A preprocessing Shapley value-based approach to detect relevant and disparity prone features in machine learning

Guilherme Dean Pelegrina*
guilherme.pelegrina@mackenzie.br
School of Applied Sciences - University of Campinas (UNICAMP)
Limeira, São Paulo, Brazil
Mackenzie Presbyterian University (UPM)
São Paulo, São Paulo, Brazil

Miguel Couceiro*
miguel.couceiro@loria.fr
Université de Lorraine, CNRS, LORIA
Vandœuvre-lès-Nancy, Lorraine
France
INESC-ID, Instituto Superior Técnico, Universidade de Lisboa
Lisbon, Portugal

Leonardo Tomazeli Duarte*
leonardo.duarte@fca.unicamp.br
School of Applied Sciences - University of Campinas (UNICAMP)
Limeira, São Paulo, Brazil

ABSTRACT

Decision support systems became ubiquitous in every aspect of human lives. Their reliance on increasingly complex and opaque machine learning models raises transparency and fairness concerns with respect to unprivileged groups of people. This motivated several efforts to estimate importance of features towards the models' performance and to detect unfair/disparate decisions. The latter is often dealt with by means of fairness metrics that rely on performance metrics with respect to predefined features that are considered protected (salient features such as age, gender, ethnicity, etc.) and/or sensitive (such as education, /occupation, banking information). However, such an approach is subjective (as fairness metrics depend on the choice features), there may be other features that lead to unfair (disparate) decisions and that may ask for suitable interpretations.

In this paper we focus on the latter issues and propose a statistical preprocessing approach that is inspired by both the Hilbert-Schmidt independence criterion and Shapley values to estimate feature importance and to detect disparity prone features. Unlike traditional Shapley value-based approaches, we do not require trained models to measure feature importance or detect disparate results. Instead, it focuses on data and statistical criteria to measure the dependence of feature distributions. Our empirical results show that features with the highest dependence degrees with the label vector are also the ones with the highest impact on the model performance. Moreover, our empirical results indicate that this relation enables the detection of disparity prone features.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by classification**.

*All authors contributed equally to this research.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658905

KEYWORDS

Hilbert-Schmidt independence criterion, Shapley value, fairness measure, disparity detection

ACM Reference Format:
Guilherme Dean Pelegrina, Miguel Couceiro, and Leonardo Tomazeli Duarte. 2024. A preprocessing Shapley value-based approach to detect relevant and disparity prone features in machine learning. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658905

1 INTRODUCTION

Machine Learning (ML) models are now massively used to tackle a wide variety of real-world problems [19, 53]. Such models have thus a direct impact into people lives, and there are increasing transparency and fairness concerns that practitioners should be aware of when deploying ML methods into real applications [2, 9]. One may consider as a transparent method the ones whose result can be understood. For example, in recidivism risk prediction [8], one should be able to interpret how each individual characteristic is contributing towards the model's output. However, some practical situations require the use of complex methods which may be rather opaque and whose results may be hard to understand. To deal with such issues, several model-agnostic approaches have been proposed in the literature, *e.g.*, those based on surrogate models and on Shapley values [36, 48, 58]. Besides transparency, one also expects the algorithm to be fair with respect to sensitive or protected groups of people, for instance, according to their gender [13] or race [8]. The fairness concern has become an important discussion in machine learning. Several techniques have been developed to mitigate disparate results. Examples include pre-processing [11, 47, 52], in-processing [3, 66] and post-processing [26, 28, 40, 43] algorithms (see also [6, 38, 61] for recent surveys and further discussions on fairness related issues). Moreover, to analyze ethical disparities provided by a ML model, one should understand the results achieved by such models. This asks for the interpretation of feature contributions towards disparate outcomes [10, 22, 34]. A common point in the aforementioned approaches to explain the performance of ML models or the potentially disparate results is that they are generally based on a trained model. In other words, one must firstly train the model and, secondly, determine the features that contribute to the predicted outcome and/or the disparate result. However, training

a model before analysing features contributions can be computationally expensive, especially when using coalition based feature importance values, *e.g.*, Shapley values [36]. Indeed, as the complexity of computation of Shapley values increases exponentially with the number of features[1], the retraining of ML models could be practically infeasible.

In [46] the authors addressed the issue of detecting features and subfeatures that are likely to entail disparate outcomes, possibly disadvantageous for minorities and/or protected groups. Such situations may be due to imbalanced feature distributions, or to features that are prone to systematically dividing instances into subgroups with disparate outcomes, and that this division may not be rooted in acceptable ethical principles. The authors of [46] refer to such features (or subfeatures) as *disparity prone features*, *i.e.*, (sub)features that entail disparate (unfair) outcomes. To verify such disparities among different (sub)groups, it is common to train an ML model, *e.g.,* a classifier, and compute a pertaining fairness metric that measures the difference or the ratio between the performances for the different (sub)groups [6, 25, 26, 38, 65]. This process may be computationally costly depending on the ML model that is adopted, and raises the question of how to detect possible disparities before the training step. To tackle these issues, [46] proposed a statistical approach based on the Hilbert-Schmidt independence criterion (HSIC) [24, 63]. The HSIC essentially measures the dependence between vectors or matrices, and it has been used as a feature selection method [56, 57] in classification and regression tasks. The working hypothesis in [46] was that, if a feature that divides instances into different groups has a high HSIC with the label vector, then it will entail disparate outcomes among the (sub)groups discriminated by this feature. The authors also observed that the HSIC or its normalized variant (NOCCO) simply indicate *the features' potential to entail disparities, and features with high HSIC should then be analyzed by practitioners and domain experts to decide whether they should be considered as sensitive, especially, when a cause-effect relation is detected.* The appealing aspect of this framework is the detection of disparity prone and proxy features at the preprocessing stage, thus avoiding the computational cost of using trained ML models to verify unfair results.

However, HSIC (or NOCCO) based approach has some limitations. Firstly, it cannot account for coalitions (combinations) of features, that may themselves entail disparate outcomes. For instance, in recent years several anti-discrimination efforts have resulted in decreased gender pay gaps in Europe. However, this tendency is reversed when combined with additional information on whether it is in the public or the private sector[2]. Secondly, it does not provide key information (*e.g.,* in the form of coalition importance) that can help interpreting and explaining the impact of such coalitions in the performance metrics (for instance, fairness metrics).

This motivated us to generalize NOCCO by taking into account these two aspects, namely, different combinations of features together with their importance to the outcomes. Although being a general framework, we only illustrate its feasibility by taking Shapley values (that compute the importance of coalitions to the outcomes) of ML models such as multi-layer perceptron and random forest. We first perform an empirical study on synthetic data to attest our hypotheses, and then experiments on well-know datasets followed by a qualitative analysis that clearly shows the correlation between NOCCO Shapley values and fairness metrics (here, overall accuracy equality).

This paper is organized as follows. In Section 2, we introduce the underlying problem considered together with the notation employed throughout the paper. We briefly recall the key notions pertaining to the HSIC and Shapley values and discuss their use in practice in Sections 3 and 4, respectively. We propose extensions to the NOCCO based approach to detect disparity prone features in Section 5, that is followed by a preliminary experiment on synthetic data in Section 6. Further empirical studies on real world data are presented in Section 7, followed by qualitative analysis of the obtained results. We conclude the paper in Section 8 where we discuss some future perspectives.

Main Contributions:

- We propose a preprocessing approach based on the NOCCO and Shapley values to address the problem of assigning feature importance measures towards the algorithm performance.
- In the case of sensitive/protected features with high dependence degree with the outcome, our proposal also allows us to detect disparity prone features without the need of a trained model.
- We present an empirical study on four datasets frequently used in the literature. The higher is the NOCCO Shapley values, the higher is the feature impact on either the model performance and the disparate outcome.
- The interaction indices indicate the presence of interaction effects between features, which can be used to detect proxies in the dataset (the case of a redundant effect) or coalitions of features whose simultaneous use increases the predictive power of the ML model (the case of a complementary effect).

2 NOTATION AND PROBLEM SETTING

In this paper, we deal with binary classification problems whose input variables are represented by matrices $\mathbf{X}_{n\times m}$, where n is the number of samples and m is the number of features. For each sample, we have the associated class $y \in \{-1, 1\}$. Let us also define G_j, $j = 1, \ldots, m$, as the feature described in the $j-$th column of $\mathbf{X}$, *i.e.,* $\mathbf{X}^{(j)}$. Very often, some features are categorical and, therefore, one should convert them into binary features[3] to be able to apply a ML model. Suppose, for instance, that a categorical feature G_j (*e.g., race*) is described by $q = 3$ categories (or subfeatures), namely $G_{j,1}$, $G_{j,2}$ and $G_{j,3}$ (*e.g., whites*, *blacks* and *Asians*). By converting this categorical feature into 3 binary variables, the novel data $\tilde{\mathbf{X}}^{(j)}_{n\times 3}$ representing $\mathbf{X}^{(j)}$ has 3 columns. The same procedure is repeated for all categorical features, which will increase the total number of features (and subfeatures) from m to $\tilde{m}$. The input variables after converting all categorical features is represented by $\tilde{\mathbf{X}}_{n\times\tilde{m}}$. The problem that we address consists in estimating the importance of features towards performance and fairness measures

[1] Surely, there are techniques to estimate the Shapley values [1, 7, 41], however, one still need to train the model several times.
[2] https://l1nq.com/FYJU8
[3] In this paper, we adopted the one-hot encoding strategy to convert the categorical features into binary variables.

before the training step. In contrast with classical feature attribution approaches, we only consider the dataset $\mathbf{X}$ and the output vector $\mathbf{y}_{n\times 1}$ to calculate the importance measures. For this purpose, one firstly needs to define a measure that evaluates the predictive power of each feature or coalition of features. Based on such values, one should then adopt a strategy to calculate the marginal predictive power of features towards the prediction $\mathbf{y}$. We discuss these two steps in the next sections.

3 HILBERT-SCHMIDT INDEPENDENCE CRITERION

Calculating the predictive power of features is an important step in feature selection process. In summary, the idea is that feature whose predictive power is insignificant could be removed from the dataset without relevant impact on the ML performance. In order to calculate such a measure, several approaches have been proposed in the literature, such as those based on the correlation coefficient or mutual information (see [29, 59, 60] for reviews of feature selection approaches in ML). Although these strategies can be easily deployed to calculate the relation between each feature $\mathbf{X}^{(j)}$ and the predicted value $\mathbf{y}$, there are some drawbacks with their use. For instance, the correlation coefficient is based on second-order moments between vectors, which is a weaker measure if the purpose is to estimate statistical dependence. Although the mutual information is stronger than the correlation coefficient in measuring statistical dependence, its calculation requires estimating probability density functions, which can be a difficult task. Moreover, the use of both measures is straightforward only to evaluate the dependence degree between two vectors. However, in our approach, one needs a strategy to evaluate the relation between any coalition of features and the output vector $\mathbf{y}$.

Recently [46, 62, 63] used another measure to estimate the dependence degree between variables, namely, the Hilbert-Schmidt independence criterion [17, 18, 24]. Differently from the correlation coefficient or the mutual information, the HSIC can be used to calculate the dependence degree between matrices[4]. The HSIC can be empirically calculated by

$$HSIC(\mathbf{X},\mathbf{y}) = \frac{\operatorname{tr}\left(\mathbf{K_X H K_y H}\right)}{(n-1)^2} = \frac{\operatorname{tr}\left(\mathbf{H K_X H K_y}\right)}{(n-1)^2}, \tag{1}$$

where $\mathbf{K_X}$ is the kernel matrix of $\mathbf{X}$, $\mathbf{K_y}$ is the kernel matrix of $\mathbf{y}$, $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{e}\mathbf{e}^T$ is the centering matrix, and $\mathbf{e}$ is the n-dimensional column vector $\mathbf{1}$. Note that, by multiplying $\mathbf{K_X}$ by $\mathbf{H}$ on both sides, one removes its columns and rows mean. Therefore, one ensure that the kernel is centered.

Although in (1) the kernels are centered, one may have an incorrect interpretation when comparing the HSIC for different entities (*e.g.,* $HSIC(\mathbf{X},\mathbf{y})$ and $HSIC(\mathbf{X}',\mathbf{y})$ where $\mathbf{X} \neq \mathbf{X}'$). Indeed, the HSIC calculation is sensitive to the dataset scale [32]. Therefore, in order to properly compare relative dependence degrees, in our analysis we consider a normalized version of HSIC called NOCCO (NOrmalized Cross-Covariance Operator) [17, 18]. It is defined as follows:

$$NOCCO(\mathbf{X},\mathbf{y}) = \operatorname{tr}\left(\mathbf{R_X R_y}\right), \tag{2}$$

where $\mathbf{R_X} = \mathbf{H K_X H}\left(\mathbf{H K_X H} + n\epsilon\mathbf{I}_n\right)^{-1}$, $\mathbf{R_y} = \mathbf{H K_y H}\left(\mathbf{H K_y H} + n\epsilon\mathbf{I}_n\right)^{-1}$, ϵ is a regularization parameter (*e.g.,* 10^{-6}) and $\mathbf{I}_n$ is a $n \times n$ identity matrix. As $NOCCO(\mathbf{X},\mathbf{y}) \in [0,1]$, $\mathbf{y}$ is independent of $\mathbf{X}$ if the NOCCO value is 0. Conversely, greater the NOCCO values, greater the dependence degree between $\mathbf{X}$ and $\mathbf{y}$.

4 SHAPLEY VALUE AS A FEATURE ATTRIBUTION METHOD

The application of the Shapley value as a feature attribution method has gained attention in the ML community in the last years [5, 39, 51, 58]. The inspiration lies on the use of the Shapley value as a solution concept in cooperative game theory [55]. Assume a set of players $M = \{1,\ldots,m\}$ that cooperate to achieve a common goal. Let us define $v(A)$ as the payoff (or gain) of a game v when players in $A \subseteq M$ act by cooperation (one generally assumes $v(\emptyset) = 0$). Given the payoff achieved by the grand coalition M, *i.e.,* $v(M)$, the Shapley value emerges as a mechanism to fairly distribute the total gain among all players $j = 1,\ldots,m$. The Shapley value of a player j is calculated as follows:

$$\phi_j = \sum_{A\subseteq M\setminus\{j\}} \frac{(m-|A|-1)!\,|A|!}{m!}\left[v(A\cup\{j\}) - v(A)\right], \tag{3}$$

where $|A|$ indicates the cardinality of subset A. It indicates the marginal contribution of player j towards the game payoff when joining all possible coalitions of players. One considers the Shapley value a fair mechanism since it satisfies some desirable properties when sharing the benefits from a game [64]. The most relevant for our analysis are the following:

Properties 1. Efficiency: The total gain $v(M)$ is divided among all players $j = 1,\ldots,m$. Mathematically, $\sum_{j=1}^{m}\phi_j = v(M)$.

Properties 2. Dummy: If a player does not contribute towards the goal (*i.e.,* its marginal contribution is null), he/she should not receive a benefit when sharing the total gain. In other words, if, for all subsets $A \subseteq M$, $v\left(A\cup\{j\}\right) = v\left(A\right)$, then $\phi_j = 0$.

Properties 3. Symmetry: In the case where two players contribute equally when joining all coalitions, they should receive the same benefit. In this case, for all $A \subset M$ $(A\setminus\{j,j'\})$ and two players j and j' such that $v\left(A\cup\{j\}\right) = v\left(A\cup\{j'\}\right)$, then $\phi_j = \phi_{j'}$.

The aforementioned properties (specially the efficiency one) brought the attention of the ML community to use the Shapley value as a feature attribution approach. It has been applied to interpret either local predictions [4, 36, 37] or global performances of machine learning models [12, 22, 33, 35, 50]. For instance, in global interpretability, the marginal contribution assigned to each feature can be used as a feature selection strategy [15, 23, 50].

Besides the marginal contribution of features, Eq. (3) can be extended to evaluate the interaction degree between pairs of features[5]. This measure, called Shapley interaction index, is defined

[4]Note that the possibility to calculate the dependence degree between matrices generalizes the use of HSIC between two matrices, two vectors or a matrix and a vector.

[5]In fact, one can extend the Shapley value calculation to any coalition of features A. However, the interpretation is clear only for singletons or pairs of features.

as follows [20, 42]:

$$I_{j,j'} = \sum_{A \subseteq M \setminus \{j,j'\}} \frac{(m-|A|-2)!\,|A|!}{(m-1)!} \times \left[v(A \cup \{j,j'\}) - v(A \cup \{j\}) - v(A \cup \{j'\}) + v(A)\right]. \quad (4)$$

The sign of $I_{j,j'}$ indicates the type of interaction between features $G_j, G_{j'}$. If $I_{j,j'} < 0$, there is a redundant effect between $G_j, G_{j'}$. This is the case of negatively correlated features, where the payoff $v(\{j,j'\}) < v(\{j\}) + v(\{j'\})$ means that both features together are not better than the sum of them individually. If $I_{j,j'} > 0$, there is a complementary effect between $G_j, G_{j'}$. This is the case of positively correlated features, where the payoff $v(\{j,j'\}) > v(\{j\}) + v(\{j'\})$ and means that the effect of both features together is better than the sum of them individually. Finally, if $I_{j,j'} = 0$, there is no interaction between $G_j, G_{j'}$. In this case, they act independently. Although the interaction index has been largely exploited in the context of multicriteria decision making problems [21], in machine learning, only few works used such a measure for the purpose of interpretability [35, 44, 48, 49]. However, it brings interesting insights about how features interacts towards the model performance.

A remark in most of the aforementioned Shapley value-based approaches is that they require a trained model in order to calculate the marginal contributions and assign importance measures to features. In this scenario, one first needs to train the model and then apply the game theory strategy. This can be computationally tricky in scenarios with a complex training step. Moreover, if it is required to train the model for all 2^m coalitions of features, this can be timely infeasible. In this paper, we propose a preprocessing approach that assigns importance measures to features and interaction indices to coalitions, *i.e.,* without the need of a trained model. We discuss our proposal in the next section.

5 PROPOSED APPROACH

The use of the HSIC (more precisely, the normalized version NOCCO) in our analysis was inspired by the work conducted in [46]. The authors proposed a preprocessing approach based on the NOCCO value to detect disparity prone features. The idea is that, greater the NOCCO value of a feature (or subfeature), greater the chance that such a feature (or subfeature) entails disparate results. However, the analysis conducted in [46] was only based on dependence measures between the vector representing a feature (or subfeature) and the vector of labels. The authors did not consider, for instance, interactions between features. In order to exploit the marginal dependence degree of features towards the trained model performance and fairness metrics, in this paper, we adopted an approach based on cooperative game theory. Instead of detecting relevant and disparity prone features based on single measures of dependence between vectors, we consider all coalitions of features when calculating the marginal importance of features.

In the game theory framework, we assume as the payoff of a coalition A the dependence measure between features in A and the output vector $\mathbf{y}$. Mathematically, we define $v(A) = NOCCO(\mathbf{X}^{(A)}, \mathbf{y}) = \text{tr}\left(\mathbf{R}_{\mathbf{X}^{(A)}}\mathbf{R}_{\mathbf{y}}\right)$, where $\mathbf{X}^{(A)}$ represents the input data whose columns are composed by features (or subfeatures after converting categorical features into binary variables) in the subset A. Note that, in [46], the authors used $v(\{j\}) = NOCCO(\mathbf{X}^{(j)}, \mathbf{y})$ for singletons $j = 1, \ldots, m$ to detect disparity prone features. In this paper, we consider the Shapley value ϕ_j as the measure indicating the relevance of feature G_j towards the trained model performance and fairness measures. The marginal contribution of feature G_j is defined as follows:

$$\phi_j^{NOCCO} = \sum_{A \subseteq M \setminus \{j\}} \frac{(m-|A|-1)!\,|A|!}{m!} \times \left[NOCCO(\mathbf{X}^{(A\cup\{j\})}, \mathbf{y}) - NOCCO(\mathbf{X}^{(A)}, \mathbf{y})\right]. \quad (5)$$

For the empty set, we assumed $v(\emptyset) = 0$, as there is no data to calculate the dependence degree with the output vector $\mathbf{y}$.

Clearly, we considered NOCCO values between any coalitions of features (or transformed data, after converting categorical features into binary ones) and the vector of labels in the Shapley value calculation. Based on the Shapley values, we may estimate relevant features towards the trained model performance before the training step. The idea is that, greater ϕ_j^{NOCCO}, greater is the impact of feature G_j in the trained model. Recall from the efficiency property that $\sum_{j=1}^m \phi_j^{NOCCO} = NOCCO(\mathbf{X}, \mathbf{y})$, *i.e.,* the dependence degree of the whole dataset $\mathbf{X}$ is decomposed by the marginal contributions of each feature G_j. Moreover, in scenarios with sensitive features, high marginal contributions will also indicate the ones that may entail ethical disparities.

We also considered in our analysis the Shapley interaction indices between features $G_j, G_{j'}$, defined by

$$I_{j,j'}^{NOCCO} = \sum_{A \subseteq M \setminus \{j,j'\}} \frac{(m-|A|-2)!\,|A|!}{(m-1)!} \times \Big[NOCCO(\mathbf{X}^{(A\cup\{j,j'\})}, \mathbf{y}) - NOCCO(\mathbf{X}^{(A\cup\{j\})}, \mathbf{y}) - NOCCO(\mathbf{X}^{(A\cup\{j'\})}, \mathbf{y}) + NOCCO(\mathbf{X}^{(A)}, \mathbf{y})\Big].$$

As will be further discussed in the numerical experiments (see Sections 6 and 7.3), while the Shapley values indicate the marginal contribution of features, the interaction indices will indicate how pairs of features interact towards the model performance. For example, in the case when the interaction index between features j and j' is negative (*i.e.,* there is a redundant effect between them), these features may be acting as proxies. In other words, the use of both features may bring similar results as in the case where only one of them is considered. Furthermore, a positive interaction index may indicate a synergistic effect on the model performance.

Algorithm 1 presents a pseudo-code of our proposal. For each coalition of features $A \subseteq M$, one firstly encodes the categorical features in A (if it is necessary). Then, we calculate the NOCCO value between $NOCCO(\mathbf{X}^{(A)}$ and the vector of outcomes $\mathbf{y}$. Once all $NOCCO(\mathbf{X}^{(A)}$ was obtained, we calculate the NOCCO Shapley values and interactions indices. The Shapley values will indicate which features are more relevant in predicting the vector of outcomes and the interaction indices will highlight redundant and/or complementary effects between pairs of features.

It is worth mentioning that, in our analysis we considered the radial basis function (RBF) kernel in the NOCCO values calculation.

Algorithm 1 (*Preprocessing approach to detect relevant and disparity prone features*)

Input: $\mathbf{X}$ and $\mathbf{y}$.
Output: Features contributions ϕ_j^{NOCCO} and interaction indices $I_{j,j'}^{NOCCO}$, $j, j' = 1, \ldots, m$.
1: **Calculate the kernel matrix of** $\mathbf{y}$: $\mathbf{K}_\mathbf{y} = kernel\,(\mathbf{y})$.
2: **Calculate the NOCCO values for all coalitions of features**:

for $A \subseteq M$ **do**
  Define the (initially empty) dataset used in NOCCO calculation: $\mathbf{X}^{(A)} = []$.
  for $j \in A$ **do**
    if G_j is either a numerical or binary feature **then**
      Update the dataset: $\mathbf{X}_{NOCCO} = \left[\mathbf{X}^{(A)}; \mathbf{X}^{(j)}\right]$.
    else
      Encode the categorical feature: $\tilde{\mathbf{X}}^{(j)} \leftarrow encode\left(\mathbf{X}^{(j)}\right)$.
      Update the dataset: $\mathbf{X}_{NOCCO} = \left[\mathbf{X}^{(A)}; \tilde{\mathbf{X}}^{(j)}\right]$.
    end if
  end for
  Calculate the kernel matrix associated with $\mathbf{X}^{(A)}$: $\mathbf{K}_{\mathbf{X}_{NOCCO}} = kernel\left(\mathbf{X}^{(A)}\right)$.
  Calculate the dependence measure: $NOCCO(\mathbf{X}^{(A)} = \text{tr}\left(\mathbf{R}_{\mathbf{X}^{(A)}}\mathbf{R}_\mathbf{y}\right)$
end for
for $j = 1, \ldots, m$ **do**
  Calculate the Shapley value: ϕ_j^{NOCCO} (Eq. (5)).
  for $j' = j+1, \ldots, m$ **do**
    Calculate the interaction index: $I_{j,j'}^{NOCCO}$ (Eq. (6)).
  end for
end for

The RBF kernel is defined by

$$K^{RBF}(h_i, h_{i'}) = e^{-\frac{1}{n}(h_i - h_{i'})^2}. \tag{6}$$

However, there exists other kernels that could be used in our formulation [54].

6 ILLUSTRATIVE EXAMPLE WITH SYNTHETIC DATA

In order to illustrate the application of our proposal in assigning an importance measure to each feature, we first consider a synthetic dataset with $n = 3000$ samples and $m = 5$ features. The dataset $\mathbf{X}$ was randomly generated as follows: **(1)** for $j \in \{1, 2, 3, 4\}$, $\mathbf{X}^{(j)} \sim \mathcal{U}(0, 1)$, were $\mathcal{U}(0, 1)$ indicates a uniform distribution in the range $[0, 1]$; **(2)** for $j = 5$, $\mathbf{X}^{(5)}$ is a copy $\mathbf{X}^{(4)}$. Before defining the class of each sample, assume the weights $\mathbf{w} = [w_1, w_2, w_3, w_4, w_5] = [0.25, 0.40, 0, 0.15, 0.20]$ and the aggregation function

$$z_i = \sum_{j=1}^{5} w_j X^{(j)} + 0.01\eta_i, \quad i = 1, \ldots, 3000, \tag{7}$$

where $\eta_i \sim \mathcal{N}(0, 1)$ is a random Gaussian noise (with zero mean and unit variance). The classes were defined as follows: $y_i = 1$ if $z_i > \hat{z}$, and $y_i = 0$, otherwise, where $\hat{z} = \frac{1}{3000}\sum_{i=1}^{3000} z_i$ is the mean value of $\mathbf{z} = [z_1, \ldots, z_{3000}]$.

The application of our proposal leads to the features importance presented in Figure 1a. The results are in accordance with the weights used to aggregate the dataset and define the classes. Feature 2, which is associated with the highest weight ($w_2 = 0405$) in the aggregation function, achieved the highest Shapley value. The contribution assigned to feature 3 was practically zero. This attests the null player property of the Shapley value, which ensures that features with no contribution into the predicted outcome would receive zero payoff. As features 4 and 5 are identical (recall that $\mathbf{X}^{(5)}$ is a copy $\mathbf{X}^{(4)}$) and given the symmetry property of the Shapley value, the obtained marginal contributions for such features were the same regardless the associated weights. Clearly, the magnitude of w_4 and w_5 impacts both ϕ_4^{NOCCO} and ϕ_5^{NOCCO}, however, both features 4 and 5 equally share their impacts on the output vector. These results highlight the novelty of our approach in comparison with [46]. As illustrated in Figure 1b, if one only considers the individual NOCCO values, redundant features will achieve higher importance measures than the correct one. Indeed, redundant features should share their total impact instead of receiving the same total impact.

Another interesting result can be seen in the interaction effects presented in Figure 1c. Note that $I_{4,5}^{NOCCO} \approx -0.3$, *i.e.,* the interaction index attests that features G_4 and G_5 are redundant in predicting the classes. Therefore, besides estimating the feature contributions towards the model prediction, our approach also indicates proxies. Although this could be verified by taking a similarity degree between features (such as the correlation coefficient), proxies may also exists between categorical features. In such a case, in contrast of similarities measures which are difficult to be calculated, our approach is able to indicate which features can be seen as proxies.

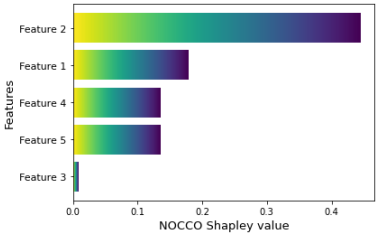
7 NUMERICAL EXPERIMENTS ON REAL DATA

In this section, we present the numerical experiments based on real datasets. Differently from the previous section where we know the aggregation function parameters, by assuming real datasets we do not know the relation between input features and the outcome vector. Given the absence of the ground truth weights to compare with the obtained feature contributions (NOCCO Shapley values), we will compare them with the Shapley values calculated after training the ML model and collecting the performance measures. Therefore, for each subset of features $A \subseteq M$, we trained the model[6] and calculated the performance measure. In this paper, we considered the overall accuracy, defined by

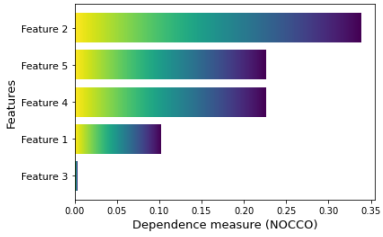
$$OA_A = \frac{TP_A + TN_A}{n}, \tag{8}$$

where TP_A and TN_A are the true positive (# of instances correctly classified as class 1) and the true negative (# of instances correctly

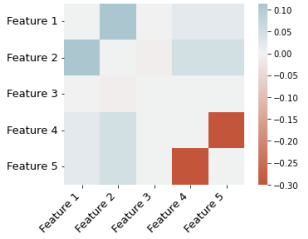
[6]In all experiments conducted in this paper, we applied a Neural Network classifier borrowed from Scikit-learn library [45] in Python (MLPClassifier), with at most 1000 iterations. We performed similar experiments with Random Forests with similar results. These are presented in the Supplementary Material. As our proposal consists in a preprocessing step independent from the training step, it is model-agnostic, that is, any ML model could be used.



(a) Features importance.



(b) NOCCO values for individual features.



(c) Interaction effects between features.

Figure 1: Results for the synthetic dataset.

classified as class -1) calculated from the trained model based on features in A. Once all OA_A, $A \subseteq M$, are calculated, one obtains the overall accuracy Shapley values as follows:

$$\phi_j^{OA} = \sum_{A \subseteq M \setminus \{j\}} \frac{(m - |A| - 1)!\,|A|!}{m!} \left[OA_{A \cup \{j\}} - OA_A \right]. \quad (9)$$

As the ϕ_j^{OA} indicates the importance measure of feature j in the trained machine, we compare it with our preprocessing proposed approach.

In some scenarios, we also aim at detecting disparity prone features. For this purpose, we compare the obtained features contributions with a fairness metric (also calculated after training the ML model). In this paper, we adopted the overall accuracy equality. As some sensitive features have more than two categories, for each sensitive feature G_j with q categories, we calculated the average disparity when splitting the samples according to all possible pairs of categories. This leads to the following overall accuracy equality considered in this paper:

$$OAE = \frac{1}{C_{q,2}} \sum_{k,k',k \neq k'} \left| \frac{TP_{G_{j,k}} + TN_{G_{j,k}}}{n_{G_{j,k}}} - \frac{TP_{G_{j,k'}} + TN_{G_{j,k'}}}{n_{G_{j,k'}}} \right|,$$

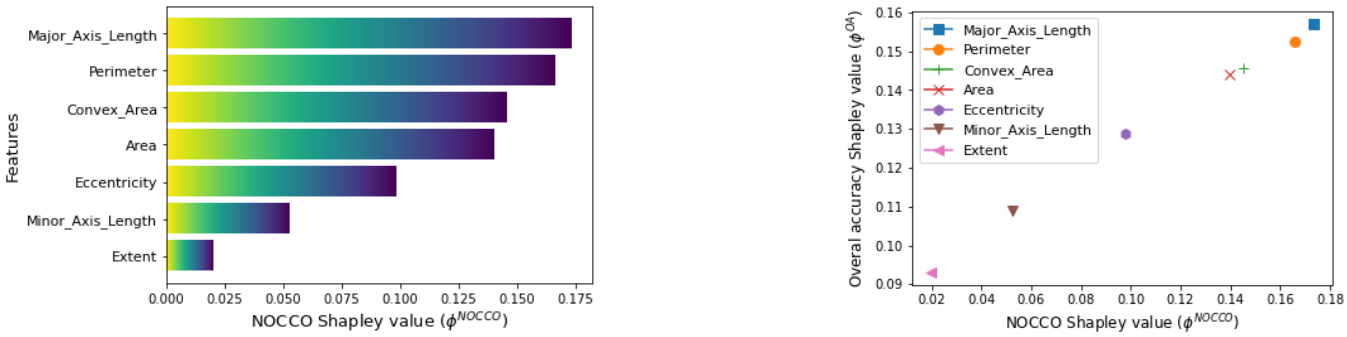
where $n_{G_{j,k}}$ and $n_{G_{j,k'}}$ are the number of samples that belong to group $G_{j,k}$ and $G_{j,k'}$, respectively, and $C_{q,2} = \frac{q!(q-1)!}{2}$ is the number of pairs of categories for feature G_j. The supplementary material as well as all datasets and codes are freely available at https://github.com/GuilhermePelegrina/NOCCO_Shapley_values.git.

7.1 Datasets

We attest our proposal based on four datasets frequently considered in the literature[7]: Rice [14], Red Wine Quality [16], COMPAS recidivism risk [8] and Adult[8]. We used the first two to estimate the relevance of features towards the model performance. For both COMPAS and Adult datasets, in addition to feature importance, we also explore the detection of disparity prone features. We briefly describe each dataset in the sequel.

- **Rice dataset**: In this dataset, the aim is to identify specie of rices (Cammeo or Osmancik). There are $n = 3810$ samples and the following $m = 7$ rices characteristics: *Area*, *Perimeter*, *Major_Axis_Length*, *Minor_Axis_Length*, *Eccentricity*, *Convex_Area* and *Extent*.
- **Red Wine Quality dataset**: This dataset contains $n = 1599$ samples of red wines described by $m = 11$ features: *fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *pH*, *sulphates* and *alcohol*. For each wine, we also have an score (from 0 to 10) indicating its quality. The purpose is to classify each wine as good or a bad one. In our analysis, we consider as good (resp. bad) wines the ones with score greater than (resp. at most) 5.
- **COMPAS dataset**: In this dataset, the aim is to classify individuals as a potential criminal recidivist. Each individual (for a total of $n = 6167$) is characterized by the following $m = 8$ features: *sex* (male or female), *age_cat* (age category - less than 25, between 25 and 45 or greater than 45), *race* (African Americans, Caucasians or others), *juv_fel_count* (number of juvenile felony), *juv_misd_count* (number of juvenile misdemeanor), *juv_other_count* (number of others infractions), *priors_count* (number of priors) and *c_charge_degree* (type of the charge degree - felony or misdemeanor). Although only race and sex are typically considered as sensitive features, in our experiments, we also evaluate ethical disparities associated with age category.
- **Adult dataset**: This dataset is composed by $n = 45222$ people described by $m = 12$ features (after removing missing values and grouping some categories - see [31] for further details), namely *age*, *workclass*, *educational-num* (educational degree), *marital-status* (married, never married or other), *occupation*, *relationship* (husband, not in family, other relative, own child, unmarried or wife), *race* (split among Indian-Eskimo, Asian-Pacific Islander, black, white or other), *gender* (male or female), *capital-gain*, *capital-loss*, *hours-per-week* and *native-country* (US or non-US). For each person, the goal is to predict whether he/she makes over 50K a year. Age, gender and race are considered the sensitive features within this dataset. However, we also evaluate unfair results entailed by marital status, relationship and native country.

[7] See the supplementary material for more results in other datasets.
[8] https://archive.ics.uci.edu/ml/datasets/adult



(a) Features importance. **(b) Relation between ϕ^{NOCCO} and ϕ^{OA}.**

Figure 2: Results for the Rice dataset.

7.2 Estimating the importance of features towards the model performance

As a first analysis, we aim at assigning an importance measure to each feature which will be used to evaluate how relevant they are in predicting the output vector. For this analysis, we considered the Red Wine Quality and Rice datasets. The results are discussed in the sequel.

Rice dataset. We present the obtained NOCCO Shapley values in Figure 2a. Note that, while the *Major_Axis_Length* and *Perimeter* have high marginal contribution towards the whole dependence degree, the impact of *Extend* is very low. The scatter plot in Figure 2b shows a comparison between the NOCCO Shapley values and the overall accuracy Shapley values. We can see a clear relation between both measures (almost perfect, in this case), with a Spearman's rank correlation $\rho \approx 1$. Therefore, we attested our hypothesis that by using marginal dependence measures in the preprocessing step, we can estimate the contribution of features towards the trained model (*i.e.,* in a postprocessing step). The time used to compute all dependence measures was 620.00 seconds, which indicates an average of 4.84 seconds for each coalition of features[9].

Red Wine Quality dataset. Figure 3 presents the results for the Red Wine Quality dataset. As can be seen in Figure 3a, although all features have relevant Nocco Shapley values, volatile acidity, sulphates and alcohol (mainly the latter one) are the ones with the highest contributions towards the dependence degree. The relation between ϕ^{NOCCO} and ϕ^{OA} is depicted in Figure 3b. As in the previous experiment, we also see an strong relation between both importance measures ($\rho = 0.91$). Moreover, in this dataset, the time used to compute all dependence measures was 1700 seconds (*i.e.*, an average of 0.83 seconds for each coalition of features).

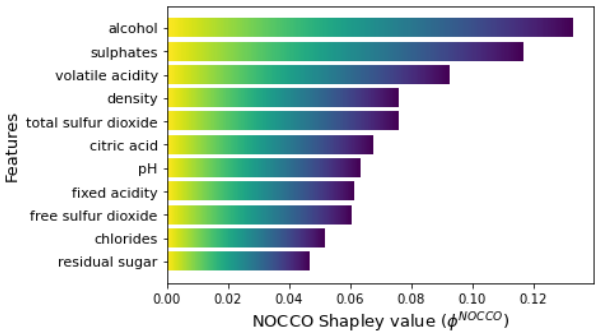
7.3 Detecting disparity prone features

Similarly as in the previous analysis, in this second set of experiments we also calculate the NOCCO Shapley values and evaluate their relation with features importance assigned after training the ML model. However, as in the following datasets we have ethical concerns associated with sensitive features, we also highlight how the NOCCO Shapley values can be used to detect disparity prone features.

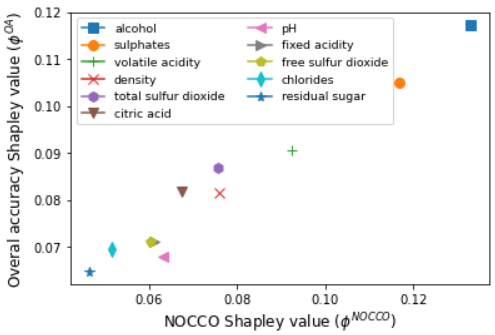
COMPAS dataset. The results for the COMPAS dataset are presented in Figure 4. Note from Figure 4a that the number of priors, age category and race are the features with higher NOCCO Shapley values. In Figure 4c, we attest the relation between NOCCO and overall accuracy Shapley values ($\rho = 0.93$). Note that some features frequently considered as sensitives, such as age category and race, are important in predicting recidivism risk. We confirm the hypothesis that such features are prone to entail disparate results in Figure 4d, where there is a clear relation between the NOCCO Shapley values and the overall accuracy equality ($\rho \approx 1$). Greater the Shapley value of a sensitive feature, greater the disparity when splitting the dataset based on such a feature. Indeed, both age category and race are associated with the highest inequalities in this dataset. On the other hand, it is interesting to remark that although gender is assumed as a sensitive feature, its NOCCO Shapley value is very low and it does not entail disparity results.

Other relevant findings in the COMPAS dataset are depicted in Figure 4b. The interaction indices indicate that there are interactions effects between pairs of features. The complementary effect between *age_cat* and *priors_count* can be explained by how the use of both features simultaneously contributes towards predicting recidivism risk (indeed, $0.2977 = NOCCO(\mathbf{X}^{(\{age\_cat,priors\_count\})}, \mathbf{y}) > NOCCO(\mathbf{X}^{(\{age\_cat\})}, \mathbf{y}) + NOCCO(\mathbf{X}^{(\{priors\_count\})}, \mathbf{y}) = 0.0754 + 0.1706 = 0.2460$). If one only considers the number of priors, one would classify as possible recidivists the individuals with at least a predetermined number of priors. However, by including the age category into the classification task, one may define different thresholds for the number of priors according to the age. For instance, for a certain numbers of priors, younger individuals can be classified as recidivists while older ones will only be for a greater number of priors. On the other hand, the use of *race* with *priors_count* is not better than both of them individually (indeed, $0.2079 = NOCCO(\mathbf{X}^{(\{race,priors\_count\})}, \mathbf{y}) < NOCCO(\mathbf{X}^{(\{race\})}, \mathbf{y}) + NOCCO(\mathbf{X}^{(\{priors\_count\})}, \mathbf{y}) = 0.0812 + 0.1706 = 0.2518$). An explanation for this negative effect lies in the average number of priors of each race, which are 4.24, 2.29 and 1.98

[9]Computations performed on a laptop Intel Core i7-8565U, CPU 1.80 GHz, 8.00 GB RAM, Python 3.9.



(a) Features importance.



(b) Relation between ϕ^{NOCCO} and ϕ^{OA}.

Figure 3: Results for the Red Wine Quality dataset.

for African Americans, Caucasians and other races, respectively. Therefore, there is a degree of redundancy between *race* with *priors_count*, African Americans are likely to have more priors than Caucasians which are likely to have more priors than other races. The algorithm spent 2470.40 seconds to compute all dependence measures. This constitutes to an average time of 9.65 seconds for each coalition.

Adult dataset. We present the results for the Adult dataset in Figure 5. The NOCCO Shapley values described in Figure 5a indicates that *relationship* and *capital-gain* are the features with more contributions towards the total dependence degree. The relation between ϕ^{NOCCO} and ϕ^{OA} is presented in Figure 5c. Although the points does not compose a straight line[10], we note a positive relation between both importance measures ($\rho = 0.91$). A positive relation is also achieved between the NOCCO Shapley values and the fairness measure, with a Spearman's rank correlation $\rho = 0.94$ (see Figure 5d). Among the features frequently considered as sensitive (age, gender and race), age is the one with the highest chance of entailing ethical disparities (which was confirmed by the associated OAE).

It is important to highlight the negative interaction index achieved for features *marital-status* and *relationship* in Figure 5b. Indeed, both features bring very similar information. For instance, being a husband or a wife (relationship categories) implies that the person is married (marital status category). Therefore, as pointed out by the interaction index, both features have a relevant degree of redundancy and can be considered as proxies. Moreover, as *relationship* is related to *gender* (note that being a husband or a wife generally indicates a man or woman, respectively) and marital status is related to age (we tend to have more aged married people than younger ones), one should be aware of such features when training the model. Although they are not typically assumed as sensitive ones [31], as they bring sensitive information, they can also entail disparate results. With respect to the computational time, all dependence measures were computed in 18100.00 seconds, that is, an average of 4.41 seconds for each coalition.

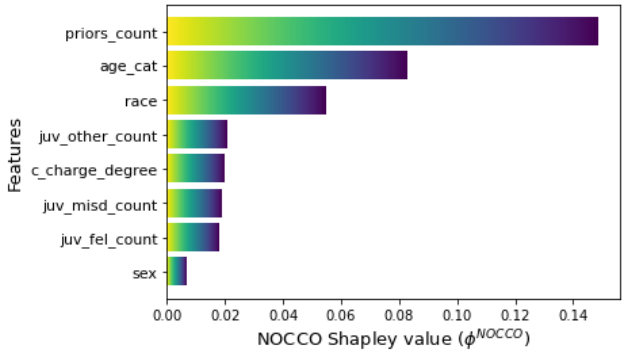
[10]It is worth highlighting that our approach is based on statistics measures extracted from the dataset. Moreover, the training step is not deterministic. Therefore, both characteristics explain the deviation from a perfect relation between the NOCCO and the overall accuracy Shapley values.

8 CONCLUSION AND PERSPECTIVES

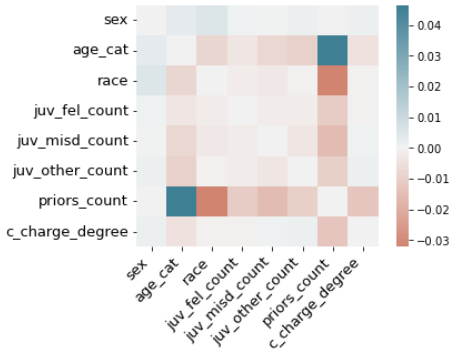
In this paper, we proposed a preprocessing approach to detect relevant and disparity prone features based on a normalized version of the Hilbert-Schmidt independence criterion and the Shapley values. A novelty of our proposal in comparison with existing methods is that we do not need a trained model to estimate feature contributions towards the model performance. Therefore, we reduce this effort when conducting feature importance analysis. Indeed, in the empirical experiments, we attested the relation between our proposal and methods based on a trained model. Greater the NOCCO Shapley value, greater is the feature contributions towards the model performance. Moreover, we also proposed to use this marginal dependence degree to evaluate disparity prone features. As also attested by the experimental results, our proposed approach can be used to detect such features. Similarly as with the algorithm performance, sensitive features with high marginal dependence degree with the vector of labels are the ones that entail high disparate outcomes.

Besides evaluating features contributions, our proposal also indicates feature interactions. As could be noted from the COMPAS dataset, the positive interaction indices between the numbers of priors and the age category suggests that there is a complementary effect between such features and, therefore, the use of both simultaneously, brings relevant information to the classifier predictive power. Moreover, from the Adult income dataset, we could highlight the use of our proposal as a mechanism to investigate the presence of proxies. Indeed, although marital status is not assumed as a sensitive feature, in this dataset, its redundant effect with gender indicates that it is acting as a proxy.

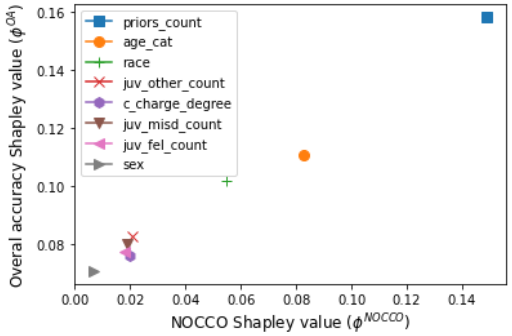
As future work it would be interesting to consider further combinations of features and their impact on fairness metrics. Indeed, we looked at disparities from a global perspective, but a fine grained analysis may reveal hidden disparities. For instance, it could be the case that gender disparities are overlooked when looking at the European population as a whole, but that they become apparent when focusing on different ethnic groups. Also, it would be interesting to compare our utility based method (here, focused on the HSIC and Shapley values) with other approaches assessing marginal contributions, such as in [27] where optimal transport distances are used instead, as well as evaluate the robustness of the method to noisy



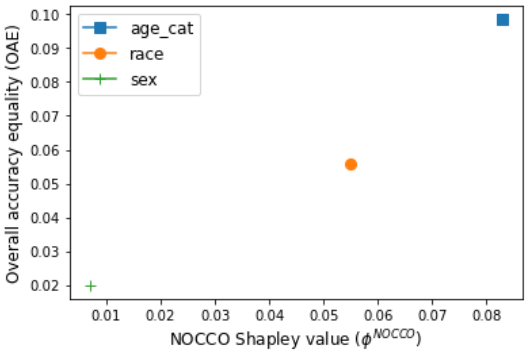
(a) Features importance.



(b) Interaction effects between features.

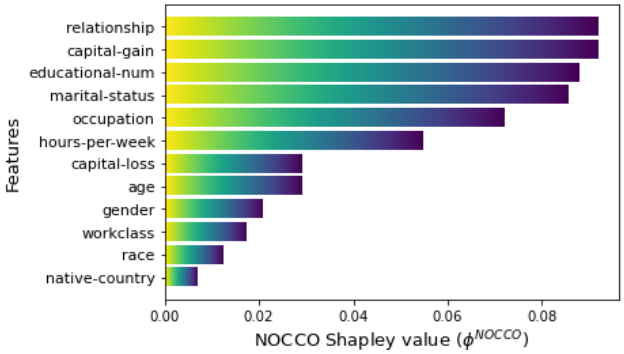


(c) Relation between phi NOCCO and phi OA.



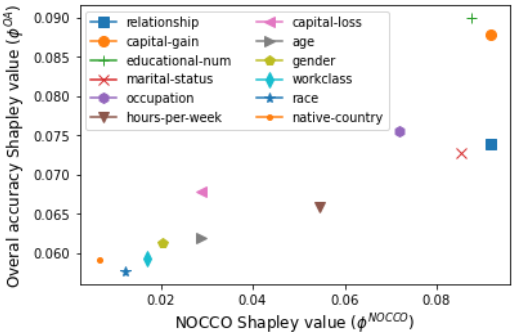
(d) Relation between phi NOCCO and OAE.

Figure 4: Results for the COMPAS dataset.

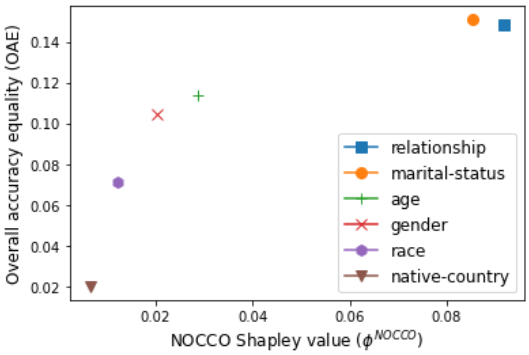


(a) Features importance.

(b) Interaction effects between features.



(c) Relation between phi NOCCO and phi OA.



(d) Relation between phi NOCCO and OAE.

Figure 5: Results for the Adult dataset.

or out-of-sample data. Another potentially interesting extension is to adapt our framework to other data types such as audio data[11] where performance disparities have been detected in models such as automatic speech recognition (ASR) systems [30][12]. Furthermore, our approach can be naturally applicable to multiclass classification settings, and this constitutes a topic of ongoing research.

ACKNOWLEDGMENTS

This work was partially done while G. Pelegrina was visiting the Orpailleur team of LORIA (UMR 7503). The research of M. Couceiro was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation program under GA No 95221. G. Pelegrina and L. Duarte thank the grants #2020/09838-0 (BI0S - Brazilian Institute of Data Science), #2020/10572-5 and #2021/11086-0, São Paulo Research Foundation (FAPESP), for the financial support. L. T. Duarte would also like to thank the National Council for Scientific and Technological Development (CNPq, Brazil) for the financial support.

REFERENCES

[1] Kjersti Aas, Martin Jullum, and Anders Løland. 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence* 298 (2021), 103502. https://doi.org/10.1016/j.artint.2021.103502

[2] Behnoush Abdollahi and Olfa Nasraoui. 2018. Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems. In *Human and Machine Learning. Human–Computer Interaction Series.*, J. Zhou and F. Chen (Eds.). Springer, Cham, 21–35. https://doi.org/10.1007/978-3-319-90403-0_2

[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.

[4] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. 2022. Counterfactual Shapley Additive Explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Seoul, Republic of Korea, 1054–1070. https://doi.org/10.1145/3531146.3533168

[5] Guilherme Alves, Maxime Amblard, Fabien Bernier, Miguel Couceiro, and Amedeo Napoli. 2021. Reducing Unintended Bias of ML Models on Tabular and Textual Data. In *8th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2021, Porto, Portugal, October 6-9, 2021*. IEEE, 1–10.

[6] Guilherme Alves, Fabien Bernier, Miguel Couceiro, Karima Makhlouf, Catuscia Palamidessi, and Sami Zhioua. 2023. Survey on Fairness Notions and Related Tensions. *EURO Journal on Decision Processes* 11 (2023), 100033. https://doi.org/10.1016/j.ejdp.2023.100033

[7] Marco Ancona, Cengiz Öztireli, and Markus Gross. 2019. Explaining deep neural networks with a polynomial time algorithm for Shapley values approximation. In *36th International Conference on Machine Learning*, Vol. 97. PMLR, 272–281.

[8] Julia. Angwin, Jeff. Larson, Surya. Mattu, and Lauren. Kirchner. 2016. Machine Bias - ProPublica. \https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[9] Nuno Antunes, Leandro Balby, Flavio Figueiredo, Nuno Lourenco, Wagner Meira, and Walter Santos. 2018. Fairness and Transparency of Machine Learning for Trustworthy Cloud Services. In *Proceedings - 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W 2018)*. IEEE, Luxembourg, Luxembourg, 188–193. https://doi.org/10.1109/DSN-W.2018.00063

[10] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. *ArXiv ID: 2010.07389* (2020). http://arxiv.org/abs/2010.07389

[11] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[12] Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. 2019. Visualizing the Feature Importance for Black Box Models. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science*, G. Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim (Ed.). Vol. 11051. Springer, Cham, 655–670. https://doi.org/10.1007/978-3-030-10925-7_40

[13] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. 2021. The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing: Overview and techniques. *IEEE Signal Processing Magazine* 38, 6 (2021), 39–49.

[14] Ilkay Cinar and Murat Koklu. 2019. Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering* 7, 3 (2019), 188–194. https://doi.org/10.18201/ijisae.2019355381

[15] Shay Cohen, Gideon Dror, and Eytan Ruppin. 2007. Feature Selection via Coalitional Game Theory. *Neural Computation* 19 (2007), 1939–1961. https://doi.org/10.1162/neco.2007.19.7.1939

[16] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* 47, 4 (2009), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

[17] Kenji Fukumizu, Francis R. Bach, and Arthur Gretton. 2007. Statistical Consistency of Kernel Canonical Correlation Analysis. *Journal of Machine Learning Research* 8 (2007), 361–383.

[18] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2007. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20 (NIPS)*.

[19] Javid Ghahremani nahr, Hamed Nozari, and Mohammad Ebrahim Sadeghi. 2021. Artificial intelligence and Machine Learning for Real-world problems (A survey). *International Journal of Innovation in Engineering* 1, 3 (2021), 38–47. https://doi.org/10.59615/ijie.1.3.38

[20] Michel Grabisch. 1997. Alternative Representations of Discrete Fuzzy Measures for Decision Making. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 05, 05 (1997), 587–607. https://doi.org/10.1142/S0218488597000440

[21] Michel Grabisch. 2016. *Set Functions, games and capacities in decision making*. Springer International Publishing, Switzerland. https://doi.org/10.1007/978-3-319-30690-2

[22] Przemyslaw A. Grabowicz, Nicholas Perello, and Aarshee Mishra. 2022. Marrying Fairness and Explainability in Supervised Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Seoul, Republic of Korea, 1905–1916. https://doi.org/10.1145/3531146.3533236

[23] Alex Gramegna and Paolo Giudici. 2022. Shapley Feature Selection. 1 (2022), 72–80. https://doi.org/10.3390/fintech1010006

[24] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings 16*. Springer Berlin Heidelberg, Singapore, 63–77.

[25] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. AAAI Press, New Orleans, Lousiana, USA, 51–60.

[26] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.

[27] Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. 2023. LAVA: Data Valuation without Pre-Specified Learning Algorithms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

[28] Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence* 294 (2021), 103459. https://doi.org/10.1016/j.artint.2021.103459

[29] Sotiris B. Kotsiantis. 2011. Feature selection for machine learning classification problems: A recent overview. *Artificial Intelligence Review* 42, 1 (2011), 157–176. https://doi.org/10.1007/s10462-011-9230-1

[30] Ajinkya Kulkarni, Anna Tokareva, Mohammed Rameez Qureshi, and Miguel Couceiro. 2024. The Balancing Act: Unmasking and Alleviating ASR Biases in Portuguese. In *EACL 2024 LT-EDI WorkShop*. St. Julians, Malta. https://hal.science/hal-04436147

[31] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 3 (2022), 1–59. https://doi.org/10.1002/widm.1452

[32] Zhu Li, Adrián Pérez-Suay, Gustau Camps-Valls, and Dino Sejdinovic. 2022. Kernel dependence regularizers and Gaussian processes with applications to algorithmic fairness. *Pattern Recognition* 132 (2022), 108922. https://doi.org/10.1016/j.patcog.2022.108922

[33] Stan Lipovetsky and Michael Conklin. 2001. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry* 17, 4 (2001), 319–330. https://doi.org/10.1002/asmb.446

[34] Scott M. Lundberg. 2020. Explaining quantitative measures of fairness. In *Fair & Responsible AI Workshop CHI2020*.

[11] https://ai.meta.com/datasets/casual-conversations-v2-dataset/

[12] https://biasinai.github.io/

[35] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[36] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). 4765–4774.

[37] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2, 10 (2018), 749–760. https://doi.org/10.1038/s41551-018-0304-0

[38] Ninareh. Mehrabi, Fred. Morstatter, Nripsuta. Saxena, Kristina. Lerman, and Aram. Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019). http://arxiv.org/abs/1908.09635

[39] Luke Merrick and Ankur Taly. 2020. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. In *Machine Learning and Knowledge Extraction. CD-MAKE 2020. Lecture Notes in Computer Science*, A. Holzinger, P. Kieseberg, A. Tjoa, and E. Weippl (Eds.), Vol. 12279. Springer, Cham, 17–38. https://doi.org/10.1007/978-3-030-57321-8_2

[40] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2021. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Virtual Event, Canada, 386–400. https://doi.org/10.1145/3442188.3445902

[41] Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. 2022. Sampling Permutations for Shapley Value Estimation. *Journal of Machine Learning Research* 23 (2022), 1–46.

[42] T. Murofushi and S. Soneda. 1993. Techniques for reading fuzzy measures (III): interaction index. In *9th fuzzy system symposium* (Sapporo, Japan). 693–696.

[43] Preetam Nandy, Cyrus Diciccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. 2022. Achieving Fairness via Post-Processing in Web-Scale Recommender Systems*. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Seoul, Republic of Korea, 715–725. https://doi.org/10.1145/3531146.3533136

[44] Neel Patel, Martin Strobel, and Yair Zick. 2021. High dimensional model explanations: An axiomatic approach. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Virtual event, Canada, 401–411. https://doi.org/10.1145/3442188.3445903

[45] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[46] Guilherme Dean Pelegrina, Miguel Couceiro, and Leonardo Tomazeli Duarte. 2023. A statistical approach to detect disparity prone features in a group fairness setting. *AI and Ethics* (2023). https://doi.org/10.1007/s43681-023-00363-9

[47] Guilherme Dean Pelegrina and Leonardo Tomazeli Duarte. 2023. A Novel Approach for Fair Principal Component Analysis based on Eigendecomposition. *IEEE Transactions on Artificial Intelligence* (2023), 1–12. https://doi.org/10.1109/TAI.2023.3298291

[48] Guilherme Dean Pelegrina, Leonardo Tomazeli Duarte, and Michel Grabisch. 2023. A k-additive Choquet integral-based approach to approximate the SHAP values for local interpretability in machine learning. *Artificial Intelligence* 325 (2023), 104014. https://doi.org/10.1016/j.artint.2023.104014

[49] Guilherme Dean Pelegrina, Leonardo Tomazeli Duarte, and Michel Grabisch. 2023. Interpreting the contribution of sensors in blind source extraction by means of Shapley values. *IEEE Signal Processing Letters* 30, 1 (2023), 878–882. https://doi.org/10.1109/LSP.2023.3295759

[50] Guilherme Dean Pelegrina and Sajid Siraj. 2022. Shapley value-based approaches to explain the robustness of classifiers in machine learning. *ArXiv ID: 2209.04254* (2022). http://arxiv.org/abs/2209.04254

[51] Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. 2022. The Shapley Value in Machine Learning. In *IJCAI International Joint Conference on Artificial Intelligence*. Vienna, Austria, 5572–5579. https://doi.org/10.24963/ijcai.2022/778

[52] Samira Samadi, Uthaipon Tantipongpipat, Jamie Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The price of fair PCA: One extra dimension. In *Advances in Neural Information Processing Systems*. 10976–10987.

[53] Iqbal H. Sarker. 2021. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science* 2, 3 (2021), 160. https://doi.org/10.1007/s42979-021-00592-x

[54] Bernhard Schölkopf, Alexander J. Smola, and Francis Bach. 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* MIT press, Cambridge, MA, USA.

[55] Lloyd S. Shapley. 1953. A value for n-person games. In *Annals of mathematics studies: Vol. 28. Contributions to the theory of games, Vol. II*, W. Kuhn and A. W. Tucker (Eds.). Princeton University Press, Princeton, 307–317.

[56] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. 2012. Feature selection via dependence maximization. *Journal of Machine Learning Research* 13 (2012), 1393–1434.

[57] Le Song, Alex Smola, Arthur Gretton, Karsten M. Borgwardt, and Justin Bedo. 2007. Supervised feature felection via dependence estimation. In *Proceedings of the 24th international conference on Machine learning*. 823–830.

[58] Mukund Sundararajan and Amir Najmi. 2020. The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*. PMLR, 9269–9278. arXiv:1908.08474

[59] Jiliang Tang, Salem Alelyani, and Huan Liu. 2014. Feature selection for classification: A review. In *Data Classification: Algorithms and Applications* (1st ed.), Charu C. Aggarwal (Ed.). CRC Press, 37–64. https://doi.org/10.1201/b17320

[60] B. Venkatesh and J. Anuradha. 2019. A review of Feature Selection and its methods. *Cybernetics and Information Technologies* 19, 1 (2019), 3–26. https://doi.org/10.2478/CAIT-2019-0001

[61] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ACM/IEEE International Workshop on Software Fairness*. IEEE, 1–7. https://doi.org/10.1145/3194770.3194776

[62] Hao Wang, Yijie Ding, Jijun Tang, and Fei Guo. 2020. Identification of membrane protein types via multivariate information fusion with Hilbert–Schmidt independence criterion. *Neurocomputing* 383 (2020), 257–269. https://doi.org/10.1016/j.neucom.2019.11.103

[63] Tinghua Wang, Xiaolu Dai, and Yuze Liu. 2021. Learning with Hilbert–Schmidt independence criterion: A review and new perspectives. *Knowledge-Based Systems* 234 (2021), 107567. https://doi.org/10.1016/j.knosys.2021.107567

[64] Eyal Winter. 2002. The Shapley value. In *Handbook of game theory with economic applications*. Vol. 3. Elsevier, 2025–2054. https://doi.org/10.1016/S1574-0005(02)03016-3

[65] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.

[66] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.