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

Anonymous

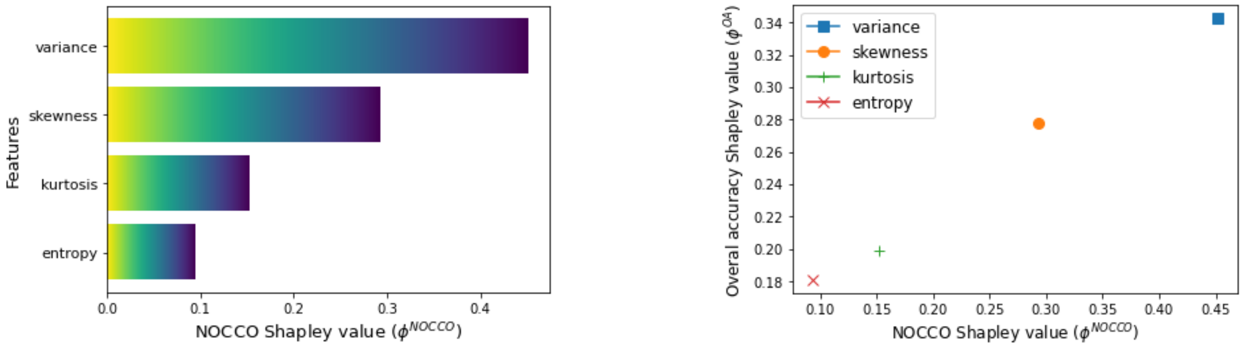
1 Supplementary datasets and results

Below we present further results on other datasets not appearing in the paper, namely,

- Banknote Authentication dataset[1]: The goal in this dataset is to detect genuine or forged banknotes. There are 1372 images described by the following features: *variance*, *skewness*, *kurtosis* and *entropy*.
- Diabetes Pima Indian dataset[2]: This dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases, and aims to diagnostically predict whether a patient has diabetes, based on certain measurements included. All patients here are females at least 21 years old of Pima Indian heritage.
- Raisin dataset[3]: This dataset consists of characteristics of Kecimen and Besni raisin varieties grown in Turkey obtained with CVS. It contains 7 morphological features, namely, *Area*, *MajorAxisLength*, *MinorAxisLength*, *Eccentricity*, *ConvexArea*, *Extent*, *Perimeter* and *Class*, extracted from images of 900 raisin grains, including 450 pieces from both varieties.
- Law School Admission Council (LSAC) dataset[4]: In this dataset, the task is to predict whether a candidate (from a total of 23.726) would succeed in the bar exam. For each candidate, one has 11 features: *decile1b*, *decile3*, *lsat*, *ugpa*, *zygpa*, *zgpa*, *fulltime*, *fam_inc*, *male*, *race* and *tier*.

1.1 Banknotes dataset

Figure 1 presents the results for the Banknotes dataset. The higher contribution is assign to the *variance* while *entropy* achieved the lower contribution (see Figure 1a). The relation between the NOCCO and the overall accuracy Shapley values is depicted in Figure 1b. Clearly, there is a positive relation between them.



(a) Features importance.

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

Figure 1: Results for the Banknotes dataset.

[1] https://archive.ics.uci.edu/dataset/267/banknote+authentication

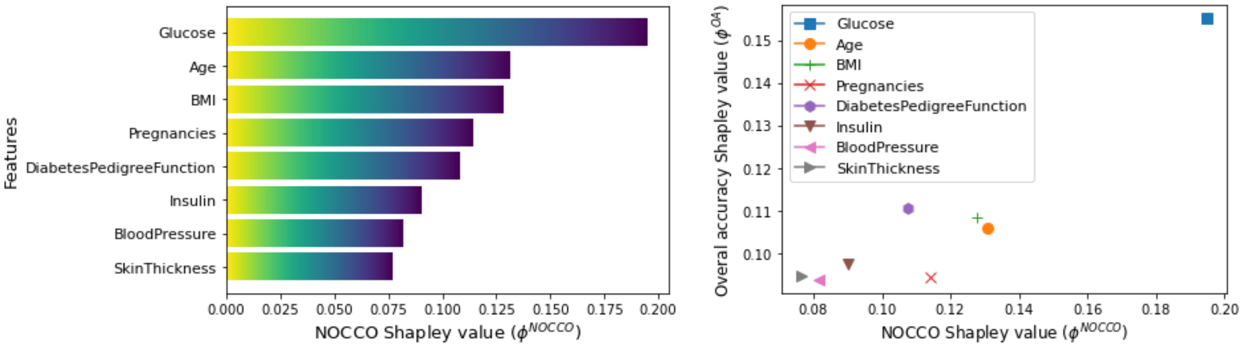
[2] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[3] https://archive.ics.uci.edu/dataset/850/raisin

[4] https://archive.lawschooltransparency.com/reform/projects/investigations/2015/documents/NLBPS.pdf

1.2 Diabetes Pima Indian dataset

In Figure 2 we present our findings for the diabetes Pima Indian dataset. As showed in Figure 2a, *glucose* has a strong marginal impact in predicting diabetes. A relation between our proposal and the Shapley values obtained after training the machine learning model can be seen in Figure 2b. Although there are some deviations from a perfect correlation between both measures, one may note a positive correlation.



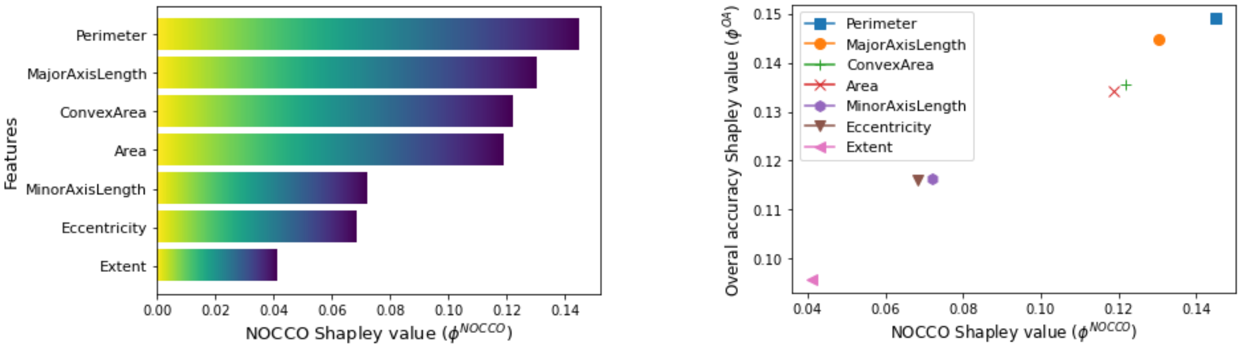
(a) Features importance.

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

Figure 2: Results for the Diabetes Pima dataset.

1.3 Raisin dataset

Figure 3 shows the result for the Raisin dataset. Features such as *Perimeter*, *MajorAxisLength*, *ConvexArea* and *Area* have relevant contributions towards the model performance (see Figure 3a). Moreover, a positive relation can be seen in Figure 3b, where greater the NOCCO Shapley value, greater the overall accuracy Shapley value.



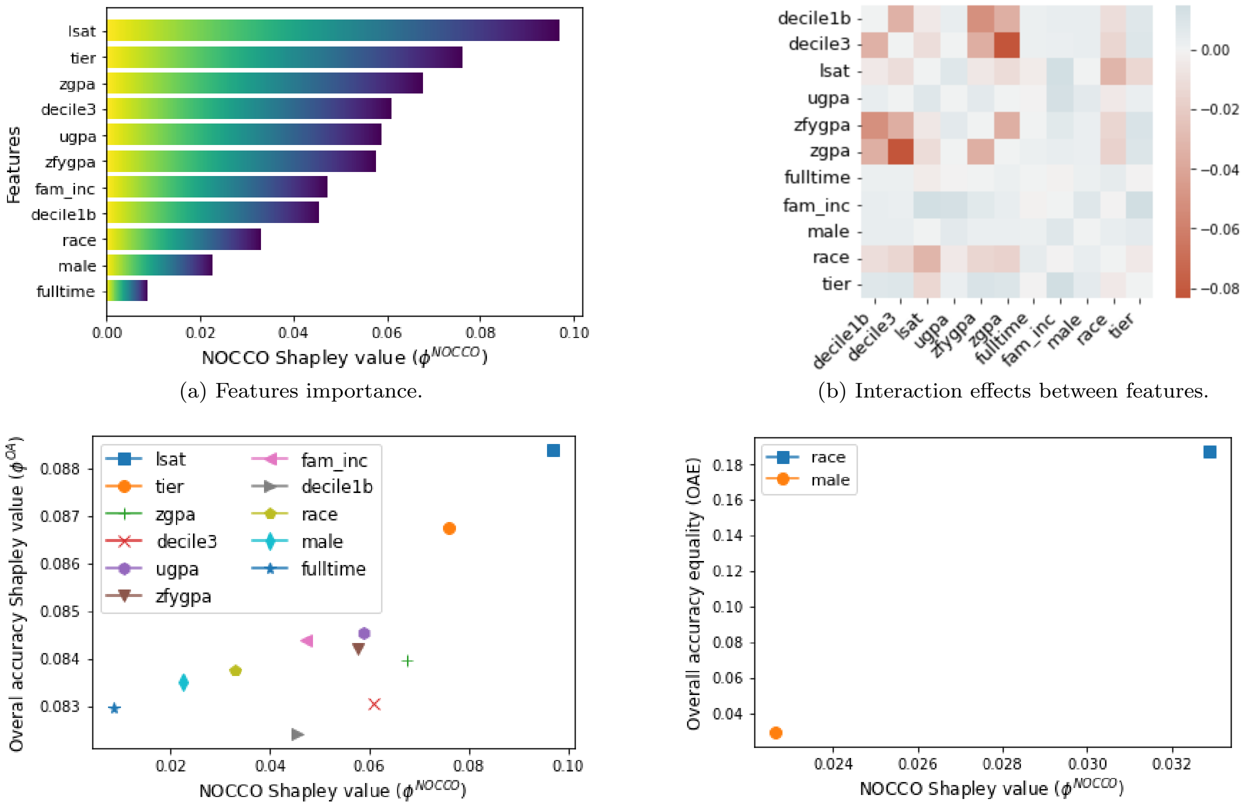
(a) Features importance.

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

Figure 3: Results for the Raisin dataset.

1.4 LSAC dataset

The application of our framework on the LSAC dataset leads to the results presented in Figure 4. Note that in Figure 4a the highest and the lowest contributions towards predicting success in the bar exam are assigned to the *lsat* and *fulltime*, respectively. Also, in Figure 4b some redundancies among features associated with school performances (*decile1b*, *decile3*, *zygpa* and *zgpa*) can be observed. In Figure 4c we show the comparison between NOCCO and overall accuracy Shapley values, where we also find a positive correlation between such a measures. The relation between sensitive features (*race* and *male*) is depicted in Figure 4d. Note that *race*, which has the highest marginal dependence degree with the model performance (among the sensitive features), entails more disparity than *male*. Therefore, we also attest the use of the proposed NOCCO Shapley values to detect disparity prone features.



(a) Features importance.

(b) Interaction effects between features.

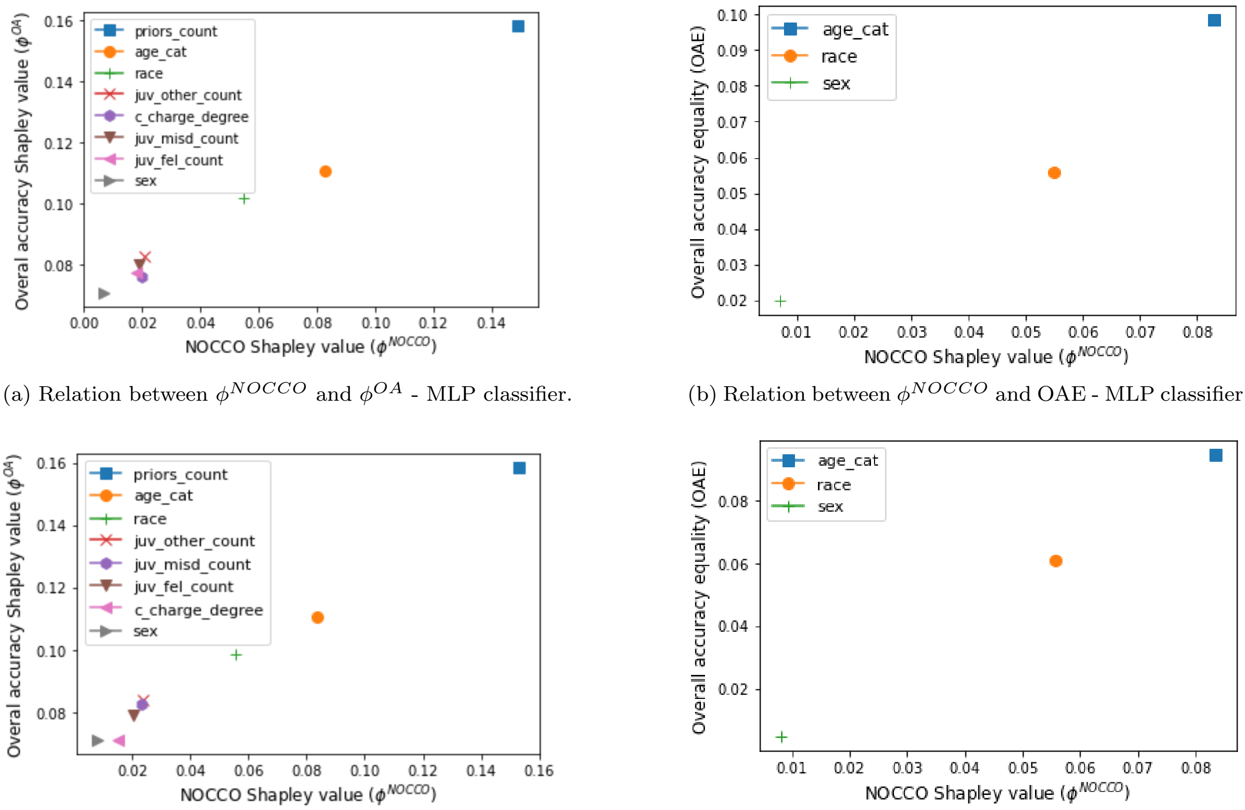
(c) Relation between ϕ^{NOCCO} and ϕ^{OA}.

(d) Relation between ϕ^{NOCCO} and OAE.

Figure 4: Results for the LSAC dataset.

2 Comparison of NOCCO with MLP and Random Forests on the COMPAS dataset

We also experimented with other classical models, namely, random forests, to indicate the model-agnostic character of our proposed framework. Figure 5 presents the results obtained by both MLP and random forests on the COMPAS dataset. The comparative analysis of Figures 5a and c, and of Figures 5b and d, show similar behaviour of MLP and Random Forest with respect to the relations between ϕ^{NOCCO} and ϕ^{OA} and between ϕ^{NOCCO} and OAE.



(a) Relation between ϕ^{NOCCO} and ϕ^{OA} - MLP classifier.

(b) Relation between ϕ^{NOCCO} and OAE - MLP classifier.

(c) Relation between ϕ^{NOCCO} and ϕ^{OA} - Random Forest classifier.

(d) Relation between ϕ^{NOCCO} and OAE - Random Forest classifier.

Figure 5: Results for the COMPAS dataset.