ALEKSANDRA ZANIEWSKA

e-mail: 183534@student.ue.wroc.pl

ORCID: 0000-0002-2355-9312

Wroclaw University of Economics
and Business

FAIRNESS IN MACHINE LEARNING – BIAS IDENTIFICATION AND REDUCTION

JEL Classification: Y80

Abstract: The author investigates the problem of biases occuring in machine learning algorithms, and the strategies for their identification and mitigation. The biases are classified into three main categories: bias in data, bias in algorithms and bias generated by users. The German credit data set used in this article comes from the UCL Machine Learning Repository and represents credit risk assigned to the applicants applying for credit from the bank. The two machine learning algorithms: Random Forest and XGBoost are trained on this data set, and they are then analysed for the presence of gender bias. Subsequently, pre-processing mitigation bias techniques are used to minimize the impact of gender bias. It is identified that both algorithms have bias present and the False Negative Rate for females is the most common problem. The mitigation strategies help reduce bias but do not reduce them completely.

Keywords: machine learning, classification algorithms, bias identification, bias mitigation.

1. Introduction

The importance of technology in our lives is changing dramatically, on the one hand it allows easy communication across all the continents and time zones, makes businesses more efficient through process improvement, personalised marketing campaigns and remote working, which allows the company to source the best talents across the word. On the other hand, there are also negative consequences from the rapid digitalisation, resulting in: social disconnection, cyber-attacks, and ethical issues regarding data privacy. At the same time, computational algorithms are being used increasingly to make decisions or support decision making processes for businesses and governmental institutions. These decisions, such as credit approval, job applications, recidivism assessment most of the time have a notable impact on people's lives. Those choices can, in many circumstances, have an out of proportion influence on a distinct group of people represented in populations. On account of these, recently bias and algorithmic fairness have attracted a lot of attention from researchers and business representatives. Researchers have proven many widely used algorithms are prone to bias and unfairness towards certain groups, for example the COMPAS algorithm used to support the decision process of US Courts by predicting the risk of recidivism – the tendency of the criminal to reoffend.

Biases can occur in many different settings and are ever present in our society. The important focus is the ability to acknowledge their presence, assess the risk factor where they may occur and research their origin to understand the impact they are having. Once bias presence is known, efforts can be made to minimalize it, reduce it, or even highlight the problem to ensure it is considered during any decision-making process. As a society we cannot allow the presence of bias to be overlooked and ignored as this will deepen the social injustice and create discriminatory behaviours.

The structure of the paper is as follows. The second section introduces the definition of the bias and bias classification based on literature review. The third section addresses the mitigation strategies for bias reduction based on the literature review. In the fourth section the practical application of some of the methods of bias identification and mitigation are applied to the case study data set which is a German Credit application data. The reduction of the bias reduces the risk for the bank, as it minimizes the potential of a wrong lending decision or the consequences of commercial discrimination. It will also benefit the customer and provide a better level of service to those who are less of a risk and yet may be assessed otherwise as a result of bias. The aim of the article is to answer the following questions:

1. What type of bias are present in machine learning algorithms?
2. How to identify and mitigate the bias in case study data set?

The following research methods have been used in this article: analysis of the literature on the subject, analysis of practical documentation, scientific experiment

2. The essence of bias in machine learning

The Cambridge dictionary describes the term 'bias' as the activity of unfairly favouring or opposing a certain person or item as a result of allowing personal beliefs to affect your judgment. An alternative definition identifies bias as predistortion or prejudice towards or against one individual or a group, especially in an unjust manner (Baer, 2019). In the context of machine learning algorithms bias exist if the average of all predictions deviates from the right answer in a systematic manner. This can occur based on two factors – human error or algorithm error (Baer, 2019). The systematic error can significantly damage the reputation of businesses and public institutions, but it is particularly harmful to the individuals affected. As an example, a bias found in artificial intelligence technologies was in translation software and the dataset used to train these algorithms. Researchers found that during an article translation from English to other languages such as Spanish or Finnish most of the pronouns are converted from female to male and gender-neutral words are more likely to be translated to male pronouns (Criado-Perez, 2020). These findings prove how extremely important it is to detect and remove any biases from Machine Learning algorithms as they can cause irreparable damage and lower public trust in these technologies. To measure and prevent biases, researchers and businesses introduced tools which assess the algorithms for the unfairness, particularly focusing on any

bias toward certain social groups. This is significant in helping policy makers and scientists assess models which are being put into use. Bias can occur in three places during the process of building, training, and using machine learning algorithms. Firstly, the model can be influenced by the bias of the engineer constructing the model, the person can subconsciously influence the outcome of the model with his/her own prejudgments. Secondly, the data being used to train the model may contain multiple unfairnesses for a number of reasons, for example a sampling error. Finally, the model user can influence the algorithm or create the negative influence often without realisation it has been done.

3. Bias classification

There are many options to classify bias identification in Machine Learning, in this paper the taxonomy provided by Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2021) is reviewed, taking into account the most important represented groups. The bias will be split into three categories based on the part of the process they relate to: 1) data, 2) algorithm, or 3) user:

1. Bias in Data. These biases are found in data that is used to train Machine Learning (ML) algorithms.

a. *Reporting bias.* Reporting bias occurs during the collection or calculation of the variables and labels, then fed into the ML model to make a prediction. Often, they are present if measurement methods vary across different groups or if the accuracy of the measurements is low or disproportionate across the groups (Suresh & Guttag, 2021).

b. *Omitted variable bias.* When one or more variables are excluded from the model, this is referred to as omitted variable bias. The model will attribute the missing values to the present variables, this will usually contribute to false prediction.

c. *Representational bias and sampling bias*. Representational bias occurs during a process of sampling the data from a population while the data is being collected. The sample can create an unrealistic representation of the population, quite often under sampling certain groups of the population (Mehrabi et al., 2021).

d. *Aggregation bias.* In the data we observe aggregation bias when assumptions and generalisations are made regarding the whole observation, without considering specific features or requirements for specific groups or individuals.

2. Bias in algorithms. Bias in algorithms occurs when the outcome of the algorithm impacts the user actions, thoughts or beliefs.

a. *Algorithmic bias.* Algorithmic bias arises when the bias is entirely generated from the learning process of the model, through choice of statistical function and optimalization, without the impact of any bias from the input data.

b. *User interaction bias and popularity bias.* Both biases are mostly generated through any interface and Web services. The bias arises when the user is made to choose certain content or is presented with certain information without being made aware of the choice itself.
c. *Evaluation bias.* Evaluation bias appear when the model performance is evaluated and an incorrect benchmark is set, or when the algorithm is fit around the benchmark to achieve the highest performance and is therefore misrepresenting the population (Suresh & Guttag, 2021).

3. Bias generated by users. These biases are often represented in the data generated by users which could be affected by bias from their own believes or presented to them through interaction with the algorithm.

a. *Historical bias.* When the pre-existing inaccurate beliefs impact the data generation process, the historical bias occurs. For example, this type of bias was identified in an engine search when looking for female CEOs, they were only represented by 5% of the images (Mehrabi et al., 2021).
b. *Population bias.* Population bias appears when the targeted population differs from the actual audience, due to social, economic, or geographic characteristics.
c. *Social bias and behavioural bias.* Social bias occurs if one's judgment is impacted by general social believes or standards, this could especially be present between social classes. Similarly, when behaviour of others or their communication style leads to misunderstandings (Mehrabi et al., 2021).
d. *Content production bias.* Content production biases are generated by the social and economic background of people creating and sharing the content, their beliefs and underlying bias impacts others interacting with the content.

4. Bias mitigation strategies

The challenge of addressing bias in AI and machine learning is currently a topic growing in popularity within research communities. The fairness of the algorithms is being widely studied and different approaches to mitigate bias are being implemented. The approach to mitigating bias proposed by Mehrabi et al. (2021) divides the mitigation strategies into three categories:

1. *Pre-processing.* This strategy attempts to alter the data in order to remove inherent prejudice. This method is only an option if the algorithm can change the training data. Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian (2014) developed an approach which does not involve changes to the training labels and does not include the removal of protected variables. The method is based on the disparate impact also known as the 80% rule. The rule specifies that the protected group's selection rate must be at least 80% of the non-protected group's selection rate. This method of removing disparate impact and the possibility of predicting based on the remaining attributes, resulting in extremely different outcomes for

distinct groups relating to protected characteristics, should be avoided. The change also needs to preserve a possibility of making a prediction. The method proposed by Feldman et al. preserves valuable information about two distributions (in the event of binary classification) but at the same time makes the distinguishing between classes more difficult. The algorithms iterate through the non-protected variables and return repair features. Merging the repaired features with the original dataset generates an unbiased dataset. Another strategy presented by Calmon, Wei, Ramamurthy, & Varshney (2017) proposes an optimization strategy which involves making changes with a view to mitigate the bias. The objective is to establish a mapping P (Xˆ, Yˆ |X, Y, D). The D representing the protected variable, X is a non-protected variable and Y is presented as an outcome of the prediction. The following approaches are proposed to satisfy the mapping:

a. *Discrimination control.* This approach focuses on reducing the transformed results of Yˆ dependency on the discriminating factor D, as represented in the conditional distribution p(Yˆ|D) (Calmon, Wei, Ramamurthy, & Varshney, 2017).
b. *Distortion control.* Distortion control promoted fairness of the individual feature. This method limits the mapping to prevent large changes in the mapping. Distortion control ensures the values from the original observation and the accompanying outcome from the initial dataset are very similar to the values in the converted dataset.
c. *Utility preservation.* Utility preservations aim to make sure that the modified model does not deviate too much from the one trained on the source dataset. In order to achieve this, the distribution of transformed variables and original features need to be statistically close.

2. *In-processing.* The in-processing strategy involves making changes to the algorithm learning process to eliminate a present bias. One of the methods of in-processing mitigation is adversarial debiasing. This method increases prediction accuracy while training a model, at the same time minimizing the capability of identifying the protected attribute amongst the forecasted attributes (Bellamy et al., 2018).

3. *Post-processing.* This method can be used if neither of the strategies mentioned above are possible, for example if we are using the black-box model. After processing, the approach involves making adjustments to the test dataset (the dataset which has not been presented to the algorithm during the training process) (Mehrabi et al., 2021). Equalized odds postprocessing and calibrated equalized odds postprocessing strategies both focus on making changes to the output labels.

5. Case study application

Financial institutions and banks have, for a considerable time, been researching the credit risk and bank lending decisions and the impact of these decisions and choices. For a bank, credit is one of the biggest risk factors and despite all the efforts it's proven

a problematic risk to mitigate. Simultaneously, developments in information technology and data science have made it possible to gather more data in a shorter period of time and with a lower cost associated to the process.This way, the methods to provide deeper and more detailed data analysis have been created, which in turn has led to the creation of more secure financial services (Angelini, Di Tollo, & Roli, 2008). The data set chosen for this case study comes from UCL Machine Learning Repository and details credit requests made at the banks in Germany. When a loan application is submitted to the financial institution, the decision needs to be made based on information provided by the applicant, as to whether the loan request should be approved or declined. The bank will assign each applicant to one of these two categories: high risk or low risk. The data set has been modified to convert some of the qualitative attributes into the numerical variables and some of the attributes were dropped. The data set used contains 1000 entry points and 15 attributes presented in Table 1.

Table 1. Description of attributes present in the dataset

No.	Attribute	Description
1	status	status of existing checking account
2	duration	duration in months
3	credit_history	history of credits taken out and promptness in repayments
4	purpose	grounds on which credit wish to be taken
5	amount	total quantity of the credit request
6	savings	amount of money available in savings accounts
7	employment_duration	full length of current employment
8	sex	gender of the applicant
9	present_residence	number of years in present residence
10	property	possessions owned by the applicant
11	age	the age of the applicant
12	housing	housing arrangements
13	job	employment status of the applicant
14	foreign_worker	status confirming applicant is working abroad
15	credit_risk	confirmation if applicant status has been considered risky or not

Source: own study.

Baseline model

To investigate the performance differences and investigate if the algorithms show signs of bias, a series of experiments is planned as follows. First the baseline model of Random Forest and XGBoost performance are investigated for bias presence in regard to the protected variable age, present in German credit dataset. Next, the methods of preventing bias impacting algorithms are applied to both methods and comparisons made to evaluate the effect. In terms of performance both models will

bevaluated based on accuracy measure (ACC) which is calculated by dividing the number of correctly made predictions by all the samples in the dataset and the F1 score which is calculated by accounting for both the precision and recall of the model.

Bias identification and mitigation approaches to the case study

Literature presents many ways to identify bias in the classification problem. In this paper the author elected to focus on one protected variable which is gender, therefore further analysis will be focused on gender bias. The method of assessing the bias will be based on classification metrics. Taking into account the correctly made predictions: True Positive, True Negatives and wrongly made classifications False Positive and False Negatives as proposed by Zafar, Valera, Gomez Rodriguez, & Gummadi (2015), and Hardt, Price, & Srebro (2016). The actual values and predicted values are assessed based on predicted variable and measures are compared between males and females. In the first step we will assess the data set and any significant imbalances present, which could lead to model unfair predictions. In the second step we will calculate the classification metrics for both genders and algorithms used in a previous classification as proposed by Viglid & Johansson (2021). Almost always we cannot completely discard the presence of the bias, most of the data set will be impacted by it even if the protected variable is not present. To measure the impact, the 80% rule proposed by Feldman et al. (2014) and used by Viglid & Johansson (2021) will be applied. To apply the 80% method, the classification metrics for females will be divided by the classification metrics for men. If the results follow the 80% rule, we should expect a value in a lower limit of 0.8. If the ratio between the measures is significantly above the limit value, there is a high probability of the classifier being biased in terms of this metric. It is recommended to use multiple mitigation methods if the risk of biased classification is high.

In the second step, classification metrics are calculated for both previously used algorithms: Random Forest and Extreme Gradient Boost, the parameters of the algorithms are unchanged. The calculation is performed separately for females and males. The results of classification metrics, including accuracy, are presented in Table 2 for both classification algorithms.

Table 2. Classification metrics grouped by gender

		TPR	FPR	TNR	FNR	Accuracy (%)
Random Forest	male	0.93	0.72	0.28	0.07	73
	female	0.76	0.54	0.46	0.24	69
XGBoost	male	0.92	0.65	0.35	0.08	75
	female	0.78	0.48	0.52	0.22	71

Source: own study.

Random Forest algorithm performs with a higher accuracy of 73% for men, as compared to 69% for women. The model predicts the positive classification better for men as opposed to women, this means the algorithms are more likely to correctly

classify men who are actually low risk group in terms of credit application to a low--risk category. Women have significantly higher scores for True Negative Rate and False Negative Rate; therefore, they are more likely to be correctly classified as high risk but also more likely to be falsely classified as high risk even if in reality they are a low-risk category.

The XGBoost algorithm is following the same patterns as Random Forest in its classification. The accuracy of the model for both women and men are higher than in Random Forest, however, the accuracy for women is lower than that for men. Although XGBoost performs very well in truly classifying men into low--risk category, the algorithm has significantly higher number of false negative classifications for women in comparison to men. This misclassification creates a higher risk for women, who are more likely to score high risk and be potentially rejected for credit application even if their actual credit score is low risk.

The next step involves calculating the relationship between male and female classifications to apply the 80% rule. The results of the calculation are presented in Table 3 for both algorithms. The False Negative Rate has the highest results for both algorithms. The True Positive Rate is the only metric falling within the 80% rule.

Table 3. Relationship scores between classification metrics (80% rule)

	TPR	FPR	TNR	FNR
Random Forest	0.81	0.75	1.64	3.67
XGBoost	0.85	0.74	1.48	2.66

Source: own study.

Having identified the presence of gender bias in the credit application data set the next step involves the selection of mitigation strategies. The author decided to focus in this article on the pre-processing strategies to mitigate bias.

The first strategy proposed in the research paper by Feldman & Peake (2021), implies fairness connected to lack of awareness. This method suggests not using the protected attributes to make a classification, on that basis the predicated variable cannot directly impact the algorithms. On the other hand, it is important to highlight that some variables can be indirectly impacted by gender bias, for example the number of women in senior executive roles. Table 4 represents the classification metrics for Random Forest and XGBoost when the sex variable is removed.

Table 4. Classification metrics grouped by gender after protected variable removal

		TPR	FPR	TNR	FNR	Accuracy (%)
Random Forest	male	0.95	0.73	0.27	0.05	75
	female	0.70	0.54	0.46	0.30	63
XGBoost	male	0.92	0.65	0.35	0.08	75
	female	0.80	0.50	0.50	0.20	71

Source: own study.

After removal of the gender variable from the data set the accuracy for XGBoost remained unchanged, however, the accuracy metrics for Random Forest have increased for the data set representing male information and decreased for the data set which has female information. There is no significant change in patterns between positive and negative classifications between the groups, the false negative classifications are still dominant in female represented information data.

In the next step the relationship between the classification metrics is calculated and the 80% rule applied to evaluate if any change in scores can be observed after the protected variable has been removed. The results of the calculation are shown in Table 5.

Table 5. Relationship between classification metrics after protected variable removal

	TPR	FPR	TNR	FNR
Random Forest	0.74	0.75	1.67	6.02
XGBoost	0.87	0.77	1.42	2.49

Source: own study.

After removal of the gender variable, the False negative rate has dropped slightly for XGBoost model, however, increased significantly for Random Forest. The negative metrics are still far away from being within the 80% rule. Based on this calculation the removal of the protected variable failed to completely remove the gender bias in this case study. However, the author suspects the unbalanced representation in the data set could be impacting the predictions.

The second pre-processing method of mitigating gender bias to be applied in this paper was proposed by Feldman et al. in 2014 and focuses on minimizing the disparate impact. The method is applied in pre-processing and improves fairness by changing the original data distribution (Feldman & Peake, 2021). To apply this method, the author decided to use open-source Python package AIF360 which is based on the IBM model. The tool used is Disparate Impact Remover, which will change values of features to increase fairness. Table 6 presents the classification metrics after using the AIF 360 for both machine learning methods.

Table 6. Classification metric after using AIF 360

		TPR	FPR	TNR	FNR	Accuracy (%)
Random Forest	male	0.97	0.67	0.33	0.03	81
	female	0.80	0.48	0.52	0.20	69
XGBoost	male	0.91	0.61	0.39	0.09	78
	female	0.68	0.52	0.48	0.32	61

Source: own study.

After applying the Disparate Impact Remover from AIF 360 package, we can observe the increase in the accuracy for male classification. However, we can still see the disproportionate differences between positive and negative classification for females and males. Table 7 presents the relationship between the classification metrics while using the AIF 360 method. The observed change is not significant; however, we can see the improvement in True Negative Rate for Extreme Gradient Boosting, now falling within the 80% bracket. In spite of this the False Negative ratio is still very high for both algorithms.

Table 7. Relationship between the classification metrics for AIF360

	TPR	FPR	TNR	FNR
Random Forest	0.83	0.71	1.57	6.70
XGBoost	0.75	0.86	1.22	3.63

Source: own study.

The author decided to combine two methods of pre-processing as the third method of mitigating bias. The first pre-processing method involves the use of the Synthetic Minority Oversampling Technique (SMOT) to manage the imbalance of the data set. After this problem is addressed, the AIF360 is used once again to see if the combined methodology has managed to mitigate the bias in this data set. The decision to combine these two methods is based on the fact that the data is significantly imbalanced, and this could prevent the methods working as they are designed to. Results of the classification metrics for the combined method of SMOT and AIF360 are shown in Table 8. The significant improvement in accuracy for the female data set is observed in both Random Forest and XGBoost models, moreover the accuracy for males has not shown any significant drop. In addition, the True Positive rate has increased for the female data set and the differences between false negative rates are significantly smaller compared to the previously used methods in this article.

Table 8. Classification metrics after using SMOT and AIF360

		TPR	FPR	TNR	FNR	Accuracy (%)
Random Forest	male	0.95	0.75	0.25	0.05	77
	female	0.90	0.14	0.86	0.10	86
XGBoost	male	0.91	0.61	0.39	0.09	78
	female	0.88	0.17	0.83	0.12	88

Source: own study.

The combined method produced results closest to the optimum outcome, achieved so far. The values for XGBoost and Random presented in Table 9, represent the metrics closest to fully mitigating the bias in this data set. The values of FNR from

both algorithms are very close to the 80% threshold. The XGBoost achieved better results for making less biased classifications with this method. With this highest success rate, this method highlights the importance of a balanced data set and a full representation of the population. If it is not possible to balance the data set or if the balancing would have a negative impact on the classification, it is important to communicate and seek alternate resolutions to mitigate the bias as the outcome can be harmful to the individual, as shown with the German credit data set example.

Table 9. Relationship of classification metrics after using SMOT and AIF360

	TPR	FPR	TNR	FNR
Random Forest	0.94	0.19	3.44	2.12
XGBoost	0.97	0.28	2.14	1.35

Source: own study.

6. Conclusions

The development of artificial intelligence and machine learning has had a very positive impact on everyday life. Furthermore, the progression of machine learning has improved business processes significantly by allowing companies to automate many of the repetitive tasks. This has also been implemented to support decision making or make direct decisions. While making these suggestions or decisions machine learning algorithms can be affected by bias, which can act as a prejudice towards one or more of the groups represented in the data, working for or against the group. This paper provided a detailed classification of biases which can occur in machine learning. The three main categories account for bias generated by algorithms, bias generated by user and bias generated by data. The presence of any one category does not discount the presence of bias from other categories. Since the multiple sources can affect an algorithm, it is extremely important to identify if bias is present in a dataset during the pre-processing stages and if the algorithm outcome has been affected by bias. When bias is identified, it is necessary to investigate further the reason behind it and to make every effort to mitigate it, even if the cost is loss in the performance measure of the models. The mitigation strategy depends strongly on the type of bias and the way it is impacting the algorithm.

The author identified the presence of bias and implemented appropriate mitigation strategies in the German Credit Application Data set. The limitation of the data set is the age of it, which could make it less relevant as a result of social--economic changes in society. The author decided to focus on one protected variable present in the data – gender. The first observation was the high imbalance of the genders represented in the data set, females were strongly underrepresented, making up less than 50% of the data set. This proved the presence of representation bias and

presented the risk of the algorithm being affected by it. The second observation was that females are more likely to be classified to high-risk category even if their actual class label is low risk. This wrongful classification could be particularly harmful and can have a direct impact on the rejection of the credit application. In this article the author used the pre-processing mitigation to address the gender bias present in the German credit data set. The strategies used removed the protected variable from the data set completely to check if this could improve the bias, however, the effect was not as expected. The author believes this could be partially due to the gender variable also indirectly affecting other variables present in the data set. The second strategy involved using the AIF 360 algorithm which was constructed to mitigate bias and serve as open library in Python. For the third strategy the author decided to first address the imbalance of the data set by synthetically creating additional samples for the underrepresented class, in this case females, and then applied the AIF 360 algorithm to the data set. This strategy has proven the most successful with the closest gap between the predictions for two genders.

References

Angelini, E., Di Tollo, G., & Roli, A. (2008). A neural network approach for credit risk evaluation. *The Quarterly Review of Economics and Finance*, *48*(4), 733-755. DOI: 10.1016/j.qref.2007.04.001

Baer, T. (2019). *Understand, manage, and prevent algorithmic bias: A guide for business users and data scientists*. Berkeley, CA: Apress. DOI: 10.1007/978-1-4842-4885-0

Bellamy, R. K. E. et al. (2018). *AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*. https://doi.org/10.48550/arXiv.1810.01943

Calmon, F. P., Wei, D., Ramamurthy, K. N., & Varshney, K. R. (2017). *Optimized data pre-processing for discrimination prevention*. https://doi.org/10.48550/arXiv.1704.03354

Criado-Perez, C. (2020). *Invisible women: Exposing data bias in a world designed for men*. London: Vintage.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2014). *Certifying and removing disparate impact* (KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining). https://doi.org/10.1145/2783258.2783311

Feldman, T., & Peake, A. (2021). *End-to-end bias mitigation: Removing gender bias in deep learning.*

Hardt, M. Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning*. https://doi.org/10.48550/arXiv.1610.02413

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *EAAMO '21: Equity and access in algorithms, mechanisms, and optimization*. New York: Association for Computing Machinery. https://doi.org/10.1145/3465416.3483305

Vigild, D., & Johansson, L. (2021). *Identifying and mitigating bias in machine learning models*.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2015). *Fairness constraints: Mechanisms for fair classification*. https://doi.org/10.48550/arXiv.1507.05259

UCZCIWOŚĆ W UCZENIU MASZYNOWYM – IDENTYFIKACJA I REDUKCJA UPRZEDZEŃ

Streszczenie: Autorka badała, dlaczego w algorytmach uczenia maszynowego występują błędy systematyczne i jakie są strategie ich identyfikacji i łagodzenia. Błędy dzielą się na trzy główne kategorie: stronniczość danych, stronniczość algorytmów i stronniczość generowaną przez użytkowników. Dane użyte w tym artykule pochodzą z UCL Machine Learning Repository i reprezentują ryzyko kredytowe przypisane wnioskodawcom ubiegającym się o kredyt w banku. Dwa algorytmy uczenia maszynowego: Random Forest i XGBoost, są uczone na tym zbiorze danych i analizowane pod kątem obecności uprzedzeń związanych z płcią. Następnie stosuje się wstępne techniki łagodzenia uprzedzeń, aby zminimalizować ich wpływ. Stwierdzono, że oba algorytmy mają błąd systematyczny, a najczęstszym problemem jest wskaźnik wyników False Negative dla kobiet. W artykule zastosowano następujące metody badawcze: analiza literatury przedmiotu, analiza dokumentacji praktycznej, eksperyment naukowy na danych ze studium przypadku.

Słowa kluczowe: uczenie maszynowe, algorytmy, klasyfikacja, identyfikacja stronniczości, łagodzenie wpływu stronniczości.