Measuring and Mitigating Bias in Bank Customers Data with XGBoost, LightGBM, and Random Forest Algorithm

Berliana Shafa Wardani, Siti Sa'adah, Dade Nurjanah
Informatics, Telkom University, Bandung, Indonesia

ARTICLE INFO

Article history:

Received February 23, 2023
Revised March 11, 2023
Published March 14, 2023

Keywords:

Bank customers data;
Bias;
Fairness;
Bias mitigation;
Bias detection

ABSTRACT

To retain its customers, Portuguese banking institutions carry out direct marketing in the form of telephone calls to conduct marketing so that customers subscribe to the bank's term deposits. This research was conducted with bank customer data from a Portuguese banking institution that implemented agent acquisition. The problem is that the large amount of bank customer data can cause data diversity which allows the results of agent acquisition to be unfair so that the features in the data must really be considered in the acquisition process. For example, gender inequality in data can cause decision results to be skewed to one group so that other groups are disadvantaged. Thus, a bias detection and mitigation algorithm is needed to achieve fairness so as to produce better predictive results. AI fairness 360 (AIF 360) is a toolkit that provides bias detection and mitigation algorithms. The bias mitigation algorithm in AIF 360 is divided into three processes, namely reweighing and learning fair representation at the pre-processing stage, debunking and debasing hostility at the in-processing stage, and classification of equalized odds and reject options at the post-processing stage. The output of this study is a comparison of the calculation of bias detection with different impacts (DI) and statistical parity differences (SPD) before and after mitigation. The adversarial debiasing algorithm performs better than others with DI 0.943, SPD -0.004, and also increases the AUC score by 0.015%. Doing this research can help predict customer deposits in Portuguese banking institutions more fairly.

Corresponding Author:

Berliana Shafa Wardani, Informatics, Telkom University, Bandung, Indonesia
Email: berlianashafa@student.telkomuniversity.ac.id

1. INTRODUCTION

In the current era, where technology is developing rapidly and is used in various aspects, one of which is in the financial sector, such as the bank. Portuguese banking institution builds customer loyalty, one of which is by conducting a direct marketing campaign via phone calls whether clients will subscribe to time deposits. With a large number of Portuguese banking institution customers, this can create opportunities for unfair acquisitions even with Machine learning (ML) involved in decision-making [1]. It is very impossible to ensure that current Artificial Intelligence (AI) models do not reflect existing bias due to the complexity of these models and their reliance on Big Data [2]. Hence, the need to ensure that automated decision-making is not biased has been a topic discussion in the AI community [3]. The term "bias" was introduced by Mitchell (1980) which means the basis for choosing a generalization (hypothesis) over another individual or group without regard to belonging in a certain group [4]. Bias can appear due to protected attributes or demographic features [5] such as location [6], gender [4], [7], [8], race [4], [8]–[11], and also age [12], [13] can cause bias in the data that lead individuals or groups to be harmed [14]. Several potential demographic features in bank customers data can caused bias, such as age, job, marital, and education. Fairness is a situation where data or decision is considered fair and there is no discrimination between individuals or groups. Nothing is truly fair in making decisions even AI can lead us to fatal outcomes and miss understanding decisions [15], ML, human decisions, and historical data [16] can lead us to biased results, therefore it is necessary to mitigate the bias. Referring to the bank customers data, fewer clients weren't subscribed the term deposit, this indicates that fairness has not been achieved. Bias can be very dangerous if not overcome because it can give wrong prediction results. In research conducted by T. Burch (2015) there is a feature that causes bias, namely race, where criminals are dominated by the black race so that it can lead to decisions that criminals are black. Here it causes the black race to be harmed. Referring to bank customer data, the fewer customers who do not subscribe to term deposits signifies that fairness has not been achieved. Based on this problem, it is necessary to observe the causes of bias in this banking dataset and to detect and mitigate bias to produce predictions that are fairer and do not harm certain groups.

There had been some prior studies on bias and fairness, including [2], [4], [17]–[19]. In 2022, Mishrakye *et al.* conducted a study regarding the bias that exists in attribute names that correlate with protected attributes. As a result, the features that have more correlation with the protected attribute are significantly biased [2]. in the same year Mosteiro *et al.* conducted bias mitigation research with ML models on mental health datasets. As a result, reweighing show a disparate impact of 0.869% and prejudice remover shows a disparate impact of 0.886%, which means that both algorithms perform well to mitigate bias [18]. In 2021, Kozodoi *et al.* implemented fairness in the banking industry through the use of credit scoring, which will fairly anticipate a person's decision to apply for a loan by mitigating the bias. According to Kozodoi *et al.*, the post-processing techniques (reject option classification) were the most cost-effective ways to increase fairness.

To mitigate bias, there is a toolkit called AI Fairness 360 (AIF 360) which is the latest and complete toolkit because AIF 360 also provides an algorithm for bias detection [17], [19]. Bias detection in the Portuguese banking institution's customers dataset was carried out using the disparate impact (DI) and statistical parity difference (SPD) methods because these methods are simple to compute. After bias detection, bias mitigation is done by pre-processing, in-processing, and post-processing. The results of DI and SPD before and after going through the mitigation process will be compared to see which method is most effective in carrying out mitigation on this dataset. This research can overcome the bias in the data and produce fair predictions for Portuguese banking institution clients to subscribe to term deposits. The contribution of this research is certainly different from previous studies related to bias and fairness because this study uses the bank customers dataset which can help provide recommendations on what features can cause bias so that bank agents can be acquired fairly and a different comparison of bias detection and mitigation methods from AIF 360. In this study, tree-based algorithms were used, namely XGBoost, LightGBM, and random forest.

2. METHODS

In this study, several stages of the process were carried out to achieve bias detection results before and after mitigation. The process in Fig. 1 begins with collecting datasets. Then feature selection is carried out on the existing dataset. At this stage protected attributes, privileged group, and unprivileged group are determined. After that, the detection process can be done by disparate impact (DI) and statistical parity difference (SPD). The next process is the process of mitigating bias by applying pre-processing, in-processing, and post-processing. Feldman *et al.*, (2010) provided a preprocessing method that does not change the training labels but alters each attribute such that the marginal distributions based on the subsets of that attribute with a particularly sensitive value are all equal [20]. A pre-processing stage consists of reweighing and learning fair representation. To reduce biases, the in-processing technique modifies the loss function during model training [17]. There is a prejudice remover and adversarial debiasing in in-processing. In order to remove bias after training, the post-processing algorithms change the output predictions. Reject option classification and equalized odds compensate post-processing. Following the mitigation phase, bias detection is done to assess the effectiveness of the mitigation strategy. The performance model can be seen by comparing the AUC score before and after mitigation.

The entire detection and mitigation process is carried out with the AI Fairness 360 toolkit [17], [19]. AI Fairness 360 (AIF 360) was developed and became one of the open-source toolkits by IBM that can be used to detect, understand, and mitigate bias in algorithms [21], [22]. In Fig. 1, you can see the research method's flowchart and each method is explained in the following subsections.

2.1. Data Collection

This study used direct marketing campaigns (phone calls) of a Portuguese banking institution dataset from www.kaggle.com. This dataset consists of 42.639 clients bank with 17 features. Table 1 provides a summary of the information.

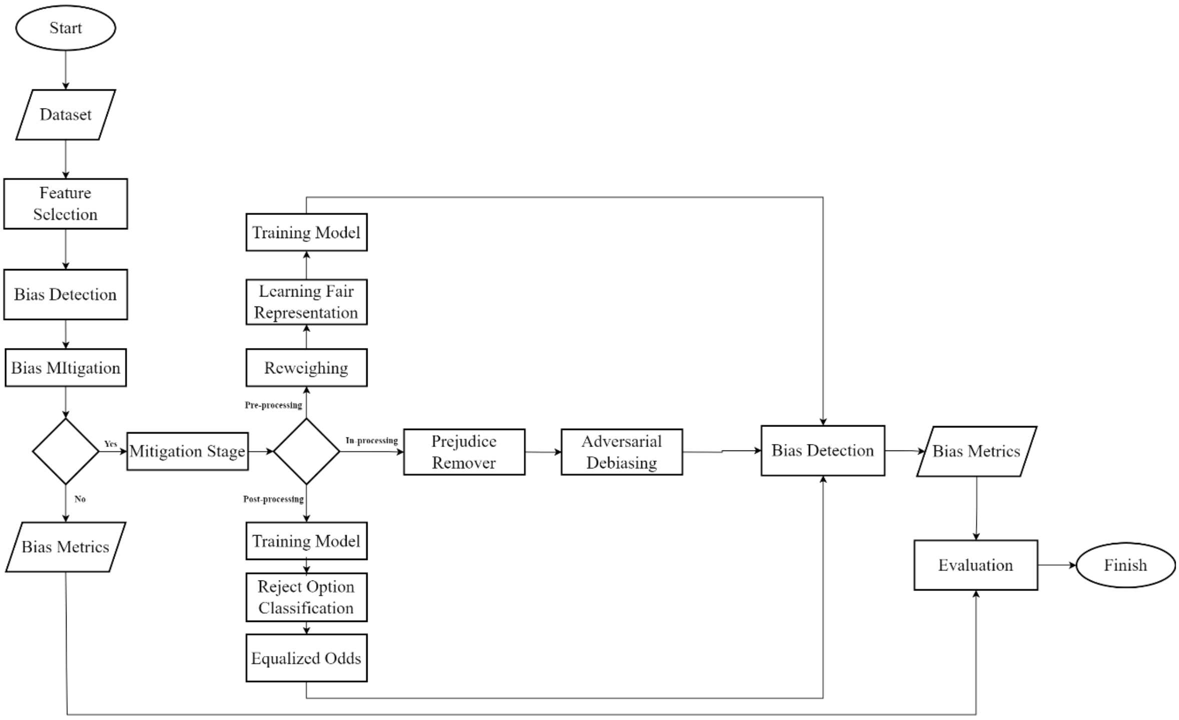

Fig. 1. Research Method's Flowchart

Table 1. Bank Customers Data

ID	age	job	marital	education	default	balance	…	term_deposit
0	58	management	married	tertiary	no	2143	…	no
1	44	technician	single	secondary	no	29	…	no
2	33	entrepreneur	married	secondary	no	2	…	no
3	47	Blue-collar	married	unknown	no	1506	…	no
4	33	unknown	single	unknown	no	1	…	no
…	…	…	…	…	…	…	…	…

This dataset has several features, the features with their distribution figures can be seen in Fig. 2. In research related to bias mitigation, it is necessary to choose which features are included in the demographic features and is classified as protected attributes. Features such as age, job, marital, and education can be features that are included in demographic features, this is because these features have population characteristics that can be used to identify various groups so that decisions can be made in favor of a particular group. The age feature in this study will be used as a protected attribute that has an uneven age distribution. In the term_deposit graph which is a label in this dataset, the number of customers who subscribe to deposits at this bank is very small.

2.2. Feature Selection

At this stage, dropping feature was performed according to [23] that dropping features from the dataset can effectively change the fairness model. Several features in the bank customers data are demographic features that can be categorized into protected attributes. The protected attribute needs to be considered because it can indicate bias. The protected attribute is grouped into two categories, namely the privileged group and the unprivileged group. Privileged and unprivileged groups must be equally involved in the process of mitigating bias to produce justice [24]. The features on the protected attribute have many different values, to classify into privileged and unprivileged groups is calculated the potential percentage is as shown in (1)

$$Potential\ percentage = \frac{(Y = 1\ |Unprivileged\ Individual)}{Total\ Individual} 100 \tag{1}$$

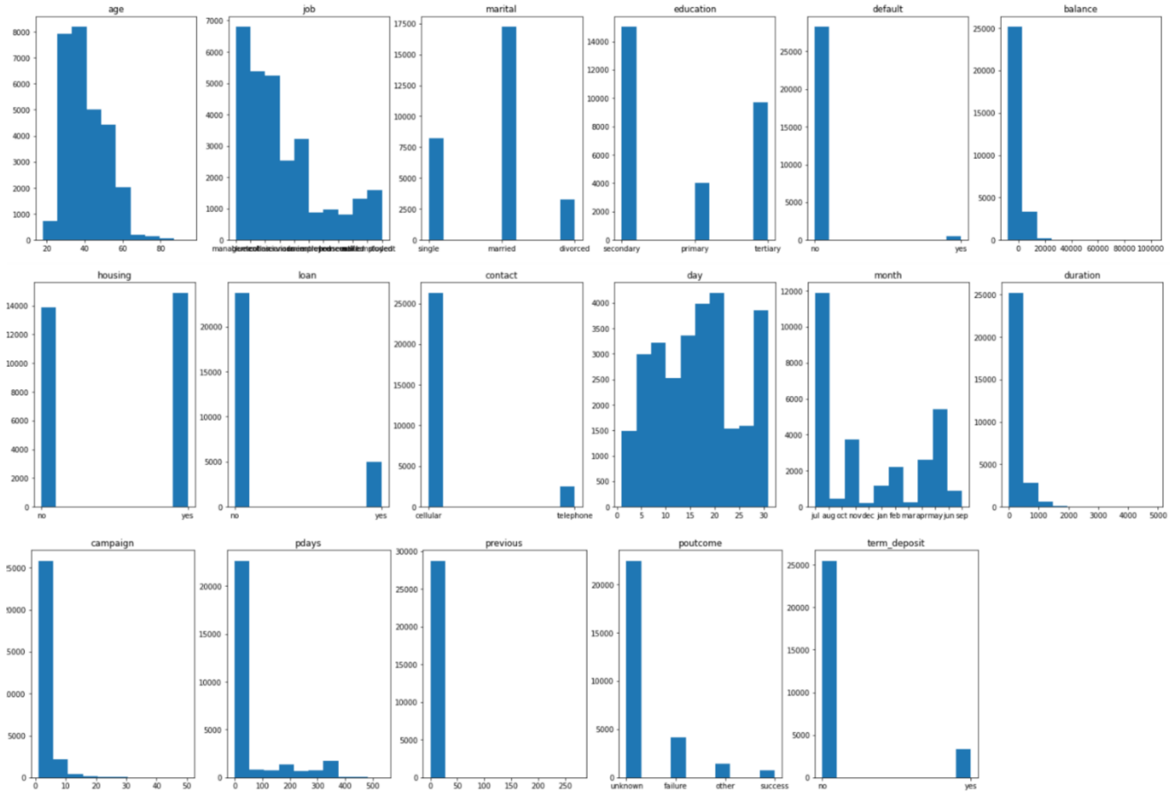

Fig. 2. Deployment of the features used

Y in (1) is a label/target, Y is positive if it has a value of 1 and negative if it has a value of 0. Based on [12], [13], the age feature is one of the demographic features that is potential to cause bias. Therefore, in this study applied the age feature as a protected attribute. The age feature divided into several age ranges and calculated the potential percentage using (1) to classify into privileged group and unprivileged group.

2.3. XGBoost

XGBoost stands for eXtreme Gradient Boosting is one of the new scalable and efficient tree-based algorithms that has gained popularity in the field of data classification [25], [26]. XGBoost can be a linear model solver or a learning tree algorithm. This algorithm can be used for regression, classification, or ranking functions, but recently it turned out to be a very effective method in data classification [26]. The package of the XGBoost algorithm is equipped with several features, such as input type, speed, sparsity, customization, and performance. XGBoost is also a scalable machine-learning technique that uses tree boosting to avoid overfitting [26]. XGBoost is based on a weighted quantile sketch (approximate tree learning for merging and pruning operations) and a sparsity-aware function (focused on zero or missing values) [27].

$$\mathrm{O} = \sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{L}(y_i, F(x_i)) + \sum_{\mathrm{k=1}}^{\mathrm{t}} R(f_k) + \mathrm{C} \tag{2}$$

XGBoost can prevent overfitting by calculating the objective function in equation (2) [28]. $R(f_k)$ is the regularization term at the k iteration time and C is a constant. The XGBoost divides the trees into levels or by depth and the tree structures are grown through repeated splits [28].

2.4. LightGBM

LightGBM is a gradient-boosting framework based on a decision tree [29] to increase model efficiency and reduce memory usage. This algorithm is designed to be as efficient as possible, with several advantages, such as faster training speed and higher efficiency, can handle large amounts of data, and has a minimal memory usage [30], [31]. The base classifiers (decision trees) were generated throughout the training process, and weight parameters were computed for each classifier in iterations.

$$f_m(X) = \partial_1 f_1(X) + \partial_2 f_2(X) + \cdots + \partial_m f_m(X) \tag{3}$$

All of the base classifiers and their weights were then integrated to create the final model, which may be described as an equation (3) [32]. From equation (3), $f_m(X)$ means the base classifier and ∂_m means the weight parameter of each classifier [32].

2.5. Random Forest

The random forest algorithm is a supervised model that combines output from an ensemble of decision trees [33]. Random forest is accurate, does not require feature scaling, or categorical feature encoding, and requires little parameter tuning. Random forests are very well used in classification or regression, outlier detection, grouping, and interpreting data sets. Random forest can control over-fitting and improve the predictive accuracy [34]. It is one of the most used algorithms due to its simplicity, flexibility, versatility, and easy-to-use supervised machine learning algorithms [35]. First, each tree is trained on the bootstrap subset hen the features for which splitting is performed at each node are not selected from all possible features, but only from a random subset and finally generate prediction [35]. To see the process random forest can be seen in Fig. 3.

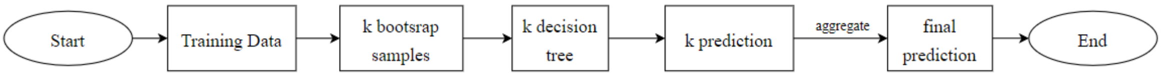

Fig. 3. The process of Random Forest

2.6. AUC Score

In measuring model performance, AUC calculations are carried out. ROC is commonly used to visualize the performance of binary classification. The ROC curve is a plot of the true positive rate (y) and the false positive rate (x) in each classification. In this curve there is an area under the curve or often called AUC (Area Under Curve) which is indicated by the blue shaded area in Fig. 4. AUC is the percentage of the area under the curve. AUC itself has limits or constraints for evaluating the model. A model that has AUC = 1 has a good model, and if the model has AUC = 0, the model is said to be bad.

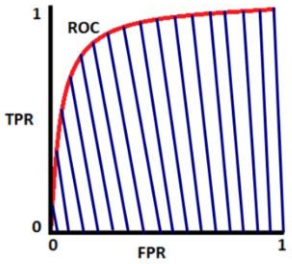

Fig. 4. AUC and ROC

To calculate AUC, formula (4) can be used. TPR or True Positive Rate is the likelihood that a real positive will result in a positive test. FPR (False Positive Rate) is the chance that a test will incorrectly reject the null hypothesis. It evaluated a test's accuracy.

$$AUC = ROC - \left(\int_0^1 TPR(FPR)dFPR \right) \tag{4}$$

2.7. Disparate Impact

Disparate Impact (DI) is one of the measurements for evaluating fairness. The concept of DI is to compare the proportion between unprivileged and privileged individuals or groups of those with positive labels [36]. Unprivileged are groups that are not benefited and privileged are groups that benefit from protected attributes [24].

A simple example illustration is shown in Fig. 5. The disparate impact calculation in (5) has a range of (0, ∞) [20]. Y is a label/target that have a binary value while D is a protected attribute. Based on US law and

Feldman, Michael, *et al.*, (2015) said that disparate impact has the 80% rule, it means a result of less than 0.8 indicates bias. Results of more than 1 mean that the unprivileged are more profitable than the privileged themselves, this refers to a negative bias condition. A measurement with a value of 0.8 indicates the creation of fair conditions, while a measurement with a value of 1 indicates demographic parity (creation of group fairness). The closer to 1, the fairer the result [37].

$$Disparate\ Impact = \frac{\Pr\,(Y = 1|D = unprivileged)}{\Pr(Y = 1|\, D = privileged)} \tag{5}$$

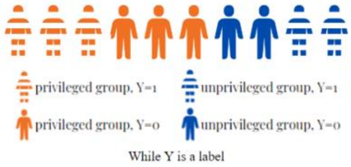

Fig. 5. Privileged and Unprivileged group in a protected attribute

2.8. Statistical Parity Difference

A technique for measuring fairness called statistical parity difference (SPD) reduces the percentage of people in two groups, the privileged group and the unprivileged group who receive favorable results [28]. The formula of the statistical parity difference is shown in (6)

$$SPD = \Pr\,(Y = 1|D = unprivileged) - \Pr(Y = 1|\, D = privileged) \tag{6}$$

Pr in (6) is the probability of D, D is a protected attribute that have value unprivileged and privileged group, and Y is a label/target. The results of calculating the statistical parity difference formula can be said to be fair if the results are closer to 0 [37].

2.9. Bias Mitigation

Bias mitigation is a process to remove unwanted bias in data. The bias mitigation in AIF 360 is divided into 3 stages, namely pre-processing, in-processing, and post-processing. Pre-processing is the bias mitigation stage before the model is trained, while post-processing is the bias mitigation stage after the model is trained. In AIF 360 in-processing, the model training process occurs simultaneously with the bias mitigation process. The pre-processing used reweighing and learning fair representation mitigation algorithms. For in-processing, the prejudice remover and adversarial debiasing algorithms are used. For post-processing, the reject option classification and equalized odds algorithms are used. Each process on bias mitigation (pre-processing, in-processing, and post-processing) will be carried out separately.

2.10. Reweighing

The Reweighing algorithm was explained by Faisal Kamiran and Toon Calders, (2012) that reweighing is a bias mitigation technique in pre-processing stage for giving different weights to each combination (group, label) to ensure fairness before classification. The reweighing process will not change the value of a feature or label [3]. In reweighing, it will be assumed that the discrimination or bias will be eliminated to 0 while maintaining positive class probabilities. To calculate the weight, the following formula of reweighing can be used:

$$w(x) = \frac{Pexp\ (s = x(s) \wedge class = x(class))}{Pobs\ \big(s = x(s) \wedge class = x(class)\big)} \tag{7}$$

In (7), $Pexp$ denotes the probability of expectation and Pobs denotes the probability of observation. The s symbol indicates a protected attribute, or protected attributes, such as gender, age, and other demographic features. While class is a label or output feature of the dataset.

Fig. 6 is the flow of the reweighing process. Data that has defined labels and protected attributes are converted into binary label datasets. A binary label dataset is a form of data that can be run on existing methods in AIF 360. After the data is changed, the reweighing process is executed after which the data that has gone through the mitigation process will be trained using XGBoost, LightGBM, and random forest.

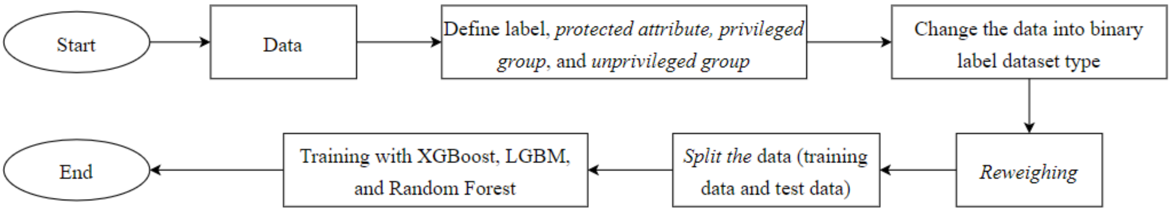
Start
Data
Define label, *protected attribute*, *privileged group*, and *unprivileged group*
Change the data into binary label dataset type
Reweighing
Split the data (training data and test data)
Training with XGBoost, LGBM, and Random Forest
End

Fig. 6. The Reweighing Process

2.11. Learning Fair Presentation

Learning fair representations find latent representations that encode data but ignore information about protected attributes. R. Zemel, Y. *et al.*, (2013) mentioned the two main objectives of learning fair representation, namely group fairness and individual fairness. Group fairness ensures that the overall proportion of members on the protected attribute who receive a positive or negative classification is identical to the proportion of the population as a whole [38]. Individual fairness is a condition that any two similar individuals must be classified together [38].

In the process of learning fair representation data will be partially taken as training data. In Fig. 7 the training data (Xo) will go through a fair representation process so that it has an output Z. Then Z will go through a prediction process so that there is a predicted output (Y) where Y is expected to be a fair prediction result.

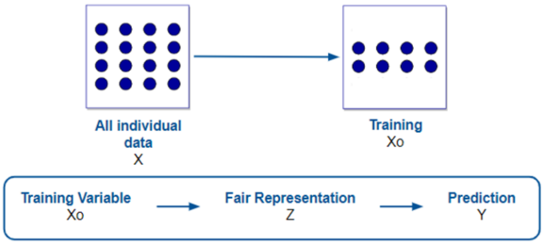

Fig. 7. Following the division of the data into training data, a procedure is used to generate Y predictions

Y is a binary variable (0/1) that represents individual classification results while Z is a variable that represents group fairness. After that, the training data will be studied by the system (8).

$$L = Az\,.\,Lz \; + \; Ax\,.\,Lx \; + \; Ay\,.\,Ly \tag{8}$$

Az, Ax, Ay in (8) are parameters that set the tradeoff of the desired system. The bias detection results will be fair when the loss results at Lz, Lx, and Ly are getting smaller.

2.12. Prejudice Remover

Fig. 8 is a simple illustrative example of high school student enrollment data. The protected attribute is in the form of gender with the unprivileged group being females and the privileged group being males. The labels in this illustration are either accepted (Y=1) or not accepted (Y=0). Because of the data in Fig. 8, females who try to apply in the science field are more often rejected, this is due to previous prejudice. Machine learning creates stereotypes that can lead to false prejudices. This situation is called prejudice bias, that the training data we have already contained (human) prejudices, including implicit racial, gender, or ideological prejudices [22].

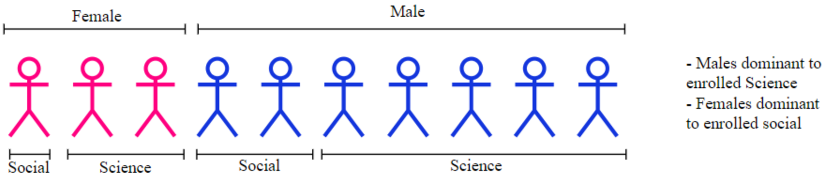
Female
Male
Social
Science
Social
Science
- Males dominant to enrolled Science
- Females dominant to enrolled social

Fig. 8. Prejudice Example

This study applied indirect prejudice. From T. Kamishima, *et al.*, (2012), indirect prejudice gives Y predictions that depend on the protected attribute. Indirect prejudice was chosen, because indirect prejudice applies the red lining effect (that ignoring sensitive features or protected attributes is not effective). The prejudice remover focuses on classifying and forming regularization with the logistic regression method. The flow of the prejudice remover process can be seen in Fig. 9.

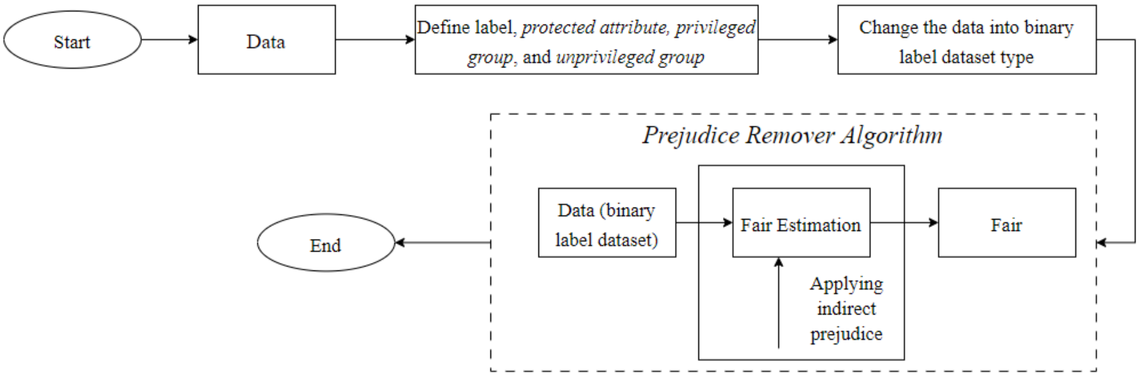
Start
Data
Define label, *protected attribute, privileged group*, and *unprivileged group*
Change the data into binary label dataset type
Prejudice Remover Algorithm
End
Data (binary label dataset)
Fair Estimation
Fair
Applying indirect prejudice

Fig. 9. Prejudice Remover Process

2.13. Adversarial Debiasing

Adversarial debiasing used adversarial training to eliminate bias from the model's latent representations. Let's say Z is a protected characteristic for which you want to avoid discrimination, like that based on race or age. Because Z can occasionally be connected with other traits, simply removing Z is insufficient. It is intended to stop the model from picking up input representations that rely on Z. To do this, a model is developed that concurrently predicts label Y and stops adversaries who were jointly trained from predicting Z.

According to the theory, the adversarial model can quickly recover and forecast Z using the original model's X representation if it contains information about Z (such as race) encoded in it. On the contrary, if the adversary is unsuccessful in obtaining any knowledge about Z, we must be successful in learning an input representation that does not heavily rely on the protected characteristic. Therefore in Fig. 10, the data will go through two processes f and a. f is a predictive function where Y = f(g(x)). The prediction of the result depends on the input data g(x), while a is an adversarial function where Z = a(g(x)) (protected attribute is predicted by the adversarial function). The model used in adversarial debiasing is a gradient-based model.

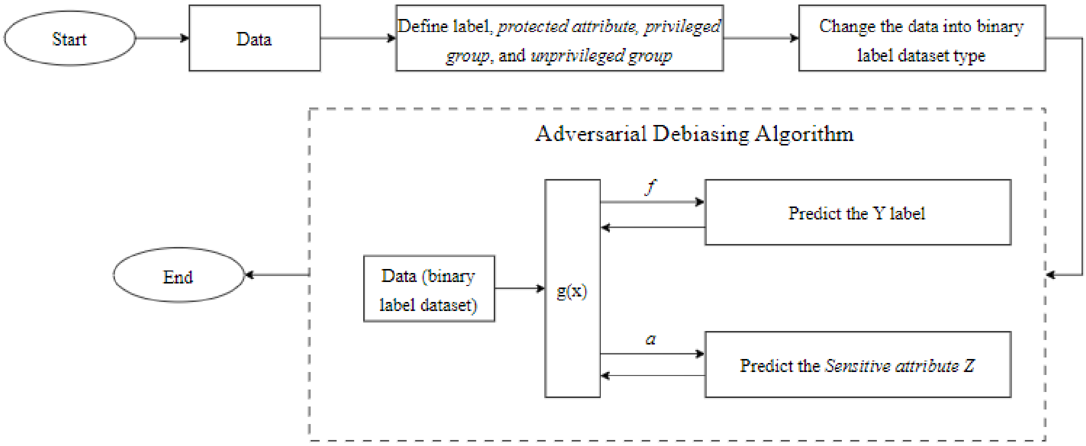
Start
Data
Define label, *protected attribute, privileged group*, and *unprivileged group*
Change the data into binary label dataset type
Adversarial Debiasing Algorithm
f
Predict the Y label
End
Data (binary label dataset)
g(x)
a
Predict the *Sensitive attribute Z*

Fig. 10. Adversarial Debiasing Process

The attached formula (9) is a gradient-based model at the target predicting stage (f) while formula (10) is used at the adversarial stage (a) in Fig. 10. Where b is a learnable scalar, $w2$ is a learnable vector, and σ is a sigmoid function. From Zhang *et al.* (2018) the benefits of adversarial debiasing include generality (may be used for many definitions of fairness), model agnostics (can be used to basic or complicated prediction models using the gradient-based technique), and Optimality (converge to satisfy the fairness definition).

$$y = \sigma(w1.x + b) \tag{9}$$

$$z = w2.[s, sy, s(1-y)] + b \tag{10}$$

Regarding prejudice remover, Thoshihiro Kashima, *et al.*, (2012) said that avoiding sensitive features on bias is not enough to overcome bias, while adversarial debiasing tries to obscure the presence of sensitive attributes. Through these differences, one can compare which method is better for mitigating bias in in-processing.

2.14. Reject Option Classification

In the reject option classification, there is an assumption that the most discrimination or bias occurs if the model has a prediction close to the decision limit of the classification threshold [39]. So, if the prediction model has the highest results, then the model must be modified [37].

Judging from Fig. 11, with a classification limit of 0.5, if the model prediction is 0.81 or 0.1, the model has clear predictive results (including bad/good) but for 0.51 or 0.49, the model is uncertain about the category or predictive results it has. So, by treating regions that have low predictive results from the classifier for reduced discrimination and rejecting their predictions, it is expected to reduce the bias in model predictions. The reject option classification (ROC) process flow can be seen in Fig. 12.

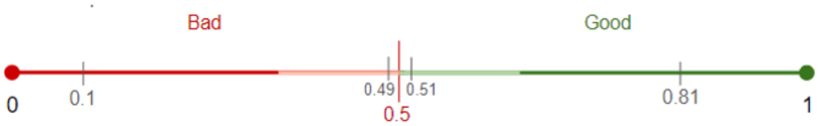

Fig. 11. Example of classification results

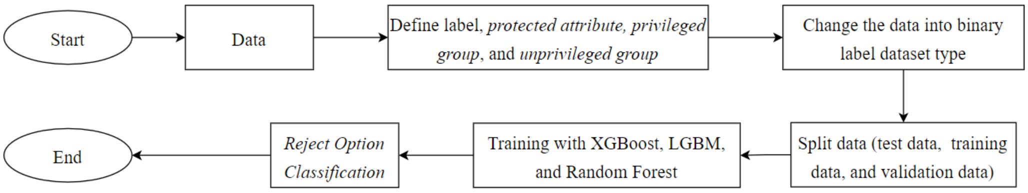
Start
Data
Define label, *protected attribute, privileged group*, and *unprivileged group*
Change the data into binary label dataset type
Split data (test data, training data, and validation data)
Training with XGBoost, LGBM, and Random Forest
Reject Option Classification
End

Fig. 12. Reject Option Classification Process

2.15. Equalized Odds

Equalized Odds is a post-processing method known as equalized odds post-processing solves a linear program to discover probabilities that can be used to modify output labels in order to maximize equalized odds. Taken from Hardt *et al.*, (2016) the conditions of True Positive Rate (TPR) and False Positive Rate (FPR) must be equal which can be written within (11) [40].

$$Pr\{\hat{Y} = 1 | A = 0, Y = y\} = Pr\{\hat{Y} = 1 | A = 1, Y = y\}, y \in \{0,1\} \tag{11}$$

From (11), it can be learned that for the result y = 1, the constraint equalizes false positive rates and for y = 0, it demands that Ŷ have equal true positive rates across the two demographics A = 0 and A = 1 while y is a label/target. Equalized odds are achieved if the equation in (11) has the same result on the right and left sides. The process of equalized odds showed in Fig. 13.

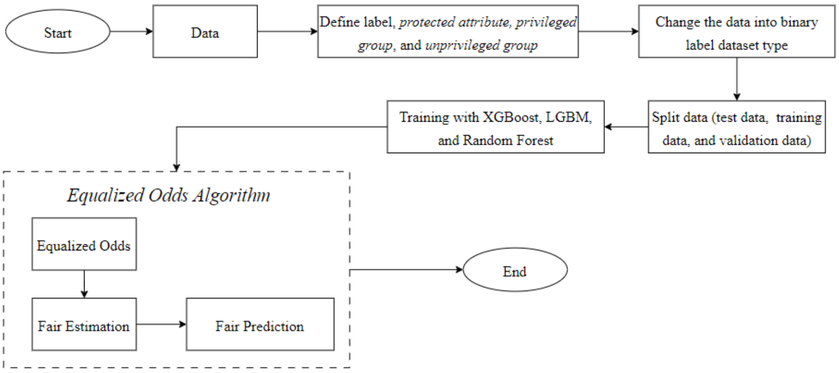
Start
Data
Define label, *protected attribute, privileged group*, and *unprivileged group*
Change the data into binary label dataset type
Split data (test data, training data, and validation data)
Training with XGBoost, LGBM, and Random Forest
Equalized Odds Algorithm
Equalized Odds
Fair Estimation
Fair Prediction
End

Fig. 13. Equalized Odds Proces

The flowchart in Fig. 13 showed that the original data will be trained with the baseline classifier in this study, namely XGBoost, LightGBM, and random forest. After the data is trained, equalized odds will be applied by calculating the formula (11) in the data so that it produces a fair prediction.

2.16. Evaluation

Pre-processing, in-processing, and post-processing outcomes of disparate impact and statistical parity differences before and after mitigation are compared throughout the evaluation stage. The findings of the detection bias for each approach will be compared to the baseline to determine whether method is more successful in this experiment and whether the bias mitigation is successful.

3. RESULTS AND DISCUSSION

Experimental results and experimental discussion are attached to this subsection. The bias mitigation experiment on the bank customers data was carried out in 3 stages, namely pre-processing, in-processing, and post-processing. The measurement results are biased in the form of calculations of the disparate impact (DI) and statistical parity difference (SPD), as well as the AUC score to see the model performance. The experimental results after and before mitigation in pre-processing can be seen in Table 2, in-processing in Table 3 and Table 4, as well as post-processing in Table 5. Age feature used as a protected attribute in the experiments.

Table 2. Bias Detection Result After Mitigation in Pre-processing

Pre-processing	**XGBoost**			**LightGBM**			**Random Forest**		
	AUC	**DI**	**SPD**	**AUC**	**DI**	**SPD**	**AUC**	**DI**	**SPD**
Original Data	0.653	1.010	0.001	0.709	1.027	0.002	0.666	0. 931	-0.005
Reweighing	0.487	2.131	0.280	0.503	2.212	0.277	0.492	2.355	0.361
Learning Fair Representation	0.508	0.737	-0.186	0.503	1.586	0.104	0.506	1.583	0.104

Based on Table 2, the reweighing and learning fair representation didn’t perform well in bank customers data calculated by DI and SPD. For DI scores on reweighing and learning fair representation (LightGBM and random forest), all are above 1 (fair limit) which indicates that the mitigation process was not executed properly. Also learning fair representations that were trained using XGBoost produced a DI that decreased from the baseline, but this figure was also outside the fair range (0.8). Meanwhile, the SPD in both methods has increased from the baseline, which keeps the SPD value away from 0. The model performance was getting worse by the decreased AUC score.

Table 3 and Table 4 will be compared to see which in-processing method is better. In-processing there is no need for a classifier, such as XGBoost, random forest, and LightGBM for model training, because the classifier is already included. Therefore, the results of bias mitigation in in-processing will be compared with the classifiers owned by the prejudice remover and adversarial debiasing methods. Bias mitigation was not successful in the prejudice remover algorithm for DI and SPD. The baseline data that was trained with logistic regression produced a disparate impact value of 0.873. Actually, this can be said to be fair, but it would be fairer if it was closer to 1. After that, the model that was mitigated using the prejudice remover produced a value of 1.357. This shows that the model is increasingly biased after mitigation. The same thing was also detected by the SPD. The baseline value of -0.005 increased to 0.015 after mitigation. The performance of the model showed by the AUC score decreased 0.013%.

Table 3. Bias Detection Result After Mitigation with Prejudice Remover

In-Processing	**AUC**	**DI**	**SPD**
Baseline (Logistic Regression)	0.604	0.873	-0.005
Prejudice Remover	0.591	1.357	0.015

Table 4. Bias Detection Result After Mitigation with Adversarial Debiasing

In-Processing	**AUC**	**DI**	**SPD**
Baseline (Without debiasing)	0.631	2.969	0.094
Adversarial Debiasing	0.647	0.943	-0.004

Debiasing is a process of reducing bias. The adversarial debiasing algorithm has a debiasing parameter that can be set to true/false. Adversarial debiasing compares adversarial debiasing with false debiasing parameters and adversarial debiasing with true debiasing parameters. The adversarial debiasing algorithm successfully mitigates the bias calculated by both DI and SPD. DI which initially had a value of 2.969 changed to 0.943 after mitigating the bias, indicating that the model experienced a reduced bias. In the SPD, the value decreased from 0.094 to -0.004 after the bias mitigation process. The SPD value is getting closer to 0 which indicated fair. Adversarial debiasing also has good model performance (AUC increased).

The reject option classification unsuccessfully mitigated the bias. The DI value in XGBoost and LightGBM has significantly increased by around 2% and in the random forest decreased to 0.690% which can be said to be far to the fair limit (Table 5). However, even though the SPD results have increased, especially in XGBoost, it can be said that it is still close to 0.

The Equalized Odds Algorithm fails to handle bias, both those calculated by DI and SPD. At the average DI value, the model becomes more biased with a value of 1.119, while SPD also increased. With the failure of bias mitigation, the AUC at equalized odds decreased.

Table 5. Bias Detection Result After Mitigation in Post-processing

Post-processing	**XGBoost**			**LightGBM**			**Random Forest**		
	AUC	**DI**	**SPD**	**AUC**	**DI**	**SPD**	**AUC**	**DI**	**SPD**
Original Data	0.653	1.010	0.001	0.709	1.027	0.002	0.666	0. 931	-0.005
Reject Option Classification	0.526	3.159	0.023	0.520	3.633	0.018	0.585	0.690	-0.011
Equalized Odds	0.610	1.034	0.004	0.657	1.230	0.029	0.627	1.093	0.011

Based on the previous experimental results, it was found that the methods that succeeded in mitigating bias was adversarial debiasing showed in Table 6. At the pre-processing stage, the reweighing and learning fair representation algorithm was not effective in reducing the bias. The learning fair representation method failed in learning because the loss results for Lz, Lx, and Ly do not get smaller, this indicates that fair conditions have not been achieved. In-processing consists of prejudice remover and adversarial debiasing. Prejudice remover method, the prejudice applied to the data fails to produce fair estimation and caused poor bias mitigation results. Whereas adversarial debiasing reduces bias quite effective, DI improve from 2.069 to 0.943 and SPD close to 0. Adversarial debiasing also Increased the AUC score from 0.631 to 0.647. In post-processing, the reject option classification and equalized odds algorithms failed to handle unbalanced data, resulting in poor bias mitigation. DI results in both methods are further away from the fair limit. As for SPD, it has increased but is still close to 0. The best result from this study is bias mitigation by adversarial debiasing.

Table 6. Experiment Analysis Result

Mitigation Algorithm		**Classiffier**	**Success to mitigate the bias? (Yes/No)**
Pre-Processing	Reweighing	XGBoost	No
		LightGBM	No
		Random Forest	No
	Learning Fair Representation	XGBoost	No
		LightGBM	No
		Random Forest	No
In-Processing	Prejudice Remover	-	No
	Adversarial Debiasing	-	Yes
Post-Processing	Reject Option Classification	XGBoost	No
		LightGBM	No
		Random Forest	No
	Equalized Odds	XGBoost	No
		LightGBM	No
		Random Forest	No

AI and machine learning play a massive role in causing bias. Mitigation bias can be applied to help acquire bank customers more fairly. Based on research [18], prejudice remover is the best in reducing bias at the in-processing stage, while in this study it is proven that adversarial is better at the in-processing stage. In addition, previous studies [2], [4], [17]–[19] have not used XGBoost, LightGBM, and random forest which are tree-based algorithms that are effective and have good performance in training a model.

4. CONCLUSION

Based on the results of research that has been done, when doing mitigation, there is a tradeoff between model performance and a fair model, this is evidenced by the decreasing AUC score in the model. In pre-processing and post-processing stages, bias failed to handle caused a bad result for DI and SPD. At in-processing, adversarial debiasing showed good performance. The results of the bias detection are close to the fair score, DI with 0.943 and SPD with -0.004. In addition, the adversarial debiasing showed an increase in AUC of 0.015%. The methods that fail to mitigate bias in this case study do not mean that the method is wrong, but it can be concluded that this bank customers data is suitable for mitigation at the in-processing stage with the adversarial debiasing method. By achieving fairness in the bank customers data, it can help determine which clients would subscribe to the term deposit by considering the existing features and can obtain Portuguese banking institution clients more evenly and hide existing demographic features. For future research, hope that all bias mitigation methods will be equally suitable for handling unbalanced datasets thereby increasing fairness while preserving the performance of the model itself.

REFERENCES

[1] J. Chakraborty, S. Majumder, Z. Yu, and T. Menzies, “Fairway: A way to build fair ML software,” in *ESEC/FSE 2020 - Proceedings of the 28th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 654–665, 2020, https://doi.org/10.1145/3368089.3409697.

[2] E. Mishraky, A. ben Arie, Y. Horesh, and S. M. Lador, “Bias detection by using name disparity tables across protected groups,” *Journal of Responsible Technology*, vol. 9, p. 100020, 2022, https://doi.org/10.1016/j.jrt.2021.100020.

[3] M. Vega-Gonzalo and P. Christidis, “Fair Models for Impartial Policies: Controlling Algorithmic Bias in Transport Behavioural Modelling,” *Sustainability (Switzerland)*, vol. 14, no. 14, 2022, https://doi.org/10.3390/su14148416.

[4] U. Peters, “Algorithmic Political Bias in Artificial Intelligence Systems,” *Philos Technol*, vol. 35, no. 2, 2022, https://doi.org/10.1007/s13347-022-00512-8.

[5] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, “Fairness under unawareness: Assessing disparity when protected class is unobserved,” in *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 339–348, 2019, https://doi.org/10.1145/3287560.3287594.

[6] K. Martinus and B. Reilly, “To boundary or not: The structural bias of ‘fair representation’ in rural areas,” *J Rural Stud*, vol. 79, pp. 136–144, 2020, https://doi.org/10.1016/j.jrurstud.2020.08.039.

[7] L. Doornkamp, L. D. van der Pol, S. Groeneveld, J. Mesman, J. J. Endendijk, and M. G. Groeneveld, “Understanding gender bias in teachers’ grading: The role of gender stereotypical beliefs,” *Teach Teach Educ*, vol. 118, Oct. 2022, https://doi.org/10.1016/j.tate.2022.103826.

[8] A. Misch, Y. Dunham, and M. Paulus, “The developmental trajectories of racial and gender intergroup bias in 5- to 10-year-old children: The impact of general psychological tendencies, contextual factors, and individual propensities,” *Acta Psychol (Amst)*, vol. 229, 2022, https://doi.org/10.1016/j.actpsy.2022.103709.

[9] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, “Demographic Bias in Biometrics: A Survey on an Emerging Challenge,” *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020, https://doi.org/10.1109/tts.2020.2992344.

[10] K. J. J. Lee and P. Setoh, “The developmental trajectories of racial categorization and explicit racial biases in Singapore,” *Acta Psychol (Amst)*, vol. 229, 2022, https://doi.org/10.1016/j.actpsy.2022.103694.

[11] M. Pryce-Miller, C. R. Pennington, E. Bliss, A. Airey, and A. Garvey, “The lived experiences of racial bias for Black, Asian and Minority Ethnic students in practice: A hermeneutic phenomenological study,” *Nurse Educ Pract*, vol. 66, 2023, https://doi.org/10.1016/j.nepr.2022.103532.

[12] B. Zhang, S. Zhang, J. Feng, and S. Zhang, “Age-level bias correction in brain age prediction,” *Neuroimage Clin*, vol. 37, 2023, https://doi.org/10.1016/j.nicl.2023.103319.

[13] D. Neal, J. L. Morgan, R. Kenny, T. Ormerod, and M. W. Reed, “Is there evidence of age bias in breast cancer health care professionals’ treatment of older patients?,” *European Journal of Surgical Oncology*, 2022, https://doi.org/10.1016/j.ejso.2022.07.003.

[14] J. M. Zhang and M. Harman, “‘Ignorance and Prejudice’ in software fairness,” in *Proceedings - International Conference on Software Engineering*, pp. 1436–1447, 2021, https://doi.org/10.1109/ICSE43902.2021.00129.

[15] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, “Addressing bias in big data and AI for health care: A call for open science,” *Patterns*, vol. 2, no. 10, 2021, https://doi.org/10.1016/j.patter.2021.100347.

[16] M. Clavero, A. García-Reyes, A. Fernández-Gil, E. Revilla, and N. Fernández, “On the misuse of historical data to set conservation baselines: Wolves in Spain as an example,” *Biol Conserv*, vol. 276, p. 109810, 2022, https://doi.org/10.1016/j.biocon.2022.109810.

[17] S. Raza, “A machine learning model for predicting, diagnosing, and mitigating health disparities in hospital readmission,” *Healthcare Analytics*, vol. 2, p. 100100, 2022, https://doi.org/10.1016/j.health.2022.100100.

[18] P. Mosteiro, J. Kuiper, J. Masthoff, F. Scheepers, and M. Spruit, “Bias Discovery in Machine Learning Models for Mental Health,” *Information (Switzerland)*, vol. 13, no. 5, 2022, https://doi.org/10.3390/info13050237.

[19] P. Cerrato, J. Halamka, and M. Pencina, “A proposal for developing a platform that evaluates algorithmic equity and accuracy,” *BMJ Health and Care Informatics*, vol. 29, no. 1, 2022. https://doi.org/10.1136/bmjhci-2021-100423.

[20] S. A. Friedler, S. Choudhary, C. Scheidegger, E. P. Hamilton, S. Venkatasubramanian, and D. Roth, “A comparative study of fairness-enhancing interventions in machine learning,” in *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 329–338, 2019, https://doi.org/10.1145/3287560.3287589.

[21] M. P. Fernando, F. Cèsar, N. David, and H. O. José, "Missing the missing values: The ugly duckling of fairness in machine learning," *International Journal of Intelligent Systems*, vol. 36, no. 7, pp. 3217–3258, 2021, https://doi.org/10.1002/int.22415.

[22] R. K. E. Bellamy *et al.*, "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM J Res Dev*, vol. 63, no. 4–5, 2019, https://doi.org/10.1147/JRD.2019.2942287.

[23] S. Biswas and H. Rajan, "Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness," in *ESEC/FSE 2020 - Proceedings of the 28th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 642–653, 2020, https://doi.org/10.1145/3368089.3409704.

[24] D. Pessach and E. Shmueli, "Improving fairness of artificial intelligence algorithms in Privileged-Group Selection Bias data settings," *Expert Syst Appl*, vol. 185, 2021, https://doi.org/10.1016/j.eswa.2021.115667.

[25] B. Li *et al.*, "A molecular classification of gastric cancer associated with distinct clinical outcomes and validated by an XGBoost-based prediction model," *Mol Ther Nucleic Acids*, vol. 31, pp. 224–240, 2023, https://doi.org/10.1016/j.omtn.2022.12.014.

[26] X. Y. Liew, N. Hameed, and J. Clos, "An investigation of XGBoost-based algorithm for breast cancer classification," *Machine Learning with Applications*, vol. 6, p. 100154, 2021, https://doi.org/10.1016/j.mlwa.2021.100154.

[27] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, "The effect of feature extraction and data sampling on credit card fraud detection," *J Big Data*, vol. 10, no. 1, 2023, https://doi.org/10.1186/s40537-023-00684-w.

[28] R. Szczepanek, "Daily Streamflow Forecasting in Mountainous Catchment Using XGBoost, LightGBM and CatBoost," *Hydrology*, vol. 9, no. 12, 2022, https://doi.org/10.3390/hydrology9120226.

[29] G. Csizmadia, K. Liszkai-Peres, B. Ferdinandy, Á. Miklósi, and V. Konok, "Human activity recognition of children with wearable devices using LightGBM machine learning," *Sci Rep*, vol. 12, no. 1, 2022, https://doi.org/10.1038/s41598-022-09521-1.

[30] I. U. Khan *et al.*, "A Proactive Attack Detection for Heating, Ventilation, and Air Conditioning (HVAC) System Using Explainable Extreme Gradient Boosting Model (XGBoost)," *Sensors*, vol. 22, no. 23, Dec. 2022, https://doi.org/10.3390/s22239235.

[31] G. Logeswari, S. Bose, and T. Anitha, "An Intrusion Detection System for SDN Using Machine Learning," *Intelligent Automation and Soft Computing*, vol. 35, no. 1, pp. 867–880, 2023, https://doi.org/10.32604/iasc.2023.026769.

[32] W. Cai, R. Wei, L. Xu, and X. Ding, "A method for modelling greenhouse temperature using gradient boost decision tree," *Information Processing in Agriculture*, vol. 9, no. 3, pp. 343–354, 2022, https://doi.org/10.1016/j.inpa.2021.08.004.

[33] G. Gazzola and M. K. Jeong, "Dependence-biased clustering for variable selection with random forests," Pattern Recognit, vol. 96, 2019, https://doi.org/10.1016/j.patcog.2019.106980.

[34] S. R. Polamuri, K. Srinivas, and A. Krishna Mohan, "Stock market prices prediction using random forest and extra tree regression," International Journal of Recent Technology and Engineering, vol. 8, no. 3, pp. 1224–1228, 2019, https://doi.org/10.35940/ijrte.C4314.098319.

[35] R. K. Paul *et al.*, "Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India," PLoS One, vol. 17, no. 7, 2022, https://doi.org/10.1371/journal.pone.0270553.

[36] A. Ashokan and C. Haas, "Fairness metrics and bias mitigation strategies for rating predictions," *Inf Process Manag*, vol. 58, no. 5, 2021, https://doi.org/10.1016/j.ipm.2021.102646.

[37] S. Radovanovic, B. Delibasic, and M. Suknovic, "Do we Reach Desired Disparate Impact with In-Processing Fairness Techniques?," *in Procedia Computer Science*, vol. 214, no. C, pp. 257–264, 2022, https://doi.org/10.1016/j.procs.2022.11.173.

[38] A. Z. Jacobs, "Measurement and fairness," *in FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021, https://doi.org/10.1145/3442188.3445901.

[39] F. Kamiran, S. Mansha, A. Karim, and X. Zhang, "Exploiting reject option in classification for social discrimination control," *Inf Sci (NY)*, vol. 425, pp. 18–33, 2018, https://doi.org/10.1016/j.ins.2017.09.064.

[40] D. Wang, Y. Mao, G. Oatley, and V. Yogarajan, "Data and model bias in artificial intelligence for healthcare applications in New Zealand," *Frontiers in Computer Science*, vol. 4, 2022, https://doi.org/ 10.3389/fcomp.2022.1070493.

BIOGRAPHY OF AUTHORS

Berliana Shafa Wardani, is a Telkom University informatics student pursuing a bachelor's degree in informatics. Her particular area of study in artificial intelligence is machine learning. Email: berlianashafa@student.telkomuniversity.ac.id

Siti Sa'adah, studied at Telkom University's Informatics Faculty for her Bachelor's and Master's degrees. She began working as a lecturer at her alma mater in 2009 and has since developed an interest in the fields of artificial intelligence and machine learning, data mining, and financial computing. Email: sitisaadah@telkomuniversity.ac.id

Dade Nurjanah defended her doctoral dissertation on the use of social computing and artificial intelligence for learning in 2013 at the University of Southampton. Dade is actively teaching undergraduate courses, including programming algorithm, knowledge representation, and algorithm complexity analysis, and also graduate courses, including introduction to social computing and research methodology. Email: dadenurjanah@telkomuniversity.ac.id.