Fairness Analysis via Different ML Models on the UCI's Adult Dataset

Utku Acar
Ozyegin University
utku.acar@ozu.edu.tr

***Abstract*—The "Adult" dataset from UCI contains vital demographic information—age, education, occupation, race, gender, and income. Utilizing this data for algorithmic decision-making in employment may inadvertently perpetuate historical biases and lead to potential discrimination. Sensitive attributes like race, gender, and marital status within the dataset may perpetuate biased patterns, erroneously linking specific demographics with higher income levels.**

Instead of finding the 'biased' features manually, we can also make the models find the biased features themselves by iterating through the features while making binary classification. Biased models risk producing discriminatory outcomes, exacerbating societal disparities, and hindering equitable opportunities. These repercussions impact individual livelihoods and contribute to broader societal inequalities, eroding trust in automated decision systems. Fairness concerns with the dataset extend beyond statistical metrics, encompassing ethical, social, and legal dimensions. Rectifying fairness isn't solely about metric enhancement; it's about ensuring equal opportunities, mitigating biases, and fostering a fairer society. Addressing fairness in this dataset is a call for social responsibility and ethical data utilization in decision-making processes, significantly influencing individuals' lives.

I. Problem Statement

The UCI "Adult" dataset comprises diverse demographic attributes crucial for algorithmic decision-making, notably in employment contexts. However, employing this dataset for automated decision-making systems poses significant challenges related to fairness, bias, and discrimination. Sensitive attributes such as race, gender, and marital status in the data raise concerns about perpetuating biased patterns, potentially perpetuating historical inequalities and yielding unfair outcomes.

The impact of biased models extends beyond statistical measurements, encompassing ethical, societal, and legal ramifications. Biases in algorithmic decision-making systems have the potential to exacerbate societal disparities, deny equitable opportunities, and erode trust in automated processes. A comprehensive approach beyond quantitative metrics is necessary, focusing on ethical use, fairness, and societal impact to rectify biases and promote equitable decision-making.

Addressing fairness issues in the "Adult" dataset demands a holistic perspective that goes beyond numerical metrics to champion fairness, social responsibility, and ethical data practices in decision-making processes, with far-reaching implications for individuals' lives.

II. Introduction

A. Evaluation of Fairness Metrics

Assessing fairness in machine learning models involves measuring various metrics to ensure equitable treatment across different groups within a dataset. This analysis employed several fairness metrics to gauge disparities in model predictions across sensitive attributes.

- **Equalized Odds Difference (EOD)**: Measures the difference in error rates between groups, ensuring equal predictive accuracy. EOD focuses on the discrepancy in predictive performance, highlighting differences in error rates between sensitive groups for each feature. Higher EOD values indicate significant disparities in predictive accuracy across these groups.
- **Predictive Parity (PP)**: Aims to achieve fairness in predictions by ensuring that the predictive accuracy remains equal across different sensitive groups. It emphasizes the equality in true positive and false positive rates across these groups.
- **Statistical Parity (SP)**: Evaluates whether predicted outcomes are consistent across different sensitive groups (e.g., gender or race). It aims for an equal distribution of positive outcomes among these groups, irrespective of other factors.
- **Disparate Impact (DI)**: Measures the difference in outcomes or predictions between different groups. It evaluates whether certain groups receive more favorable outcomes and aims to detect and mitigate biases based on sensitive attributes.
- **True Positive Rate (TPR0)**: TPR0 represents the True Positive Rate for the non-sensitive group, indicating the proportion of correctly predicted positive cases among all actual positive cases within this specific group.
- **False Positive Rate (FPR0)**: FPR0 stands for the False Positive Rate in the non-sensitive group, representing the proportion of actual negative cases incorrectly predicted as positive within this particular group.
- **True Positive Rate (TPR1)**: TPR1 signifies the True Positive Rate for the sensitive group, denoting the proportion of correctly predicted positive cases among all actual positive cases within this specific group.
- **False Positive Rate (FPR1)**: FPR1 denotes the False Positive Rate in the sensitive group, representing the

proportion of actual negative cases incorrectly predicted as positive within this particular group.

1) Interpreting the Metrics

The computed values for metrics including EOD, PP, SP, DI, TPR, and FPR were assessed across various features. These metrics were utilized to evaluate disparities in model predictions concerning gender, education, occupation, and other attributes. Higher values of these metrics suggest potential bias or unfairness in the model's predictions, necessitating attention to mitigate these disparities.

The detailed breakdown of these fairness metrics and their corresponding values for different features is provided in the respective tables and visual representations in the subsequent sections.

III. Tasks

A. Task 1: Discussion on the Significance of Addressing Fairness Issues

The "Adult" dataset from UCI, designed for predicting whether an individual's income exceeds $50K/year based on census data, poses critical challenges that necessitate a thorough examination beyond mere quantitative assessments. This dataset, encompassing diverse demographic attributes like age, education level, occupation, race, gender, and native country, holds immense significance in algorithmic decision-making, particularly within employment contexts. However, its utilization demands a comprehensive approach toward fairness to mitigate potential biases and discriminatory outcomes.

The dataset's multifaceted nature extends beyond statistical figures, delving into ethical, societal, and legal dimensions. The inclusion of sensitive attributes such as race, gender, and marital status raises profound concerns about perpetuating historical biases and exacerbating inequalities if not handled meticulously. Biased patterns within the dataset could lead to unfair associations between specific demographics and income levels, affecting individuals' opportunities and reinforcing societal disparities.

Rectifying fairness issues in this dataset is not merely a quantitative exercise; it's a call for ethical responsibility and societal impact assessment. A failure to address biases within this dataset could not only produce discriminatory outcomes but also erode trust in algorithmic decision-making systems. Ethical considerations, encompassing fairness, equity, and social impact, are pivotal in ensuring that the use of this dataset aligns with principles of fairness and societal well-being.

In essence, treating fairness issues in the "Adult" dataset requires a holistic approach that transcends numerical metrics, focusing on ethical considerations, societal implications, and the promotion of equitable decision-making practices to avoid perpetuating biases and fostering a fairer and more inclusive society.

B. Task 2: Finding the Blackbox Model and Calculating KPIs based on Fairness

We aim to find an effective black-box classification model to predict the binary target variable. Following the model implementation using Python and TensorFlow/Keras, we want to evaluate this model's fairness by calculating KPIs based on its predictions.

The Python code employs several steps:

- Data Preprocessing: Handling missing values, encoding categorical features, and splitting the dataset into training, validation, and test sets.
- Model Definition: A sequential neural network model comprising input, hidden, and output layers using the rectified linear unit (ReLU) and sigmoid activation functions.
- Model Training: The compiled model is trained on the training data for a set number of epochs using the binary cross-entropy loss function and the Adam optimizer, with early stopping and model checkpointing callbacks.
- Model Evaluation: The trained model is evaluated on the test set to compute loss and accuracy metrics, followed by the selection of the best-performing model based on accuracy and loss.

Now, to assess the fairness of this black-box classifier, we need to calculate fairness-related KPIs using its predictions. These fairness metrics can be calculated using the black-box model's predictions on sensitive attributes and comparing the outcomes across different demographic groups.

The process involves assessing and quantifying potential biases in the model's predictions to ensure fairness and equity across diverse groups in the dataset.

C. Task 3: Global Surrogate Models

1) Global Surrogate Model: Logistic Regression

III-C1.1 Creation of the Surrogate Model

A Logistic Regression model is trained using the predictions of the Neural Network model on the validation set as features and the original labels as the target variable. This model serves as a surrogate to interpret the predictions of the black-box model.

III-C1.2 Evaluation of Surrogate Model Performance

The accuracy of the Logistic Regression surrogate model is assessed by comparing its predictions against the actual labels in the test set. Additionally, the performance metrics of fairness, including Equalized Odds Difference, Demographic Parity Difference, and Equalized Odds Ratio, are computed to analyze the model's fairness across different features.

III-C1.3 Interpretation of Model Coefficients

The coefficients obtained from the trained Logistic Regression model are interpreted to understand the importance of different features in predicting the target variable based on the black-box model's predictions.

2) Global Surrogate Model: Decision Tree

III-C2.1 Development of the Surrogate Decision Tree

A Decision Tree classifier is trained using the predictions of the Neural Network model on the validation set as features and the original labels as the target variable. This model aims to provide an interpretable representation of the black-box model's predictions.

III-C2.2 Assessment of Surrogate Decision Tree Performance

The accuracy and fairness metrics of the Decision Tree surrogate model are evaluated by comparing its predictions against the actual labels in the test set. Fairness metrics like Equalized Odds Difference, Demographic Parity Difference, and Equalized Odds Ratio are computed to gauge the model's fairness concerning various features.

III-C2.3 Visualization of the Surrogate Decision Tree

If feasible, a visualization of the trained Decision Tree model may be included to depict the hierarchy and decision rules learned by the surrogate model for interpretability.

D. Task 4: Comparison of Fairness Metrics between Black-Box and Surrogate Models

1) Fairness Evaluation of Logistic Regression Surrogate Model

The logistic regression and decision tree surrogate model were built to interpret the predictions of the black-box neural network model. Various fairness metrics were calculated to assess the disparities and biases present in the surrogate model's predictions. The fairness metrics include Equalized Odds Difference (EOD), Predictive Parity (PP), Statistical Parity (SP), Disparate Impact (DI), True Positive Rate (TPR), and False Positive Rate (FPR) for different sensitive features.

The results of the fairness evaluation showcased significant disparities in certain features, highlighted by high EOD values, indicating potential bias in the surrogate model's predictions across groups associated with these features. The comparison was made against the original black-box model to understand how the surrogate model differed in terms of fairness metrics.

2) Comparison of Fairness Metrics

A detailed comparison was conducted between the fairness metrics obtained from the black-box neural network model and its logistic regression and decision tree surrogate model. The differences in EOD, PP, SP, Disparate Impact, TPR, and FPR for various features were analyzed to understand the extent of the differences in predictive performance across sensitive groups. This comparison aimed to elucidate whether the surrogate model mitigated or exacerbated fairness issues present in the original black-box model.

E. Task 5: Comparison of Interpretable Models with Surrogate Models

1) Task 5.1: Comparison of Interpretable Logistic Regression Model with Surrogate Model

In this task, we aim to compare the performance and fairness metrics of an interpretable logistic regression model against its surrogate counterpart.

Approach:

- Trained an interpretable logistic regression model on the preprocessed dataset.
- Evaluated the model's predictions on the test set.
- Calculated fairness metrics including Equalized Odds Difference, Predictive Parity, Statistical Parity, Disparate Impact, True Positive Rate (TPR), and False Positive Rate (FPR).

2) Task 5.2: Comparison of Interpretable Decision Tree Model with Surrogate Model

This task involves comparing the performance and fairness metrics of an interpretable decision tree model against its surrogate counterpart.

Approach:

- Trained an interpretable decision tree model on the preprocessed dataset with a specified maximum depth.
- Evaluated the model's predictions on the test set.
- Calculated fairness metrics such as Equalized Odds Difference, Predictive Parity, Statistical Parity, Disparate Impact, True Positive Rate (TPR), and False Positive Rate (FPR).

The comparison between interpretable models (logistic regression and decision tree) and their surrogate counterparts provides insights into the trade-offs between model transparency, accuracy, and fairness across different demographic features.

IV. NOTE:

You can find more detailed tables and figures about the data and the model representations in the Jupyter Notebook named "DS530_V_Project_Utku_Acar.ipynb"

V. RESULTS

A. Task 2: Black-Box Model Training and Evaluation

The neural network model was trained and evaluated on the dataset, with the following results:

- **Run 1:**
  - Test loss: 0.3359006643295288
  - Test accuracy: 0.8313066363334656
- **Run 2:**
  - Test loss: 0.3357697129249573
  - Test accuracy: 0.8363917469978333
- No significant improvement from the previous run was observed. The training stopped at iteration 2.

This accuracy which is 83 percent enough for being a successful model with a significantly low loss value which is 0.33 and you can see the KPI results for every feature in a dataset sorted with decreasing EOD values Table I below.

The top five elements in Table I, including individuals from native countries like *Iran*, *Thailand*, and those with higher education levels such as *Prof-school* graduates, *Doctorates*, and *Masters*, exhibit significantly higher *EOD* values, reflecting a stronger predictive performance. In contrast, the bottom five elements, representing workclasses like *Private*, specific native countries like the *Philippines* and the *United States*, and a distinct race category, *Asian-Pac-Islanders*, display notably lower *EOD* values, suggesting comparatively weaker predictive outcomes in this context.

	Feature	EOD	PP	SP	DI	TPR0	FPR0	TPR1	FPR1
0	native-country_Iran	0.702	0.404	0.169	5.925	0.298	0.056	1.0	0.333
1	native-country_Thailand	0.702	0.103	0.0	0.0	0.298	0.057	1.0	0.0
2	education_Prof-school	0.672	0.656	0.066	15.136	0.279	0.048	0.774	0.72
3	education_Doctorate	0.446	0.483	0.108	9.239	0.287	0.054	0.63	0.5
4	education_Masters	0.409	0.372	0.152	6.565	0.235	0.046	0.644	0.3
64	workclass_Private	0.047	0.064	0.486	0.486	0.321	0.092	0.287	0.045
65	native-country_Philippines	0.035	0.037	0.842	0.842	0.299	0.057	0.333	0.048
66	education_Assoc-voc	0.034	0.014	0.686	1.457	0.301	0.055	0.267	0.081
67	native-country_United-States	0.023	0.019	0.692	1.446	0.321	0.04	0.298	0.058
68	race_Asian-Pac-Islander	0.022	0.012	0.933	1.022	0.298	0.056	0.32	0.058

TABLE I
COMPARISON OF KPIS: NEURAL NETWORK SUMMARY

B. Task 3: Defining Global Surrogate Models

Global Surrogate Models train on the output of the model that they surrogate to try to understand the behavior of the model and then make predictions of the outcome due to this training. Generally, these models are more interpretable models than black-box models such as Logistic Regression or Decision Tree.

1) Logistic Regression as Global Surrogate Model

The table (II) illustrates various KPIs derived from a Logistic Regression model, showcasing diverse performance metrics for different features and providing insights into the model's predictive capabilities.

- Education-based features such as 'Prof-school' and 'Doctorate' exhibit relatively higher values in EOD (0.723 and 0.436, respectively), indicating a stronger predictive performance for individuals with these educational backgrounds. These values imply that the model predicts well for these categories, suggesting their potential importance in predictive accuracy.
- Distinct native countries like 'France,' 'Iran,' and 'Thailand' portray differing predictive performances. While 'France' and 'Thailand' exhibit lower values across most metrics, suggesting weaker predictive capabilities for individuals from these countries compared to others, 'Iran' shows moderate values across most KPIs, indicating a somewhat stronger predictive performance compared to 'France' and 'Thailand.'
- The 'Private' workclass demonstrates lower performance across various metrics like EOD (0.054) and SP (0.495), suggesting a comparatively weaker predictive outcome for this category. Similarly, the 'Asian-Pac-Islander' race category also displays relatively lower values across several metrics, implying weaker predictive performance for this racial group.
- Notably, the 'United-States' native country representation depicts moderate predictive performance with values closer to the mean across different KPIs.

These KPIs offer insights into the model's predictive tendencies concerning different demographic or categorical groups. Educational backgrounds like 'Prof-school' and 'Doctorate' exhibit stronger predictive performances, while certain native countries, work classes, and racial categories present weaker predictive capabilities, highlighting potential biases or disparities in the model's predictions across these groups.

	Feature	EOD	PP	SP	DI	TPR0	FPR0	TPR1	FPR1
0	education_Prof-school	0.723	0.691	0.073	13.69	0.318	0.057	0.806	0.78
1	native-country_France	0.665	0.639	0.0	0.0	0.335	0.067	1.0	0.0
2	native-country_Iran	0.663	0.389	0.199	5.024	0.337	0.066	1.0	0.333
3	native-country_Thailand	0.663	0.088	0.0	0.0	0.337	0.067	1.0	0.0
4	education_Doctorate	0.436	0.468	0.128	7.782	0.327	0.064	0.63	0.5
64	native-country_Italy	0.067	0.055	0.0	0.0	0.338	0.067	0.333	0.0
65	workclass_Private	0.054	0.075	0.495	0.495	0.369	0.108	0.32	0.053
66	education_Assoc-voc	0.039	0.004	0.816	1.225	0.339	0.066	0.3	0.081
67	native-country_United-States	0.026	0.026	0.628	1.591	0.358	0.043	0.336	0.069
68	race_Asian-Pac-Islander	0.023	0.013	0.936	1.012	0.337	0.067	0.36	0.067

TABLE II
COMPARISON OF KPIS: LOGISTIC REGRESSION SUMMARY

2) Decision Tree as Global Surrogate Model

The table III presents various KPIs derived from a Decision Tree model, showcasing diverse performance metrics for different features and offering insights into the model's predictive capabilities.

- Education-related categories like 'Prof-school' and 'Doctorate' display relatively higher values in EOD (0.715 and 0.397, respectively), indicating a stronger predictive performance for individuals with these educational backgrounds. These values suggest the model's proficiency in predicting outcomes for these categories, implying their potential importance in predictive accuracy.
- Distinct native countries such as 'Iran,' 'Thailand,' and 'France' showcase varying predictive performances. While 'Iran' and 'Thailand' display similar values across most metrics, suggesting similar predictive capabilities for individuals from these countries, 'France' shows comparatively lower values across several metrics, implying weaker predictive performance for individuals from this country compared to 'Iran' and 'Thailand.'
- Categorical features like 'sex' and 'workclass_Private' demonstrate lower performance across various metrics, suggesting comparatively weaker predictive outcomes for these categories in the Decision Tree model.
- Racial categories like 'Asian-Pac-Islander' also display relatively lower values across several metrics, indicating weaker predictive performance for this racial group in the context of the model.
- Notably, 'United-States' as a native country representation depicts moderate predictive performance with values closer to the mean across different KPIs.

These KPIs offer valuable insights into the model's predictive tendencies concerning different demographic or categorical groups. Educational backgrounds like 'Prof-school' and 'Doctorate' exhibit stronger predictive performances, while certain native countries, categorical features, and racial categories present weaker predictive capabilities, suggesting potential biases or disparities in the model's predictions across these groups.

	Feature	EOD	PP	SP	DI	TPR0	FPR0	TPR1	FPR1
0	education_Prof-school	0.715	0.721	0.06	16.733	0.266	0.045	0.871	0.76
1	native-country_Iran	0.71	0.406	0.164	6.097	0.29	0.055	1.0	0.333
2	native-country_Thailand	0.71	0.106	0.0	0.0	0.29	0.055	1.0	0.0
3	education_Doctorate	0.397	0.442	0.117	8.525	0.28	0.053	0.593	0.45
4	native-country_France	0.377	0.406	0.0	0.406	0.29	0.055	0.667	0.0
64	sex	0.052	0.077	0.3	3.33	0.298	0.022	0.29	0.074
65	workclass_Private	0.044	0.057	0.502	0.502	0.302	0.088	0.285	0.044
66	native-country_Cuba	0.041	0.039	0.603	1.659	0.291	0.055	0.25	0.091
67	race_Asian-Pac-Islander	0.033	0.039	0.624	1.603	0.29	0.054	0.32	0.087
68	native-country_United-States	0.032	0.001	0.901	0.987	0.321	0.056	0.289	0.055

TABLE III
COMPARISON OF KPIS: DECISION TREE SUMMARY

C. Task 4: Comparison of Neural Networks and Global Surrogate Models

Comparing the KPIs between Neural Networks and surrogate models like Logistic Regression and Decision Trees (as shown in Tables IV and V, respectively), exposes notable disparities in percentage differences across key performance indicators.

1) Comparison of KPIs Neural Network and Logistic Regression as Global Surrogate Model

	NN	LR Sur	Diff LR Sur	% Diff LR Sur
EOD	0.254	0.281	0.027	10.674
PP	0.128	0.143	0.015	11.472
SP	0.176	0.185	0.009	5.321
DI	2.533	2.188	-0.345	-13.604
TPR0	0.293	0.331	0.038	12.919
FPR0	0.056	0.066	0.01	17.581
TPR1	0.219	0.244	0.024	11.086
FPR1	0.071	0.079	0.008	11.057

TABLE IV
COMPARISON OF KPIS: NEURAL NETWORK VS LOGISTIC REGRESSION SURROGATE

When examining the Neural Network against the Logistic Regression Surrogate model, significant disparities in Table IV metrics such as FPR0, showcasing a deviation of 17.581%, indicate substantial differences in predictive outcomes. These differences could suggest potential disparities in model performances concerning specific subgroups or classes within the dataset, raising fairness concerns.

2) Comparison of KPIs Neural Network and Decision Tree as Global Surrogate Model

	NN	DT Sur	Diff DT Sur	% Diff DT Sur
EOD	0.254	0.244	-0.011	-4.146
PP	0.128	0.125	-0.003	-2.603
SP	0.176	0.181	0.006	3.141
DI	2.533	2.053	-0.48	-18.933
TPR0	0.293	0.287	-0.007	-2.255
FPR0	0.056	0.054	-0.001	-2.408
TPR1	0.219	0.224	0.005	2.061
FPR1	0.071	0.072	0.001	1.898

TABLE V
COMPARISON OF KPIS: NEURAL NETWORK VS DECISION TREE SURROGATE

Contrasting the Neural Network against the Decision Tree Surrogate model unveils a notably lower DI at -18.933% in Table V. This significant difference emphasizes a considerable variance in fairness-related metrics between these models. It hints at disparities in predictive equity across different demographic or categorical groups present in the dataset.

Such disparities in percentage differences among surrogate models may indicate varying predictive abilities, possibly affecting fairness and equity in model outcomes. These observations emphasize the importance of thorough fairness assessments when deploying machine learning models, especially when surrogate models exhibit substantial discrepancies compared to the baseline Neural Network model.

D. Task 5: Finding Interpretable Models and Comparing Global Surrogate Model Counterparts

The interpretable models are generally assumed to have lower accuracy but higher explainability compared to black-box models. In specific scenarios where the data has a lower number of features like 5 or lower and a lower number of samples like a couple of thousand, the accuracy of the black-box models can converge to have similar values with the interpretable models. In our "Adult" dataset we have a much larger number of features, 97 in total which largely consists of the One hot encoded feature for transforming categorical features to binary features to be able to process them with the NN models. So, we should not expect them to give better results compared with the surrogate counterparts. In this part, we trained Logistic Regression and Decision Tree models and got the result accordingly.

1) Logistic Regression as Interpretable Model

The table (VI) offers an overview of various KPIs derived from a Logistic Regression model, reflecting diverse performance metrics for different features and providing insights into the model's predictive capabilities.

- Native countries like 'France,' 'Iran,' and 'Thailand' exhibit varying predictive performances within the model. 'France' and 'Thailand' showcase higher values across most metrics compared to 'Iran,' indicating comparatively stronger predictive capabilities for individuals from these countries. Conversely, 'Iran' demonstrates moderate values across most KPIs, suggesting somewhat weaker predictive performance compared to 'France' and 'Thailand.'
- Educational background features such as 'Prof-school' display relatively higher values in EOD (0.625), indicating a stronger predictive performance for individuals with this educational background. This suggests the model's proficiency in predicting outcomes for individuals with a 'Prof-school' educational status, highlighting its importance in predictive accuracy.
- Distinct categories like the 'White' race and specific occupations like 'Tech-support' and 'Local-gov' manifest varying predictive capabilities. The 'White' race exhibits lower values across several metrics, implying weaker predictive performance for this racial group in the context of the model. Similarly, occupations such as 'Tech-support' and 'Local-gov' also display relatively lower values, indicating comparatively weaker predictive outcomes for individuals in these occupational categories.
- Categorical features like 'sex' and 'workclass_State-gov' present moderate predictive performance, displaying values closer to the mean across different KPIs.

This analysis illuminates the model's predictive tendencies across diverse demographic, categorical, and occupational groups. While certain native countries and educational backgrounds show stronger predictive performances, specific racial groups, occupations, and categorical features demonstrate weaker predictive capabilities, implying potential biases or disparities in the model's predictions across these groups.

	Feature	EOD	PP	SP	DI	TPR0	FPR0	TPR1	FPR1
0	native-country_France	0.758	0.674	0.0	0.674	0.242	0.044	1.0	0.0
1	native-country_Iran	0.756	0.423	0.13	7.706	0.244	0.043	1.0	0.333
2	native-country_Thailand	0.756	0.123	0.0	0.0	0.244	0.044	1.0	0.0
3	education_Prof-school	0.625	0.626	0.053	18.747	0.223	0.035	0.742	0.66
4	native-country_England	0.508	0.324	0.26	3.85	0.242	0.043	0.75	0.167
64	race_White	0.081	0.043	0.512	1.951	0.171	0.024	0.252	0.047
65	occupation_Tech-support	0.065	0.035	0.386	0.386	0.247	0.044	0.182	0.017
66	workclass_Local-gov	0.039	0.054	0.529	1.889	0.241	0.041	0.28	0.078
67	sex	0.038	0.064	0.336	2.976	0.219	0.019	0.249	0.058
68	workclass_State-gov	0.023	0.009	0.853	1.173	0.244	0.043	0.267	0.051

TABLE VI
COMPARISON OF KPIS: LOGISTIC REGRESSION MODEL SUMMARY

2) Decision Tree as Interpretable Model

The table (VII) outlines an array of KPIs derived from a Decision Tree model, offering insights into its predictive capabilities based on different features.

- Significant disparities in predictive performance are observed among various features. For instance, 'Thailand' and 'France' as native countries showcase differing predictive performances. 'Thailand' displays higher values across most metrics compared to 'France,' suggesting a stronger predictive capability for individuals from 'Thailand' within the model's context.
- Educational backgrounds like 'Prof-school' demonstrate distinct predictive performances. 'Prof-school' exhibits relatively higher values in EOD (0.653), indicating a stronger predictive performance for individuals with this educational background. Conversely, 'Doctorate' shows moderate values across most KPIs, implying a comparatively weaker predictive performance compared to 'Prof-school.'
- Occupational categories such as 'Exec-managerial' display relatively lower values across several metrics, implying weaker predictive outcomes for individuals within this occupational category.
- Categorical features like 'sex' and 'marital-status_Separated' present moderate predictive performance, with values closer to the mean across different KPIs.

The model's predictions exhibit variations across demographic, educational, and occupational categories, suggesting varying levels of predictive accuracy. While some features demonstrate stronger predictive capabilities, others manifest comparatively weaker predictive performances, hinting at potential biases or disparities in the model's predictions across these categorized groups.

	Feature	EOD	PP	SP	DI	TPR0	FPR0	TPR1	FPR1
0	native-country_Thailand	0.834	0.143	0.0	0.0	0.166	0.035	1.0	0.0
1	education_Prof-school	0.653	0.646	0.039	25.558	0.144	0.027	0.71	0.68
2	native-country_France	0.501	0.443	0.0	0.0	0.166	0.035	0.667	0.0
3	education_Doctorate	0.417	0.394	0.073	13.612	0.157	0.033	0.444	0.45
4	occupation_Exec-managerial	0.344	0.223	0.118	8.484	0.076	0.02	0.42	0.166
64	native-country_Philippines	0.057	0.024	0.0	0.0	0.168	0.035	0.111	0.0
65	workclass_Private	0.041	0.05	0.393	0.393	0.194	0.065	0.153	0.026
66	marital-status_Separated	0.037	0.053	0.0	0.0	0.167	0.037	0.2	0.0
67	sex	0.035	0.05	0.267	3.747	0.158	0.013	0.169	0.048
68	native-country_United-States	0.023	0.001	0.88	0.948	0.189	0.037	0.166	0.035

TABLE VII
COMPARISON OF KPIS: DECISION TREE MODEL SUMMARY

3) Linear Regression vs Surrogate Linear Regression

Table VIII showcases a comparison of KPIs between a Logistic Regression (LR) model and a Surrogate Logistic Regression (LR Sur), revealing insights into their performance differences.

- Significant disparities emerge across various metrics between the LR and LR Sur models. Notably, in Statistical Parity (SP), the LR model displays a considerably lower value (0.122) compared to the LR Sur model's 0.185, resulting in a substantial -0.063 difference and a notable -33.813% percentage difference. This significant SP discrepancy indicates a notable difference in correctly identifying true negatives, favoring the LR Sur model.
- Similarly, the LR sur model outperforms the LR model in True Positive Rate for class 0 (TPR0). With the LR model exhibiting a TPR0 value of 0.242 and the LR Sur model showing a higher value of 0.331, the difference of -0.089 and the -26.961% percentage difference suggest the LR Sur model's better capture of true positives for class 0.

- Furthermore, disparities in Equal Odds Rate (EOD) and Predictive Parity (PP) between both models suggest varying predictive accuracies across different outcomes. The LR model generally demonstrates better performance in these metrics compared to the LR Sur model, as indicated by the negative differences and considerable percentage differences.

These differences in KPIs between the LR and LR Sur models underscore varying predictive capabilities and performance across multiple evaluation metrics, implying potential trade-offs in predictive accuracy between the original and surrogate logistic regression models.

	LR Sur	LR	Diff LR	% Diff LR
Accuracy	0.834	0.838	0.004	0.477
EOD	0.281	0.258	-0.024	-8.454
PP	0.143	0.131	-0.012	-8.313
SP	0.185	0.122	-0.063	-33.813
DI	2.188	1.857	-0.332	-15.163
TPR0	0.331	0.242	-0.089	-26.961
FPR0	0.066	0.043	-0.022	-33.933
TPR1	0.244	0.214	-0.03	-12.308
FPR1	0.079	0.065	-0.014	-17.394

TABLE VIII
COMPARISON OF KPIS: LOGISTIC REGRESSION VS SURROGATE LOGISTIC REGRESSION

4) Decision Tree vs Surrogate Decision Tree

The table (IX) presents a comparative analysis of various KPIs between a Decision Tree (DT) model and a Surrogate Decision Tree (DT Sur) model, offering insights into performance disparities.

- Significant variations exist across several metrics between the DT and DT Sur models. The DT Sur model demonstrates notably higher values for Statistical Parity (SP) and True Positive Rate for class 0 (TPR0) compared to the DT model. For instance, the SP of the DT Sur model is substantially higher at 0.181 versus the DT model's 0.113, indicating a -0.068 difference and a significant -37.482% percentage difference, suggesting a decline in identifying true negatives.
- Contrarily, in metrics such as Disparate Impact (DI), the DT Sur model outperforms the DT model, displaying a lower value (2.053) versus the DT model's 2.268. This difference of 0.215 and a 10.458% percentage difference suggests that the DT Sur model has a better ability to treat more equally between different classes compared to the DT model.

The observed differences in KPIs between the DT and DT Sur models underscore varying strengths and weaknesses in their predictive capabilities across different evaluation metrics, implying trade-offs and disparities in performance between the original and surrogate decision tree models.

	DT Sur	DT	Diff DT	% Diff DT
Accuracy	0.836	0.832	-0.004	-0.48
EOD	0.244	0.178	-0.065	-26.845
PP	0.125	0.088	-0.037	-29.875
SP	0.181	0.113	-0.068	-37.482
DI	2.053	2.268	0.215	10.458
TPR0	0.287	0.164	-0.122	-42.621
FPR0	0.054	0.035	-0.019	-35.43
TPR1	0.224	0.131	-0.093	-41.411
FPR1	0.072	0.05	-0.022	-30.805

TABLE IX
COMPARISON OF KPIS: DECISION TREE VS SURROGATE DECISION TREE

VI. CONCLUSION

This project has comprehensively explored how fairness concerns evolve based on different models and approaches, extensively detailed through tables. The analysis reveals that black box models tend to exhibit higher KPIs compared to the Surrogate Decision Tree while demonstrating lower KPIs when compared to the Surrogate Logistic Regression. Interpretable models display a significant decrease in KPIs in contrast to their surrogate of NN counterparts which is desirable behavior. The findings across these methodologies consistently suggest that biases within the "Adult" dataset are more likely to be associated with Ethnicity (specifically Eastern ethnicity) and higher education levels such as 'Prof' rather than gender, occupation, or work class, impacting the likelihood of exceeding an income threshold of $50k/year.