Author(s): Saarela, Mirka

Title: On the relation of causality- versus correlation-based feature selection on model fairness

Year: 2024

Version: Published version

Please cite the original version:

Saarela, M. (2024). On the relation of causality- versus correlation-based feature selection on model fairness. In SAC '24 : Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (pp. 56-64). ACM. https://doi.org/10.1145/3605098.3636018

On the relation of causality- versus correlation-based feature selection on model fairness

Mirka Saarela
University of Jyväskylä
Jyväskylä, Finland
mirka.saarela@jyu.fi

ABSTRACT

As machine learning models are used increasingly in the educational domain, ensuring that they are fair and do not discriminate against certain groups or individuals is imperative. Although there are a few recent attempts to ensure fairness in these models, the majority of fairness literature tends to overlook the feature selection (FS) process despite its critical role as one of the foundational steps in the machine learning pipeline. Moreover, traditional FS methods identify features by examining the correlational relationships between predictive features and the target variable without seeking to uncover causal connections between them. To address these issues, we compare for four openly available datasets—two educational ones and two benchmark datasets regularly used in the fairness literature—the impact of these two different ways of FS (i.e., causality- versus correlation-based) on the performance and fairness of the resulting models. Our results show that causality-based FS generally leads to fairer models, while the models built after correlation-based FS manifest higher performance.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; *Causal reasoning and diagnostics*; • **Applied computing** → Education.

KEYWORDS

Feature Selection, Causality, Markov Blanket, IPCMB, Machine Learning Fairness

ACM Reference Format:
Mirka Saarela. 2024. On the relation of causality- versus correlation-based feature selection on model fairness. In *The 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24), April 8–12, 2024, Avila, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3605098.3636018

1 INTRODUCTION

Ensuring inclusive and equitable quality education and promoting lifelong learning opportunities for all is one of the key sustainability development goals [44]. However, access to superior educational resources remains skewed in favor of more privileged learners. Furthermore, global assessments highlight a concerning shortage of educators, leaving them increasingly strained and fatigued. The onset of the COVID-19 pandemic and subsequent school closures have further exacerbated these challenges [23, 36, 39]. Artificial intelligence (AI) in education can be part of the solution to overcome these problems: It can unburden teachers, enable learners and educators to access specialized materials well beyond textbooks in multiple formats that bridge time and space, and deliver quality education for all, personalized and at scale [38]. With this, the application of machine learning algorithms is necessary to extract beneficial information from large educational data sets with multiple modalities, support automatic decision-making, and provide appropriate content to the learners. For instance, personalized learning systems can provide instructions in mixed-ability learning groups, chatbots can provide students with detailed and timely feedback on their writing products, and automated assessments can free teachers from some repetitive work and give them more room to support their students [21, 47].

However, despite the remarkable results achieved by current educational AI models, ensuring their fairness remains a significant challenge [5, 29]. Fairness, in this context, pertains to treating individuals or groups equitably, without any bias or favoritism based on their inherent or acquired characteristics, particularly within decision-making processes [30, 37]. As in other domains, educational applications and tools driven by machine learning algorithms carry the potential for ethical challenges [1]. Not only can bias in the real world "creep" into AI systems [37]. Even if the underlying data itself is unbiased, the behavior of algorithms can exhibit bias based on certain design choices [30]. Moreover, unlike low-stakes applications, such as a Netflix model predicting movie preferences, deploying machine learning models in education often involves high-stakes decisions, such as determining a student's admission to a study program or eligibility for a scholarship [41]. A recent review on fairness in educational models points out that ethical challenges have been identified in various dimensions, encompassing student attributes like race, ethnicity, nationality, gender, native language, urbanicity, parental educational background, and socioeconomic status [5]. Thus, with the proliferation of AI and automated machine learning models in education, it has become crucial to prioritize the fairness of these models.

Feature selection (FS) is one of the most commonly applied preprocessing or data-transformation/-generation techniques in the machine learning pipeline before training a model [19]. It involves the task of pinpointing and choosing a subset of input features that hold the highest relevance to the target variable. Using it offers numerous advantages, including aiding in data visualization and comprehension, decreasing the need for extensive measurement and storage, reducing training and processing times, and mitigating the challenges of high-dimensional data to enhance prediction accuracy [18, 28]. However, despite it being such an essential step in the machine learning pipeline, according to Galhotra et al. (2022) [16],

SAC '24, April 8–12, 2024, Avila, Spain

ACM ISBN 979-8-4007-0243-3/24/04.
https://doi.org/10.1145/3605098.3636018

the majority of the fairness literature neglects it. In addition, traditional FS methods identify features by examining the correlational relationships between predictive features and the class variable without seeking to uncover causal connections between them [48]. Thus, an intriguing topic to explore is how the causality-based versus correlation-based FS relates to the performance and fairness of the resulting models.

Especially in the educational domain, recent machine learning fairness articles emphasize the importance of using causal algorithms as future work [9, 29]. Typically, correlations merely indicate the co-occurrence of features without capturing their causal relationships with the class variable, but research indicates that incorporating causal features in FS for classification can offer two significant potential advantages: First, incorporating causal features can enhance the resilience and robustness of classification models as causal relationships indicate the fundamental mechanism behind the class variable, making them consistent across various settings or environments [4, 34, 40]. Second, causal features can potentially enhance the explanatory power of classification models [32]. Correlations only capture the simultaneous occurrence of features and the class variable. Consequently, the selected features often fail to provide a compelling explanation for predictions. For instance, a strong correlation between a pupil's height and their mathematics skills might be observed in one primary school. This observation might suggest that height is a significant predictive feature of a pupil's mathematics skills. However, it clearly is not a reasonable explanation for mathematical skills. In reality, factors like age serve as more plausible and understandable causes of mathematical skills, and such predictors will also be more robust if the model is applied in another school.

To address these issues, this article analyzes four openly available datasets—two specifically from the educational domain and two well-known benchmark datasets from the general fairness literature—and examines how causality- versus correlation-based FS relates to the performance and fairness of the resulting models. More specifically, we compare the traditional correlation-based filter FS technique with a specific causality-based filter FS algorithm, which came out on top in a recent evaluation of causality-based FS algorithms [48].

2 THEORETICAL FOUNDATIONS

In this section, we provide the theoretical foundations for the empirical experiments. First, we explain the main concepts of FS. Second, we give a short introduction to the discovery of the Markov blanket, which is needed for causality-based FS. Third, we summarize current fairness metrics that can be used to assess the fairness of machine learning models.

2.1 Feature selection

FS involves identifying and selecting a subset of input features that exhibit the highest relevance to the target variable. FS techniques currently in use can be divided into three main categories: wrappers, embedded methods, and filters [15, 18, 28]. Wrapper methods conduct an exhaustive search through potential combinations of features. They evaluate each subset by employing the target learning algorithm as a black box [25]. Wrapper FS approaches can be computationally demanding, as model training and cross-validation must be performed for each feature subset, and the results are tailored to a specific model. Embedded methods carry out FS intrinsically as part of the training process and are typically designed for specific learning algorithms. In comparison to wrappers, embedded methods can offer several efficiency advantages. They make optimal use of available data without the need to split the training data into separate sets for training and validation. They also reach a solution more swiftly by avoiding the need to retrain a predictor entirely for each variable subset under investigation [18]. Embedded methods, which have an integrated FS mechanism as part of the predictive model construction process, encompass techniques such as decision trees and ensemble machine learning methods, with random forests [6] and gradient boosting [12] being the most prominent examples.

Compared to these other two FS types that rely on specific predictive models, filter methods operate independently of predictive models. They share a similar search approach with wrappers, but instead of evaluating against a predictor, they use a basic filter as a preprocessing step. Consequently, filters operate independently of the chosen predictor. Because of their model independence, filter methods offer swift processing speeds and do not exhibit bias towards particular predictive models. As high-dimensional data become more prevalent, filter methods are garnering increased attention. Traditional filter FS methods rely on correlations (see Section 2 in [18] for an overview), whereas emerging and successful filter methods are based on causality [48].

Causality-based FS aims to pinpoint the Markov Blanket (MB) associated with a class variable, with the goal of constructing predictive models that are both more interpretable and robust [17, 46]. The MB provides insight into the local causal relationship between the class variable and the features within it. As explained in more detail below (Section 2.2), because all other features are probabilistically independent of the class variable when conditioned on its MB, the MB of a class variable represents the theoretically optimal subset of features for classification [26, 31].

2.2 Markov blanket

The concept of Markov blanket (MB) in a Bayesian network was developed by Pearl [31]. A Bayesian network is a visual tool that succinctly illustrates a combined probability distribution across a set of random variables using a directed acyclic graph adorned with conditional probability tables [31]. These tables detail the probability distribution of a node based on any instantiation of its parent nodes. As a result, the graph conveys qualitative insights about the random variables, such as conditional independence properties. Meanwhile, the associated probability distribution, which aligns with these properties, offers a numerical portrayal of how the variables are interrelated. The probability distribution and the graph of a Bayesian network are linked by the Markov condition, which asserts that a node is conditionally independent of its non-descendants when given knowledge of its parents.

Definition 1. *(Faithfulness): A Bayesian network G and a joint distribution P are faithful to each other if every conditional independence implied by G and the Markov condition is also reflected in P [31].*

The MB of a variable within a Bayesian network comprises its parents (direct causes), children (direct effects), and spouses (other parents of these children). Given a target variable T and with the faithfulness assumption (Definition 1) in place, the MB of T is unique, and it becomes straightforward to extract it from the associated Bayesian network within a given application domain [15, 46]. More specifically, this means that by conditioning on the MB of a class variable T in a dataset, all the remaining features are conditionally independent of T. Thus, the MB of the class variable is theoretically optimal for FS [2, 3, 46, 49, 50].

Given that possessing complete knowledge of the $MB(T)$ is sufficient to ascertain the probability distribution of the class variable T, rendering the values of all other variables redundant, the process of inducing $MB(T)$ can be classified as a causal FS filter procedure [15, 46, 48]. Nonetheless, this necessitates having the Bayesian network pre-established. Conventionally, we must initially learn the desired Bayesian network to ascertain the MB of a specific variable. Hereby, one distinguishes between the global and the local learning of the network. Global learning refers to learning the whole Bayesian network. Local learning refers to discovering only the local structure around the target variable T or any other specific variable of interest [2, 3]. Generally, the process of structure learning for Bayesian networks is recognized as an NP-complete problem. Hence, several algorithms have been invented to deduce the MB without the prerequisite of having the entire Bayesian network pre-constructed, thereby significantly diminishing the complexity of time and computing resources.

In a recent review, several of these MB discovery algorithms were assessed in their capability to act as causality-based FS techniques [48]. The algorithm that came off best was the Iterative Parent-Child based search of MB (IPC-MB) algorithm by Fu and Desmarais [13]. The IPC-MB algorithm uses the following procedure to find the parent-child (PC) set of the target variable T: First, all features are the candidate PC of T. Second, conditional independence tests check each feature in the candidate PC of T level by level of the cardinality of the conditioning sets, starting with an empty set. Third, the local search is repeated given all found candidates, which not only recognizes false positives but also candidate spouses. The IPC-MB algorithm came off as the optimal choice in several comparisons of local MB discovery algorithms, excelling in terms of robustness, speed, data utilization, and information retrieval [13, 14, 48].

In this paper, we will compare this causality-based FS (i.e., IPC-MB) with a standard correlation-based FS with regard to the performance and fairness of the resulting models. As illuminated above, the causality-based FS algorithm should select the theoretically optimal features. Another advantage of learning the MB is that it also gives the number of optimal features automatically, while for other FS algorithms, one usually has to provide the number of features one wishes to select.

2.3 Fairness Metrics

In the history of constructing and implementing machine learning models in education, the main priority has frequently been to maximize the overall performance of these models. This is particularly evident in typical educational classification tasks, such as endeavors to identify as many at-risk students as possible. However, there has been a conspicuous shortage of attention given to guaranteeing the fairness of these models [29]. As machine learning models are used increasingly in the educational domain, ensuring that they are fair and do not discriminate against certain groups or individuals is imperative.

There is no universally accepted notion of fairness. Several different notions of fairness exist, and many metrics are used to measure these different notions [45]. As recent reviews have given excellent overviews of fairness notions (see [30] for a comprehensive overview of fairness concepts in machine learning in general, and [22] for one specifically tailored to the educational domain), we refrain from repeating those and only explain what is needed for understanding our experiments.

In our experiments, we concentrated on group fairness. Following Saxena et al. [37], we judged fairness as the absence of any favoritism towards a specific group in the context of the machine learning decision-making process. More specifically, we measured fairness between groups labeled as a and b—determined by the sensitive attribute group membership, as explained below in Section 3. Our evaluation was based on the nine quantitative metrics outlined in Table 1. The combination of these metrics allowed us to holistically assess the fairness of the machine learning models.

3 EXPERIMENTAL SETUP

This section explains the overall experimental setup. First, we describe the used datasets, including references to the original articles, dataset sizes, and the features used as sensitive attributes in the experiments. Second, we describe our general analysis and evaluation pipeline, including the model and hyperparameter optimization processes. All experiments were performed in Python version 3.12. Moreover, we used the pyCausalFS[1] toolbox [48, 49] to find the causal features in the datasets, and the holistic AI tool[2] for assessing the fairness of the models.

3.1 Datasets

We used four openly available datasets: two educational ones, that is, the Open University Learning Analytics Dataset (OULAD) [27] and the Portuguese Secondary School Math Performance (here referred to as PSSMP) dataset [8], and two commonly used benchmark datasets in the fairness literature: one from the social/law area, that is, the Correctional Offender Management Profiling for Alternative Sanction (COMPAS) dataset [24], and one from the financial domain, that is, the German credit (here referred to as GERCRE) dataset [10].

3.1.1 Open university learning analytics dataset. The OULAD [27] data originates from courses taught at the Open University in the United Kingdom and consists of five tables with information from 24,806 students, their interactions in the virtual learning environment, assessments, courses, and registrations. We used the binary information of whether a student passed a course as the target variable and the disability status (*disability*) of the student as the sensitive attribute. To ensure reproducibility of the results, we preprocessed the data following a public repository.[3]

[1] https://github.com/wt-hu/pyCausalFS
[2] https://holisticai.readthedocs.io/en/latest/
[3] https://github.com/gogoladzetedo/Open_University_Analytics

Table 1: Fairness metrics used in this study. DV refers to the desired value (i.e., the value that would mean fairness was achieved according to the metric).

Metric	DV	Formula	Interpretation
Statistical Parity	0	$sr_a - sr_b$	Fairness is achieved if the probability of a specific prediction is not dependent on sensitive group membership.
Disparate Impact	1	sr_a/sr_b	Very similar to statistical parity but computes the ratio instead, meaning fairness is achieved if this metric equals 1.
Four Fifths Rule	1	$4/5$	Uses statistical parity with a result considered fair if the ratio exceeds 80% for all groups.
Cohen D	0	$sr_a - sr_b/std_{pool}$	Normalised statistical parity. 0.2 is considered a small effect size, 0.5 is medium, and 0.8 is considered large.
2SD Rule	0	$Ztest(sr_a - sr_b)$	Z-test statistic for the difference in success rates. Fairness is achieved if the computed value is between -2 and 2, indicating no statistically significant difference in success rates.
Equality of Opportunity Difference	0	$tpr_a - tpr_b$	Difference in true positive rates for $group_a$ and $group_b$, considered fair if it achieves 0, range between -1 and 1.
False Positive Rate Difference	0	$fpr_a - fpr_b$	Difference in false positive rates between $group_a$ and $group_b$, considered fair if it achieves 0, range between -1 and 1.
Average Odds Difference	0	$0.5 * (fpr_a - fpr_b + tpr_a - tpr_b)$	Difference in average odds between $group_a$ and $group_b$, considered fair if it achieves 0, range between -1 and 1.
Accuracy Difference	0	$accuracy_a - accuray_b$	Difference in the accuracy of predictions for $group_a$ and $group_b$, considered fair if it achieves 0, range between -1 and 1.

3.1.2 Portuguese secondary school math performance dataset. The Portuguese secondary school mathematics performance (PSSMP) dataset [8] consists of 649 students from two Portuguese secondary schools. We used pass/fail of the students' final math grade (feature *G3*) as a binary classification task and removed the first and second grades because they are highly correlated with the target, making the prediction task trivial if included. Moreover, we one-hot-encoded all categorical features in the dataset. As the sensitive attribute, we used the students' gender (*sex*).

3.1.3 Correctional offender management profiling for alternative sanctions dataset. The COMPAS software assesses the likelihood of an individual committing another crime. Judges rely on COMPAS to determine whether to grant release to an offender or maintain their incarceration. A scrutiny of the software revealed a bias against African Americans: COMPAS tends to exhibit higher rates of incorrect positive predictions for African-American offenders compared to Caucasian offenders, falsely indicating a greater risk of re-offending [30]. Because of that, the COMPAS dataset [24] has become a well-known benchmark dataset in the fairness literature. It entails data about 6,172 individuals. We designated the target class as the indication of whether a person commits a crime in the following two years or not (*Two_yr_Recidivism*). Additionally, we identified the sensitive attribute (*race*) as the information regarding whether this person is African-American.

3.1.4 German credit dataset. The German credit dataset, here referred to as GERCRE, serves as a benchmark dataset frequently employed in the machine learning fairness literature (see, e.g., [20, 33, 43] for a selection of articles published after 2020). Recently, this fairness benchmark dataset faced criticism [11] due to the use of gender as the sensitive attribute in several machine learning articles (including those mentioned, i.e., [20, 33, 43]), despite the absence of a specific coding for gender in the data. Instead, the dataset includes only the combined feature of sex and marital status. Despite its age and recent scrutiny regarding its suitability for assessing fairness in machine learning models [11], we opted to use the GERCRE dataset for illustrative and comparative purposes. The dataset comprises information on 1,000 individuals. For the sensitive attribute, we selected the *sex and marital status* column, designating divorced/separated males as the sensitive group. Our target class was determined by creditworthiness, where one denotes credit-worthy, and zero signifies not credit-worthy.

3.2 Analysis pipeline and performance evaluation

Our goal was to compute the effect of two different ways of FS (causality- versus correlation-based) in comparison to no FS on the performance and fairness of the resulting machine learning models. In order to reduce the risk of getting results by chance and to increase stability and robustness, we implemented two nested

cross-validation loops. The outer stratified five-fold cross-validation loops over each experimental dataset, always using four folds for training and one for testing. Within each division, the training set is used to

(1) build a model (another five-fold cross-validation grid-search is employed to select the best hyperparameters, as explained in Section 3.2.1 below) using all features (i.e., without FS),
(2) perform causality-based FS and build a model using the same model-building function as in (1) but by using only the p selected causal features, and
(3) perform correlation-based FS to select exactly as many features p as the causality-based FS selected and build a model using the same model-building function as in (1) and (2) but by using only the p selected correlational features.

Hereby, the *IPC-MB* FS algorithm (see Section 2.2) of the *pycausalFS* toolbox is employed to get the causal features in (2), and the *f_classif* is employed to get the correlation-based features in (3). As explained in Section 2, the causality-based FS also readily returns the number of optimal features p, and hence, in our algorithmic pipeline, we use the p most important features from the correlation-based FS that by default returns only an ordering of feature importances. After model building and FS on the train set, the model performance and fairness are evaluated on the selected features of the respective hold-out test set. To further decrease the risk of getting results by chance, the model building and performance and fairness assessing process is repeated ten times for each kind of FS (i.e., 1 - without FS, 2 - causality-based FS, 3 - correlation-based FS). Finally, the performance and fairness results on the 50 different runs and test sets (i.e., ten repetitions for each five-fold split) are averaged for each kind of FS.

3.2.1 Models and hyperparameter optimization. Initially, we tried several different classification model types (logistic regression, support vector machines, multilayer perceptron, random forest), but since random forest [6] consistently gave the best performance in the initial tests, we used only this model class for the final pipeline and evaluation to improve comparability and simplicity. To select the best parameters, each training set of the outer cross-validation loop was split further into five folds to select the best hyperparameters. A random forest model with the current hyperparameters was trained on each fold, while the objective function of each step was to increase the performance on the hold-out datasets of each fold. For hyperparameter optimization of the random forest model, we implemented a grid-search over the *max_depth*, the *max_features*, the *min_samples_leaf*, and the *min_samples_split* of the trees in the forest. For classification performance evaluation, we used five common metrics: Accuracy, balanced accuracy, precision, recall, and the F1-score.

4 RESULTS

Table 2 summarizes the performance results, and Table 3 summarizes the fairness results for the OULAD dataset. As described in Section 3.2, the tables report the averages (mean and standard deviation) over the 50 different runs and test sets (i.e., ten repetitions for each five-fold split). Tables 4 and 5 summarize these results for the PSSMP dataset, Tables 6 and 7 for the COMPAS, and Tables 8 and 9 for the GERCRE dataset, respectively.

A first observation that can be made from the results is that generally, FS decreased model performance and fairness. *Without FS* shows the best results in most cases. However, there are a few notable exceptions to this general observation: For the PSSMP dataset, correlational-FS yielded the best performance according to all metrics (one possible reason for this is that there may be adverse features in the dataset that the correlation-based FS correctly did not select), and for the OULAD dataset, the causality-based FS often gave better results in terms of model fairness (indicating that the causality-based FS correctly not selected features that could be linked to the sensitive attribute).

For the two FS types, we bolded for each row the FS that gave the best results. As shown in the Tables 2–9, typically, *causality-based FS* performed better when measuring fairness and *correlation-based FS* performed better when measuring classification performance. Although there were a few exceptions, generally, the models built after correlation-based FS outperformed those built after causality-based FS in terms of accuracy, balanced accuracy, precision, recall, and F1-score. Moreover, on average, the models built after causality-based FS more often had fairness metrics value closer to the desired value of the respective metric (see Table 1 for an overview of all employed fairness metrics and their respective desired values).

Correlation-based FS is a technique that focuses on finding and selecting the most relevant features from a dataset. Causality-based FS aims to identify the MB of a class variable to build more interpretable and robust predictive models. Thus, using the causal features for similar data points in another environment or at a later time (for example, by updating the already quite old GERCRE dataset with current information) would probably lead to better results, although the performance of the correlation-based selected features is better for the existing data.

Another result worth pointing out that possibly affects the application of causality-based FS algorithms in real-world applications is that the causality-based FS took significantly longer than the correlation-based FS. For example, even for the relatively small PSSMP dataset, the computation of the causal features with the *pyCausalFS* toolbox took on average 0.266 seconds per training set fold, while the computation of the correlational features on the same folds took only 15.625 milliseconds on average. Thus, causality-based FS is not recommendable if computing resources are an issue.

5 DISCUSSION

Machine learning is now being applied in a diverse array of decision-making contexts, many of which carry significant consequences for both individuals and society at large. While this technology holds the promise of mitigating undesirable elements of human decision-making, there is a valid apprehension that biases present in the data and inaccuracies in the model can result in decisions that unfairly disadvantage groups with a history of discrimination. Consequently, the research community has begun to explore methods to guarantee that the models we train do not render decisions that exhibit unfairness concerning sensitive attributes [7].

Until now, the fairness literature has largely overlooked the FS step in the machine learning pipeline [16], and several fairness AI in education articles have pointed out the need for more causal

Table 2: Performance metrics for the OULAD test sets for the features selected on the training sets (mean and standard deviation over five-fold cross-validation and ten repetitions). The best average FS (causal- or correlation-based) is bolded for each metric.

	Without FS		Causal-based FS		Correlation-based FS	
Metric	mean	std	mean	std	mean	std
Accuracy	0.793	0.001	0.790	0.001	**0.790**	0.001
Balanced accuracy	0.796	0.002	0.792	0.001	**0.792**	0.001
Precision	0.755	0.001	0.749	0.001	**0.752**	0.001
Recall	0.834	0.003	**0.834**	0.002	0.827	0.001
F1-Score	0.792	0.002	**0.789**	0.001	0.788	0.001

Table 3: Fairness metrics for the OULAD test sets for the features selected on the training sets (mean and standard deviation over five-fold cross-validation and ten repetitions). The best average FS (causal- or correlation-based) is bolded for each metric.

	Without FS		Causal-based FS		Correlation-based FS	
Metric	mean	std	mean	std	mean	std
Statistical Parity	0.14	0.01	**0.12**	0.01	0.12	0.00
Disparate Impact	1.37	0.03	**1.28**	0.02	1.29	0.01
Four Fifths Rule	0.73	0.02	**0.78**	0.01	0.78	0.01
Cohen D	0.29	0.02	**0.24**	0.01	0.24	0.01
2SD Rule	6.61	0.38	**5.44**	0.28	5.47	0.21
Equality of Opportunity Difference	0.09	0.01	0.08	0.01	**0.07**	0.01
False Positive Rate Difference	0.10	0.01	**0.07**	0.00	0.08	0.00
Average Odds Difference	0.10	0.01	0.07	0.01	**0.07**	0.00
Accuracy Difference	-0.01	0.00	**-0.00**	0.00	-0.01	0.00

Table 4: Performance metrics for the PSSMP test sets for the features selected on the training sets (mean and standard deviation over five-fold cross-validation and ten repetitions). The best average FS (causal- or correlation-based) is bolded for each metric.

	Without FS		Causal-based FS		Correlation-based FS	
Metric	mean	std	mean	std	mean	std
Accuracy	0.886	0.022	0.722	0.033	**0.903**	0.041
Balanced accuracy	0.876	0.025	0.639	0.056	**0.902**	0.063
Precision	0.923	0.018	0.750	0.038	**0.952**	0.066
Recall	0.906	0.019	0.881	0.011	**0.906**	0.050
F1-Score	0.914	0.017	0.810	0.017	**0.926**	0.030

Table 5: Fairness metrics for the PSSMP test sets for the features selected on the training sets (mean and standard deviation over five-fold cross-validation and ten repetitions). The best average FS (causal- or correlation-based) is bolded for each metric.

	Without FS		Causal-based FS		Correlation-based FS	
Metric	mean	std	mean	std	mean	std
Statistical Parity	-0.00	0.15	**0.02**	0.08	0.03	0.04
Disparate Impact	1.01	0.17	**1.04**	0.11	1.04	0.04
Four Fifths Rule	0.89	0.08	0.94	0.07	**0.96**	0.03
Cohen D	-0.03	0.46	**0.05**	0.19	0.09	0.11
2SD Rule	-0.13	1.97	**0.22**	0.83	0.40	0.50
Equality of Opportunity Difference	-0.02	0.11	0.00	0.03	**0.00**	0.02
False Positive Rate Difference	0.02	0.25	**0.06**	0.14	0.09	0.09
Average Odds Difference	0.00	0.16	**0.03**	0.08	0.04	0.06
Accuracy Difference	-0.04	0.04	**-0.04**	0.02	-0.05	0.04

algorithms [9, 29]. Thus, the goal of this article was to assess the impact of correlational versus causal FS on the resulting machine-learning models. The theoretical superiority of causality-based FS has been discussed in several works [31, 48, 49], but to our knowledge, no direct comparisons to classical correlation-based FS have been performed.

Table 6: Performance metrics for the COMPAS test sets for the features selected on the training sets (mean and standard deviation over five-fold cross-validation and ten repetitions). The best average FS (causal- or correlation-based) is bolded for each metric.

	Without FS		Causal-based FS		Correlation-based FS	
Metric	mean	std	mean	std	mean	std
Accuracy	0.688	0.005	0.586	0.007	**0.670**	0.004
Balanced accuracy	0.679	0.004	0.581	0.005	**0.660**	0.004
Precision	0.682	0.012	0.549	0.011	**0.666**	0.009
Recall	0.588	0.007	0.517	0.024	**0.551**	0.013
F1-Score	0.631	0.002	0.532	0.008	**0.603**	0.007

Table 7: Fairness metrics for the COMPAS test sets for the features selected on the training sets (mean and standard deviation over five-fold cross-validation and ten repetitions). The best average FS (causal- or correlation-based) is bolded for each metric.

	Without FS		Causal-based FS		Correlation-based FS	
Metric	mean	std	mean	std	mean	std
Statistical Parity	-0.26	0.02	**-0.15**	0.06	-0.29	0.05
Disparate Impact	0.49	0.02	**0.69**	0.12	0.45	0.06
Four Fifths Rule	0.49	0.02	**0.69**	0.12	0.45	0.06
Cohen D	-0.56	0.04	**-0.32**	0.13	-0.62	0.11
2SD Rule	-9.45	0.55	**-5.50**	2.17	-10.37	1.73
Equality of Opportunity Difference	-0.25	0.05	**-0.15**	0.06	-0.28	0.04
False Positive Rate Difference	-0.20	0.03	**-0.13**	0.07	-0.22	0.07
Average Odds Difference	-0.22	0.02	**-0.14**	0.06	-0.25	0.05
Accuracy Difference	0.02	0.04	**0.02**	0.02	0.02	0.03

Table 8: Performance metrics for the GERCREtest sets for the features selected on the training sets (mean and standard deviation over five-fold cross-validation and ten repetitions). The best average FS (causal- or correlation-based) is bolded for each metric.

	Without FS		Causal-based FS		Correlation-based FS	
Metric	mean	std	mean	std	mean	std
Accuracy	0.757	0.028	0.715	0.026	**0.727**	0.019
Balanced accuracy	0.629	0.040	0.547	0.034	**0.622**	0.051
Precision	0.763	0.021	0.721	0.015	**0.766**	0.035
Recall	0.948	0.015	**0.967**	0.018	0.883	0.066
F1-Score	0.845	0.016	**0.826**	0.016	0.819	0.016

Table 9: Fairness metrics for the GERCRE test sets for the features selected on the training sets (mean and standard deviation over five-fold cross-validation and ten repetitions). The best average FS (causal- or correlation-based) is bolded for each metric.

	Without FS		Causal-based FS		Correlation-based FS	
Metric	mean	std	mean	std	mean	std
Statistical Parity	0.10	0.08	**-0.06**	0.01	0.07	0.04
Disparate Impact	1.14	0.12	**0.94**	0.01	1.09	0.06
Four Fifths Rule	0.88	0.10	**0.94**	0.01	0.92	0.05
Cohen D	0.30	0.24	-0.27	0.03	**0.17**	0.09
2SD Rule	0.95	0.79	-0.83	0.12	**0.53**	0.30
Equality of Opportunity Difference	0.14	0.06	**-0.03**	0.02	0.12	0.01
False Positive Rate Difference	-0.07	0.25	-0.14	0.06	**-0.12**	0.11
Average Odds Difference	0.04	0.12	-0.09	0.02	**-0.00**	0.05
Accuracy Difference	0.17	0.11	**0.14**	0.18	0.17	0.05

Our results showed that, mostly, the causality-based FS led to fairer models than the correlation-based FS. However, there is a trade-off, as the performance of the resulting models was usually better after correlation-based FS. Causal FS aims to identify the MB of a class variable to build more interpretable and robust predictive models. Correlation-based FS is a technique that focuses on finding and selecting the most relevant features from a dataset. Since the correlation between two variables is a less stringent criterion compared to independence, it is logical to question why there is not much work on causal algorithms and machine learning [40], and why fairness algorithms and standards are typically framed in terms of correlations. One pragmatic justification is that, as discussed in Section 4, computing correlations is significantly more straightforward than estimating independence. While correlation is a descriptive statistic demanding relatively few assumptions for calculation, establishing independence necessitates the application of inferential statistics, which can generally be quite complex and computationally expensive [22, 42].

5.1 Limitations and future research

Our study opens avenues for further exploration and refinement. Firstly, our analysis was anchored in the utilization of the prevalent correlation-based filter FS, and the best-performing causality-based filter FS. Exploring additional FS techniques and types, such as wrappers, could prove worthwhile. Another aspect deserving attention is the sensitivity of our outcomes to the choices embedded in different algorithms. While our approach involved a meticulous grid search, coupled with the aggregation of averages across multiple iterations for result stability, the impact of diverse hyperparameter search spaces warrants investigation.

Furthermore, we focused on specific sensitive attributes, namely disability, gender, race, and marital status, aiming to encompass various factors where individuals may face discrimination. However, it is essential to note that even if an algorithm is deemed fair regarding one attribute, this does not necessarily extend to others. Future work should also investigate the effect of FS on additional sensitive attributes. Finally, an intriguing avenue for future research lies in delving into the comparative effects of causality-versus correlation-based FS on the interpretability and quality of subsequent AI models and explanations. Quality metrics such as explanation robustness and fidelity [35] could serve as valuable benchmarks in evaluating these effects.

ACKNOWLEDGEMENTS

This work was supported by the Otto A. Malm Foundation and the Academy of Finland (project no. 356314).

REFERENCES

[1] Selin Akgun and Christine Greenhow. 2021. Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics* 2, 3 (2021), 1–10.
[2] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. 2010. Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. *Journal of Machine Learning Research* 11, 7 (2010), 171–234.
[3] Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. 2010. Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: analysis and extensions. *Journal of Machine Learning Research* 11, 8 (2010), 235–284.
[4] Susan Athey. 2017. Beyond prediction: Using big data for policy problems. *Science* 355, 6324 (2017), 483–485.
[5] Ryan S Baker and Aaron Hawn. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education* 32 (2021), 1–41. Issue 4.
[6] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.
[7] Silvia Chiappa and William S Isaac. 2019. A causal Bayesian networks viewpoint on fairness. In *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data: 13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Vienna, Austria, August 20-24, 2018, Revised Selected Papers*. Springer International Publishing, Cham, 3–20.
[8] Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance.
[9] Oscar Blessed Deho, Chen Zhan, Jiuyong Li, Jixue Liu, Lin Liu, and Thuc Duy Le. 2022. How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology* 53, 4 (2022), 822–843.
[10] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[11] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2074–2152.
[12] Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational statistics & data analysis* 38, 4 (2002), 367–378.
[13] Shunkai Fu and Michel C Desmarais. 2008. Fast Markov blanket discovery algorithm via local learning within single pass. In *Advances in Artificial Intelligence: 21st Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2008 Windsor, Canada, May 28-30, 2008 Proceedings 21*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 96–107.
[14] Shunkai Fu and Michel C Desmarais. 2008. Tradeoff analysis of different Markov blanket local learning approaches. In *Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, May 20-23, 2008 Proceedings 12*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 562–571.
[15] Shunkai Fu and Michel C Desmarais. 2010. Markov blanket based feature selection: a review of past decade. In *Proceedings of the world congress on engineering*, Vol. 1. Newswood Ltd. Hong Kong, China, WCE, London, UK, 321–328.
[16] Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. 2022. Causal feature selection for algorithmic fairness. In *Proceedings of the 2022 International Conference on Management of Data*. ACM, New York, NY, USA, 276–285.
[17] Isabelle Guyon, Constantin Aliferis, et al. 2007. Causal feature selection. In *Computational methods of feature selection*. Chapman and Hall/CRC, Boca Raton, Fla, USA, 79–102.
[18] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.
[19] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, Springer New York, NY.
[20] Yuzi He, Keith Burghardt, and Kristina Lerman. 2020. A geometric solution to fair representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 279–285.
[21] Wayne Holmes and Ilkka Tuomi. 2022. State of the art and practice in AI in education. *European Journal of Education* 57, 4 (2022), 542–570.
[22] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. Association for Computing Machinery, New York, NY, USA, 49–58.
[23] Eric Jotkoff. 2022. NEA survey: Massive staff shortages in schools leading to educator burnout; alarming number of educators indicating they plan to leave profession. https://www.nea.org/about-nea/media-center/press-releases/nea-survey-massive-staff-shortages-schools-leading-educator.
[24] Julia Kirchner, Surya Angwin, Jeff Mattu, and Lauren Larson. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.
[25] Ron Kohavi and George H John. 1997. Wrappers for feature subset selection. *Artificial intelligence* 97, 1-2 (1997), 273–324.
[26] Daphne Koller, Mehran Sahami, et al. 1996. Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning* (Bari, Italy) *(ICML'96)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 284–292.
[27] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. 2017. Open university learning analytics dataset. *Scientific data* 4, 1 (2017), 1–8.
[28] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: A data perspective. *ACM computing surveys (CSUR)* 50, 6 (2017), 1–45.
[29] Lin Li, Lele Sha, Yuheng Li, Mladen Raković, Jia Rong, Srecko Joksimovic, Neil Selwyn, Dragan Gašević, and Guanliang Chen. 2023. Moral Machines or Tyranny

of the Majority? A Systematic Review on Predictive Bias in Education. In *LAK23: 13th International Learning Analytics and Knowledge Conference*. Association for Computing Machinery, New York, NY, USA, 499–508.

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[31] Judea Pearl. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, Burlington, Massachusetts, USA.

[32] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books, Penguin UK.

[33] Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. 2021. Fair Bayesian optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 854–863.

[34] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, Cambridge, MA 02142.

[35] Mirka Saarela and Lilia Geogieva. 2022. Robustness, stability, and fidelity of explanations for a deep skin cancer classification model. *Applied Sciences* 12, 19 (2022), 9545.

[36] Mirka Saarela, Ville Heilala, Päivikki Jääskelä, Anne Rantakaulio, and Tommi Kärkkäinen. 2021. Explainable Student Agency Analytics. *IEEE Access* 9 (2021), 137444–137459.

[37] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 99–106.

[38] Andreas Schleicher. 2020. The impact of COVID-19 on education insights from education at a glance 2020.

[39] Andreas Schleicher. 2023. *PISA 2022 Insights and Interpretations*. Technical Report. OECD.

[40] Bernhard Schölkopf. 2022. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. Association for Computing Machinery, New York, NY, USA, 765–804.

[41] Lele Sha, Mladen Raković, Angel Das, Dragan Gašević, and Guanliang Chen. 2022. Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education. *IEEE Transactions on Learning Technologies* 15, 4 (2022), 481–492.

[42] Rajen D. Shah and Jonas Peters. 2020. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics* 48, 3 (2020), 1514 – 1538. https://doi.org/10.1214/19-AOS1857

[43] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. 2020. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 358–364.

[44] United Nations. 2023. The 17 Sustainable Development Goals. https://www.un.org/en/sustainable-development-goals. Accessed: 2023-17-09.

[45] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. Association for Computing Machinery, New York, NY, USA, 1–7.

[46] Hao Wang, Zhaolong Ling, Kui Yu, and Xindong Wu. 2020. Towards efficient and effective discovery of Markov blankets for feature selection. *Information Sciences* 509 (2020), 227–242.

[47] Mary E Webb, Andrew Fluck, Johannes Magenheim, Joyce Malyn-Smith, Juliet Waters, Michelle Deschênes, and Jason Zagami. 2021. Machine learning for human learners: opportunities, issues, tensions and threats. *Educational Technology Research and Development* 69, 4 (2021), 2109–2130.

[48] Kui Yu, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. 2020. Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–36.

[49] Kui Yu, Lin Liu, and Jiuyong Li. 2021. A unified view of causal and non-causal feature selection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 4 (2021), 1–46.

[50] Kui Yu, Lin Liu, Jiuyong Li, Wei Ding, and Thuc Duy Le. 2019. Multi-source causal feature selection. *IEEE transactions on pattern analysis and machine intelligence* 42, 9 (2019), 2240–2256.