A GUI for the Fair & Explainable Selective Classifier IFAC

Daphne Lenders[1,*], Roberto Pellungrini[1] and Fosca Giannotti[1]

[1]Scuola Normale Superiore, Pisa, Italy

Abstract

In this paper, we present a Graphical User Interface (GUI) for IFAC, a selective classification model that refrains from making decisions in case they are uncertain or unfair. Since IFAC makes use of explainable-by-design methods to detect potentially unfair decisions, our GUI visualizes these explanations to let users understand the reason for abstention. We demonstrate how users can interpret the explanations, allowing them to contextualize, validate, and challenge the detected bias patterns.

Keywords

Selective Classification, Bias Audit, Explainable AI, Human-in-the-loop

1. Introduction

Ensuring fairness in automated decision-making (ADM) systems has been a longstanding challenge, particularly in high-stakes domains such as hiring and lending. The goal is to design classifiers that do not discriminate based on demographic group membership, and do not perpetuate their training data's bias against vulnerable population groups, like women or people of color.

Initial efforts to build fair systems often focused on group fairness metrics, such as demographic parity, optimizing models to satisfy these constraints while maintaining predictive accuracy [1, 2]. However, recent research has highlighted the limitations of this approach: blindly optimizing for a predefined fairness metric can obscure the real-world impact on individual decision subjects and fail to capture the nuances of how unfair decisions are made and corrected [3, 4, 5]. A more effective strategy is to first analyze how discrimination manifests in a specific decision-making task before directly addressing the bias where it occurs [5]. To implement such an approach successfully, human experts with ethical training and appropriate domain knowledge must be actively involved in bias audits and decision reviews. This call is also echoed by emerging AI regulations, mandating that AI-driven decision-making processes remain overseeable, interpretable, and subject to human intervention.

One possible approach to enable human oversight is given by the recently proposed selective classification framework IFAC [6]. Just like any selective classifier IFAC can abstain from predictions in case they are uncertain, but additionally, IFAC also abstains from making potentially unfair predictions. By making use of explainable-by-design techniques to uncover potential cases of unfairness, a human-in-the-loop can review these instances and their explanation, to make more well-informed decisions on them. In this paper, we present a GUI behind this selective classification framework. We highlight the methods behind the discrimination discovery and show how the interface can assist users in further understanding and addressing a classifier's discriminatory behavior.

2. Background & General Intuition

Our GUI visualizes the instances whose original predictions were rejected by IFAC, due to *uncertainty* or *unfairness.* For instances that got rejected for the latter reason, users can view an explanation of why the original prediction was deemed as unfair. This explanation consists of a global part, displaying which

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3d World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey

[*] Corresponding author.

daphne.lenders@sns.it (D. Lenders); roberto.pellungrini@sns.it (R. Pellungrini); fosca.giannotti@sns.it (F. Giannotti)
0000-0002-9839-9077 (D. Lenders)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4017/paper_33.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

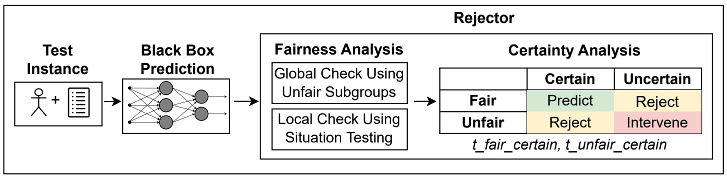


Figure 1: Basic Intuition behind IFAC's rejection mechanism

larger at-risk subgroup an instance was a part of, and a local part, showing an individual discrimination analysis for the instance. In this section, we describe the basic intuition behind the rejection framework and both types of fairness analysis. To do so we make use of the *folktables dataset* as our running example. This dataset contains sensitive information about peoples' gender and race, as well as some neutral characteristics, like their occupation and working hours. The associated task is to predict a person's income level (high vs. low). Classification models developed for this task typically favour the group of *white men.* They have higher positive decision rates and make fewer False Negatives and more False Positives for this group compared to other demographics.

2.1. Selective Classification

We describe a dataset as a triplet $(\mathbf{L}, \mathbf{S}, Y)$, where $\mathbf{L}$ represents the legally-grounded features and takes values in $\mathcal{L} \subseteq \mathbb{R}^{d_l}$; $\mathbf{S}$ refers to the sensitive attributes and takes values in $\mathcal{S} \subseteq \mathbb{R}^{d_s}$; Y is the binary target variable, with domain $\mathcal{Y} = \{0, 1\}$. We use $\mathbf{X}$ to describe the pair of legally grounded and sensitive features $(\mathbf{L}, \mathbf{S})$. Hence, in our running example, Y is the income level, that needs to be predicted based on $\mathbf{L}$, which includes features like education level and working hours, and $\mathbf{S}$, which are gender and race.

To find a mapping between the feature space of $\mathbf{X}$ and $\mathcal{Y}$, we can learn a classification model $h(\mathbf{x})$, minimizing some empirical risk function. To prevent h from discriminating based on features in $\mathbf{S}$ and to increase its predictive accuracy, the selective classification framework behind IFAC proposes to learn some abstention mechanism over its predictions. We denote the selective function, determining which of hs predictions are kept as g. In the case of IFAC, g considers the *fairness* and *uncertainty* of predictions. IFAC measures uncertainty through the prediction probability $v(\mathbf{x}) = P(Y = y|\mathbf{X} = \mathbf{x})$ outputted by h for predicted label $h(\mathbf{x}) = y$. Depending on whether the predictions are considered fair/unfair and certain/uncertain, there are four different scenarios that IFAC must deal with, as shown in Figure 1. The easiest case is when a prediction is both deemed fair and certain, in which case the prediction can be kept. Predictions that are fair yet uncertain get rejected, in line with the classical selective classification framework. In the case of prediction unfairness, there are two scenarios: if a prediction is both unfair and uncertain, and hence there are double reasons to doubt h's original decision, a fairness intervention is performed and h's original label is flipped. In case the prediction is unfair yet certain, human expertise is required to assess this prediction, and IFAC rejects it.

To prevent IFAC from rejecting all predictions, a user-defined *coverage* parameter determines the minimum amount of predictions that should be made. Additionally, IFAC takes a fairness-weight parameter, that denotes the ratio of rejections that can be made out of unfairness concerns, and how much room should be left for rejecting (fair but) uncertain predictions. These parameters serve to tune two separate thresholds - *t_fair_certain* and *t_unfair_certain* - that respectively determine at which prediction probabilities fair and unfair predictions should be viewed as certain/uncertain, to consequentially keep, reject, or intervene on the predictions. For full details behind tuning these parameters we refer to the original paper behind IFAC [6]

2.2. At-Risk Subgroups

The first step in IFACs fairness assessment is identifying population subgroups at risk of discrimination by a classifier h, for which the methodology of *discriminatory association rule mining* is adopted [7].

Let us assume access to a dataset of realizations, $\mathcal{D}$, that consists of the features $\mathbf{X} = (\mathbf{L}, \mathbf{S})$

- A specific realization of a single feature $\subset \mathbf{X}$ is called an *item*.
- An *itemset*, denoted by I, is a combination of multiple items, that can be decomposed into (I_L, I_S) where I_L is an itemset consisting of only legally grounded features, and I_S one consisting of only sensitive features
- A *transaction*, denoted by T, represents an itemset corresponding to one instance in $\mathcal{D}$, where each feature is assigned exactly one value.
- We say T *verifies* itemset (I_L, I_S) if $(I_L, I_S) \subseteq T$.

For example, in the *folktables* dataset, consider the feature `race`. A specific realization, such as `(race=Black)`, is an item. An itemset is a combination of multiple items, such as `(race=Black, education=Masters)` and can be decomposed into $I_S =$ `(race=Black)` and $I_L =$ `(education=Masters)` One single row from the dataset can be called a transaction.

To learn associations between the data's features and the decision outcome in $\mathcal{D}$ we can extract decision rules of the form $(I_L, I_S) \rightarrow Y$. The support of a decision rule regarding $\mathcal{D}$ is calculated as

$$supp_{\mathcal{D}}\left((I_L, I_S) \rightarrow Y\right) = supp_{\mathcal{D}}((I_L, I_S), Y) \quad \text{with} \quad supp_{\mathcal{D}}(I) = \frac{|\{T \in \mathcal{D} : I \subseteq T\}|}{|\mathcal{D}|}$$

, where || is the cardinality operator. Further, the confidence of a rule is defined as

$$conf_{\mathcal{D}}((I_L, I_S) \rightarrow Y) = \frac{supp_{\mathcal{D}}((I_L, I_S), Y)}{supp_{\mathcal{D}}((I_L, I_S))}$$

Finally, IFAC assesses how problematic a decision rule is by measuring the impact of the sensitive features in I_S on Y, through the Selective Lift (*slift*) [7]. In this paper, we use the definition of slift *by difference*, which measures how the confidence of a rule decreases when negating its sensitive part.

$$slift_{\mathcal{D}}\left((I_L, I_S) \rightarrow Y\right) = conf_{\mathcal{D}}\left((I_L, I_S) \rightarrow Y\right) - conf_{\mathcal{D}}\left((I_L, \neg I_S) \rightarrow Y\right) \tag{1}$$

Computing $conf_{\mathcal{D}}(I_L, \neg I_S) \rightarrow Y$ requires one to take the confidence of all the transactions that verify I_L but do not verify I_S. If the slift of some rule exceeds some user-defined threshold, we can describe the itemset (I_L, I_S) as an *at-risk* subgroup of the data.

Example : Consider the decision rule `(race = Black, education = Masters` $\rightarrow$ `income = low)` with a confidence of 0.9. If we find that the rule `(`$\neg$`race = Black, education = Masters` $\rightarrow$ `income = low)` has a confidence of 0.3, the slift is 0.6. This relatively high measure can indicate that the subgroup `(race = Black, education = Masters)` are at risk of discrimination.

2.3. Individual Discrimination

If an instance falls under any subgroup at risk of discrimination, IFAC also performs Situation Testing, an explainable-by design method to determine fair treatment on a local level [8]. Given some individual instance $\mathbf{x}_i$, Situation Testing searches $\mathcal{D}$ for its k most similar instances from the favoured and non-favoured group, which we denote respectively as $\mathcal{K}_{tr}^{r}$ and $\mathcal{K}_{tr}^{nr}$. Recall that in the *folktables* dataset we see *white men* as favoured, and all other demographic groups as non-favoured. To define $\mathbf{x}_i$'s individual discrimination score, we compute the difference in positive decision ratios between $\mathcal{K}_{tr}^{r}$ and $\mathcal{K}_{tr}^{nr}$. Hence:

$$disc(\mathbf{x}_i) = \frac{|\{j \in \mathcal{K}_{tr}^{r} : y_j = 1\}|}{k} - \frac{|\{j \in \mathcal{K}_{tr}^{nr} : y_j = 1\}|}{k} \tag{2}$$

If this discrimination score exceeds some user defined threshold, $\mathbf{x}_i$ is deemed to be treated unfairly.

3. Methodology

The GUI behind IFAC visualizes the instances in a decision task, their predictions, as well as IFACs decisions to keep, reject or intervene on these predictions. In case IFAC rejects or intervenes on predictions based on unfairness concerns, the GUI also visualizes the explanations behind these rejections; i.e. the at-risk subgroups these instances belong to and the outcome of their individual discrimination analysis.

Currently, the GUI is only available as a prototype meant to explore the rejections of an IFAC model trained on the *folktables dataset*. However, the platform is built to be extensible and flexible to various classification tasks Before describing the (visual) components of the tool, we shortly outline the classification model used for building this prototype and explain how we ran the global and local discrimination analysis on it [1].

Classification Model After preprocessing the dataset, we split the initial dataset into a training part (n=9600), two validation sets (*val_1, val_2* both with n=3600) and a test set (n=1200).

For our classification model, we train a Random Forest Classifier using the default sklearn hyperparameters. The GUI visualizes the model's predictions on the test set, while in the background both *val_1* and *val_2* are used for the discrimination analysis as described in the next paragraphs.

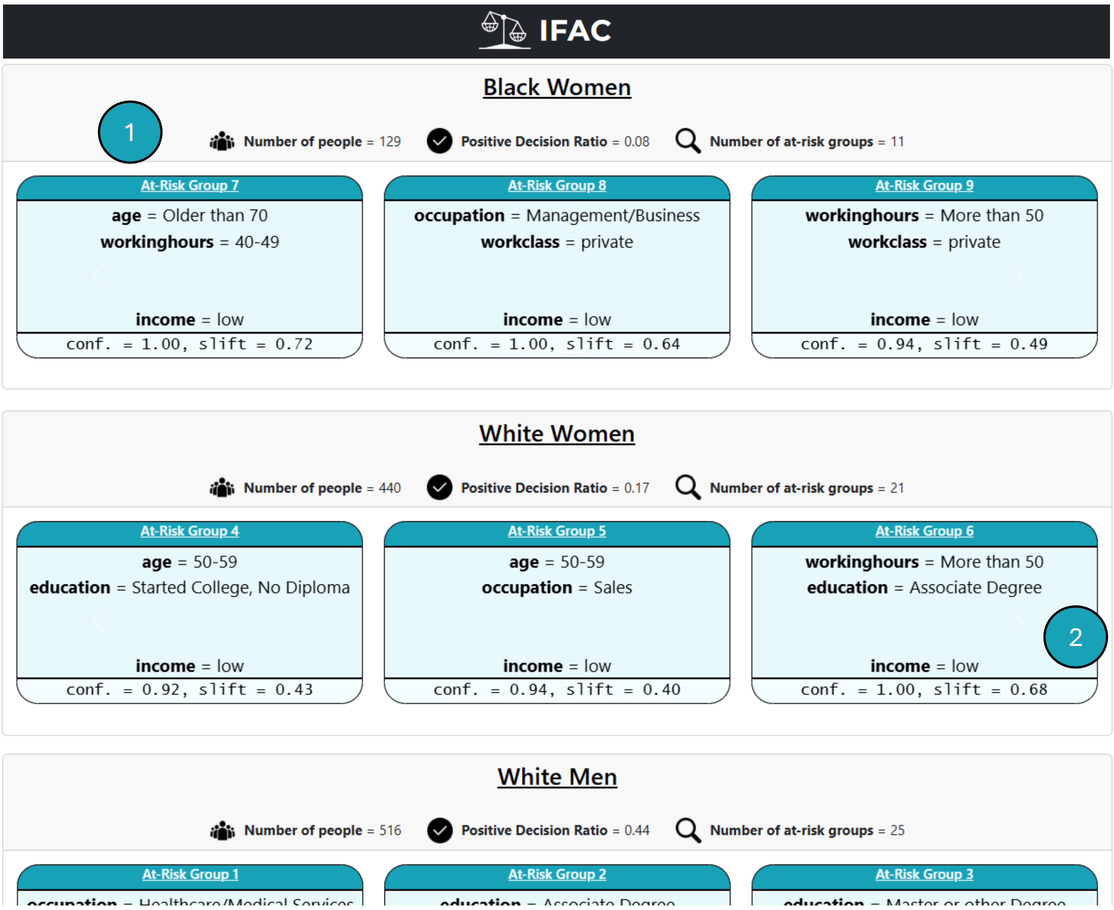
At-Risk Subgroups The at-risk subgroups that are visualized in the GUI are extracted after applying the initial random forest model on *val_1*. To display at-risk groups for each single-axis and intersectional demographic group, we split *val_1* according to each sensitive feature value and their combination. In our case, using sensitive attributes race (black or white) and gender (male or female), we end up splitting the data according to `(race = white)`, `(race = black)`, `(sex = male)`, `(sex = female)`, `(race = white, sex = male)`, `(race = white, sex = female)`, `(race = black, sex = male)`, `(race = black, sex = female)`. On the data belonging to each of these groups, we seperately apply the apriori algorithm to mine decision rules of the form $(I_L, I_S) \rightarrow Y$, where Y represents the classifier's h decision outcome (i.e. people's income) which is either *high* or *low*. Since we assume the group of *white men* to be favoured, we only extract rules with $Y = high$ for them, while for all other demographic groups that are potentially discriminated, we only select rules with $Y = low$. Using equation 1, we compute the slift for each of the associations, and filter rules with $slift > 0.4$. Further, we assess statistical significance with a Z-test and only retain rules with $p < 0.01$

Situation Testing For all of the test instances that belong to one of the identified *at-risk* subgroups, we also compute an individual discrimination score as described in section 2.3. To compute these scores, we search *val_2* for each instance's top 5 nearest neighbors from both the favoured and non-favoured group, and compute their difference in positive decision ratio. If an instance's *disc_score* exceeds 0.2, we view its prediction as unfair.

Learning Rejection Thresholds IFAC deems instances unfair if they fall under an at-risk subgroup and if they are individually discriminated against. How many of these instances can then be rejected, depends on the *coverage*, which we set at 0.8, meaning that 20% of the test-set instances can be rejected. Moreover, we set IFAC's fairness weight to 0.9, meaning that out of those rejected instances 90% should get rejected out of unfairness concerns and the remainder should be rejected solely because of uncertainty. Based on these two parameters we tune the two thresholds (*t_fair_certain* and *t_unfair_certain*) on the *val_2*, such that

$$(h,g)(\mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{if } Fair(\mathbf{x}) \text{ and } \upsilon(\mathbf{x}) => \textit{t\_fair\_certain} \\ \text{abstain} & \text{if } Fair(\mathbf{x}) \text{ and } \upsilon(\mathbf{x}) < \textit{t\_fair\_certain} \\ \text{flip} & \text{if } \neg Fair(\mathbf{x}) \text{ and } \upsilon(\mathbf{x}) < \textit{t\_unfair\_certain} \\ \text{abstain} & \text{if } \neg Fair(\mathbf{x}) \text{ and } \upsilon(\mathbf{x}) >= \textit{t\_unfair\_certain} \end{cases}$$

[1] for full information we refer to our github: https://github.com/daphnetje/IFAC_GUI



Black Women

Number of people = 129 Positive Decision Ratio = 0.08 Number of at-risk groups = 11

At-Risk Group 7
age = Older than 70
workinghours = 40-49
income = low
conf. = 1.00, slift = 0.72

At-Risk Group 8
occupation = Management/Business
workclass = private
income = low
conf. = 1.00, slift = 0.64

At-Risk Group 9
workinghours = More than 50
workclass = private
income = low
conf. = 0.94, slift = 0.49

White Women

Number of people = 440 Positive Decision Ratio = 0.17 Number of at-risk groups = 21

At-Risk Group 4
age = 50-59
education = Started College, No Diploma
income = low
conf. = 0.92, slift = 0.43

At-Risk Group 5
age = 50-59
occupation = Sales
income = low
conf. = 0.94, slift = 0.40

At-Risk Group 6
workinghours = More than 50
education = Associate Degree
income = low
conf. = 1.00, slift = 0.68

White Men

Number of people = 516 Positive Decision Ratio = 0.44 Number of at-risk groups = 25

At-Risk Group 1
occupation = Healthcare/Medical Services

At-Risk Group 2
education = Associate Degree

At-Risk Group 3
education = Master or other Degree

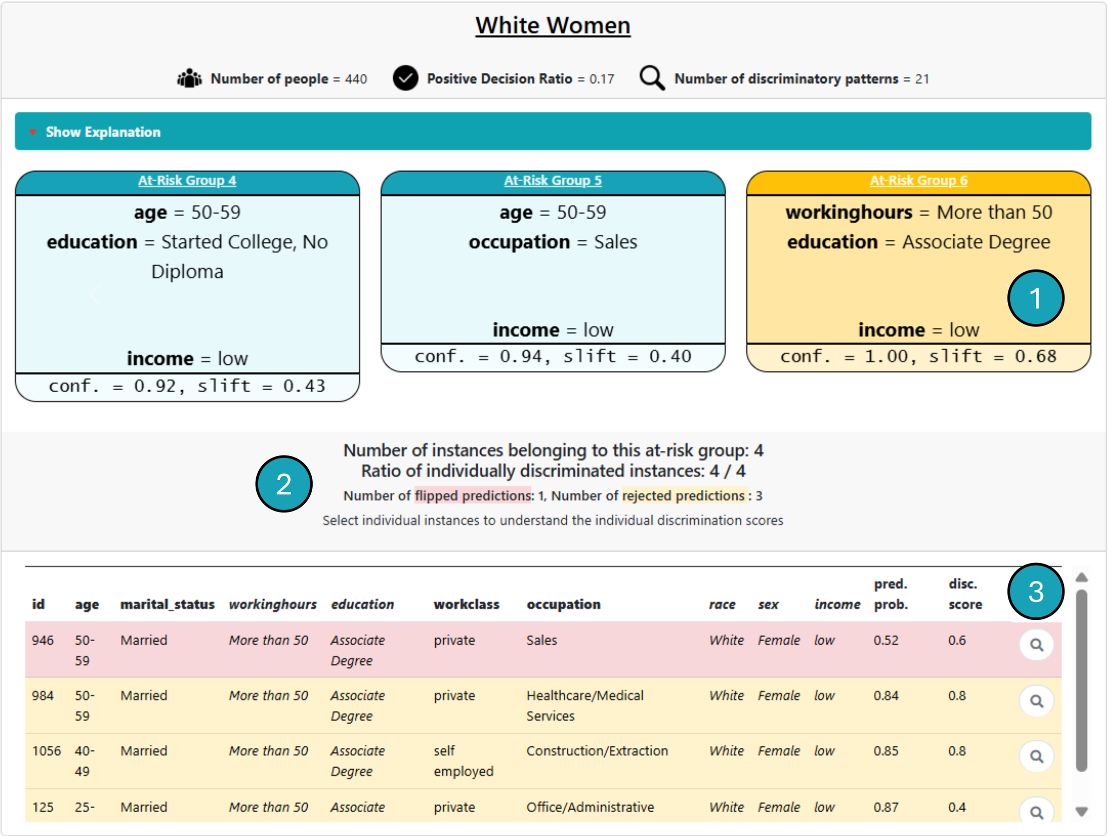
Figure 2: The starting page: for each single-axis and intersectional demographic group some statistical information, like the number of instances and positive decision ratio within the group, is displayed **(1)**. For each demographic group, also the at-risk subgroups are visualized in a slide carousel as seen in **(2)**.

4. The Tool Through the Eyes of a User

Now that we have described all the theoretical background and methodology behind our GUI, we are going to describe each of its components and how a user can interact with them.

4.1. Inspecting Single-Axis and Intersectional Demographic Groups

The first thing a user sees when opening the GUI are the different single-axis and intersectional groups of the data, along with some statistical information and the at-risk subgroups, based on which IFAC makes its rejections. In Figure 2 a fragment of this starting screen is visualized. Based on the statistical information displayed, a user can quickly assess how the group of white men is being favoured for this decision task, as their positive decision ratio is 44%, considerably higher than for white and black women. The at-risk groups within each demographic group are displayed inside a slide carousel, that a user can browse through to understand where the biggest fairness concerns lay. Here users can click on a specific at-risk group, like for instance *group #6* within white women. This group consists of white women, with an associate degree, working more than 50 hours a week. As indicated by the confidence measure, they receive a low-income prediction 100% of the time. This group could be of interest as their high education level and amount of working hours, would intuitively be associated with high incomes. This is further confirmed by the slift of 0.68, indicating that white men with the same degree and working hours are only associated with a low income 32% of the time. In the next section, we visualize the interface after a user has selected this at-risk group.



White Women

Number of people = 440 ✓ Positive Decision Ratio = 0.17 Number of discriminatory patterns = 21

Show Explanation

At-Risk Group 4

age = 50-59
education = Started College, No Diploma

income = low
conf. = 0.92, slift = 0.43

At-Risk Group 5

age = 50-59
occupation = Sales

income = low
conf. = 0.94, slift = 0.40

At-Risk Group 6

workinghours = More than 50
education = Associate Degree

income = low
conf. = 1.00, slift = 0.68

Number of instances belonging to this at-risk group: 4
Ratio of individually discriminated instances: 4 / 4
Number of flipped predictions: 1, Number of rejected predictions : 3
Select individual instances to understand the individual discrimination scores

id	age	marital_status	*workinghours*	*education*	workclass	occupation	*race*	*sex*	*income*	pred. prob.	disc. score	
946	50-59	Married	*More than 50*	*Associate Degree*	private	Sales	*White*	*Female*	*low*	0.52	0.6	🔍
984	50-59	Married	*More than 50*	*Associate Degree*	private	Healthcare/Medical Services	*White*	*Female*	*low*	0.84	0.8	🔍
1056	40-49	Married	*More than 50*	*Associate Degree*	self employed	Construction/Extraction	*White*	*Female*	*low*	0.85	0.8	🔍
125	25-	Married	*More than 50*	*Associate*	private	Office/Administrative	*White*	*Female*	*low*	0.87	0.4	🔍

Figure 3: The interface upon selecting a specific at-risk group to inspect. **(1)** The selected at-risk group is highlighted in yellow. In Table (2) all instances belonging to the selected group are displayed, along with their individual discrimination score, their associated prediction probabilities, and IFACs decision to reject or flip the predictions. **(3)** Users can click on the magnifying glass to view the individual discrimination analysis

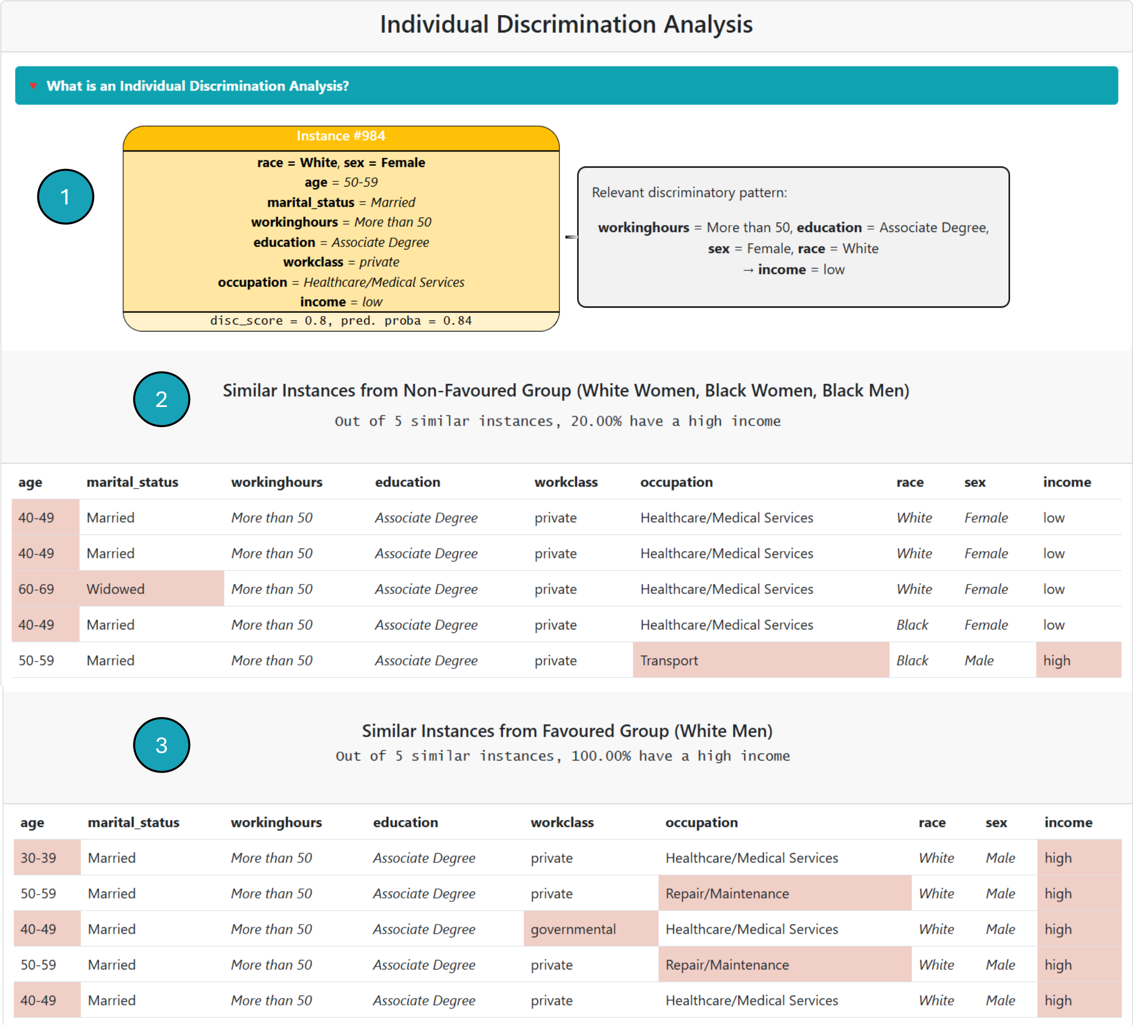
4.2. Selecting an At-Risk Subgroup

Figure 3 shows the interface for the at-risk group of white women with an associate degree working over 50 hours per week. The table lists 4 instances, all flagged as individually discriminated by IFAC. One instance, highlighted in red, is suggested for a prediction flip due to its high discrimination score (0.6) and uncertain prediction probability (0.52), while the other instances are rejected for having high discrimination scores, yet high prediction certainties. Users can click the magnifying glass to view additional details, helping them better assess the rejected predictions and potentially override them. One interesting instance to inspect is, e.g., with id = 984: a married woman, aged between 50 and 59, working in medical services. Since being a bit older, having a stable relationship and working in a secure sector should correlate with higher incomes, a user might be surprised by their original low-income prediction, and might want to view why IFAC rejected this it.

4.3. Checking Individual Discrimination

Figure 4 shows the interface after a user selects a rejected instance, revealing its individual discrimination analysis. This instance has a discrimination score of 0.8—indicating an 80% difference in the positive decision ratios between similar favoured and non-favoured instances. Two tables display these similar instances separately, with orange-highlighted cells marking feature values that differ from the instance in question.

Notably, most non-favoured instances differ in age from the selected instance, while some favoured instances vary in age, workclass, and occupation. The visual highlights of these features, serve as reminders that even highly similar instances may differ in meaningful ways, potentially affecting the



Individual Discrimination Analysis

What is an Individual Discrimination Analysis?

Instance #984

race = White, **sex** = Female
age = *50-59*
marital_status = *Married*
workinghours = *More than 50*
education = *Associate Degree*
workclass = *private*
occupation = *Healthcare/Medical Services*
income = *low*
`disc_score = 0.8, pred. proba = 0.84`

Relevant discriminatory pattern:

workinghours = More than 50, **education** = Associate Degree, **sex** = Female, **race** = White → **income** = low

Similar Instances from Non-Favoured Group (White Women, Black Women, Black Men)

`Out of 5 similar instances, 20.00% have a high income`

age	marital_status	workinghours	education	workclass	occupation	race	sex	income
40-49	Married	*More than 50*	*Associate Degree*	private	Healthcare/Medical Services	*White*	*Female*	low
40-49	Married	*More than 50*	*Associate Degree*	private	Healthcare/Medical Services	*White*	*Female*	low
60-69	Widowed	*More than 50*	*Associate Degree*	private	Healthcare/Medical Services	*White*	*Female*	low
40-49	Married	*More than 50*	*Associate Degree*	private	Healthcare/Medical Services	*Black*	*Female*	low
50-59	Married	*More than 50*	*Associate Degree*	private	Transport	*Black*	*Male*	high

Similar Instances from Favoured Group (White Men)

`Out of 5 similar instances, 100.00% have a high income`

age	marital_status	workinghours	education	workclass	occupation	race	sex	income
30-39	Married	*More than 50*	*Associate Degree*	private	Healthcare/Medical Services	*White*	*Male*	high
50-59	Married	*More than 50*	*Associate Degree*	private	Repair/Maintenance	*White*	*Male*	high
40-49	Married	*More than 50*	*Associate Degree*	governmental	Healthcare/Medical Services	*White*	*Male*	high
50-59	Married	*More than 50*	*Associate Degree*	private	Repair/Maintenance	*White*	*Male*	high
40-49	Married	*More than 50*	*Associate Degree*	private	Healthcare/Medical Services	*White*	*Male*	high

Figure 4: Interface upon selecting an individual instance. In **(1)** the instance and its individual discrimination score are displayed. In **(2)** and **(3)** respectively, the 5 most similar instances from the non-reference and reference group are visualized in a table. Table cells marked in orange, indicate that the given instance differs in its feature value , from the instance in question.

discrimination score.

Ultimately, this explanation behind IFAC's rejection serves to assist humans expert in deciding whether the evidence of discrimination is strong enough to override the original low-income prediction. In this case, experts must decide whether high-income predictions for similar white men are justified by differences in age and occupation or whether they indicate unfair treatment. At this point, they can also seek additional details for the affected instance, such as their exact occupation within the healthcare sector, to assess whether this should deserve a high income. These considerations, underscore the essential role of human domain experts within the selective classification framework of IFAC. While computational methods can effectively identify at-risk subgroups and potential individual discrimination, these should be viewed as decision-support tools rather than definitive arbiters of fairness. Visualizing IFAC's rejected instances along with the explanations behind them, can guide experts into understanding unfairness issues and making more just predictions.

5. Conclusion & Future Work

In this paper, we introduced a prototype-GUI, that visualizes instances that were rejected by the selective classification algorithm IFAC. Behind all of IFACs unfairness-based rejections, it visualizes the explanations of why predictions are seen as unfair, making use of explainable-by-design methods of discriminatory association rule mining [7] and situation testing [8]. Through a practical scenario, we demonstrated how the GUI assists users in reviewing rejected instances while underscoring the indispensable role of human expertise in contextualizing, interpreting, and, when necessary, challenging the underlying patterns of bias.

Despite its potential, the GUI remains a prototype with room for further development. First, the current implementation is limited to a single decision task: income prediction on the folktables dataset. To evaluate the practical usability of the tool and its impact in the real world, future iterations should extend its functionality to support any classification task.

Additionally, the GUI currently serves only as a tool to view rejected predictions, without allowing users to modify them or explore the impact of corrective actions. Incorporating an interactive intervention feature, that enables users to adjust decisions and observe how fairness metrics evolve, would transform the GUI into an active bias mitigation solution.

Acknowledgments

This work was funded by the European Union under TANGO "It takes two to tango: a synergistic approach to human-machine decision making", Grant Agreement 101120763, CUP E53C23001170006. The work was also supported by the PNRR-M4C2-Investimento 1.3, Partenariato Esteso PE00000013-“FAIR-Future Artificial Intelligence Research” Spoke 1 “Human-centered AI”, funded by the European Commission under the NextGeneration EU programme. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to perform grammar checks. All generated content was reviewed by the authors, who take full responsibility for this publication.

References

[1] T. Kamishima, S. Akaho, J. Sakuma, Fairness-aware learning through regularization approach, in: 2011 IEEE 11th international conference on data mining workshops, IEEE, 2011, pp. 643–650.

[2] C. Wadsworth, F. Vera, C. Piech, Achieving fairness through adversarial learning: an application to recidivism prediction, arXiv preprint arXiv:1807.00199 (2018).

[3] D. Lenders, T. Calders, Real-life performance of fairness interventions-introducing a new benchmarking dataset for fair ml, in: Proceedings of the 38th ACM/SIGAPP symposium on applied computing, 2023, pp. 350–357.

[4] M. Favier, T. Calders, Cherry on the cake: fairness is not an optimization problem, Machine Learning 114 (2025) 160.

[5] S. Goethals, T. Calders, D. Martens, Beyond accuracy-fairness: Stop evaluating bias mitigation methods solely on between-group metrics (2024).

[6] D. Lenders, A. Pugnana, R. Pellungrini, T. Calders, D. Pedreschi, F. Giannotti, Interpretable and fair mechanisms for abstaining classifiers, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2024, pp. 416–433.

[7] D. Pedreschi, S. Ruggieri, F. Turini, Measuring Discrimination in Socially-Sensitive Decision Records, SIAM, 2009, pp. 581–592.

[8] B. L. Thanh, S. Ruggieri, F. Turini, k-nn as an implementation of situation testing for discrimination discovery and prevention, in: KDD, ACM, 2011, pp. 502–510.