Bias in Adult Income Dataset

Submitted by: Group 2

Course: DATA-6550

1. Introduction

Artificial intelligence and machine learning systems are increasingly deployed in high-stakes decision-making contexts that directly affect people's lives, from automated resume screening and credit scoring to healthcare risk assessment and criminal justice predictions. However, mounting evidence demonstrates that these systems frequently perpetuate and amplify existing societal biases. In 2018, Amazon disbanded an AI recruiting tool that systematically downgraded women's resumes. In 2019, Apple Card faced regulatory scrutiny when its algorithm offered women lower credit limits than men with similar financial profiles. These failures share a common root cause: machine learning models learn patterns from historical data, and when that data reflects historical discrimination through sampling bias, historical inequalities, underrepresentation of minorities, or measurement proxies the resulting models encode and automate that discrimination at scale.

At the heart of algorithmic bias lies training data. This project investigates these challenges through detailed analysis of the Adult Income dataset taken from UCI Machine Learning Repository is one of the most widely used datasets in machine learning research and education.

Through comparative analysis of four classification models (Logistic Regression, Decision Tree, Random Forest, SVM), we quantify bias across multiple fairness metrics, evaluate accuracy-fairness trade-offs, and propose concrete mitigation strategies at data, model, and process levels. Our findings demonstrate that the dataset exhibits severe gender imbalance (67.5% male), racial imbalance (86% White), and corresponding income disparities (males 2.7x more likely to earn >$50K), which all tested models learn and perpetuate to varying degrees, with demographic parity violations ranging from 2.4% (SVM) to 19.2% (Decision Tree). The primary objectives of this analysis are:

1. Identify and quantify known biases in the dataset, particularly gender and racial bias in income distribution

2. Analyze demographic distributions to uncover potential sampling and representation biases

3. Evaluate how these biases affect machine learning model performance across multiple algorithms

4. Compare fairness metrics across different models and propose mitigation strategies

2. Dataset

The Adult Income dataset extracted from the 1994 U.S. Census Bureau by researchers Ronny Kohavi and Barry Becker, this dataset contains 48,842 individual records with 14 demographic and employment features (age, education, occupation, race, gender, hours worked) used to predict whether annual income exceeds $50,000. We selected this dataset because it has been used in hundreds of academic papers and countless educational courses, making it representative of how bias propagates through the ML ecosystem. More critically, it contains explicit protected attributes (gender, race) from a known historical period (1994), allowing us to directly measure how historical discrimination manifests in data and propagates through modern algorithms. Moreover, the analysis of the dataset's demographic composition reveals significant imbalances in terms of gender and racial distribution:

Gender Distribution: Males constitute 67.5% of the dataset (32,975 records) while females represent only 32.5% (15,867 records). This 2:1 ratio indicates substantial sampling bias that may affect model generalization.

Racial Composition: The dataset is overwhelmingly White (86.0%), with Black individuals representing 9.3%, Asian-Pacific Islander 2.9%, American-Indian-Eskimo 1.0%, and Other 0.8%. This severe imbalance limits the model's ability to learn patterns for minority groups and may lead to disparate performance.

Furthermore, the income distribution across demographic groups reveals stark disparities based on gender and race which are discussed below:

By Gender: Males have a 31.2% rate of earning above $50K compared to females at 11.4%, representing a 2.7x disparity. This gap reflects historical wage inequality and occupational segregation patterns from the 1994 labor market.

By Race: Asian-Pacific Islander individuals show the highest high-income rate (28.3%), followed by White (26.2%), Other (12.7%), Black (12.6%), and American-Indian-Eskimo (12.2%). These disparities reflect complex historical factors including immigration patterns, educational access, and systemic discrimination.

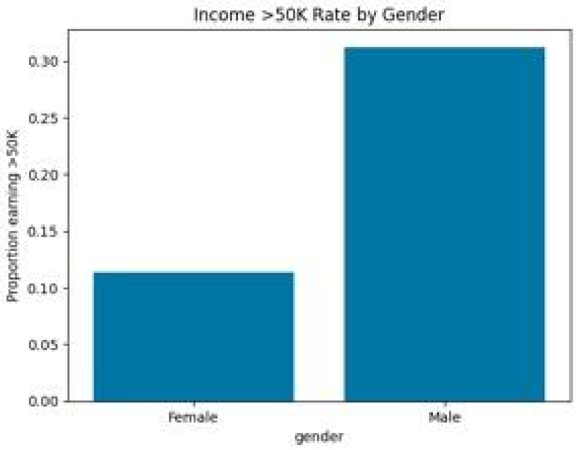

Figure 1: Income >$50K Rate by Gender

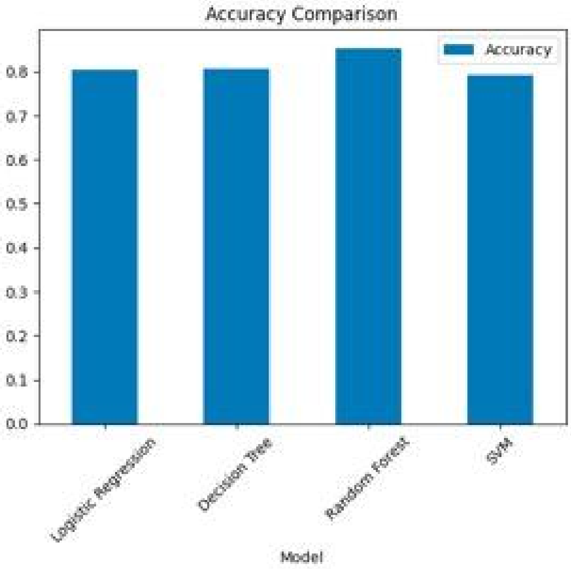

Figure 2: Income >$50K Rate by Race

3. Methodology

Our analysis began with comprehensive data preprocessing steps. Missing values, indicated by '?' in the original dataset, were handled through examination of their distribution across features. Categorical variables were encoded using label encoding to enable machine learning algorithms to process the data. Attention was paid to the binary encoding of gender (0 for Female, 1 for Male) and the income target variable (0 for ≤$50K, 1 for >$50K). The dataset was well-structured with minimal missing data, though the presence of missing occupation and work class information was noted for approximately 7% of records.

We conducted extensive exploratory data analysis to understand the distribution of features and identify potential biases. This included analyzing income distributions across gender, race, occupation, and education levels using visualization techniques and statistical summaries. Cross-tabulations between protected attributes (gender, race) and income outcomes revealed substantial disparities that warranted deeper investigation. The stacked bar chart analysis clearly demonstrated the disproportionate concentration of females in the low-income category.

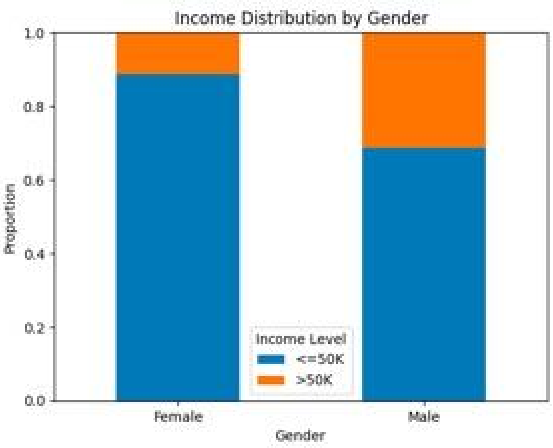

Figure 3: Income Distribution by Gender

We implemented and compared four classification algorithms to evaluate both predictive performance and fairness characteristics:

Logistic Regression: A linear model providing interpretable coefficients, trained with LBFGS solver and maximum 1000 iterations.

Decision Tree: A non-linear model using recursive partitioning, offering high interpretability through rule-based decisions.

Random Forest: An ensemble method combining multiple decision trees to improve generalization and reduce overfitting.

Support Vector Machine (SVM): A maximum-margin classifier seeking the optimal decision boundary in feature space.

The dataset was split into training (80%) and testing (20%) sets using stratified sampling to maintain class proportions. All models were implemented using scikit-learn with default hyperparameters to establish baseline performance before optimization.

To evaluate bias beyond simple accuracy, we computed two key fairness metrics:

Demographic Parity Difference: Measures the difference in positive prediction rates between males and females. A value of 0 indicates perfect demographic parity, while larger absolute values indicate greater disparate impact. This metric assesses whether the model's predictions are equally distributed across protected groups.

True Positive Rate (TPR) Difference: Also known as Equal Opportunity difference, this measures the gap in true positive rates between males and females. It assesses whether individuals who truly earn >$50K are equally likely to be correctly identified regardless of gender. Lower values indicate fairer treatment of qualified individuals across groups.

4. Analysis and Results

All four models achieved competitive accuracy scores, with Random Forest performing best overall. However, accuracy alone masks significant fairness concerns:

Model	Accuracy	Demographic Parity Diff	TPR Diff
Logistic Regression	80.4%	0.119	0.181
Decision Tree	80.6%	0.192	0.060
Random Forest	**85.3%**	0.178	0.049
SVM	79.3%	**0.024**	**0.047**

Table 1: Model Performance and Fairness Metrics

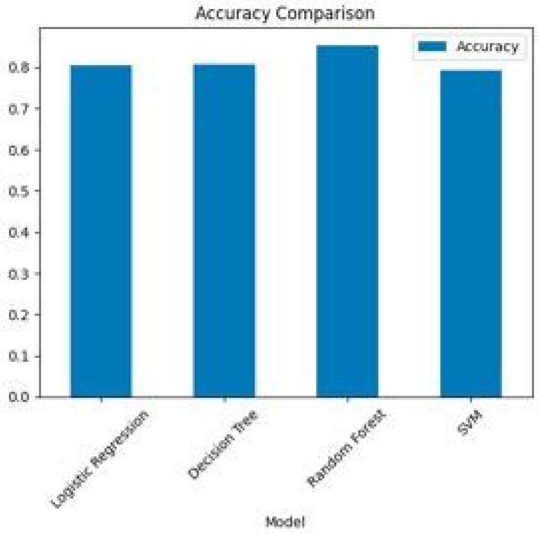

Figure 4: Model Accuracy Comparison

The fairness analysis reveals a critical trade-off between accuracy and fairness. Random Forest achieves the highest accuracy (85.3%) but exhibits substantial demographic parity violation (0.178), meaning males are 17.8 percentage points more likely to be predicted as high earners than females. In contrast, SVM demonstrates the best fairness performance with only 2.4% demographic parity difference, though at the cost of lower accuracy (79.3%).

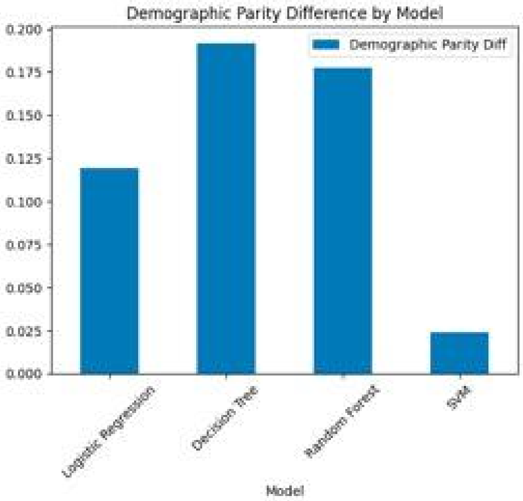

Figure 5: Demographic Parity Difference by Model

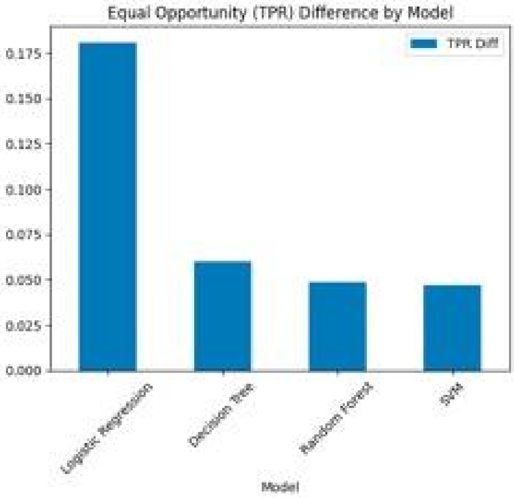

Figure 6: Equal Opportunity (TPR) Difference by Model

Decision Tree shows the worst demographic parity (0.192), suggesting it learns and amplifies the gender bias most strongly. However, it performs well on Equal Opportunity (TPR difference of 0.060), indicating that among true high earners, it treats males and females relatively equally. This demonstrates that different fairness metrics can tell different stories about model behavior.

The gender bias analysis confirms that all models systematically favor males in income predictions. Even SVM, the fairest model, exhibits measurable bias. This stems from the underlying data distribution where males have substantially higher representation in high-income occupations and educational levels. The models learn these patterns and apply them predictively, effectively encoding historical discrimination into automated decisions.

The TPR difference analysis reveals that Logistic Regression shows the largest Equal Opportunity violation (0.181), meaning that among individuals who truly earn >$50K, males are 18.1 percentage points more likely to be correctly identified than females. This type of bias is particularly concerning in applications like hiring or promotion decisions, where it means qualified female candidates are more likely to be overlooked.

The dataset exhibits clear patterns of occupational segregation by gender. Males are overrepresented in higher-paying technical and managerial roles (Exec-managerial, Prof-specialty, Craft-repair), while females are concentrated in lower-paying service and administrative positions (Adm-clerical, Other-service, Sales). This occupational distribution directly contributes to the income disparity observed and represents a form of structural bias embedded in the labor market data. When machine learning models learn from this data, they inadvertently encode these historical patterns of discrimination, potentially perpetuating them in future predictions.

The severe racial imbalance in the dataset (86% White) creates multiple problems. First, the models have insufficient data to learn accurate patterns for minority groups, leading to higher error rates for these populations. Second, the intersection of race and gender creates compound disadvantage minority women face the most severe underrepresentation and likely experience the largest prediction errors. Third, the high-income rates show Asian-Pacific Islander individuals at 28.3% and White at 26.2%, substantially higher than Black (12.6%) and American-Indian-Eskimo (12.2%), reflecting historical economic inequalities that models will learn and potentially amplify.

5. Discussion

Our analysis reveals fundamental challenges in building fair machine learning systems. We observed a clear accuracy-fairness trade-off: Random Forest achieved the highest accuracy (85.3%) but with substantial bias (17.8% demographic parity difference), while SVM demonstrated better fairness (2.4% demographic parity difference) at the cost of lower accuracy (79.3%). This tension requires context-dependent decisions high-stakes domains like lending or hiring may justify accepting lower accuracy for fairer outcomes. Moreover, fairness itself is not singular: Decision Tree showed poor demographic parity (19.2%) yet good equal opportunity (6.0% TPR difference), demonstrating that different fairness metrics can conflict. Data scientists must select criteria aligned with their application equal opportunity for employment contexts, demographic parity for resource allocation.

Deployment of models trained on this data carries serious risks. Even when gender is removed as a feature, occupation, education, and hours worked serve as proxies, enabling systematic discrimination. The temporal mismatch between 1994 data collection and current deployment is equally problematic women labor force participation and occupational distribution have changed substantially over three decades, meaning models trained on historical data risk perpetuating outdated discrimination and violating equal opportunity laws. Ethically, including protected attributes enables direct discrimination, but removing them leads to "fairness through unawareness" where bias persists through proxies. Data scientists must critically examine whether certain prediction tasks should be automated at all, whether available data is appropriate for intended use, and what safeguards prevent harm. The Adult Income dataset's widespread use in ML research has likely led to deployment of biased models in real systems, underscoring that our responsibility extends beyond accuracy to ensuring algorithms do not perpetuate historical injustices.

6. Bias Mitigation Strategies

Addressing bias requires interventions at multiple levels. At the **data level**, we can balance the dataset through resampling and reweighting underrepresented groups (females, racial minorities), supplement 1994 census data with contemporary sources like American Community Survey data to reflect current demographics, and engineer features that capture legitimate socioeconomic predictors while reducing proxy effects (e.g., education-occupation combinations rather than raw occupation categories).

At the **model level**, we can implement fairness constraints during training using libraries like Fairlearn or AIF360 to enforce demographic parity or equalized odds our results suggest SVM-like fairness could be achieved while maintaining Random Forest-like accuracy with appropriate constraints. Adversarial debiasing trains models with components that penalize predictions correlated with protected attributes, forcing the main model to produce predictions that don't reveal gender or race. Threshold optimization applies different decision thresholds for different demographic groups post-hoc, maintaining overall accuracy while improving fairness without retraining.

At the **process level**, establish regular bias audits measuring model performance across demographic subgroups using multiple fairness metrics (demographic parity, equal opportunity, calibration), with dashboards tracking these metrics over time. Involve representatives from affected communities through focus groups to ensure fairness definitions align with lived experiences and identify harms that technical metrics might miss. Document dataset limitations, known biases, and model behavior through model cards and datasheets, providing clear explanations of fairness trade-offs to enable informed deployment decisions.

We recommend a **hybrid approach**: (1) start with data augmentation using recent, balanced datasets to reduce historical and sampling bias at the source, (2) apply fairness constraints during training targeting demographic parity ≤ 5% and TPR difference ≤ 5%, (3) use threshold optimization to fine-tune fairness without sacrificing excessive accuracy, and (4) implement ongoing quarterly fairness auditing with immediate intervention if metrics deteriorate. This multi-layered strategy addresses bias at data, model, and process levels, maximizing the likelihood of achieving both accurate and fair predictions while maintaining accountability.

7. Conclusion

This analysis of the Adult Income dataset using four machine learning algorithms revealed substantial gender and racial bias across all models. Males are predicted to earn above $50,000 at rates 2.7 times higher than females, reflecting severe dataset imbalances (67.5% male, 86% White). Our evaluation exposed a fundamental accuracy-fairness trade-off: Random Forest achieved highest accuracy (85.3%) with significant bias (17.8% demographic parity difference), while SVM demonstrated best fairness (2.4%) at lower accuracy (79.3%). Critically, fairness is multifaceted Decision Tree showed poor demographic parity (19.2%) yet good equal opportunity (6.0% TPR difference) requiring context-aligned metric selection guided by ethical and legal requirements.

We propose actionable mitigation strategies: data augmentation, fairness-constrained learning, threshold optimization, and ongoing auditing. However, technical solutions alone are insufficient. Meaningful progress requires organizational commitment to fairness, stakeholder engagement, transparency about trade-offs, and willingness to forgo deployment when fairness cannot be assured. As AI systems increasingly influence employment, credit, and housing decisions, identifying and mitigating bias becomes a fundamental ethical obligation. The data we use carries the imprint of history we must ensure our algorithms build a more equitable future rather than reinforcing an unjust past.