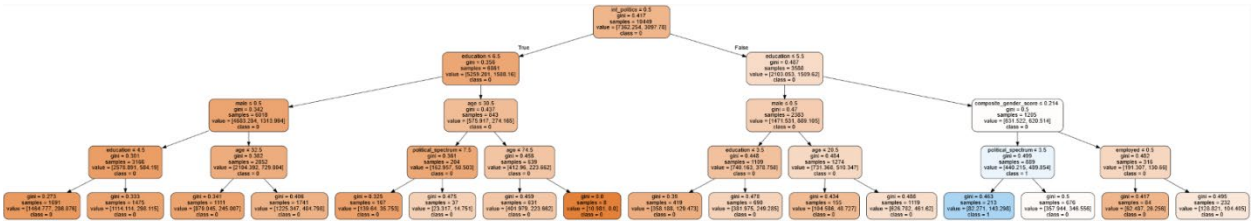
4.1. Comparing results from alternative correcting models

Decision trees yield two main results. First, they identify key factors predicting participation in community activities by mapping successful paths and analyzing how these factors influence outcomes under various conditions. For instance, education might impact participation only after specific thresholds are met. This method effectively captures complex, conditional patterns. Second, the analysis examines how societal gender bias influences participation, highlighting the significance of bias under different scenarios and thresholds. For example, a low gender bias (e.g., 0.2 on a 0–1 scale) may influence participation in one case, while a higher bias (e.g., 0.8) is required in another. Integrating these two sets of findings reveals how gender biases affect participation in Latin America and the Caribbean and identifies effective strategies for mitigating them to enhance inclusivity.

Figure 2 illustrates the key factors predicting participation while accounting for societal gender bias (baseline model). The decision tree reveals that an individual is likely to participate if they are first interested in politics, have at least completed secondary school, report a low level of gender bias (below 0.214), and identify with left-wing political views.

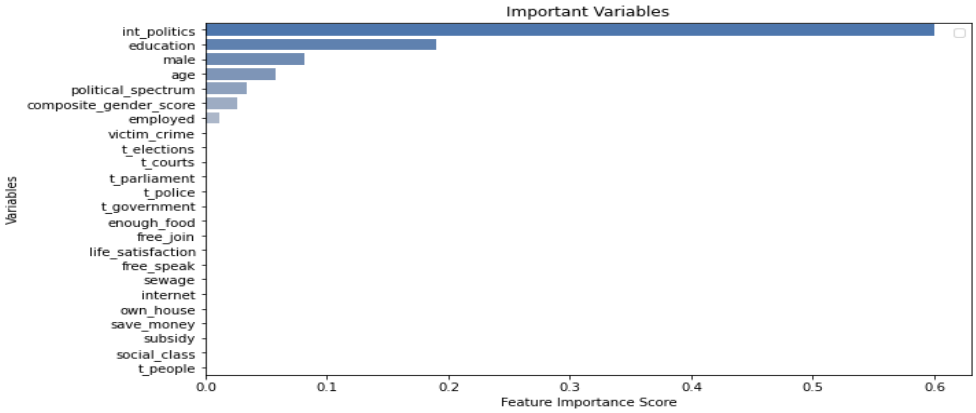
The relative importance analysis shows that gender bias is one of the few variables that significantly affects participation, though its impact is modest. It accounts for approximately 5 percent of the variation in participation, while interest in politics explains over 60 percent and education contributes an additional 20 percent.

Figure 2. Baseline model: decision tree for citizen participation (working to solve a community problem) accounting for gender bias, LAC 2023.



Source Author's elaboration using the function DecisionTreeClassifer in the Scikit-Learn 1.2.2 library in Python 3.11.
Note: The color palette of the nodes indicates the class to which the majority of the samples at each node belong (blue captures class 1 while orange captures class 0). The Gini index measures the impurity or disorder in a node. Samples refer to the number of observations that are classified in each node. Value tells how many observations at the given node fall into each category or class. In our case, we have two classes: working for a community problem or not.

Figure 3. Feature importance for the baseline model for citizen participation (working to solve a community problem) accounting for gender bias, LAC 2023.



Author's elaboration using Scikit-Learn 1.2.2 library in Python 3.11.
Note: The graph shows the feature importance scores of predictor variables in predicting the target outcome (work for a community problem).[4]

Table 1 presents the results from the alternative models analyzed. The findings consistently show that gender bias remains a significant factor, underscoring its persistent influence across all models. In the baseline model, the societal gender bias score is 0.214, reflecting a moderate impact on participation. Even when gender variables are excluded, as seen in the gender

[4] The variables are defined as follows: *int_politics* refers to the respondent's interest in politics; *education* captures the level of education; *male* indicates the respondent's sex (1 = male, 0 = female); *age* represents the respondent's age; *political_spectrum* measures self-placement on the political spectrum; *composite_gender_score* reflects perceptions of gender bias; *employed* denotes employment status; *victim_crime* indicates whether the respondent was a crime victim; *t_elections*, *t_courts*, *t_parliament*, *t_police*, and *t_government* measure trust in elections, courts, parliament, police, and government, respectively; *enough_food* captures food insecurity; *free_join* and *free_speak* measure perceptions of freedom to join organizations and freedom of speech; *life_satisfaction* captures overall life satisfaction; *sewage* and *internet* reflect access to sewage infrastructure and the internet; *own_house* indicates house ownership; *save_money* represents the ability to save money; *subsidy* captures receipt of subsidies; *social_class* measures perceived social class, and *t_people* reflects trust in people.

BLINDING and gender DEPROXING models, bias continues to play a role, indicating that algorithmic biases persist despite the absence of explicit gender data.

The SCALING machismo model, which simulates a society with half the level of machismo currently observed, reduces the gender bias score, demonstrating that data adjustments can mitigate bias, though not fully eliminate it. This suggests that even in a less machista Latin American society, gender bias—while reduced—would still influence citizens' decisions to participate. Similarly, when the SHIFTING machismo distribution model is applied to simulate a less machista society, the gender bias score remains a significant factor affecting participation.

Table 1 A comparative assessment of gender bias correcting models, LAC 2023.

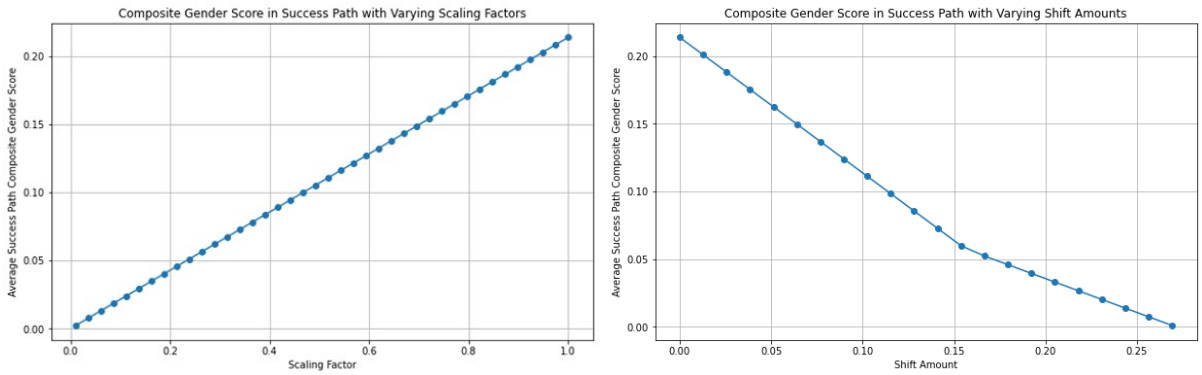
Model	Success path variables	Is societal gender bias a relevant variable?	Score at which societal gender bias is relevant	Which gender bias is addressed?
BASELINE	Interest in politics, education, gender, age, gender bias, political orientation, employment	YES	0.214	Societal gender bias
[1] Gender Blinding (no gender variables)	Same as baseline	YES	0.214	Algorithm
[2] Gender Deproxing (no gender nor proxies)	Same as baseline	YES	0.214	Algorithm
[3] WOMEN only	Same variables as baseline but adding a new variable: victim of a crime	YES	0.214	Algorithm
[3] MEN only	Same variable as baseline but for those men who have sewage, participation is more likely if he is left-wing, while if he doesn't have sewage then the model predicts participation regardless of political orientation	YES	0.078	Algorithm
[4] Gender Switching	Same as baseline	YES	0.214	Data
[5] Scaling Machismo (Down)	Same as baseline	YES	0.107	Data
[5] Scaling Machismo (Up)	Same as baseline	YES	0.427	Data
[6] Shifting Machismo distribution	Same as baseline	YES	0.01	Data

Source: authors' estimates

Figure 4 compares the effect of systematically scaling down self-reported gender biases from 1 to 0.01 in 40 steps,[5] as well as shifting toward less bias (from 0 to 0.5 in 40 steps). The results confirm that even at lower levels of gender bias, this factor remains relevant. While its overall importance does not significantly change, the threshold at which it becomes influential shifts. This indicates that lower levels of gender bias still play a role in determining participation, but the magnitude at which it begins to influence decisions changes as the bias is reduced.

The differences between the WOMEN and MEN models highlight how predictors of participation vary by gender, underscoring the context-specific nature of algorithmic bias. For women, being a crime victim emerges as an additional factor influencing participation, while for men, political orientation and access to services like sewage are more strongly linked to participation. These findings emphasize the importance of addressing gender-specific circumstances in efforts to correct algorithmic bias. They also provide valuable insights for designing policies that effectively promote participation for both women and men, illustrating how different circumstances, in conjunction with perceptions of machismo, influence participation in distinct ways for each gender.

Figure 4. Composite gender score in success path with varying scaling and shifting factors, LAC 2023.



Author's elaboration using matplotlib library in Python 3.11.

These findings are robust across alternative participation variables. Appendix 1 presents the results of several robustness tests in the form of decision trees for citizen participation measured by signing petitions and participating in protests, as reported in Latinobarómetro (other forms of participation, such as voting in elections or electronic e-governance, are not included in the survey). The results confirm that the same key variables appear in the success path of these alternative trees. Interest in politics, education, political orientation, and age remain critical factors in explaining petition signing and protest participation. In the case of signing petitions, internet access also becomes a relevant predictor of participation. Notably, societal gender bias continues to be a relevant predictor, contributing around 10 percent to the total predictive power of the model. The threshold for relevant machismo is 0.214, which is fully consistent with the levels at which gender bias was found to influence community problem-solving. Our robustness

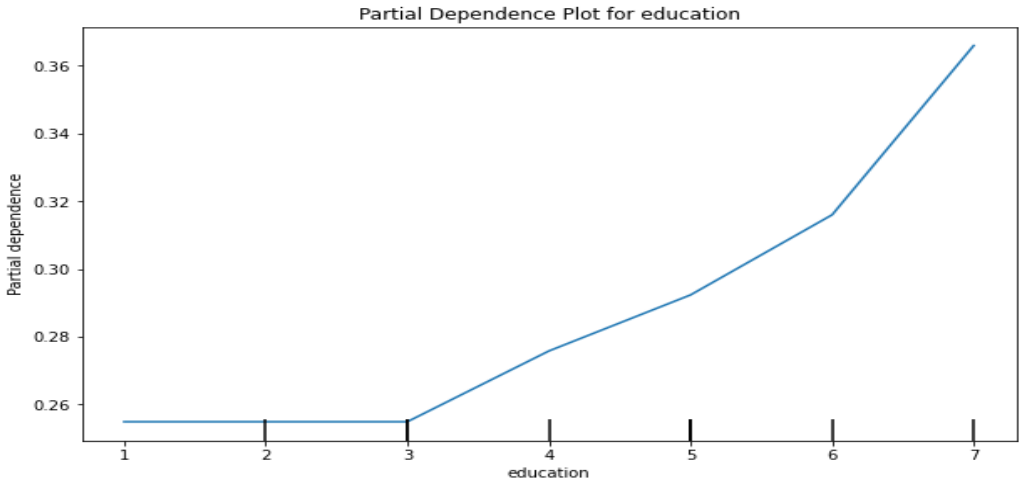
[5] For shifting factors above 0.25, the gender bias score is no longer a relevant factor. This is due to the change in the distribution of the gender bias score resulting from shifting the values, unlike when scaling is used as a strategy to simulate a less machista society.

checks confirm that PCA is the preferred method for capturing machismo. In contrast, alternative indices that aggregate the same variables used in the PCA method fail to remain significant predictors of participation in solving community problems (see Appendix 2).

4.2. The importance of nonlinearities in participation

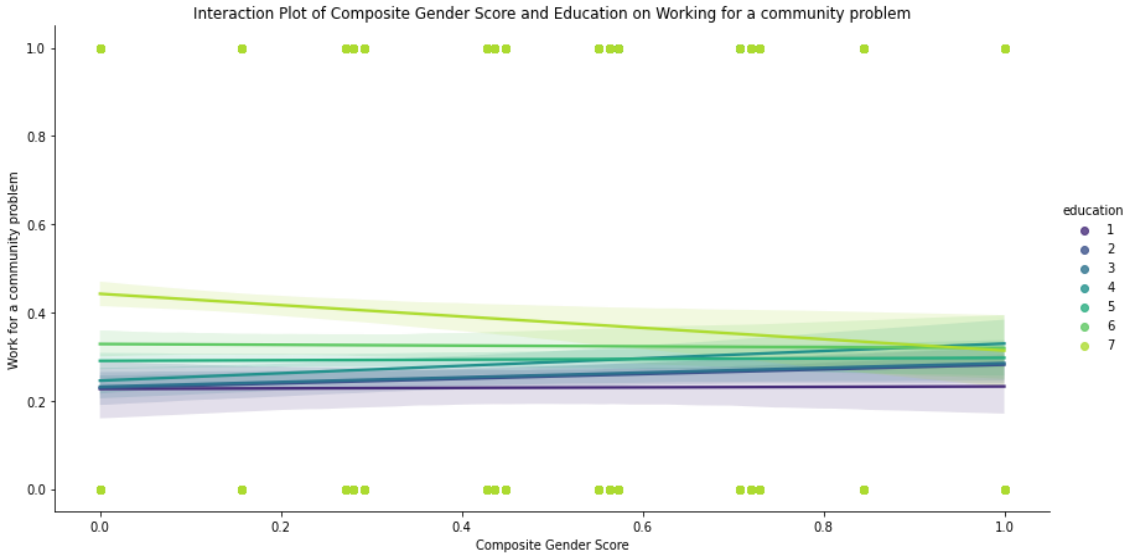
Examining how gender bias interacts with other key variables to influence citizen participation provides further insights. The decision tree from the baseline model shows that participation is predicted for individuals with political interest and relatively high levels of education. By using interaction and partial dependence plots (PDP), we can explore how education levels interact with gender bias in shaping citizen participation. Figure 5 presents a PDP for education, illustrating its impact on participation while holding all other variables constant. The plot reveals that education's influence increases with higher education levels, confirming the decision tree's findings. However, this influence is not linear. It is insignificant for individuals with less than primary school education but grows sharply for those with education beyond primary school. The highest levels of education, such as a university degree, exert the strongest influence on participation. Figure 6's interaction plot further shows that the effect of education on participation varies with gender bias, especially at higher education levels (level 7, complete university education). The influence is more pronounced for individuals with lower levels of machismo. In other words, the impact of gender bias on participation depends on the level of education, with the greatest effect observed for highly educated individuals with lower machismo scores.

Figure 5. Partial dependence plot for the effect of education on working on a community problem, LAC 2023.



Source: Author's elaboration using matplotlib library in Python 3.11.

Figure 6. Interaction plot: Effect of gender bias score and education levels on community problem involvement, LAC 2023.



Source: Author's elaboration using seaborn library in Python 3.11.

5. Discussion and conclusions

Artificial intelligence (AI) technology, which simulates human intelligence across various domains, holds transformative potential but also presents significant challenges, particularly regarding gender biases. This paper first explored how AI's implicit biases—related to gender, gender identity, and sexual orientation—can deepen societal inequalities. We defined gender bias in AI, traced its origins to both data and algorithms, and reviewed its impacts on service quality, human behavior, and democratic values.

Addressing gender bias in AI requires a multifaceted strategy that combines technical solutions with strong governance and diverse stakeholder involvement. This paper has focused on the technical side, shedding light on a two-prong hypothesis: first, that gender bias significantly affects our ability to predict citizen participation; and second, that all forms of gender bias—societal, data, and algorithmic—are crucial in shaping these predictions. Notably, although various methods to reduce data and algorithmic biases exist, there is a gap in approaches that can identify, quantify, and mitigate multiple biases simultaneously. This paper attempts to provide such an integrated analysis in addressing our hypotheses.

We build on previous research on citizen participation in Latin America and the Caribbean using machine learning techniques. Our approach innovatively explores how self-reported gender bias (societal bias) interacts with data and algorithmic biases. We account for factors such as self-reported machismo social norms, data lacking gender identifiers, gender identity switching, and the scaling and shifting of reported machismo norms. This strategy enables us to untangle the complex factors influencing citizen participation in the region (measured by involvement in community issues) while assessing the role of each type of gender bias in these predictions.

To compare results, we focus on the significance of self-reported gender bias in predicting participation and, when relevant, how machismo thresholds shift after applying bias mitigation measures. By comparing these models, we can evaluate the impact of different types of bias on

participation and examine the complex interactions between gender bias and other demographic or political factors. This analysis provides insights into how AI can reduce such biases in predicting civic engagement.

Our decision tree analysis provides two key insights into participation. First, individuals are predicted to participate if they are politically interested, have at least a secondary education, report a low level of gender bias (below 0.214), and identify with left-wing views. While it is intuitive that those with high gender bias are unlikely to participate, even if they are educated or politically interested, our findings reveal that participation is influenced by specific thresholds and a combination of conditions occurring simultaneously.

Second, decision trees expose the persistent influence of societal and algorithmic gender biases on participation. Though modest in their overall effect—5 percent of total predictive power vis-à-vis 80 percent combined with other relevant factors like education, political interest, and ideology—societal gender biases consistently shape participation. This result is true when correcting for algorithm biases (i.e., removing the gender variable in the data, deproxing variables associated with gender, or switching genders in our observations). When all those corrections are made, societal bias remains a relevant factor in predicting participation with the same threshold of low levels of machismo (0.214).

Additionally, if data biases were corrected to reflect less prevalent or lower machismo levels, civic participation would require even lower machismo thresholds. Intuitively, in more egalitarian societies, participation occurs only when individuals themselves are highly egalitarian, meaning their machismo levels are much lower than in more machismo-dominant societies. In such cases, the machismo threshold drops from 0.214 in models that control for societal and algorithmic biases to scores between 0.01 and 0.10.

A notable finding is that men must have lower levels of machismo bias to participate, assuming other conditions—such as education, political interest, and ideological commitment—are met. In other words, for individuals with equal levels of education, political interest, and strong ideological leanings, women are more likely to participate than their male counterparts. This also sheds light on the low participation rates across Latin America and the Caribbean. Men with the right conditions will only participate if they exhibit exceptionally low machismo levels, while it remains challenging to find highly educated, politically engaged, and markedly ideological women in the region.

While direct metrics for measuring bias impact are difficult to determine, focusing on shifts in machismo thresholds offers valuable insights for policy design. Large-scale reforms—such as curbing corruption, enhancing public information, strengthening the social contract, and pursuing overdue constitutional or fiscal changes—could boost citizen participation. However, our findings open the door for smaller-scale interventions also having a significant impact in boosting participation in community affairs. Measures like making community institutions more accessible, reducing bureaucratic hurdles to participation, and investing in grassroots organizations can increase engagement, especially when combined with other key drivers. Practical steps might include promoting national and community debates, fostering political and civic interest in schools, organizing open days in public institutions, improving police training to ensure protester safety, reducing signature requirements for petitions, and providing technical support to neighborhood associations, among others.

Small-scale interventions align with Banerjee and Duflo's (2011) concept of "small radical thinking" which has proven effective in other areas, such as poverty reduction. However, our findings also reaffirm a longstanding principle: silver bullets do not work. The strong predictive path for participation, after accounting for societal, data, and algorithmic biases, highlights that strategies focused on a single factor, like education or political interest, may not succeed given the complex interplay of conditions necessary for participation. This reinforces the idea that integrated, rather than isolated, interventions are needed to meaningfully increase civic participation.

Finally, effective increases in participation require not only integrated policies, but also integrated knowledge. Future research should include the analysis of emerging forms of participation, such as e-participation—participating in online surveys, using government digital transparency platforms, or adding political banners to profile pictures—in addition to traditional face-to-face participation. Additionally, research should continue to examine multiple biases simultaneously, enhancing the flexibility of predictive models to capture nonlinearities and the complexities of these participation decisions.

References

Ali, M., P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke. 2019. "Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes." Proceedings of the ACM on Human-Computer Interaction 3: 1–30.

Baeza-Yates, R. 2018. "Bias on the Web." Communications of the ACM 61(6): 54–64.

Barnabó, G., A. Fazzone, S. Leonardi, and C. Schwiegelshohn. 2019. "Algorithms for Fair Team Formation in Online Labour Marketplaces." In Companion Proceedings of the 2019 World Wide Web Conference, 484–490. New York, NY: Association for Computing Machinery.

Beauvoir, S. de. 1949. The Second Sex. London: Jonathan Cape.

Bender, E. M., and B. Friedman. 2018. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." Transactions of the Association for Computational Linguistics 6: 587–604.

Bem, S. L. 1993. The Lenses of Gender: Transforming the Debate on Sexual Inequality. Yale University Press

Blodgett, S. L., S. Barocas, H. Daumé III, and H. Wallach. 2020. "Language (Technology) Is Power: A Critical Survey of 'Bias' in NLP." arXiv:2005.14050.

Booth, B. M., L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello. 2021. "Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews." In Proceedings of the 2021 International Conference on Multimodal Interaction, 268–277.

Correll, S. , S. Benard, and I. Paik. 2007**.** Getting a job: Is there a motherhood penalty? *American Journal of Sociology*, 112(5), 1297–1338

Cornwall, A. 2003. Whose Voices? Whose Choices? Reflections on Gender and Participatory Development, World Development, 31(8): 1325–1342

Cowgill, B., and C. E. Tucker. 2020. "Algorithmic Fairness and Economics." Columbia Business School Research Paper. SSRN: https://ssrn.com/abstract=3361280 or http://dx.doi.org/10.2139/ssrn.3361280.

Cramer, H., J. Garcia-Gathright, S. Reddy, A. Springer, and R. Takeo Bouyer. 2019. "Translation, Tracks & Data: An Algorithmic Bias Effort in Practice." In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19, 1–8. New York, NY: Association for Computing Machinery.

Das, A., A. Dantcheva, and F. Bremond. 2018. "Mitigating Bias in Gender, Age, and Ethnicity Classification: A Multi-task Convolution Neural Network Approach." In Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

Dhar, P., J. Gleason, H. Souri, C. D. Castillo, and R. Chellappa. 2020. "Towards Gender-Neutral Face Descriptors for Mitigating Bias in Face Recognition." arXiv:2006.07845.

Donnelly, N., and L. Stapleton. 2021. "Digital Enterprise Technologies: Do Enterprise Control and Automation Technologies Reinforce Gender Biases and Marginalisation?" IFAC Papers Online 54: 551–556.

Eagly, A., and S. Karau, S. J. 2002. Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3), 573–598.

Eubanks, V. 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

Fajardo-Heyward, P., and J. Cuesta. 2023. Assessing the success of national human rights action plans through a political economy lens: The case of Chile. Policy Research Working Paper No. 10578. World Bank.

Feldman, T., and A. Peake. 2021. End-to-end bias mitigation: Removing gender bias in deep learning. arXiv:2104.02532.

Friedman, B., and H. Nissenbaum. 1996. Bias in computer systems. ACM Transactions on Information Systems 14: 330–347.

Garcia-Gathright, J., A. Springer, and H. Cramer. 2018. Assessing and addressing algorithmic bias - But before we get there. arXiv:1809.03332.

Gebru, T. 2020. Race and gender. In The Oxford Handbook of Ethics of AI, edited by M. D. Dubber, F. Pasquale, and S. Das. Online edition. Oxford Academic Press.

Glymour, B., and J. Herington. 2019. Measuring the biases that matter: The ethical and causal foundations for measures of fairness in algorithms. In Proceedings of the Conference on Fairness, Accountability, and Transparency, 269–278. New York, NY: Association for Computing Machinery.

Gillespie, T. 2012. Can an algorithm be wrong? Limn 2. : http://limn.it/can-an-algorithm-be-wrong/.

Gupta, M., C. Parra, and D. Dennehy. 2022. Questioning racial and gender bias in AI-based recommendations: Do espoused national cultural values matter? Information Systems Frontiers 24: 1465–1481.

Guttman, L. 1954. Some necessary conditions for common-factor analysis. *Psychometrika*, 19, 149 –161. http://dx.doi.org/10.1007/BF02289162

Haenlein, M., and A. Kaplan. 2019. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. California Management Review 61(4): 5–14.

Hicks, M. 2019. Hacking the Cis-TEM. IEEE Annals of the History of Computing 41: 20–33.

Hong, J., Z. Zhu, S. Yu, Z. Wang, H. H. Dodge, and J. Zhou. 2021. Federated adversarial debiasing for fair and transferable representations. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 617–627. New York, NY: Association for Computing Machinery.

Karimi-Haghighi, M., and C. Castillo. 2021. Enhancing a recidivism prediction tool with machine learning: Effectiveness and algorithmic fairness. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, 210–214.

Kreps, S., and D. L. Kriner. 2023. The potential impact of emerging technologies on democratic representation: Evidence from a field experiment. New Media & Society. https://doi.org/10.1177/14614448231160526.

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. An Introduction to Statistical Learning with Applications in R. Springer Science+Business Media. DOI 10.1007/978-1-4614-7138-7

Jolliffe, I.T. 2002. Principal Component Analysis. 2nd Edition, Springer-Verlag. https://doi.org/10.1007/b98835

Kabeer, N. 2005. Gender equality and women's empowerment: A critical analysis of the third millennium development goal. *Gender and Development*, 13(1), 13–24.

Kaiser, H. F. 1960. The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141–151. http://dx.doi.org/10.1177/001316446002000116

Ketchen, D. J., and C.L. Shook. 1996. The application of cluster analysis in strategic management research: An analysis and critique. Strategic Management Journal, 17(6), 441–458.

Kleven, H., C. Landais, J. Posch, A. Steinhauer, and J. Zweimüller. 2019. Child penalties across countries: Evidence and explanations. *American Economic Association Papers and Proceedings*, 109: 122-126

Latinobarometro Corporation (2023). Latinobarómetro 2023 [dataset]. Retrieved from https://www.latinobarometro.org/latContents.jsp

Lee, N. T., P. Resnick, and G. Barton. 2019. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings Research. Washington, DC: Brookings Institution.

Liu, H., W. Wang, Y. Wang, H. Liu, Z. Liu, and J. Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. arXiv:2009.13028.

Maudslay, R. H., H. Gonen, R. Cotterell, and S. Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. arXiv:1909.00871.

Morales, A., J. Fierrez, R. Vera-Rodriguez, and R. Tolosana. 2020. Sensitiveness: Learning agnostic representations with application to face images. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 2158–2164.

Nissenbaum, H., and L. D. Introna. 2000. Shaping the web: Why the politics of search engines matters. The Information Society 16(3), 169–185.

Panditharatne, M., D. Weiner, and D. Kriner. 2023. Artificial intelligence, participatory democracy, and responsive government. Expert Brief. Brennan Center for Justice. New York.

Pariser, E. 2011. The filter bubble: What the internet is hiding from you. London: Penguin Press.

Pecorari, N. and J. Cuesta. 2024 Citizen Participation and Political Trust in Latin America and the Caribbean: A Machine Learning Approach, European Journal of Development Research, 36:1227–1252.

Rahim, A., C. Mahony, and S. Bandyopadhyay. 2024. Generative artificial intelligence as an enabler for citizen engagement, governance for development. World Bank Blogs, February 12, 2024.

Ridgeway, C., and S Correll. 2004. Unpacking the gender system: A theoretical perspective on gender beliefs and social relations. *Gender and Society*, 18(4), 510–531.

Sarraf, D., V. Vasiliu, B. Imberman, and B. Lindeman. 2021. Use of artificial intelligence for gender bias analysis in letters of recommendation for general surgery residency candidates. American Journal of Surgery 222, 1051–1059.

Shrestha, S., and S. Das. 2022. Exploring gender biases in ML and AI academic research through systematic literature review. Frontiers in Artificial Intelligence 5:976838.

Singh, V. K., and C. Hofenbitzer. 2019. Fairness across network positions in cyberbullying detection algorithms. In 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 557–559.

Smith, G., and I. Rustagi. 2021. "When Good Algorithms Go Sexist: Why and How to Advance AI

Thorndike, R. L. 1953. Who belongs to the family? Psychometrika, 18(4), 267–276.

UN Women. 2024. "Artificial Intelligence and Gender Equality." Explainer. 22 May 2024. UN Women: Geneva.
Van Couvering, E. 2007. "Is Relevance Relevant? Market, Science, and War: Discourses of Search Engine Quality." Journal of Computer-Mediated Communication 12(3), 866.

Vlasceanu, M., and D. Amodo. 2022. "Propagation of Societal Gender Inequality by Internet Search Algorithms." Proceedings of the National Academy of Sciences of the United States of America, Psychological Cognitive Sciences, Vol. 119(29) e2204529119.

Wang, C., K. Wang, A. Bian, R. Islam, K. N. Keya, J. Foulde, et al. 2021. "Bias: Friend or Foe? User Acceptance of Gender Stereotypes in Automated Career Recommendations." UMBC Student Collection.

Wang, T., X. V. Lin, N. F. Rajani, B. McCann, V. Ordonez, and C. Xiong. 2020. "Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation." arXiv preprint arXiv:2005.00965. doi: 10.18653/v1/2020.acl-main.484.

Walcott, R. 2019. "The End of Diversity." Public Choice 31(2): 393–408.

Wu, W., P. Protopapas, Z. Yang, and P. Michalatos. 2020. "Gender Classification and Bias Mitigation in Facial Images." In 12th ACM Conference on Web Science, 106–114.

Young, E., J. Wajcman, and L. Sprejer. 2021. "Where Are the Women? Mapping the Gender Job Gap in AI." Policy Briefing: Full Report. The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2021-03/where-are-the-women_public-policy_full-report.pdf.

Zhao, J., T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. 2017. "Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints." arXiv:1707.09457.:1707.09457.

Appendix 1

Decision trees for alternative correcting models

Figure A1.1. Model 1: No Gender Variable (Gender Blinding)

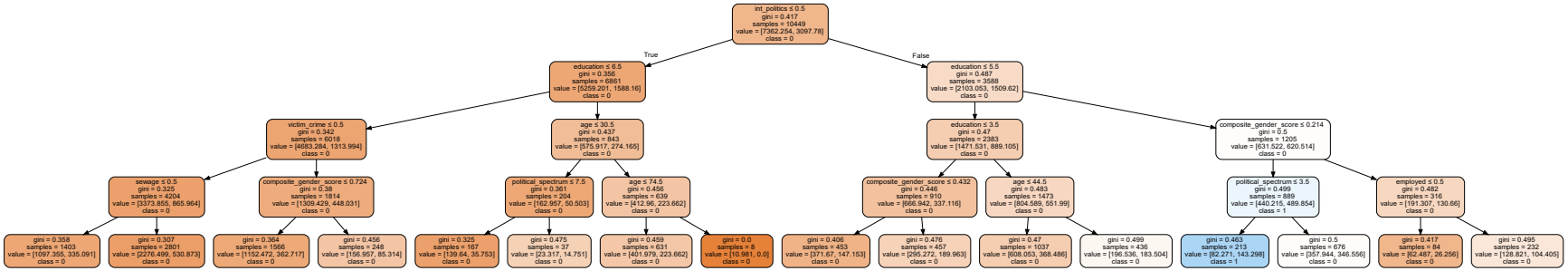


Figure A1. 2. Model 2: No Gender Variable Nor Gender Proxies (Gender Deproxing)

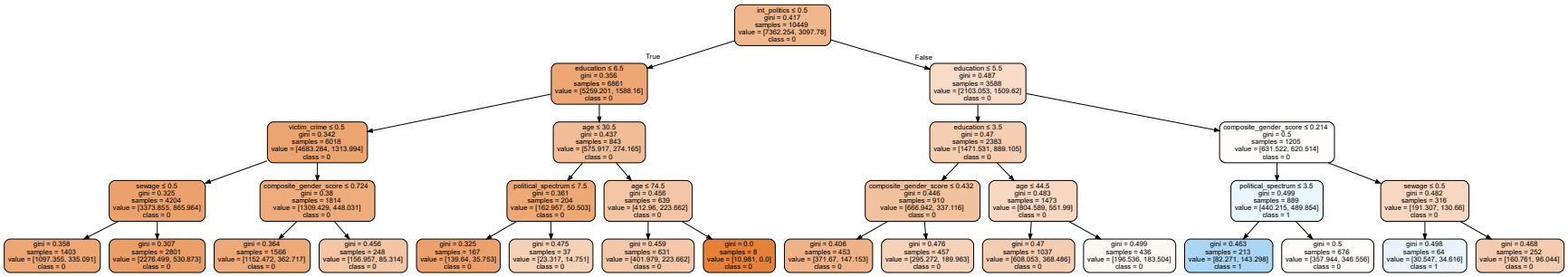
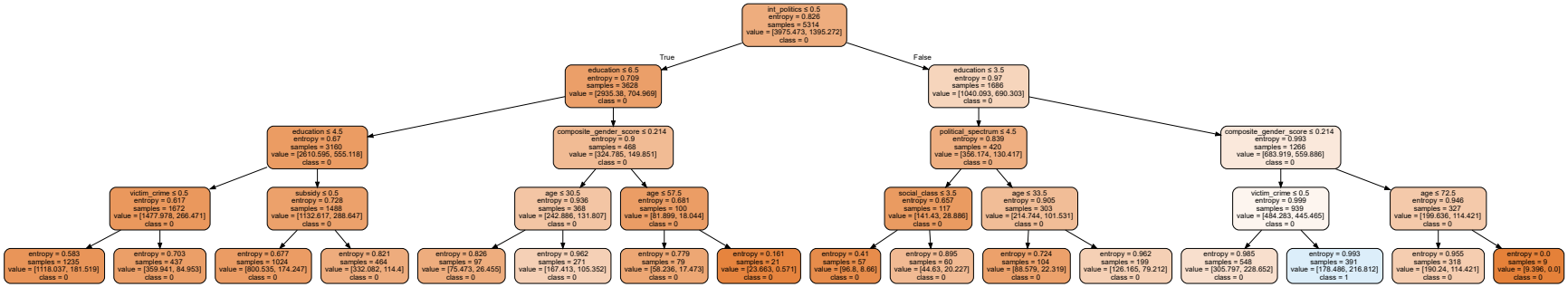


Figure A1.3. Model 3: Women and Men-only

Only women



Only men

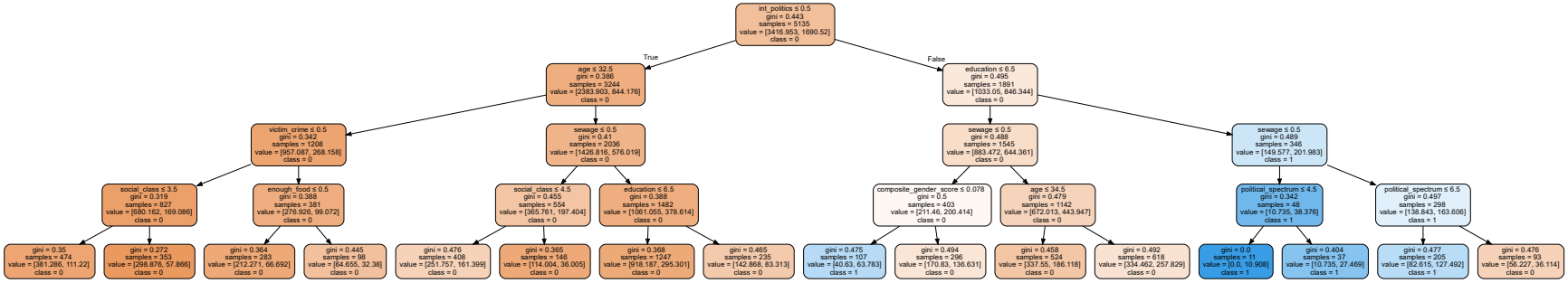


Figure A1.4. Model 4: Gender Switching

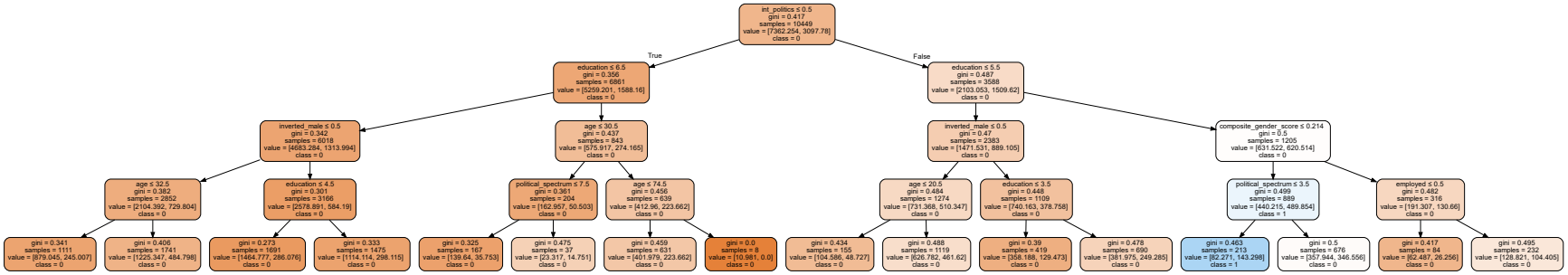


Figure A1.5. Model 5: Scaling Machismo

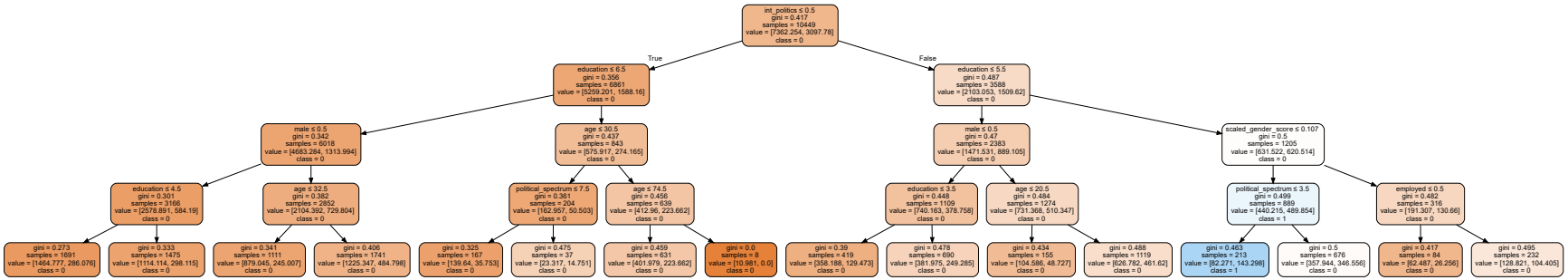
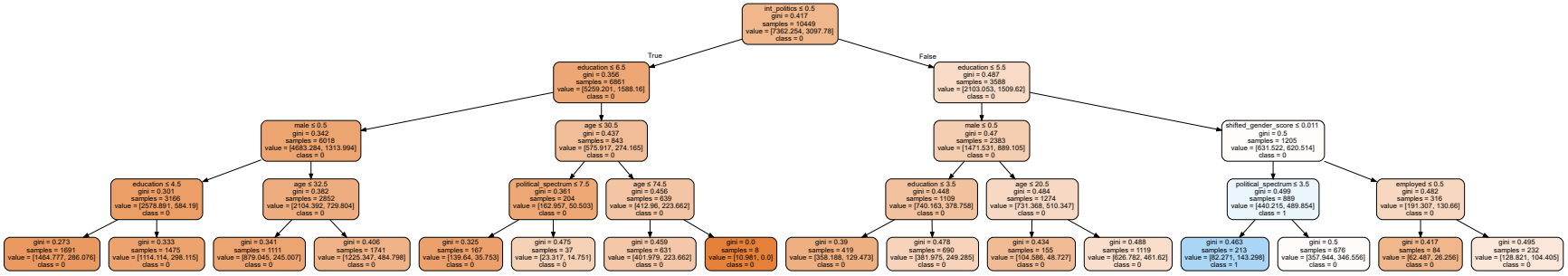


Figure A1.6. Model 6: Shifting Machismo Distribution



Source Author's elaboration using the function DecisionTreeClassifer in the Scikit-Learn 1.2.2 library in Python 3.11.

Note: The color palette of the nodes indicates the class to which the majority of the samples at each node belong (blue captures class 1 while orange captures class 0). The Gini index measures the impurity or disorder in a node. Samples refer to the number of observations that are classified in each node. Value tells how many observations at the given node fall into each category or class. In our case, we have two classes: working for a community problem or not.

Appendix 2

Robustness checks with alternative target variables

Figure A2.1. Signing a petition

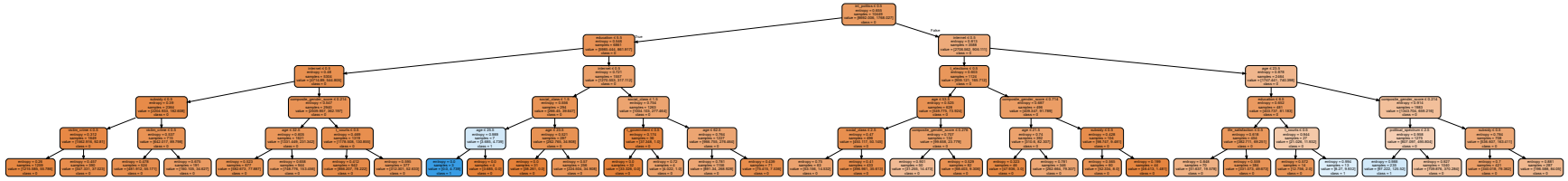
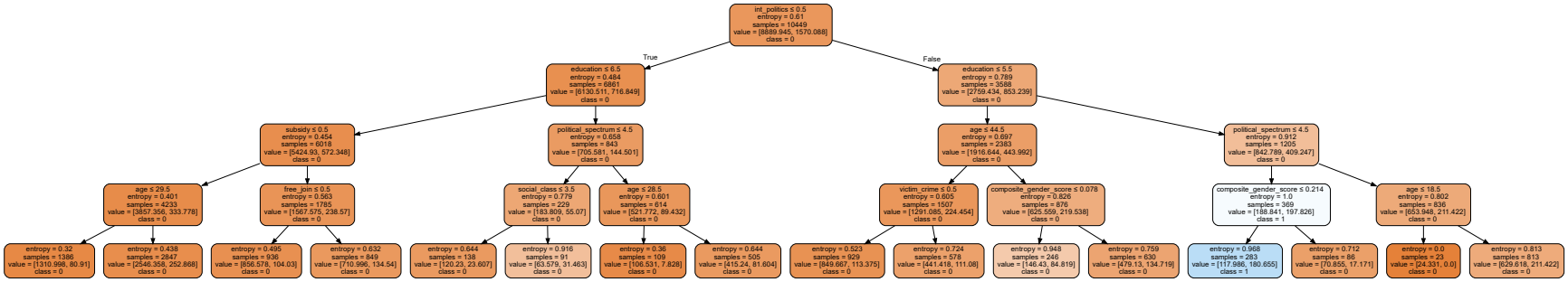


Figure A2.2. Attending demonstrations



Robustness checks using alternative configurations of the gender bias

Figure A2.3. Robustness check using gender bias variables separately instead of composite gender score

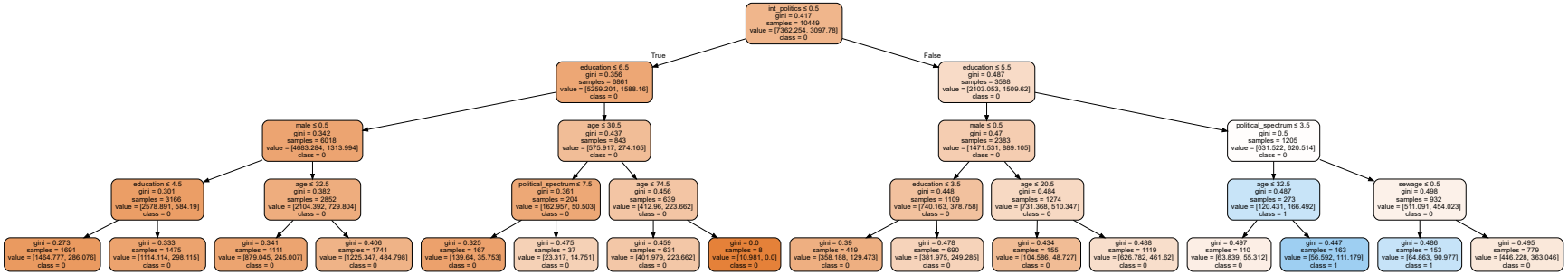
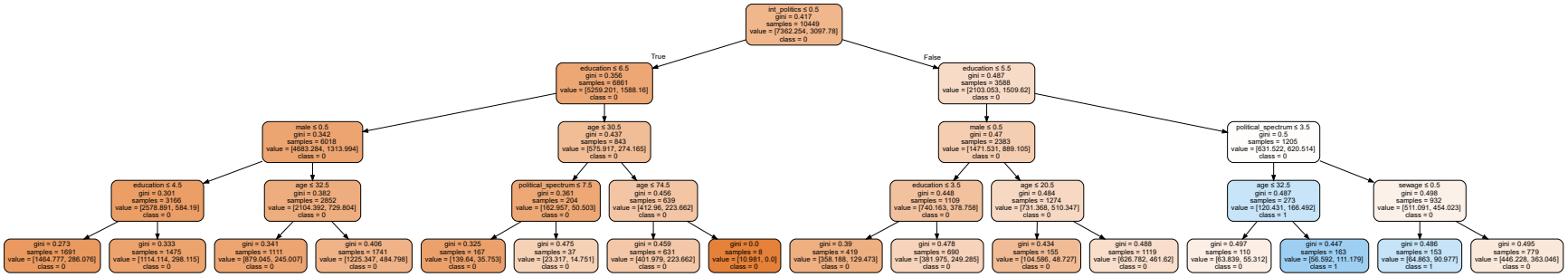


Figure A2.4. Robustness check using an indicator variable for high machismo levels (instead of a composite score)



Source Author's elaboration using the function DecisionTreeClassifer in the Scikit-Learn 1.2.2 library in Python 3.11.
Note: The color palette of the nodes indicates the class to which the majority of the samples at each node belong (blue captures class 1 while orange captures class 0). The Gini index measures the impurity or disorder in a node. Samples refer to the number of observations that are classified in each node. Value tells how many observations at the given node fall into each category or class. In our case, we have two classes: working for a community problem or not.