Exploring gender bias in misclassification with clustering and local explanations*

Aurora Ramírez[1,2][0000−0002−1916−6559]

[1] Dept. Computer Science and Numerical Analysis, University of Córdoba, Córdoba 14071, Spain
[2] Data Science and Computational Intelligence (DaSCI) Research Institute, Spain
aramirez@uco.es

Abstract. Gender bias is one of the types of bias studied in fair machine learning (ML), which seeks equity in the predictions made by ML models. Bias mitigation is often based on protecting the sensitive attribute (e.g. gender or race) by optimising some fairness metrics. However, reducing the relevance of the sensitive attribute can lead to higher error rates. This paper analyses the relationship between gender bias and misclassification using explainable artificial intelligence. The proposed method applies clustering to identify groups of similar misclassified instances between false positive and false negative predictions. These prototype instances are then further analysed using Break-down, a local explainer. Positive and negative feature contributions are studied for models trained with and without gender data, as well as using bias mitigation methods. The results show the potential of local explanations to understand different forms of gender bias in misclassification, which are not always related to a high feature contribution of the gender attribute.

Keywords: Fair machine learning · Gender bias · Explainable artificial intelligence · Clustering

1 Introduction

The rise of AI-based systems has prompted the need to analyse how their methods work and what the implications of their results are. Many automatically constructed decision models, such as those using machine learning (ML), are opaque or difficult to understand. As a result, users may distrust their predictions and lose confidence in the benefits of AI. One of the causes of distrust is the fact that ML models are not infallible. In contexts such as medicine, a wrong prediction can have serious consequences, so it is important to understand when and why a predictive model makes mistakes [13].

* Funding: GENIA project funded by the Annual Research Plan of University of Córdoba (UCOImpulsa mod., 2022). Grant PID2020-115832GB-I00 funded by MICIN/AEI/10.13039/501100011033. Andalusian Regional Government (postdoctoral grant DOC_00944).

Another aspect influencing users' perception of ML is the presence of bias, which arises not only as a consequence of data collection, but also of the training process itself. Fair ML is the field that characterises and studies bias in ML, providing metrics to detect bias and methods to mitigate biased decisions based on gender, race or age [17]. Gender bias is a particular case that has attracted attention recently due to the discovery of gender stereotyping in automatic translation [19]. Gender bias also arises in AI systems aimed at supporting biomedicine and healthcare, as the gender dimension is not taken into account in many of their algorithmic procedures [6]. Face recognition systems are another area where gender bias has recently been studied [15].

Despite the application of mitigation methods, the resulting models may still be opaque, as the problem of using black-box baseline estimators persists. Reducing the contribution of sensitive attributes (e.g. gender) during model building does not guarantee that biases have been completely eliminated. Other attributes with high correlation with the decision variable and imbalanced presence for each gender may still lead to biased decisions. In addition, improving fairness often leads to a reduction in model performance [22], as false positive and false negative rates tend to balance across categories of the sensitive attribute.

This paper proposes to jointly address both issues, i.e. misclassification and gender bias, from the perspective of explainable artificial intelligence (XAI) [7]. XAI has emerged to improve the transparency of ML-based systems by providing methods to inspect black-box models and generate explanations of their predictions [9]. Thus, XAI methods can help us to understand the nature of misclassification and to analyse the extent to which it is related to gender bias. The proposed approach first uses a clustering algorithm to identify prototype instances among those misclassified for each gender. Next, instance-level explanations are generated to analyse which features were most responsible for false positive and false negative predictions. To provide an initial validation of the approach, an experimental evaluation is presented using datasets often studied by the Fair ML community [12]. Experiments include different classification algorithms, mitigation methods and strategies for omitting gender information.

The rest of the paper is structured as follows. Section 2 introduces concepts related to Fair ML and XAI. Section 3 presents the approach based on clustering and local explanations to detect and explain gender bias. The experimental methodology is described in Section 4, while results are presented and discussed in Section 5. Section 6 concludes the paper with some lines of future work.

2 Background and related work

This section introduces and relates the various concepts surrounding this work. Firstly, the field of Fair ML is presented in Section 2.1, focusing on the classification of mitigation methods and evaluation metrics. Section 2.2 elaborates on the analysis of fairness from the XAI perspective. Approaches to explaining errors made by ML models are summarised in Section 2.3.

2.1 Fair ML and bias mitigation methods

As with human reasoning, ML algorithms are vulnerable to different types of biases that can negatively affect the fairness of their automatic decisions [17]. Although different situations may occur, an algorithm is considered unfair when it favours a particular group with respect to a sensitive (also known as protected) attribute [4]. The causes of unfair ML behaviour are often related to the presence of biases in the training data, but can also arise as a consequence of how the algorithm is trained. Mehrabi et al. have characterised the types of bias that ML systems can exhibit, associating them to the source (data, algorithm, user) where the bias originates [17]. Biases in the data are due to non-random sampling, omitted variables, lack of representation of the population or inadequate aggregations, among others. Biases in algorithms result from the choice of certain optimisation and regularisation functions, as well as preference for popular or frequent items. User-created bias occurs when training is based on user-generated data, as their behaviour on platforms is influenced by social perceptions, emotions and judgements.

Several methods have been proposed to avoid algorithmic discrimination, which typically focus on identifying privileged and unprivileged groups based on a single sensitive attribute [5]. A common classification is based on when a mitigation mechanism is introduced to detect and reduce bias [17]: 1) pre-processing techniques modify the training data by applying some transformation that eliminates the bias, 2) in-processing techniques succeed in mitigating bias during the training process, often by imposing some constraints related to the sensitive attribute, and 3) post-processing techniques review the predictions made by the ML model to reassign labels if a biased decision is detected.

Bias mitigation is assessed by fairness metrics, which look for equality in the behaviour of the ML model for the privileged and unprivileged groups [23]. Two popular metrics are statistical parity difference and equalised odds difference [12], which contrast the quality of the predictions between the groups. It should be noted that improving the ML model with respect to fairness metrics could result in a decrease in predictive performance [4], so a trade-off should be sought. In addition, a recent work has highlighted the limitations of mitigation methods that optimise fairness metrics, as the reduction achieved in dissimilarities between groups does not completely eliminate biases due to other complex relationships between attributes [5]. As part of the collection of datasets for Fair ML research, Le Quy et al. analyse the dependencies of protected attributes on all other attributes (including the target one) using Bayesian networks [12].

2.2 Explaining algorithmic fairness

Since explainable methods inspect ML models to understand their behaviour and results, some authors have applied them to analyse the extent to which the models are fair. In their study on gender bias in facial emotion recognition systems, Manresa-Yee et al. used Protodash, a XAI technique to generate prototypical examples of both male and female facial expressions [15]. The prototypes obtained show the data bias present in the source dataset, with more female images associated with expressions of sadness, fear and anger. The possibilities of XAI have also recently been explored in the context of racial bias detection [14]. Here, the authors analyse feature importance and local explanations using two post-hoc explainable methods: Integrated Gradients and SHAP. Their experiment shows that these methods were barely able to identify some of the dialectal expressions associated with racial bias in the selected dataset.

Counterfactual explanations are also deeply related to fairness, as they could be used to explain how a different value of a sensitive attribute would have changed the model's prediction. The term "counterfactual fairness" has been coined to refer to this type of analysis [16]. According to this approach, a model is said to be fair if an instance obtains the same prediction within its current group or when it belongs to another group with respect to the sensitive attribute.

2.3 Explaining errors of machine learning models

Several studies have analysed the outputs of classification models and, in particular, the erroneous decisions they made. Alirezaie et al. applied symbolic reasoning on the outputs of an image classifier in order to analyse the misidentified regions [1]. The inferred objects in the image represent the explanation, but without associating them with the classifier's decision structures. In the context of a retail application, a local XAI method has been proposed to explain the instances with the highest errors during the testing phase [13]. In this case, a threshold is set to filter the instances, after which the correlation between prediction and feature changes is studied. More recently, Rizzi et al. have applied local XAI methods on each of the misclassified instances in order to determine the features that most influence such wrong predictions [20]. The frequency with which each feature appears in the misclassified instances is then studied.

Although not focused on error explanation, the work by Kim et al. [10] is also related to this proposal. The authors present a method for selecting representative instances (called prototypical examples) to explain the behaviour of an image classifier. The prototypes are accompanied by counterexamples (called "criticisms") to provide a more complete explanation to the user. More specifically, they identify parts of the dataset that differ from the selected prototypes, using a greedy algorithm. In their experiments they compare the prototype extraction method with the k-medoids clustering algorithm.

3 Gender bias analysis with clustering and XAI methods

This section explains the proposed approach, which consists of three steps: 1) Build classifiers, 2) Identify prototypes among the misclassified instances by clustering, and 3) Generate local explanations for the prototypes.

3.1 Building classifiers

This phase includes the usual procedure for training and testing a classification model. To study gender bias, several classifiers will be considered:

- A classifier trained on the full dataset (CL_{full}). The aim is to obtain a reference model for which no special consideration of gender is assumed.
- A classifier trained on the dataset excluding the gender attribute (CL_{nogen}). The aim is to generate a classification model in which gender information is omitted, in order to understand how it influences performance and the explanation of classification errors.
- Two classifiers, CL_{fem} and CL_{mal}, trained with data samples belonging to each gender. The aim is to analyse whether gender-specific models are more accurate and present different error distributions.
- A classifier with in-processing bias mitigation (CL_{mit-in}). The aim is to reduce potential bias during the training phase by exploring how it affects model performance and, consequently, the type of misclassified instances.
- A classifier with post-processing bias mitigation ($CL_{mit-post}$). The aim is to reduce the potential bias after training to make a comparison with the previous mitigation method.

Once the classifier is built, predictions are obtained in the test partition to discern between correct and erroneous predictions. Misclassified instances are identified as false positive (FP) or false negative (FN).

3.2 Clustering for prototype identification

At this stage the error analysis process begins, which is equivalent between FP and FN. If the model is not gender-specific and the gender attribute has not been omitted, the FP and FN groups will also be split by gender. The aim is to find clusters of misclassified instances for each gender in order to detect areas of the data distribution where the classifier tends to fail. A second objective is to reduce the number of instances that would eventually be presented to a user during the explanation phase, so that the analysis focuses on the most representative ones. In addition, the prototypes should correspond to instances of the dataset in order to be realistic for the user.

Due to these constraints, the choice of clustering algorithm is not straightforward. The Affinity Propagation (AP) [8] clustering algorithm has two characteristics that make it suitable for our purpose: 1) it does not need to configure the number of clusters to be discovered and 2) it locates, for each cluster, the most representative instance, the so-called exemplar. AP relies on a "message passing" strategy between instances to choose which of them is the best exemplar for other instances. At the same time, AP determines to which exemplar each of the remaining instances should be associated.

As a result of this phase, we obtain a separation of the FP and FN instances into k clusters. Note that the value of k need not be the same in both cases, and may also vary depending on the gender subgroup. For each cluster there is also an exemplar representing the prototype within the group.

3.3 Local explanations

Each prototype identified in the previous step is explained using a local post-hoc method. More specifically, Break-down is applied in this step [21] to obtain the influence of each feature on the prediction. In Break-down, the prediction value is expressed as the sum of the attribution received for each feature, which makes it easy to interpret. Once the local explanation of each prototype is generated, the feature contributions are inspected to find those more relevant, either positively or negatively affecting the erroneous prediction.

4 Experimental methodology

This section details the experimental setup, datasets and classification algorithms used to evaluate the proposed approach. For reproducibility purposes, all experiments are available on a Zenodo repository.[3]

4.1 Research questions

The following research questions (RQ) are proposed to analyse the impact of gender bias in ML misclassification:
RQ1 *Do classifiers behave differently in terms of the number and nature of misclassified instances on the basis of gender information?* To answer this question, the subsets FP and FN of each classifier will be examined. In addition, the characteristics of the clusters returned by AP will be analysed.
RQ2 *Do local explanations of the prototypes expose a gender bias for some of the classifiers?* From the local explanations of the prototypes, the feature relevance of their associated predictions will be studied to discover whether gender is a potential cause of the erroneous prediction.

4.2 Datasets

Table 1 summarises the characteristics of the datasets used for experimentation. All of them represent tabular data. The Adult and Dutch census datasets have been studied in the field of Fair ML [4, 11, 12] and are known to exhibit gender bias. For the Adult dataset, the female subgroup is the unprivileged group, whereas the male subgroup is the unprivileged group for the Dutch census dataset. Notice that the Employee promotion dataset has not been included in previous studies, but its analysis is interesting as it has a subtle bias not directly related to the gender attribute.

The table shows the number of features and instances after removing irrelevant features (e.g. id) and missing values, respectively. The last column indicates the target variable (class). A stratified data split of 30%/70% is applied to divide into training and testing partitions.

[3] https://doi.org/10.5281/zenodo.8200196

Table 1: Characteristics of the datasets used for experimentation.

Dataset name	Acronym	Num. Feat.	Num. Inst.	Female/Male Perc.	Target
Adult	ADU	14	48832	33.15% / 66.85%	income
Dutch census	CEN	11	60420	50.10% / 49.90%	occupation
Employee promotion	EMP	12	46380	30.26% / 69.74%	is_promoted

4.3 Algorithms and evaluation metrics

The Random Forest (RF) and Gradient Boosting Tree (GBT) implementations of sklearn [18] are used as classifiers for the first phase. The clustering algorithm, AP, is also available in sklearn. Default parameters will be used for all algorithms, so that more FP and FN cases can be analysed. For the same reason, no special class imbalance mechanism is introduced in any algorithm. The performance of the classifiers is presented in terms of f-measure (F1, harmonic mean between precision and recall), false positive rate (FPR) and false negative rate (FNR). For clustering, the silhouette coefficient (SC) is calculated.

The package fairlearn [3] provides implementations of both in-processing and post-processing bias mitigation methods. More specifically, the Exponentiated Gradient method is a in-processing method that trains several classifiers using incremental values of a fairness metric as a constraint. The Threshold Optimiser post-processing method applies different thresholds to the predictions returned by a base estimator, so that a fairness metric is optimised. For both methods, the default parameters are considered, with the exception of *eps* in Exponentiated Gradient method which is set to $1/sqrt(num\_instances)$. Preprocessing mitigation methods are not applied in order to use the same dataset for all classifiers. The equalised odds difference (EOD) is the selected metric to assess fairness for all classifiers, as well as the metric internally used by the mitigation methods. This metric returns the greater of the difference of the true positive rate and the difference of the false positive rate by group.

As mentioned in Section 3.3, Break-down is the XAI method executed to generate local explanations, using the implementation available on the Dalex framework [2]. The five features that contribute most positively and the five features that contribute most negatively to prediction will be studied. This way, the analysis is focused on a small but relevant subset of features.

5 Results

This section presents and analysed the experimental results for RQ1 (Section 5.1) and RQ2 (Section 5.2).

5.1 RQ1: Analysis of misclassification

Table 2 shows the performance and fairness of the classifiers using the training strategies enumerated in Section 3.1. Tables 3 and 4 summarise the clustering results for FP and FN instances, respectively. In these tables, the columns represent the number of FP/FN, the number of clusters (CL) and the clustering quality in terms of SC for each gender and classifier. As each dataset is different in terms of privileged/unprivileged groups, the analysis is performed per dataset. Overall conclusions are discussed at the end.

Table 2: Performance and fairness metrics of the classifiers.

Dataset	Training strategy	Random Forest				Gradient Boosting Tree			
		F1	FPR	FNR	EOD	F1	FPR	FNR	EOD
ADU	CL_{full}	0.666	0.078	0.378	0.083	0.672	0.057	0.108	0.108
	CL_{nogen}	0.663	0.079	0.055	0.079	0.671	0.050	0.415	0.110
	CL_{fem}	0.604	0.025	0.478	-	0.640	0.018	0.461	-
	CL_{mal}	0.673	0.117	0.358	-	0.673	0.117	0.358	-
	CL_{mit-in}	0.666	0.078	0.378	0.083	0.655	0.049	0.424	0.031
	$CL_{mit-post}$	0.675	0.104	0.323	0.100	0.658	0.219	0.168	0.021
CEN	CL_{full}	0.834	0.191	0.160	0.226	0.846	0.178	0.149	0.249
	CL_{nogen}	0.824	0.242	0.145	0.044	0.838	0.267	0.103	0.084
	CL_{fem}	0.902	0.259	0.075	-	0.909	0.244	0.067	-
	CL_{mal}	0.705	0.159	0.311	-	0.705	0.159	0.311	-
	CL_{mit-in}	0.666	0.233	0.141	0.104	0.665	0.287	0.079	0.045
	$CL_{mit-post}$	0.811	0.238	0.174	0.045	0.820	0.247	0.151	0.004
EMP	CL_{full}	0.398	0.006	0.737	0.010	0.483	0.001	0.677	0.002
	CL_{nogen}	0.416	0.008	0.717	0.004	0.484	0.001	0.676	0.007
	CL_{fem}	0.405	0.004	0.735	-	0.442	0.001	0.712	-
	CL_{mal}	0.399	0.004	0.739	-	0.399	0.004	0.739	-
	CL_{mit-in}	0.403	0.006	0.732	0.001	0.483	0.001	0.677	0.002
	$CL_{mit-post}$	0.447	0.023	0.644	0.006	0.404	0.133	0.397	0.006

Adult. The use of mitigation methods with GBT clearly improves the fairness of the models. It is also interesting that the exclusion of the gender attribute does not profoundly influence either performance or fairness. However, training a model only with female data implies a decrease in performance, while training only with male data provides similar results to the other training strategies. A larger imbalance between FPR and FNR is observed for the female group, as the percentage of high-income females (the target variable) is lower than the percentage of high-income males. Therefore, the lower performance can be attributed to a more imbalanced class distribution in the female group.

Looking at the distribution of FP by gender (Table 3), this type of error is more frequent for males than for females. The use of the post-processing mitigation method comes at the cost of a higher number of FP for both genders, especially when GBT is used. This increase is also reflected in a greater difficulty in grouping misclassified instances with clustering: the number of clusters exceeds 100 and the silhouette coefficient can drop below 0.2. This phenomenon is also observed for classifiers trained only on male data and, to a lesser extent, for those using complete data or no gender information. More cohesive clusters are found for female data when gender information is omitted or when only female instances are used for training. This might suggest that misclassification responds to similar causes for female instances, while the greater diversity of male instances makes it more difficult to identify clusters among them.

The FN results (Table 4) are similar in terms of the distribution of classification errors between male and female instances. However, it is also difficult to identify a small number of clusters for the RF results, whereas for GBT the results are fairly stable regardless of the training strategy. In addition, the mitigation methods have reversed their tendency to misclassify instances. The in-processing method has significantly increased the number of FN compared to the post-processing method, especially for male instances when using GBT.

Table 3: Clustering of FP instances by gender.

Dataset	Training strategy	RF (female)			RF (male)			GBT (female)			GBT (male)		
		FP	CL	SC	FP	CL	SC	FP	CL	SC	FP	CL	SC
ADU	CL_{full}	115	16	0.423	749	92	0.393	69	7	0.578	498	215	0.232
	CL_{nogen}	138	9	0.555	755	284	0.232	70	9	0.565	489	150	0.292
	CL_{fem}	109	7	0.581	-	-	-	77	6	0.605	-	-	-
	CL_{mal}	-	-	-	797	507	0.158	-	-	-	797	272	0.242
	CL_{mit-in}	115	16	0.423	749	92	0.393	139	9	0.508	412	192	0.227
	$CL_{mit-post}$	184	79	0.309	972	208	0.255	956	678	0.123	1482	1125	0.106
CEN	CL_{full}	766	80	0.111	881	253	0.075	750	44	0.146	788	40	0.176
	CL_{nogen}	631	41	0.142	1459	159	0.087	627	23	0.209	1675	185	0.159
	CL_{fem}	769	27	0.193	-	-	-	725	27	0.191	-	-	-
	CL_{mal}	-	-	-	899	45	0.176	-	-	-	899	69	0.116
	CL_{mit-in}	692	71	0.091	1317	397	0.080	772	58	0.163	1707	122	0.152
	$CL_{mit-post}$	725	65	0.135	1332	476	0.129	729	22	0.191	1403	339	0.181
EMP	CL_{full}	18	4	0.467	54	7	0.275	3	2	0.522	12	3	0.470
	CL_{nogen}	29	5	0.403	68	7	0.266	2	2	-	13	3	0.535
	CL_{fem}	17	4	0.546	-	-	-	5	3	0.408	-	-	-
	CL_{mal}	-	-	-	39	5	0.327	-	-	-	39	3	0.468
	CL_{mit-in}	21	5	0.367	52	7	0.522	3	2	0.522	12	3	0.470
	$CL_{mit-post}$	72	6	0.307	217	12	0.251	483	23	0.217	1207	37	0.183

Dutch census. Clearer differences in the performance of the algorithms are observed (Table 2). Training gender-specific models is more beneficial for females than for males, whether training a RF or a GBT model. For females, the FNR is significantly reduced, while for males the FNR is higher than the FPR. In terms of mitigation methods, post-processing ensures a better balance between EOD and F1, although not taking gender information into account (CL_{nogen}) yields similar results when using RF as a classifier.

The clustering of FP (Table 3) and FN (Table 4) provides further insights. Both types of misclassification are more gender-balanced in this dataset than in the previous one, but FP are still more frequent among male instances. FN are more reduced for males, as they represent the unprivileged group. Finding a reduced number of clusters becomes a difficult task for AP, returning less than 10 clusters in a reduced number of cases. It seems that a large variety of instances for both genders appear in this dataset, making it difficult to identify prototype instances that could be used to explain gender differences.

Table 4: Clustering of FN instances by gender.

Dataset	Training strategy	RF (female)			RF (male)			GBT (female)			GBT (male)		
		FN	CL	SC	FN	CL	SC	FN	CL	SC	FN	CL	SC
ADU	CL_{full}	227	89	0.315	1097	724	0.129	259	8	0.580	1188	520	0.188
	CL_{nogen}	218	73	0.332	1107	693	0.116	261	8	0.584	1193	836	0.101
	CL_{fem}	254	100	0.317	-	-	-	245	8	0.585	-	-	-
	CL_{mal}	-	-	-	1064	902	0.074	-	-	-	1064	956	0.057
	CL_{mit-in}	227	89	0.315	1097	724	0.129	204	10	0.496	1283	705	0.154
	$CL_{mit-post}$	190	37	0.340	941	681	0.098	77	9	0.575	512	317	0.205
CEN	CL_{full}	504	23	0.156	1020	283	0.016	381	17	0.239	1031	48	0.161
	CL_{nogen}	808	39	0.131	572	12	0.319	587	21	0.161	395	9	0.417
	CL_{fem}	456	21	0.198	-	-	-	412	16	0.212	-	-	-
	CL_{mal}	-	-	-	1052	51	0.151	-	-	-	1052	136	0.086
	CL_{mit-in}	644	27	0.197	691	23	0.177	394	35	0.157	360	15	0.267
	$CL_{mit-post}$	978	28	0.156	676	53	0.193	924	165	0.175	511	43	0.087
EMP	CL_{full}	310	19	0.248	590	24	0.237	282	18	0.238	545	25	0.224
	CL_{nogen}	300	18	0.239	575	26	0.221	284	18	0.236	542	26	0.228
	CL_{fem}	289	20	0.248	-	-	-	280	19	0.246	-	-	-
	CL_{mal}	-	-	-	612	27	0.227	-	-	-	612	23	0.206
	CL_{mit-in}	305	19	0.231	589	24	0.235	282	18	0.238	545	25	0.224
	$CL_{mit-post}$	270	17	0.255	516	23	0.228	164	13	0.236	321	17	0.230

Employee promotion. The high ratio of FNR compared to FPR for all training strategies indicates that the class imbalance problem affects both genders similarly. Apparently, classifiers trained on the full data are quite fair compared to the other datasets, and classifiers that omit gender information confirm that gender is not among the most relevant attributes. Both mitigation methods seem to achieve equivalent results for both RF and GBT, with the in-processing method providing slightly better results for the trade-off between F1 and fairness.

The number of FP (Table 3) is small for all classifiers and training strategies, with the exception of the combination of GBT and the post-processing mitigation method. As seen in Table 2, this combination has the smallest difference between FPR and FNR, resulting in an increase in FP especially for males. For the rest of the training strategies, AP was able to find a cohesive clustering. Better SC values are obtained for females, but this seems to be a consequence of the lower number of FP. There is even one case (CL_{nogen} with GBT) for which clustering is not possible, as there are only two FP instances in the female group. The results are more similar between genders for the FN instances (Table 4), where the number of clusters varies between 17 and 27. In contrast to other datasets, the training strategy has less impact on the FN and how they can be clustered.

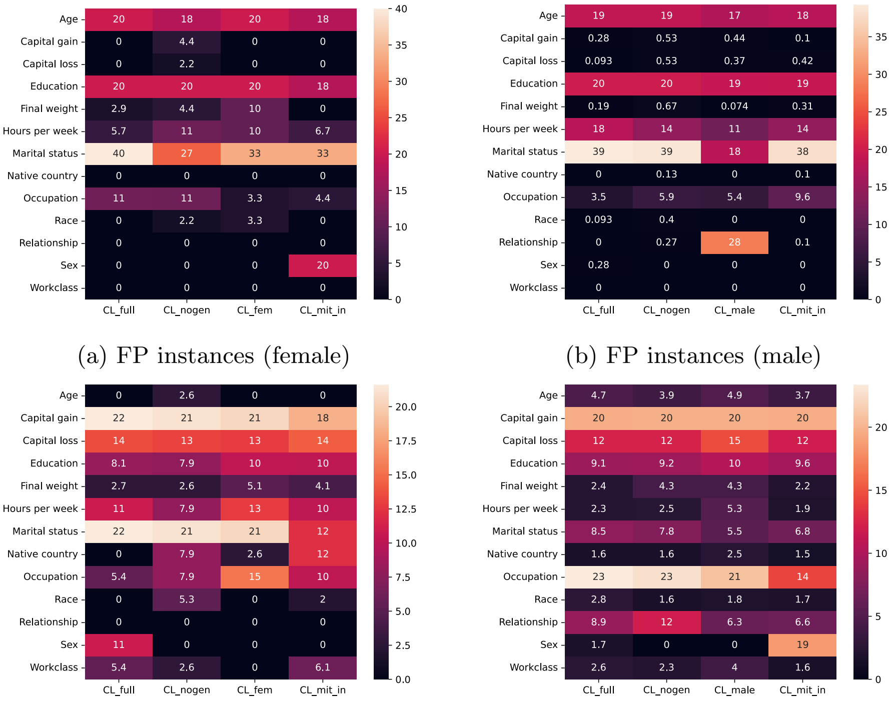
Answer to RQ1. Different training strategies have shown that classifiers can behave differently depending on how gender information is integrated into the learning process. Building gender-specific classifiers does not always guarantee fewer classification errors for the unprivileged group, as data imbalance still has a strong impact. According to the results, post-processing bias mitigation outperforms in-processing bias mitigation, but simply removing the gender attribute might even work better in some cases. The clustering of FP and FN instances with the AP algorithm has provided good results when the number of misclassified instances is less than 200. Some instability is observed when more instances need to be clustered: a similar number of misclassified instances results in a very different number of clusters, or many clusters only consist of 1 or 2 instances.

5.2 RQ2: Analysis of local explanations

Figures 1 (ADU), 2 (CEN) and 3 (EMP) show how many times (in percentage) each feature appears in the top-5 features with the highest contribution in the local explanations of the prototypes. Theoretically, each feature could appear with a maximum of 20%. While this is true for numerical attributes, the contribution of categorical attributes can be greater. In such a case, the contribution is computed as the total count of all binary attributes in which the original attribute was transformed for training. For FP instances, features with high positive contribution are considered, while for FN instances, those with high negative contribution are extracted. Due to space limitations, only the heatmaps for the GBT algorithm are included. Note that $CL_{mit-post}$ is omitted, as Break-down would use the prediction of the base estimator to generate the explanations.

Adult. Overall, we observe little variation in the relevant features for females and males in terms of FP predictions. A wider distribution of contributions for FN instances is observed for both genders, probably due to the larger number of prototypes obtained by clustering. The sex attribute was not the major contributor to FP predictions, but it did have some effect on FN predictions. Without mitigation, sex can be attributed as relevant (11% of occurrence) for erroneously predicting females as negative. When mitigation is applied, that error is shifted to males (19%). Further gender differences are observed in the FN heatmap. Females are classified as negative more often due to their marital status, while professional occupation has more impact on the decision for males.

Dutch census. The gender attribute is an important cause of misclassification in CL_{full} for females (FP) and males (FN). This confirms that males are the unprivileged group in this dataset. Although the application of in-processing bias mitigation might have reduced the influence of the sex attribute in the global



(a) FP instances (female)

(b) FP instances (male)

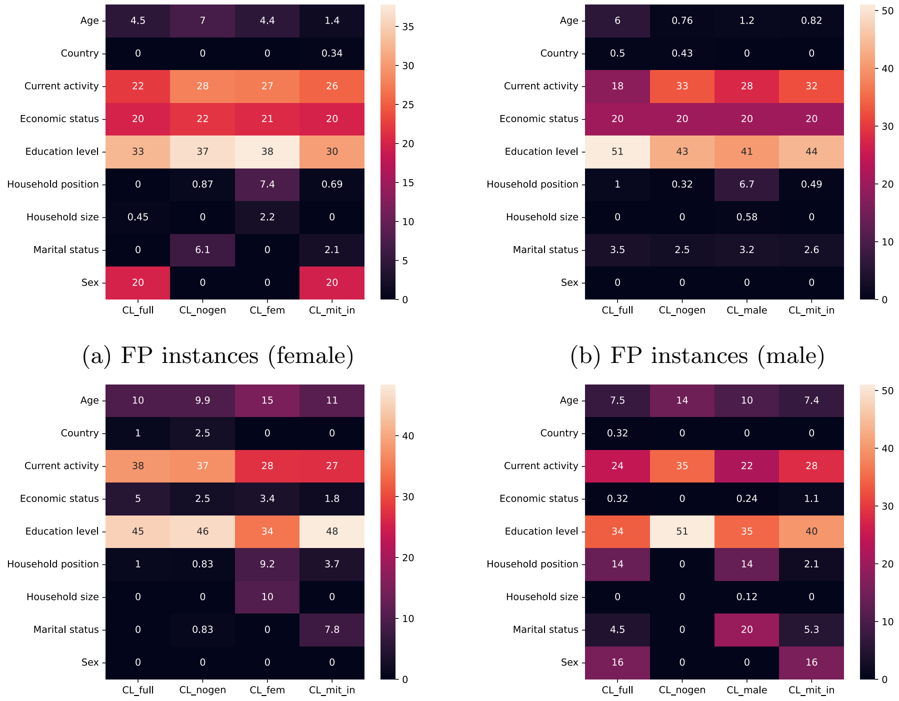
(c) FN instances (female)

(d) FN instances (male)

Fig. 1: Percentage of feature appearance in the explanations (ADU).

model, this attribute remains a reason for misclassification for several prototypes. All other attributes appear with similar occurrences in the explanations of the FN prototypes regardless of gender. However, the heatmap of FN prototypes shows a stronger influence of marital status for males when comparing the gender-specific models (CL_{fem} and CL_{male}). The explanations of the FN prototypes of the CL_{nogen} strategy also differ by gender. For males, feature contributions only focus on three attributes (age, current activity and educational level). For females, current activity and education level are also relevant, but other characteristics such as country and economic status also appear.

Employee promotion. The gender attribute does not appear among the major contributors to misclassification, irrespective of the nature of the error (FP/FN) or gender. Males and females share FP errors due to occupational aspects such as average training score, length of service and previous rating. However, FP predictions among females are more influenced by the department in which they work, especially when information on gender is omitted (CL_{fem}). This fact is even more evident for FN predictions in both genders, regardless of the training strategy applied. Exploration of the dataset reveals that the promotion rate



(a) FP instances (female)

(b) FP instances (male)

(c) FN instances (female)

(d) FN instances (male)

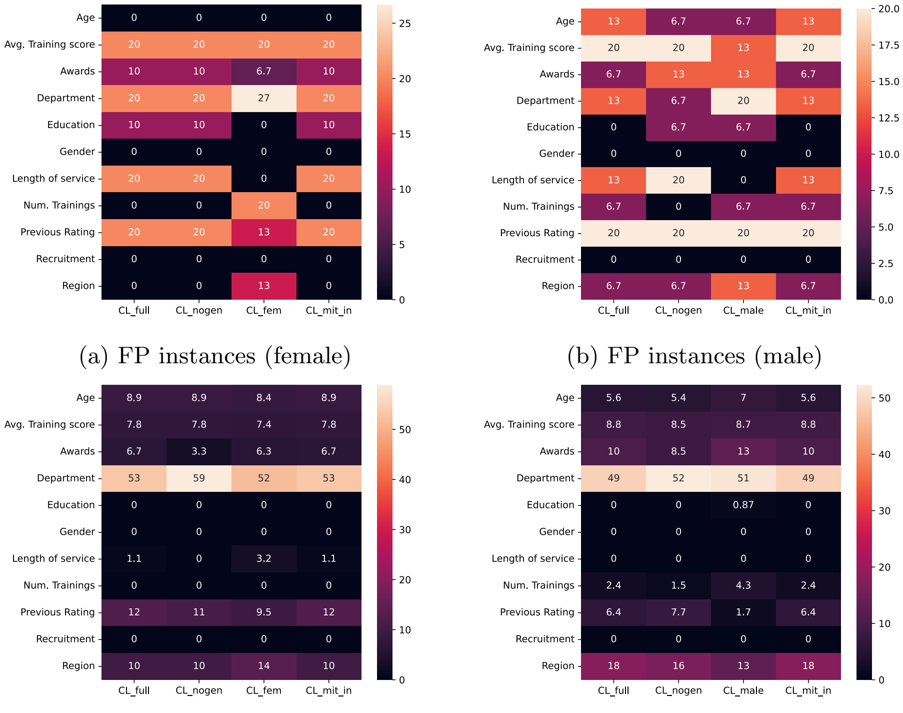
Fig. 2: Percentage of feature appearance in the explanations (CEN).

by gender can vary greatly depending on the employee's department. A similar situation occurs with the region attribute. Therefore, the prototype explanations have uncovered a gender bias not directly attributed to the sensitive feature.

Answer to RQ2. Local explanations confirm the gender bias quantified by fairness metrics. They also help to understand whether misclassification responds to different attributes within privileged and unprivileged groups. In addition, local explanations are useful for detecting additional bias patterns that escape the sensitive attribute protection adopted by mitigation methods.

6 Concluding remarks

Many ML systems are being enhanced with mitigation methods to reduce the presence of bias in the data or in the learning process. This paper has explored how to apply clustering and XAI techniques to analyse gender bias in the prediction errors made by binary classifiers. Clustering has been used to group misclassified instances, identifying prototypes of errors with similar feature values. Local explanation of such prototypes has allowed inspection of which specific features



(a) FP instances (female)

(b) FP instances (male)

(c) FN instances (female)

(d) FN instances (male)

Fig. 3: Percentage of feature appearance in the explanations (EMP).

the misclassification can be attributed to. Although gender is often considered the sensitive attribute in Fair ML, other attributes may also reveal gender bias. In such cases, the barrier between privileged and unprivileged groups is blurred, as both genders could be misclassified for similar reasons.

The proposed approach is very abstract and modular, and the application of other clustering algorithms and local explanation methods could be considered. Similarly, it could be adapted to cope with other types of data (e.g. text and images). Future work will delve into the intersection of gender bias and counterfactual fairness. Once the prototypes have been identified and explained, a natural step is to explore whether counterfactuals recommend gender-specific changes. Counterfactuals have recently been used to quantify how difficult it would be to achieve fairness by focusing on one sensitive attribute [11]. It would also be interesting to analyse whether such approaches can cope with gender biases not directly exposed by the gender attribute. Similarly, new mitigation methods might be needed to detect and correct more subtle gender biases.

References

1. Alirezaie, M., Längkvist, M., Sioutis, M., Loutfi, A.: A symbolic approach for explaining errors in image classification tasks. In: Proceedings IJCAI-ECAI Workshop on Learning and Reasoning (2018)
2. Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P.: dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. arXiv:2012.14406 (2020)
3. Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K.: Fairlearn: A toolkit for assessing and improving fairness in AI. Tech. Rep. MSR-TR-2020-32, Microsoft (2020), https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/
4. Chen, Z., Zhang, J.M., Sarro, F., Harman, M.: A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. ACM Trans. Softw. Eng. Methodol. **32**(4) (2023). https://doi.org/10.1145/3583561
5. Cheng, M., De-Arteaga, M., Mackey, L., Kalai, A.T.: Social norm bias: residual harms of fairness-aware algorithms. Data Mining and Knowledge Discovery (2023). https://doi.org/10.1007/s10618-022-00910-8
6. Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M.J., Chadha, A.S., Mavridis, N.: Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digital Medicine **3**, 81 (2020). https://doi.org/10.1038/s41746-020-0288-5
7. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., Ranjan, R.: Explainable AI (XAI): Core Ideas, Techniques, and Solutions. ACM Comput. Surv. **55**(9) (2023). https://doi.org/10.1145/3561048
8. Frey, B.J., Dueck, D.: Clustering by Passing Messages Between Data Points. Science **315**(5814), 972–976 (2007). https://doi.org/10.1126/science.1136800
9. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. **51**(5) (2018). https://doi.org/10.1145/3236009
10. Kim, B., Khanna, R., Koyejo, O.: Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability. In: Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS). p. 2288–2296. Curran Associates Inc. (2016)
11. Kuratomi, A., Pitoura, E., Papapetrou, P., Lindgren, T., Tsaparas, P.: Measuring the Burden of (Un)fairness Using Counterfactuals. In: ECML PKDD International Workshop on eXplainable Knowledge Discovery in Data Mining. pp. 402–417. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-23618-1_27
12. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. WIREs Data Mining and Knowledge Discovery **12**(3), e1452 (2022). https://doi.org/https://doi.org/10.1002/widm.1452
13. Lucic, A., Haned, H., de Rijke, M.: Why does my model fail? contrastive local explanations for retail forecasting. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*). p. 90–98. ACM (2020). https://doi.org/10.1145/3351095.3372824
14. Manerba, M.M., Morini1, V.: Exposing Racial Dialect Bias in Abusive Language Detection: Can Explainability Play a Role? In: ECML PKDD International Work-

shop on eXplainable Knowledge Discovery in Data Mining. pp. 483–497. Springer Nature Switzerland (2022). https://doi.org/10.1007/978-3-031-23618-1_32

15. Manresa-Yee, C., Ramis Guarinos, S., Buades Rubio, J.M.: Facial expression recognition: Impact of gender on fairness and expressions. In: Proceedings of the XXII International Conference on Human Computer Interaction. ACM (2022). https://doi.org/10.1145/3549865.3549904
16. Matt Kusner, Joshua Loftus, C.R., Silva, R.: Counterfactual fairness. In: 31st Conference on Neural Information Processing Systems (NIPS) (2017)
17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. **54**(6) (2021). https://doi.org/10.1145/3457607
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
19. Prates, M.O.R., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: a case study with Google Translate. vol. 32, pp. 6363–6381 (2020). https://doi.org/10.1007/s00521-019-04144-6
20. Rizzi, W., Di Francescomarino, C., Maggi, F.M.: Explainability in Predictive Process Monitoring: When Understanding Helps Improving. In: Proc. International Conference on Business Process Management (BPM). pp. 141–158. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-58638-6_9
21. Staniak, M., Biecek, P.: Explanations of Model Predictions with live and breakDown Packages. The R Journal **10**(2), 395–409 (2018). https://doi.org/10.32614/RJ-2018-072
22. Žliobaitė, I.: On the relation between accuracy and fairness in binary classification. In: Proc. ICML Workhop on Fairness, Accountability, and Transparency in Machine Learning (2015)
23. Žliobaitė, I.: Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery **31**, 1060–1089 (2017). https://doi.org/10.1007/s10618-017-0506-1