POLITECNICO DI TORINO

Master degree course in Data Science and Engineering

Master Degree Thesis

Sensitive attributes disproportion as a risk indicator of algorithmic unfairness

Supervisors
ric. Antonio Vetrò

Co-Supervisors
prof. Juan Carlos De Martin

Candidates
Federico D'Asaro
matricola: 279285

Anno accademico 2020-2021

Summary

AI is increasingly being used in highly sensitive areas such as health care, hiring, and criminal justice, so there has been a wider focus on the implications of bias and unfairness embedded in it. We know that human decision making in many areas is biased and shaped by our individual or societal biases, which are often unconscious. One may assume that using data to automate decisions would make everything fair, but we now know that this is not the case. AI bias can come in through societal bias embedded in training datasets, decisions made during the machine learning development process, and complex feedback loops that arise when a machine learning model is deployed in the real world.

Our aim is to anticipate, before applying any algorithm, unfairness phenomenon by studying balance characteristic of protected attributes such age, ethnicity, gender, education, marital status, etc. We start by replicating results of [1], thus analyzing relationships between balance measures and unfairness indices. We first evaluate balance indexes (in the interval [0,1], where 0 is imbalance while 1 is balance) as Gini, Simpson, Shannon, Imbalance ratio (IIR), Renyi, Hill on training data of 9 datasets. Then on testdata, Independence, Separation, Sufficiency and Overall Accuracy Equality (OAE) are chosen as measures of discrimination (in the interval [0,1], where 0 is fairness while 1 is unfairness) and computed with respect to the sensible attributes taken into consideration (the same used for balance assessment). In particular about Separation, it is evaluated taking into consideration both True Positive Rate (TPR) and False Positive Rate (FPR) equalities among the attribute's classes. The same goes for Sufficiecy which Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are computed.

The study is conducted on several levels: different models (LogisticRegression - LR, Support Vector Machine - SVM, K-nearest neighbors - KNN, Random Forest - RF) and variant (baseline, smote) thereof are considered.

Observations were made by analyzing separately balance indexes and unfairness ones. Further investigations were made on the relationships between the two indices to evaluate the goodness of the former as indicators of risk of discrimination. In particular, we examined how each pair balance-unfairness measures were related to each other (based on correlations and distributions among the two).

As regards balance measures, Gini and Shannon penalize disproportion less than other indexes. Furthermore they benefit of lower unfairness risk levels thresholds in terms TPR, PPV, OAE.

About unfairness, there are differences between baseline and smote variant of the algorithms: the first favours Independence and Separation (low unfairness), the second reaches lower discrimination on Sufficiency.

Looking at Unfairness distribution among two balance risk levels (with a threshold at 33%, under which imbalance is classified as 'high discrimination risk'), IIR is the index which better anticipate discrimination, it fails only on OAE. The second best performing index is Shannon which fails in FPR and NPV discrimination capabilities among the two levels of risks. Major part of these observations are robust to an extended assessment (through additional datasets) especially in correspondence of Random Forest model.

As concern correlation between balance measures and unfairness ones, Independence, TPR and PPV are the easiest to correlate with. About Independence, SVM is the model getting higher values over the four balance indexes, but it is the worst on OAE. RF performs very well on Independence and Separation, KNN on Sufficiency and OAE. A correlation comparison by attribute cardinality was carried out, and it showed that IIR takes undesired positive correlation on attributes with 8 classes.

Future work suggestions have been proposed especially as concern datasets, algorithms and measures used.

Contents

Part I

Introduction

0.1 Unfairness in ML

According to [2] nowadays, an increasing number of decisions are being controlled by artificial intelligence (AI) algorithms, with increased implementation of automated decision-making systems in business and government applications. The motivation for an automated learning model is clear – we expect algorithms to perform better than human beings for several reasons: First, algorithms may integrate much more data than a human may grasp and take many more considerations into account. Second, algorithms can perform complex computations much faster than human beings. Third, human decisions are subjective, and they often include biases. Hence, it is a common belief that using an automated algorithm makes decisions more objective or fair. However, this is unfortunately not the case since AI algorithms are not always as objective as we would expect. The idea that AI algorithms are free from biases is wrong since the assumption that the data injected into the models are unbiased is wrong. More specifically, a prediction model may actually be inherently biased since it learns and preserves historical biases. Since many automated decisions (including which individuals will receive jobs, loans, medication) can significantly impact people's lives, there is great importance in assessing and improving the ethics of the decisions made by these automated systems. Indeed, in recent years, the concern for algorithm fairness has made headlines. One of the most common examples was in the field of criminal justice, where recent revelations have shown that an algorithm used by the United States criminal justice system had falsely predicted future criminality among African-Americans at twice the rate as it predicted for white people (COMPAS).

Algorithmic bias is often discussed in machine learning, but in most cases the underlying data, rather than the algorithm, is the main source of bias. The biggest problem with machine learning models is that the training distribution does not always match the desired distribution. If the present reality puts certain individuals at a systematic disadvantage, the training data distribution is likely to reproduce that disadvantage rather than reflecting a fairer future. These decisions are reflected in the training data and subsequently baked into future machine learning model decisions.

When building machine learning models, many data scientists assume that they can just remove protected attributes (i.e., race, gender, age) to avoid unfair bias. However, there are many features that are too closely correlated

to protected attributes, which makes it easy to reconstruct a protected attribute such as ethnicity even if you drop it from your training set.

Still according to [2], the literature has indicated several causes that may lead to unfairness in machine learning:

- Biases already included in the datasets used for learning, which are based on biased device measurements, historically biased human decisions, erroneous reports or other reasons. Machine learning algorithms are essentially designed to replicate these biases;
- Biases caused by missing data, such as missing values or sample/selection biases, which result in datasets that are not representative of the target population;
- Biases that stem from algorithmic objectives, which aim at minimizing overall aggregated prediction errors and therefore benefit majority groups over minorities;
- Biases caused by "proxy" attributes for sensitive attributes. Sensitive attributes differentiate privileged and unprivileged groups, such as race, gender and age, and are typically not legitimate for use in decision making. Proxy attributes are non-sensitive attributes that can be exploited to derive sensitive attributes. In the case that the dataset contains proxy attributes, the machine learning algorithm can implicitly make decisions based on the sensitive attributes under the cover of using presumably legitimate attributes.

0.2 The study

The unfairness problem is here tackled by taking into consideration data imbalance of classification classes and sensitive attributes. In particular with respect to the first, we apply smote technique to mitigate such phenomenon and assess its impact on unfairness measures. As concern imbalance measures, these are used as upstream discrimination risk before any algorithm has been run. Multiple algorithms and datasets are employed to make several considerations.

Below we report a diagram of the levels on which actions have been taken. The scheme reports what has been done for a given run which is obtained

by splitting the development set (among the nine datasets) in training set and test set. The reading of the diagram begins from the green 'START' in correspondence of the training set. Dashed lines follows the line of reasoning applied:

1. On training set Balance measures have been computed on the sensitive attributes. All four algorithms have been applied both on row and oversampled (smote) data.

2. Trained models are tested on test data, from which unfairness measures (on the same sensitive attributes as before) are retrieved.

3. Analysis between Balance and Unfairness indexes are performed on different levels: correlation and unfairness distribution among two imbalance risk levels decided with multiple thresholds.

Solid lines are used to list different alternatives among: algorithms, balance and unfairness measures, type of analysis.

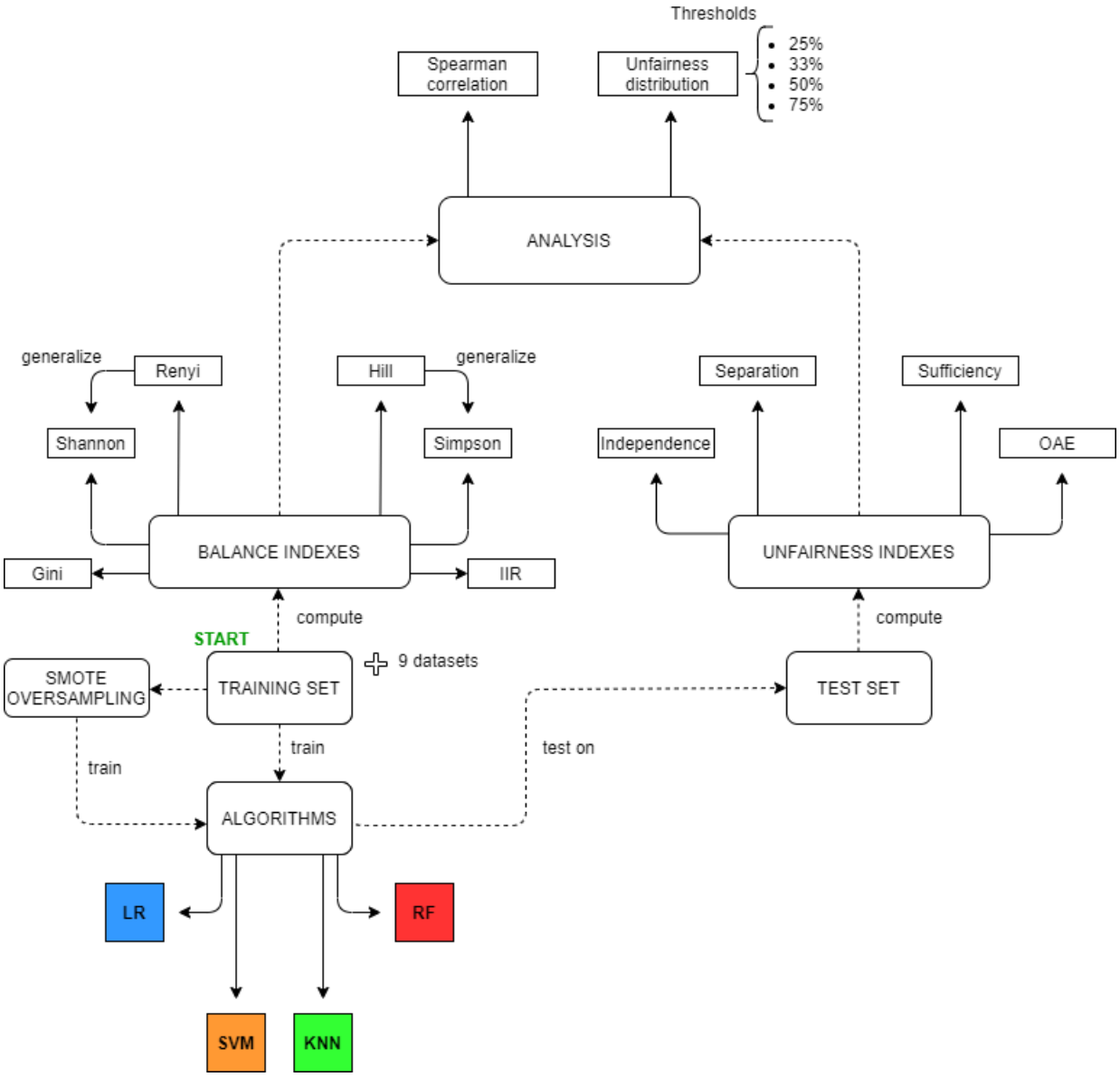


Figure 1. Study diagram

Legend:

- LR: Logistic Regression;
- SVM: Support Vector Machines;
- KNN: K-nearest neighbors;

- RF: Random Forest;
- IIR: Inverse Imbalance Ratio;
- OAE: Overall Accuracy Equality.

The described study is intended to extend [1] which, from now on, will be referred as the '**base paper**'.

Part II

Metodology

Chapter 1

Training pipeline

The datasets taken into consideration are characterized by class imbalance, generally in proportion 75/25, favouring the majority class.

Class imbalance is one of the most serious influential factors for the predictive performance of classifiers. The imbalanced data are characterized as having many more instances of certain classes than others. In this case, classifiers tend to make biased learning model that has a poorer predictive accuracy over the minority classes compared to the majority classes. This is because most standard classifier learning algorithms, such as decision tree, backpropagation neural network and support vector machines, are designed based on assumptions that the class distribution is relatively balanced and the misclassification costs are equal, classification rules that predict the minority classes tend to be rare, undiscovered or ignored [3].

To optimize models with imbalanced input data it is important to choose an appropriate evaluation metric which takes into consideration the performance of the model with respect to both the majority class and the minority one.

Numerous performance metrics have been proposed to evaluate how much a developed model is capable of distinguishing between classes. Those that are widely used include accuracy, precision, recall, F-measure, kappa statistics and AUC-ROC (area under the receiver operating characteristics curve), etc. Different performance metrics are used in different fields. For example, in the studies of rare diseases or customer churn prediction, the accuracy is

criticized as an impractical technique because of the imbalance of class distribution. Suppose we have a dataset with 1000 samples, the majority class and the minority class containing 990 samples and 10 samples, respectively. Now if a classifier classifies them all as the majority class, the accuracy will be 99%, even though the classifier missed all minority samples due to the highly imbalanced class distribution [3].

So, the precision and the recall (sensitivity and specificity) of both classes are taken into consideration by computing their harmonic mean, the F1-score. To summarize the performances of the model as a whole the macro average of the F1 is computed. It consists of the arithmetic mean of each class f1 score.

1.1 Hyperparameters tuning

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

It is a popular method because it generally results in a less biased or less optimistic estimate of the model skill than other methods, such as a simple train/test split. The estimate of the performance metric is more reliable since it is averaged over different trials on different subset of the development set. So at the end the best metric will result not only in the grater mean value but also in the most reliable (lower standard deviation).

The general procedure is as follows:

1. Shuffle the dataset randomly;
2. Split the dataset into k groups;
3. For each unique group:
   (a) Take the group as a hold out or test data set;
   (b) Take the remaining groups as a training data set;
   (c) Fit a model on the training set and evaluate it on the test set;
   (d) Retain the evaluation score and discard the model;
4. Summarize the skill of the model using the sample of model evaluation scores.

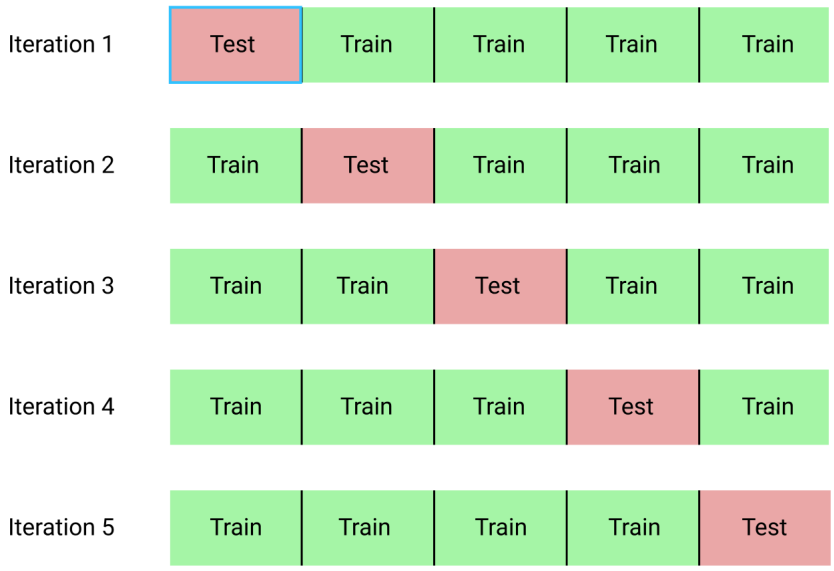


Figure 1.1. 5-fold Cross validation example. Image taken from [4]

Importantly, each observation in the data sample is assigned to an individual group and stays in that group for the duration of the procedure. This means that each sample is given the opportunity to be used in the hold out set 1 time and used to train the model k-1 times.

Since class imbalance is present, to find the optimal hyperparameters set for each algorithm used, Stratified k-fold cross-validation is used.

1.1.1 Stratified k-fold cross-validation

Here the folds are selected so that the mean response value is approximately equal in all the folds. In the case of a dichotomous classification, this means that each fold contains roughly the same proportions of the two types of class labels.

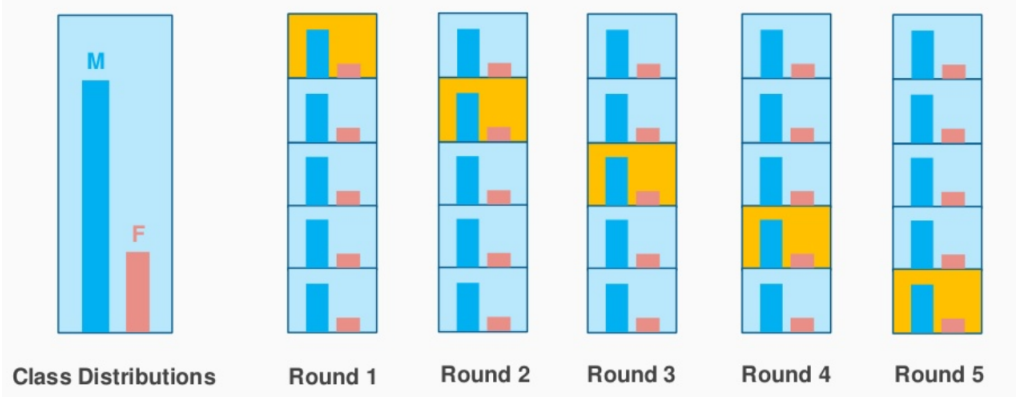


Figure 1.2. Example of 5 folds Stratified Cross Validation. Image taken from [5]

With respect to the class GridSearchCV(), according to the documentation: "If the estimator is a classifier and y is either binary or multi-class, StratifiedKFold is used. In all other cases, KFold is used."

When it was computationally feasible, Repeated Stratified k-fold cross-validation was used. It involves simply repeating the cross-validation procedure multiple times and reporting the mean result across all folds from all runs. This mean result is expected to be a more accurate estimate of the true unknown underlying mean performance of the model on the dataset. At each run, data are shuffled so that the k splits are different among the multiple runs.

1.1.2 Class weighting

Moreover all the algorithm employed have been modified to take into account the skewed distribution of the classes. This can be achieved by giving different weights to both the majority and minority classes. The difference in weights will influence the classification of the classes during the training phase. The whole purpose is to penalize the misclassification made by the minority class by setting a higher class weight and at the same time reducing

weight for the majority class.

Most of the sklearn classifier modeling libraries have an in-built parameter "class_weight" which helps to optimize the scoring for the minority class.

By default, the value of class_weight=None, i.e. both the classes have been given equal weights. Other than that, we can either give it as 'balanced' or we can pass a dictionary that contains manual weights for both the classes.

When the class_weights = 'balanced', the model automatically assigns the class weights inversely proportional to their respective frequencies.

To be more precise, the formula to calculate this is:

$$w_j = \frac{n\_samples}{n\_classes * n\_samples_j}$$

Where,

- w_j is the weight for each class(j signifies the class);
- n_samples is the total number of samples or rows in the dataset;
- n_classes is the total number of unique classes in the target;
- $n\_samples_j$ is the total number of rows of the respective class.

As an example we show how the loss Logistic Regression is modified using class weighting:

- Logistic regression log loss

$$log\ loss\ = \frac{1}{N}\sum_{i=1}^{N}[-(y_i * log(\hat{y}_i) + (1 - y_i) * log(1 - \hat{y}_i))]$$

- Logistic regression weighted log loss

$$log\ loss\ = \frac{1}{N}\sum_{i=1}^{N}[-(\omega_0(y_i * log(\hat{y}_i)) + \omega_1(1 - y_i) * log(1 - \hat{y}_i))]$$

1.2 Smote application

It is reasonable to apply an oversampling technique to help the classifier better predict on the originally less represented class. However the way how the pipeline is structured is important since putting synthetic data points into the validation or test set may lead to optimistic results, biased with respect to the original distribution. Let's say every data point from the minority class is copied 6 times before making the splits. If we did a 3-fold validation, each fold has (on average) 2 copies of each point. If our classifier overfits by memorizing its training set, it should be able to get a perfect score on the validation set. Our cross-validation will choose the model that overfits the most. So even during a k-fold cross validation, only the training set (k-1 folds) is interested by minority class upsampling, the k-th fold used for validation is maintained unchanged. In this way the upsampling procedure is applied at each iteration of the Stratified k-fold CV. Cross-validating with oversampled data may provide you with a different perspective on the classifier's ability to predict on both classes with equal importance.

The upsampling technique used is SMOTE (Synthetic Minority Oversampling Technique)

It is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement. This approach is inspired by a technique that proved successful in handwritten character recognition. They created extra training data by performing certain operations on real data. In their case, operations like rotation and skew were natural ways to perturbate training data. Synthetic examples are generated in a less application-specific manner, by operating in "feature space" rather than "data space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Synthetic samples are generated in the following way: Take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. The synthetic

examples cause the classifier to create larger and less specific decision regions as shown by the dashed lines, rather than smaller and more specific regions. More general regions are now learned for the minority class samples rather than those being subsumed by the majority class samples around them. The effect is that the algorithm generalize better [6].

- Step 1: Setting the minority class set A, for each x in A, the k-nearest neighbors of x are obtained by calculating the Euclidean distance between x and every other sample in set A.
- Step 2: The sampling rate N is set according to the imbalanced proportion. For each x in A, N examples (i.e $x_1, x_2, ..., x_n$) are randomly selected from its k-nearest neighbors, and they construct the set A_1.
- Step 3: For each example x_k in A_1(k=1, 2, 3...N), the following formula is used to generate a new example:

$$x^{'} = x + rand(0,1) * |x - x_k|$$

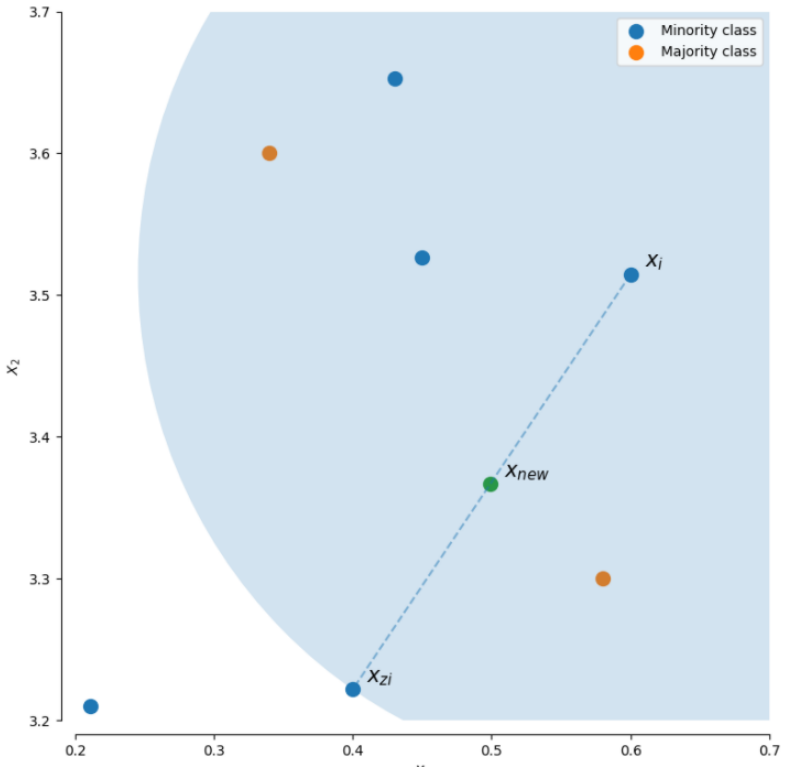


Figure 1.3. Image taken from [7]

rand(0, 1) represents the random number between 0 and 1.

Some of the datasets were composed of both continuous and categorical features. In these cases Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTE-NC) from the imbalanced-learn library, has been applied instead of SMOTE. SMOTE-NC slightly changes the way a new sample is generated by performing something specific for the categorical features. In fact, the categories of a new generated sample are decided by picking the most frequent category of the nearest neighbors present during the generation. It goeas beyond the simple SMOTE interpolation between two value of a binary attribute p[0,1] by sampling from a Bernoulli(p), so resulting in feasible values.

Chapter 2

Algorithms

2.1 Logistic Regression

Contrarily to linear regression, where the expected value of the response variable is modeled as a linear function(in the parameters) of the input covariates, in logistic regression we deal with binary classification setting. So, the response variable is supposed to have a Bernoulli distribution, which expected value is a probability in [0,1]. It can be modeled exploiting a link as the Logistic function $\sigma : \mathrm{R} \rightarrow [0,1]$.

$$p(x) = \sigma(t) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

Similar to other regression problems, we look for the Maximum Likelihood estimator for β. The log-likelihood of the data given the model is:

$$l(\beta) = \sum_{i=1}^{N} \{y_i \beta^T x_i - log(1 + e^{\beta^T x_i})\}$$

Taking the partial derivative of with respect to each component of the β vector:

$$\frac{\delta l(\beta)}{\delta \beta} = \sum_{i=1}^{N} x_i (y_i - p(x_i; \beta)) = 0$$

Since no close form solution exists, the log likelihood can be maximized by an iterative approach such as gradient ascent (with a tuned learning rate). The output of such algorithm is a parametrized logistic function fitting the data.

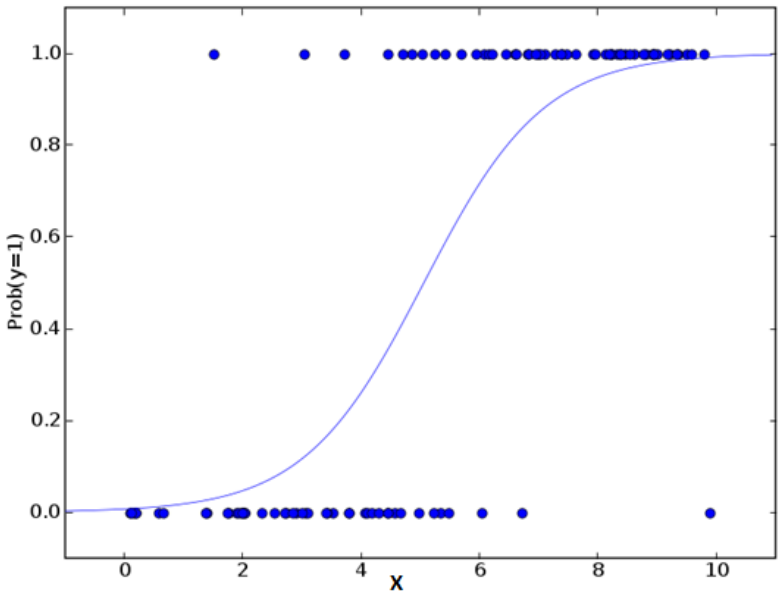


Figure 2.1. Logistic Regression. Image taken from [8]

Other than class_weight ('none', 'balanced'), the hyperparameters that were tuned to optimize a Logistic Regression model was 'C' (inverse of regularization strength) and penalty (to specify the norm used in the penalization to regularize the model). The smaller we choose 'C', the stronger regularization we get. The default value chosen by Scikit-learn is 1.0.

2.2 Support Vector Machine

Support vector machine is a supervised learning algorithm, often used for binary classification problems. This classificator tries to find a hyperplane that separates classes in feature space. Even if there are many possible ways to separate two classes, this algorithm resolves an optimization problem. Indeed, it tries to find the hyperplane that maximizes the distance between the two classes, the margin. Soft margin svm allows some points to be misclassified (to makes some problems feasible and prevent ovefitting) at training time by the introduction of slack variables. The primal optimization problem is:

$$\min_{\omega,b} \frac{1}{2}||\omega||^2 + C\sum_i \epsilon_i$$

$$subject\ to\ y_i[<\omega, x_i> + b] \geq 1 - \epsilon_i\ and\ \epsilon_i \geq 0$$

important properties of the optimal solution can be understook by looking at the dual optimization problem:

$$\begin{aligned} \max_{\alpha} -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j < x_i, x_j > + \sum_i \alpha_i \\ subject\ to\ \sum_i \alpha_i y_i = 0\ and\ \alpha_i \geq 0 \end{aligned} \tag{2.1}$$

Given α the lagrange multiplier of the first primal constraint, for the Karush-Kuhn-Tucker complementarity condition inactive constraints have null multipliers. So α>0 only for support vectors (point on the margins or which are on the wrong side of the margin). These points are more important than others and they will be the only ones that influence the hyperplane construction. C acts as regularization hyperparameter of the problem by controlling the influence of outliers in hyperplane construction (higher value brings to higher importance of these points α=C).

Since both in the optimization function and the decision function data computation occur in the form of dot product, svm can be kernelized so that also non linearly separable problems can be tackled.

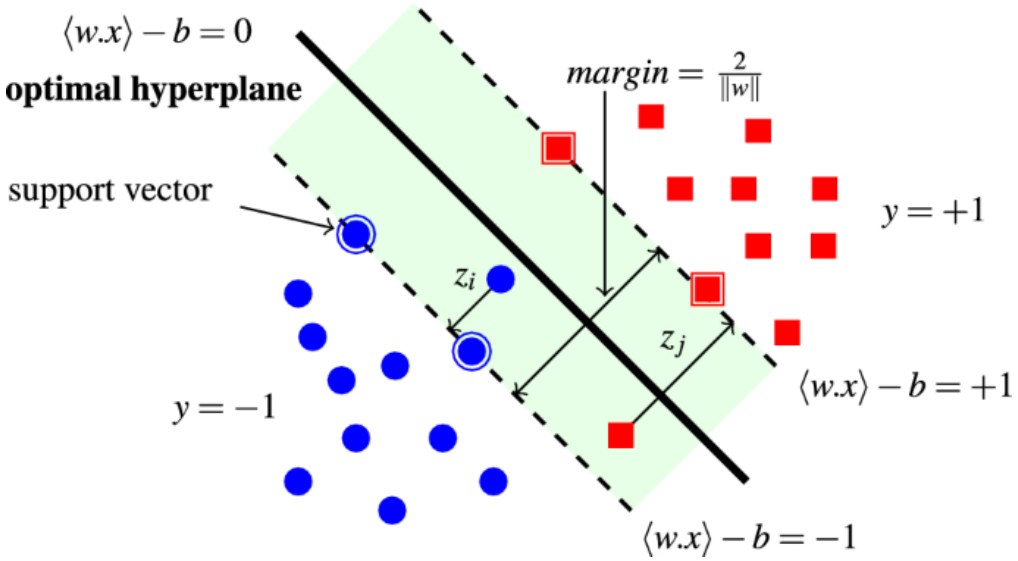


Figure 2.2. Support Vector Machine. Image taken from [9]

Other than class_weight ('none', 'balanced'), the hyperparameters that were tuned to optimize a SVM model were: 'C' and kernel (which type of kernel to use).

2.3 K-Nearest Neighbors

Many approaches attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability. One such method is the K-nearest neighbors (KNN) classifier. Given a positive integer K and a test observation x0, the KNN classifier first identifies the neighbors K points in the training data that are closest to x0, represented by N0. It then estimates the conditional probability for class j as the fraction of points in N0 whose response values equal j:

$$Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

Finally, KNN applies Bayes rule and classifies the test observation x0 to the class with the largest probability. Using k neighbors instead of only one, permits to be more robust in case of noisy problems when classes partially overlap but it makes boundaries between classes less distinct(therefore, it is an hyperparameter of the model).

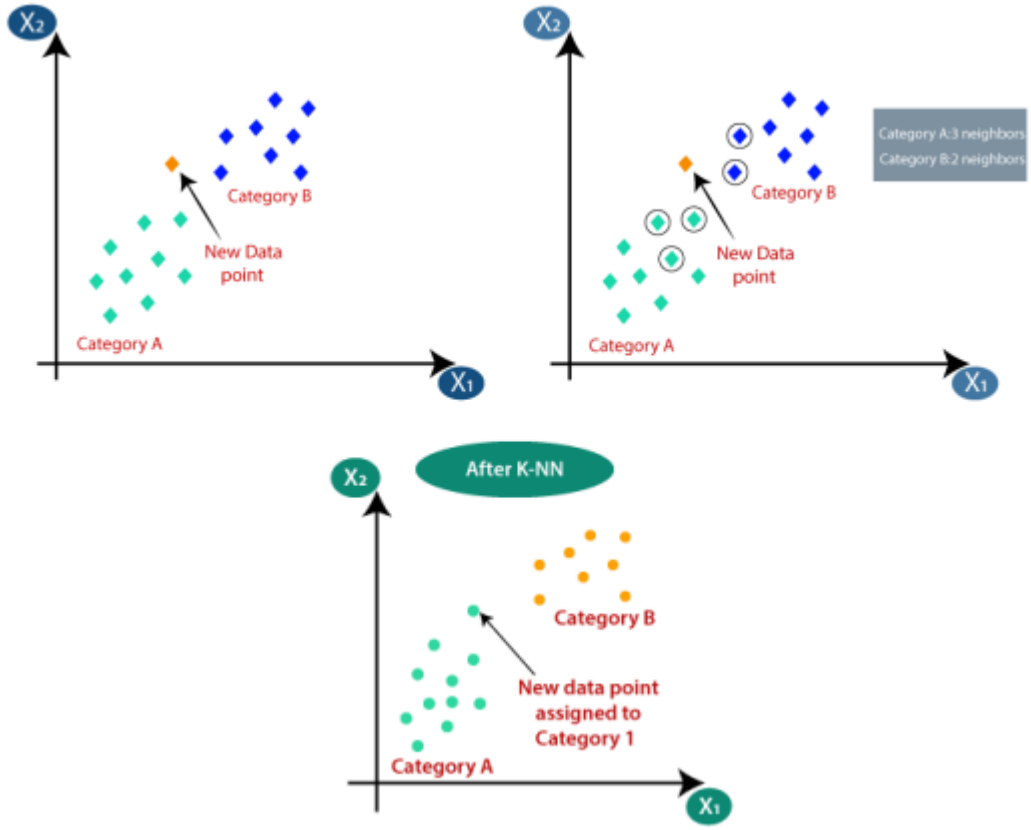


Figure 2.3. K-Nearest Neighbors. Image taken from [10]

Hence it is a non-parametric model, which simply store all the training data (it is easy to implement as long as the training set is made of a small amounts of data).

Other than class_weight ('none', 'balanced'), the hyperparameters that were tuned to optimize a KNN model were 'n_neighbors' (number of neighbors points considered to estimate conditional probability for class j of the new point) and weights (if weight points by the inverse of their distance from the new data point, or uniformly).

2.4 Random Forest

Tree-based methods involve segmenting the predictor space into a number of simple, non overlapping, regions. The procedure to form the regions consists

in a recursive binary splitting. Firstly, considering all the predictors Xj and the cut-point s for each predictor, we look for the split that minimizes Gini impurity or Entropy:

$$Gini\ impurity: \ 1 - \sum_{i=1}^{J} p_i^2$$

$$Entropy: \ -\sum_{i=1}^{J} p_i log_2(p_i)$$

The classic decision tree suffers from high variance, the output of the model varies significantly depending on the training set used to fit it. Random Forest consists of a large number of individual Decision Trees that operate as an ensemble (reducing variance). Each individual tree in the Random Forest creates its own decision and eventually, the decision with the highest number of votes becomes our model's decision.

Building Process of Random Forest:

1. Create a bootstrapped data set -same size as the original- by randomly selecting samples from original data set. It is possible to select same sample more than once.

2. Create a Decision Tree using the bootstrapped data set, by only using a random subsets of features (columns) at each split (because if there is a very strong feature, most of the trees will use this predictor in the top split).

In the Scikit-learn RandomForestClassifier, the hyperparameter 'n_estimators' defines the number of trees in the forest, which means these two steps are repeated 'n_estimators' times.

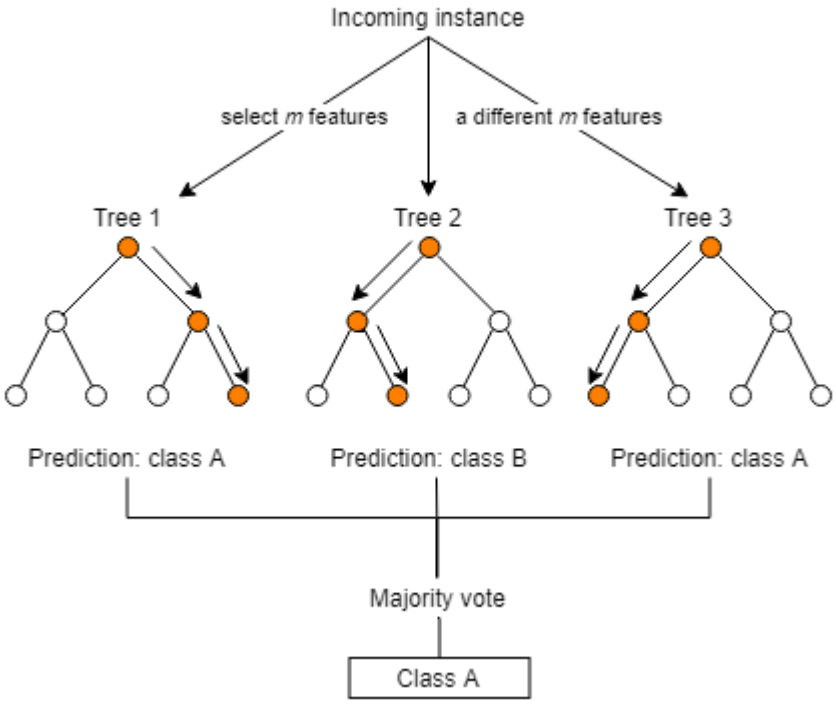


Figure 2.4. Random Forest. Image taken from [11]

Other than class_weight ('none', 'balanced'), the hyperparameters that were tuned to optimize a Random Forest model were: 'n_estimators' (number fitted trees), max_depth (the maximum depth of each tree, to prevent overfitting), max_features(function of initial number of predictors that indicates how many features randomly consider at each split as candidates to the split).

Chapter 3

Datasets

We examine nine datasets belonging to three different application domains: criminal justice systems (also juvenile), financial services, and social related topics – personal earnings and education– and a hypothetical survival situation (Titanic dataset). We sought for some variety in order to explore the potential of our approach in several fields of application of ADM systems.

Since Compas and Juvenile justice datasets are provided with black box algorithm predictions, they are excluded from the evaluation of the per-model risk indexes performances (for the four models Logistic Regression, Support Vector Machines, K-nearest neighbor, Random Forest). They are exploited to make a global and general purpose (multi-model) consideration by taking for each dataset the best performing model based on f1_macro score.

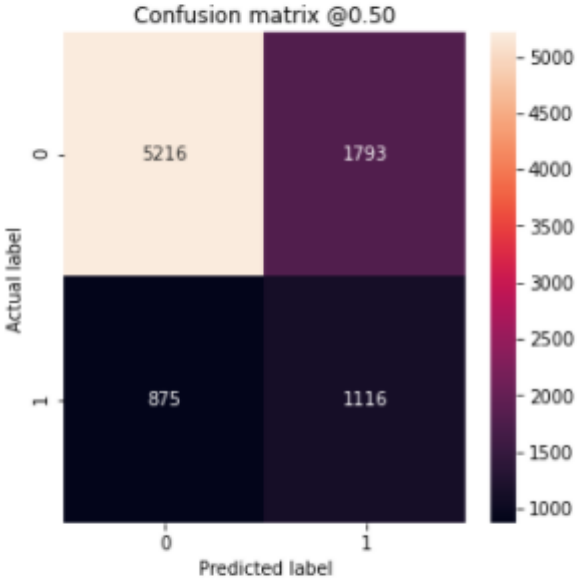
For each model, the following template is reported: The main metrics associated to the output (Recall, precision of each class, AUC), Confusion Matrix and the ROC curve.

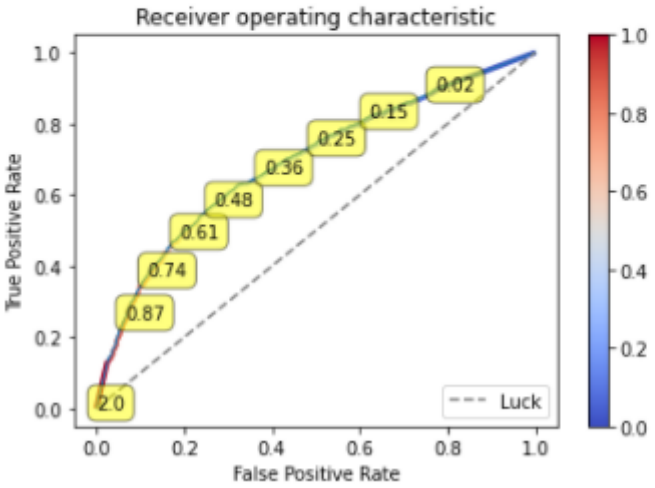
AUC is the area under ROC curve, which is the probability that a random positive instance gets a score higher than a random negative instance. An area of ½ corresponds to random guessing, and an area of 1 corresponds to perfect classification, the score vector equals the target one.

```
Best params configuration: {'n_neighbors': 9, 'weights': 'distance'}

Accuracy: 0.7035555555555556
Recall class 0 (TNR): 0.7441860465116279
Recall class 1 (TPR): 0.5605223505775992
Precision class 0: 0.8563454276801838
Precision class 1: 0.3836369886558955
F1 class 0: 0.7963358778625954
F1 class 1: 0.4555102040816326
F1_macro: 0.625923040972114

False negative rate: 0.43947764942240075
False positive rate: 0.2558139534883721
AUC score: 0.6921889335222942
```





Each dataset description is followed by baseline and SMOTE models performance in terms of macro F1 and AUC (both in the format mean(std).

3.1 Credit card default dataset

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It comes from the UCI Machine Learning Repository [12]. The dataset is composd by 25 variables, but only a subset of them, coherently with base paper, are used as predictors.

- features: 'LIMIT_BAL','PAY_0','BILL_AMT1','PAY_AMT1'
- target: 'default.payment.next.month'
- protected attributes: 'SEX','EDUCATION','MARRIAGE','AGE'

F1_MACRO

Model	Baseline	Smote
Logistic Regression	0.63(0.005)	0.617(0.005)
SVM	0.687(0.003)	0.677(0.004)
KNN	0.664(0.003)	0.618(0.008)
Random Forest	0.647(0.004)	0.638(0.003)

AUC

Model	Baseline	Smote
Logistic Regression	0.713(0.004)	0.714(0.004)
SVM	0.746(0.003)	0.746(0.003)
KNN	0.722(0.005)	0.697(0.005)
Random Forest	0.715(0.007)	0.706(0.005)

3.2 Statlog

This widely used German credit dataset from the UCI Machine Learning Repository [13] has been provided by the German professor Hans Hofmann as part of a collection of datasets from an European project called "Statlog"

and will be simply called Statlog in the following. When a bank receives a loan application, based on the applicant's profile the bank has to make a decision regarding whether to go ahead with the loan approval or not. In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes. The data are a stratified sample of 1000 credits (700 good ones and 300 bad ones) and have been collected between 1973 and 1975 from a large regional bank in southern Germany, which had about 500 branches, both urban and rural ones. As indicated with the Statlog data, one might examine misclassification cost: it has been suggested to allocate the cost for misclassifying a bad risk as good to be five times as high than the cost for misclassifying a good risk as bad, therefore we assumed cost matrix as target variable (equal to 0 or 1) and we built the predictions through a binary classification.

- Features:"Duration","Credit_history","Purpose","Credit_amount", "Savings",
  "Employment_since","Installment_rate","Other_Debtors_Guarantors","Property", "Housing","Residence_since","Other_installment_plans", "Existing_credits","Job", "People_liable_to_provide_maintenance_for", "Telephone"
- target: 'costMatrix'
- protected attributes: "Status","Sex","Foreign_Worker","Age"

F1_MACRO

Model	Baseline	Smote
Logistic Regression	0.628(0.035)	0.646(0.022)
SVM	0.597(0.027)	0.646(0.018)
KNN	0.596(0.01)	0.583(0.015)
Random Forest	0.613(0.022)	0.629(0.026)

AUC

Model	Baseline	Smote
Logistic Regression	0.737(0.025)	0.731(0.029)
SVM	0.741(0.022)	0.739(0.016)
KNN	0.636(0.013)	0.647(0.019)
Random Forest	0.732(0.02)	0.714(0.03)

3.3 Income

The extraction of these data was realized by Barry Becker from the 1994 Census database; the prediction task is to determine whether a person makes over $50, 000 a year based on that set of reasonably clean records, also known as "Census Income" dataset [14]. Thus, test.income represents the target variable, which can assume the two values <= 50K or > 50K.

- features:"workclass","occupation","capital.gain","capital.loss","fnlwgt", "hours.per.week","marital.status","relationship"
- target:'test_income'
- protected attributes: "education", "education.num","age","race","sex","native.country"

F1_MACRO

Model	Baseline	Smote
Logistic Regression	0.771(0.002)	0.753(0.001)
SVM	0.773(0.004)	0.756(0.003)
KNN	0.755(0.005)	0.733(0.003)
Random Forest	0.75(0.006)	0.741(0.005)

AUC

Model	Baseline	Smote
Logistic Regression	0.891(0.001)	0.891(0.001)
SVM	0.876(0.001)	0.882(0.004)
KNN	0.863(0.004)	0.854(0.003)
Random Forest	0.861(0.006)	0.855(0.005)

3.4 Term deposit

The Bank Marketing dataset is publicly available in the UCI repository [15], and it includes 41,188 individuals with 20 attributes. The task is to predict whether the client has subscribed to a term deposit service based on features such as:
- pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric);

- previous: number of contacts performed before this campaign and for this client (numeric);
- month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec');
- day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri');
- default: has credit in default? (categorical: 'no','yes','unknown');
- housing: has housing loan? (categorical: 'no','yes','unknown');
- loan: has personal loan? (categorical: 'no','yes','unknown').

- features:'default','housing','loan','contact','month','day_of_week','duration', 'campaign','pdays', 'previous','poutcome','emp.var.rate','cons.price.idx', 'cons.conf.idx','euribor3m','nr.employed'
- target:'y'
- protected attributes: 'age','marital','education'

F1_MACRO

Model	Baseline	Smote
Logistic Regression	0.744(0.005)	0.741(0.011)
SVM	0.717(0.006)	0.733(0.005)
KNN	0.747(0.008)	0.728(0.003)
Random Forest	0.753(0.007)	0.757(0.006)

AUC

Model	Baseline	Smote
Logistic Regression	0.931(0.005)	0.931(0.005)
SVM	0.92(0.003)	0.921(0.003)
KNN	0.903(0.005)	0.9(0.004)
Random Forest	0.935(0.003)	0.933(0.003)

3.5 Student

These two datasets contain information on student achievement in secondary education of two Portuguese schools and they have been built by using school reports and questionnaires in 2014. The attributes include student grades, as well as demographic, social and school related features. Two datasets

are provided from the UCI Machine Learning Repository [16] regarding the performance of students (not necessarily the same students) in two distinct subjects: Mathematics and Portuguese language. G3 is the target variable, which indicates the final year grade (issued at the end of the school year) between 1 and 20, corresponding to a positive grade if above 9, or negative if lower. the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful.

- features:"school","address","famsize","Pstatus","reason","internet","studytime", "failures","schoolsup", "paid","activities","nursery","higher","freetime", "goout","Dalc","Walc","health","absences","guardian","traveltime", "famsup","romantic","famrel"
- target:'G3_target'
- protected attributes: "sex","age_f","Mjob","Fjob","Medu_f","Fedu_f"

3.5.1 Math student

F1_MACRO

Model	Baseline	Smote
Logistic Regression	0.615(0.02)	0.605(0.035)
SVM	0.518(0.064)	0.67(0.029)
KNN	0.558(0.034)	0.542(0.038)
Random Forest	0.585(0.016)	0.615(0.024)

AUC

Model	Baseline	Smote
Logistic Regression	0.653(0.047)	0.645(0.045)
SVM	0.714(0.023)	0.717(0.036)
KNN	0.598(0.052)	0.586(0.046)
Random Forest	0.683(0.03)	0.692(0.045)

3.5.2 Portuguese student

F1_MACRO

Model	Baseline	Smote
Logistic Regression	0.655(0.023)	0.663(0.025)
SVM	0.554(0.039)	0.651(0.043)
KNN	0.558(0.032)	0.562(0.04)
Random Forest	0.592(0.064)	0.643(0.041)

AUC

Model	Baseline	Smote
Logistic Regression	0.785(0.038)	0.784(0.04)
SVM	0.767(0.057)	0.751(0.059)
KNN	0.625(0.045)	0.681(0.062)
Random Forest	0.813(0.018)	0.825(0.017)

3.6 Titanic

Dataset coming from [17]. The task is to predict which passengers survived the Titanic shipwreck by exploiting the following features:
- pclass: ticket class (1 = 1st, 2 = 2nd, 3 = 3rd);
- sibsp: number of siblings / spouses aboard the Titanic;
- parch number of parents / children aboard the Titanic;
- ticket: ticket number;
- fare: passenger fare;
- cabin: cabin number;
- embarked: port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Some features as Cabin and ticket number are excluded.

- features:'Pclass','SibSp','Parch','Fare','Embarked'
- target:'Survived' (yes/no)
- protected attributes: 'Sex','Age'

F1_MACRO

Model	Baseline	Smote
Logistic Regression	0.682(0.036)	0.687(0.028)
SVM	0.676(0.02)	0.673(0.032)
KNN	0.665(0.02)	0.648(0.014)
Random Forest	0.676(0.026)	0.656(0.013)

AUC

Model	Baseline	Smote
Logistic Regression	0.744(0.015)	0.745(0.014)
SVM	0.748(0.024)	0.742(0.027)
KNN	0.714(0.01)	0.694(0.024)
Random Forest	0.728(0.024)	0.715(0.018)

3.7 Communities and Crimes

Dataset coming from the UCI Repository [18]. Many variables are included so that algorithms that select or learn weights for attributes could be tested. The attribute to be predicted is Per Capita Violent Crimes. The variables included in the dataset involve the community, such as the percent of the population considered urban, and the median family income, and involving law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units. The per capita violent crimes variable was calculated using population and the sum of crime variables considered violent crimes in the United States: murder, rape, robbery, and assault. There was apparently some controversy in some states concerning the counting of rapes. These resulted in missing values for rape, which resulted in incorrect values for per capita violent crime. These cities are not included in the dataset. Many of these omitted communities were from the midwestern USA. All numeric data was normalized into the decimal range 0.00-1.00 using an Unsupervised, equal-interval binning method. The normalization preserves rough ratios of values within an attribute (e.g. double the value for double the population within the available precision except for extreme values; all values more than 3 SD above the mean are normalized to 1.00; all values more than 3 SD below the mean are nromalized to 0.00). Other than racepctblack, racePctAsian (percentage of population that is of asian

heritage) or racePctHisp (percentage of population that is of hispanic heritage) attributes were present in the dataset. Moreover, attributes disaggregated for different classes of sensitive attributes have been discarded from feature selection. Among them there were: blackPerCap (per capita income for african americans, as well as for asian heritage and native americans), MalePctDivorce (percentage of males who are divorced, same for female), PctImmigRecent (percentage of _immigrants_ who immigated within last 3 years (numeric - decimal)), PctImmigRec5 (percentage of _immigrants_ who immigated within last 5 years), PctRecImmig5 (percent of _population_ who have immigrated within the last 5 years). Only general information attributes (highly aggregated) have been retained together with data on public spending on services.

- features:'population','householdsize','pctUrban','medIncome','medFamInc', 'perCapInc','PctPopUnderPov','PersPerFam','OwnOccLowQuart','OwnOccMedVal', 'OwnOccHiQuart','RentLowQ','RentMedian','RentHighQ', 'MedRent', 'MedRentPctHousInc','MedOwnCostPctInc','MedOwnCostPctIncNoMtg', 'NumInShelters','NumStreet','LandArea','PopDens', 'PctUsePubTrans','LemasPctOfficDrugUn'

- target:'y'

- protected attributes: 'Afroamericans','12-29people'

Where agePct12t21 is the percentage of population that is 12-21 in age (numeric – decimal) and racepctblack is the percentage of population that is african american (numeric – decimal).

F1_MACRO

Model	Baseline
Logistic Regression	0.811(0.007)
SVM	0.81(0.012)
KNN	0.774(0.018)
Random Forest	0.807(0.016)

AUC

Model	Baseline
Logistic Regression	0.889(0.004)
SVM	0.894(0.007)
KNN	0.856(0.013)
Random Forest	0.89(0.004)

3.8 Compas

Data [19] contains variables used by the COMPAS algorithm in scoring criminal defendants in Broward County (Florida), along with their outcomes within two years of the decision. The original dataset contains 28 variables, among which we took sex, race and age category into account as sensitive attributes, while we assumed two year recid as target variable and the risk score as classifier R, which indicates a "recidivism degree" between 1 and 10, and can be interpreted as estimated recidivism risk if above 4, so that it represents a binary classifier. We chose the COMPAS dataset because it is well-known in the scientific communities that study measures of algorithmic bias and related mitigation strategies. It was provided by the U.S. non-profit organization ProPublica that showed that the COMPAS algorithm was distorted in favor of white individuals, whereby those who were rearrested were nearly twice as likely to be misclassified as low risk than black defendants 5 . Furthermore, the black defendants who did not get rearrested were nearly twice as likely to be misclassified as higher risk (false positive) than white defendants. The major cause was that the number of records in the dataset related to black defendants was much higher than the number of records of white defendants, as well as the number of black recidivists compared to white recidivists.

3.9 Juvenile justice

This dataset [20] consists of 4753 data and presents the statistical descriptive variables, as well as the recidivism of children and young people who completed an educational program in 2010 in Catalonia, between the date of completion of the program and the end of 2013 or the end of 2015. The dataset describes the profile of youths and also minors who had contact with juvenile justice in relation to the program done. Additional provided data are: the juvenile recidivism rate, the specific rates and the profile of the recidivist and recidivism according to the program. In particular, the SAVRY

variables present the risks of recidivism among young people, as well as their specific areas of risk and needs; among them, SAVRY_total_score indicates a "total recidivism degree" between 1 and 100. In order to assume it as binary score variable, we make reference to the COMPAS dataset, where the total percentage of moderate and high recidivism risk is around 45%, so we decide to consider the same percentage of data in the Juvenile dataset as moderate-high risk: following this line of reasoning, SAVRY total score can be interpreted as affirmative (estimated) recidivism risk if above 15. Then, we assumed reincidencia 2013 as target variable, which represents the recidivity by the end of 2013, and we examined the protected attributes sex, stranger, country of origin, area of origin, age category and age.

Chapter 4

Balance indices

Inequality Measures being adopted as risk indicators compute the disproportion within a vector values. Indices rely on the concept of diversity which can be quantified in many different ways. The two main factors taken into account when measuring diversity are:

1. Richness: The number of categories of an attribute is a measure of richness. The more categories present, the 'richer' the attribute. Attribute richness as a measure on its own takes no account of the number of individuals of each species present. It gives as much weight to those categories which have very few individuals as to those which have many individuals.

2. Evenness: it is a measure of the relative abundance of the different categories making up the richness of an area.

In addition to base paper study, another sensitive attribute of Default has been studied: Marriage attribute. Moreover, given the dependency of the risk measure on the training dataset used, indices have been computed multiple times (for 5 runs) to provide robustness of risk estimate. Hence, the five different runs involve different random split in training and test set, providing greater confidence in fairness measures computed on the test set. So, correlation between risk measures and fairness criteria is studied not only over different models, but also over different runs. With respect to base paper study, attention has been directed toward continuous attributes. In particular the attribute 'Age' is intended to be studied both in the raw form (continuous) and the categorized version by splitting it out in three intervals: 'less than 25','25-45','greater than 45'. However, as will be said, continuous

attributes have not been investigated due to the absence of a method by which to calculate unfairness on those attributes.

Six balance indexes, that widely used in the literature, have been selected. According to [1], indexes have been adjusted in order to meet three criteria:

- range in the interval [0, 1];
- share the same interpretation: the closer the measure to 1 and the higher the balance (i.e. categories have similar frequencies), and vice-versa values closer to 0 means more concentration of frequencies in few categories, thus an imbalanced distribution;
- deal with empty classes, i.e., classes that exist (potentially there could be occurrences) but are not represented in the given dataset: we decided to take into account all the classes of each selected sensitive attribute, including also the classes with zero occurrences. The motivation for this choice is that in our view a dataset that contains no instance of a given class – e.g. all males or all whites – is imbalanced.

Often, in real datasets, missing values can be found. These have not been excluded from the analysis and considered as a separate "NA" category.

4.1 Shannon index

The idea is that the more different categories there are, and the more equal their proportional abundances, the more difficult it is to correctly predict which category will be the next one. The Shannon index quantifies the uncertainty associated with this prediction. In the branch of Information Theory it is called Shannon Entropy and it is the expected value of self-information associated to a random variable that is the average level of information/uncertainty inherent in the variable's possible outcomes.
It is most often calculated as follows:

$$H = -\frac{1}{\ln(m)} \cdot \sum_{j=1}^{m} p_j \cdot ln(p_j)$$

The min value is when observations are only in one class (no uncertainty in prediction) and it is equal to 0. Max value is when observation are uniformly distributed among classes (max uncertainty in prediction) and it is equal to

ln(m), with m classes. Shannon index is a special case of a more general entropy measure, the Rényi entropy.

4.2 Rényi entropy

The Rényi entropy [21] generalizes the Hartley entropy, the Shannon entropy, the collision entropy and the min-entropy. Entropies quantify the diversity, uncertainty, or randomness of a system. The entropy is named after Alfréd Rényi. The Rényi entropy of order α , where $\alpha >= 0$ and $\alpha! = 1$, is defined as:

$$H_\alpha = \frac{1}{\ln(m)} \cdot \frac{1}{1-\alpha} \cdot \ln(\sum_{i=1}^{m} p_i^\alpha)$$

Here, X is a discrete random variable with possible outcomes in the set $A = \{x_1, x_2, .., x_n\}$ and corresponding probabilities $p_i = Pr(X = x_i)$ for $i = 1, ..., n$. The logarithm is conventionally taken to be base 2, especially in the context of information theory where bits are used. If the probabilities are $p_i = 1/n$ for all i in $1, ..., n$, then all the Rényi entropies of the distribution are equal: $H_\alpha(X) = log(n)$.
It is a non-increasing function of α. In the limit for $\alpha \to 0$, the Rényi entropy is just the logarithm of the size of the support of X. The limit for $\alpha \to 1$ is the Shannon entropy. As α approaches $+\infty$, the Rényi entropy is increasingly determined by the events of highest probability.

$$log(n) = H_0 \geq H_1 \geq H_2 \geq H_\infty$$

There are some special cases.

4.2.1 Hartley or max-entropy

Provided the probabilities are nonzero, it is the logarithm of the cardinality of the alphabet (A) of X, sometimes called the Hartley entropy of X.

$$H_0(X) = log(n) = log(A)$$

Where n is the number of categories of the attribute.

4.2.2 Collision entropy

Collision entropy, sometimes just called "Rényi entropy", refers to the case $\alpha = 2$.

$$H_2(X) = -log(\sum_{i=1}^{n} p_i^2)$$

4.2.3 Min-entropy

In the limit as $\alpha \to +\infty$, the Rényi entropy converges to the min-entropy:

$$H_\infty(X) = -log(\max_i p_i)$$

The name min-entropy stems from the fact that it is the smallest entropy measure in the family of Rényi entropies. In this sense, it is the strongest way to measure the information content of a discrete random variable. In particular, the min-entropy is never larger than the Shannon entropy.

4.3 Simpson index

This measure estimates the probability that two entities taken at random from the dataset of interest represent the same type.

$$\lambda = \sum_{i}^{R} p_i^2$$

where R is richness (the total number of types in the dataset). This equation is also equal to the weighted arithmetic mean of the proportional abundances p_i of the types of interest, with the proportional abundances themselves being used as the weights. Proportional abundances are by definition constrained to values between zero and unity, but it is a weighted arithmetic mean, hence $\lambda >= 1/R$, which is reached when all types are equally abundant.

Since mean proportional abundance of the types increases with decreasing number of types and increasing abundance of the most abundant type, λ obtains small values in datasets of high diversity and large values in datasets of low diversity. This is counterintuitive behavior for a diversity index, so often such transformations of λ that increase with increasing diversity have been used instead. The most popular are:

4.3.1 Gini–Simpson index

It is a measure of heterogeneity used in many disciplines and often discussed with different designations: examples are political polarization, market competition, ecological diversity as well as racial discrimination. Heterogeneity reflects how many different types (such as protected groups) are represented. In statistics, the heterogeneity of a discrete random variable which assumes m categories with frequency p_i (with i = 1, ..., m) can vary between a degenerate case (= minimum value of heterogeneity) and an equiprobable case (= maximum value of heterogeneity, since categories are all equally represented). This means that for a given m, the heterogeneity increases if probabilities become as equal as possible, i.e. the different protected groups have similar representations. The Gini index is computed as follows:

$$G = \frac{m}{m-1} \cdot \left(1 - \sum_{i=1}^{m} p_i^2\right)$$

Index normalized between 0 and 1. Min value: 0 (observations only in one class). Max value: (m-1)/m (max heterogeneity).

4.3.2 Inverse Simpson index

As before, we consider a discrete random variable which assumes m categories with frequency p_i where i = 1, ..., m (that is, the proportion p_i of the species i with respect to the total number of species):

$$D = \frac{1}{m-1} \cdot \left(\frac{1}{\sum_{j=1}^{m} p_j^2} - 1\right)$$

1 is subtracted since $\min(1/\lambda) = 1$ (so max = m-1).

4.3.3 Hill index

By comparing the equation used to calculate λ with the equations used to calculate true diversity, it can be seen that $1/\lambda$ (inverse Simpson) equals D_2, i.e., true diversity as calculated with q = 2. The original Simpson's index hence equals the corresponding basic sum. The true diversity [22] in a dataset is calculated by first taking the weighted generalized mean M(q-1) of the proportional abundances of the types in the dataset, and then taking

the reciprocal of this. The equation is:

$$D_q = \frac{1}{m-1} \cdot [((\sum_{i=1}^{m} p_i^q)^{\frac{1}{1-q}}) - 1]$$

In the equation, R is richness (the total number of types in the dataset), and the proportional abundance of the i-th type is p_i. The proportional abundances themselves are used as the nominal weights. The numbers D_q are called Hill numbers of order q or effective number of species. The value of q is often referred to as the order of the diversity. It defines the sensitivity of the diversity value to rare vs. abundant species by modifying how the weighted mean of the species proportional abundances is calculated. With some values of the parameter q, the value of M_{q1} assumes familiar kinds of weighted mean as special cases. In particular, q = 0 corresponds to the weighted harmonic mean, q = 1 to the weighted geometric mean, and q = 2 to the weighted arithmetic mean. As q approaches infinity, the weighted generalized mean with exponent q-1 approaches the maximum p_i value, which is the proportional abundance of the most abundant species in the dataset. Generally, increasing the value of q increases the effective weight given to the most abundant species. This leads to obtaining a larger M_{q1} value and a smaller true diversity (D_q) value with increasing q.

Previous study has shown that Gini-Simpson and Shannon Indices exhibit higher values than other two measures used, Inverse-Simpson and Imbalance Ratio. Therefore, it is worth trying to increase q for Simpson index (through true diversity) and α for Shannon (Rényi entropy).

4.4 Imbalance Ratio

It is a widely used measure made of the ratio between the highest and the lowest frequency. We take the inverse in order to normalize it in the range [0, 1] and to render it a balance measure – i.e. higher values mean balance –. In detail, the formula we adopt is:

$$IIR = \frac{\min p_j}{\max p_j}$$

4.5 Continuous attributes indices

The main problem with categorical heterogeneity measures are the assumptions of categorical data. First, categories to which one's data belong must be (A) known a priori and (B) scientifically valid. In some cases, this will be more problematic than in others.

4.5.1 Generalized entropy index

It has been proposed [23] as a measure of income inequality in a population.

$$GE(\alpha) = \begin{cases} Th_T = 1 - \frac{1}{N*\log(N)} \cdot \sum_{i=1}^{N} (\frac{x_i}{\overline{x}} \ln(\frac{x_i}{\overline{x}})), & \text{if } \alpha = 1 \\ Th_L = 1 - [-\frac{1}{N*\log(N)} \cdot \sum_{i=1}^{N} \ln(\frac{x_i}{\overline{x}})], & \text{if } \alpha = 0 \end{cases} \tag{4.1}$$

where N is the number of cases, y_i is the income for case i and α is a parameter which regulates the weight given to distances between incomes at different parts of the income distribution. For large α the index is especially sensitive to the existence of large incomes, whereas for small α the index is especially sensitive to the existence of small incomes.

For a population of N "agents" each with characteristic x, the situation may be represented by the list x_i (i $= 1, ..., N$) where x_i is the characteristic of agent i. For example, if the characteristic is income, then x_i is the income of agent i.

The second and third formula are also named respectively Theil T and Theil L [24].

If everyone has the same income, then Theil T equals 0. If one person has all the income, then Theil T gives the result ln(N), which is maximum inequality. Dividing Theil T by ln(N) can normalize the equation to range from 0 to 1. The final formula is obtained by computing 1 – Theil, to obtain 1 as max equality and 0 as imbalance.

4.5.2 Gini coefficient

It is a measure of statistical dispersion [25] intended to represent the income inequality or wealth inequality within a nation or any other group of people. It was developed by the Italian statistician and sociologist Corrado Gini. The Gini coefficient measures the inequality among values of a frequency distribution (for example, levels of income). A Gini coefficient of zero expresses

perfect equality, where all values are the same (for example, where everyone has the same income). A Gini coefficient of one (or 100%) expresses maximal inequality among values (e.g., for a large number of people where only one person has all the income or consumption and all others have none, the Gini coefficient will be nearly one). If all people have non-negative income (or wealth, as the case may be), the Gini coefficient can theoretically range from 0 (complete equality) to 1 (complete inequality); it is sometimes expressed as a percentage ranging between 0 and 100. In reality, both extreme values are not quite reached. If negative values are possible (such as the negative wealth of people with debts), then the Gini coefficient could theoretically be more than 1. Gini coefficient is defined as half of the relative mean absolute difference. The mean absolute difference is the average absolute difference of all pairs of items of the population, and the relative mean absolute difference is the mean absolute difference divided by the average $\bar{x}$, to normalize for scale. If x_i is the wealth or income of person i, and there are n persons, then the Gini coefficient G is given by:

$$G = 1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|}{2N^2 \overline{x}}$$

To obtain a value consistent we our notation (1 is max equality), 1 – G is used.

4.5.3 Pietra-Ricci index

The Pietra-Ricci [26] index is recognized as the share of the total T that should be redistributed by the people possessing more than the mean towards the people possessing less than the mean, in order to achieve perfect equality.

$$P = 1 - \frac{\sum_{i=1}^{N} |x_i - \overline{x}|}{2N\overline{x}}$$

N is the population size and $\bar{x} = \sum_i x_i / N$.

Chapter 5

Unfairness measures

In machine learning, a given algorithm is said to be fair, or to have fairness, if its results are independent of given variables , especially those considered sensitive, such as the traits of individuals which should not correlate with the outcome (i.e. gender, ethnicity, sexual orientation, disability, etc.).

If an algorithm is not operating properly the effects on people can be significant and long-lasting, such as regarding education or employment opportunities, and access to financial credit services.

In classification problems, an algorithm learns a function to predict a discrete characteristic Y, the target variable, from known characteristics X. We model A as a discrete random variable which encodes some characteristics contained or implicitly encoded in that we consider as sensitive characteristics (gender, ethnicity, sexual orientation, etc.). We finally denote by R the prediction of the classifier.

Fairness measures are properties of the joint distribution of the score, sensitive attribute, and the target variable. In other words, if we know the joint distribution of the random variables (R, A, Y), we can without ambiguity determine whether this joint distribution satisfies one of these criteria or not.

All criterion are observational, we can express them using probability statements involving the random variables at hand. Observational definitions have many appealing aspects. They're often easy to state and require only a lightweight formalism. They make no reference to the inner workings of the classifier, the decision maker's intent, the impact of the decisions on the population, or any notion of whether and how a feature actually influences the outcome. We can reason about them fairly conveniently as we saw earlier. In principle, observational definitions can always be verified given samples from the joint distribution—subject to statistical sampling error.

Most statistical measures of fairness rely on the following metrics, which are best explained using a confusion matrix – a table that is often used in ML to describe the accuracy of a classification model. Rows and columns of the matrix represent instances of the actual and predicted classes, respectively. For a binary classifier, both predicted and actual classes have two values: positive and negative.

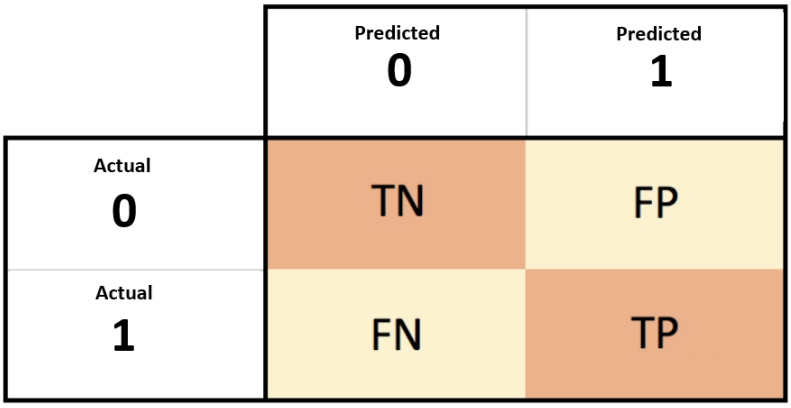


Figure 5.1. Basic representation of a confusion matrix. Image taken from [27]

Cells of the confusion matrix [28] help explain the following definitions:

1. True positive (TP): a case when the predicted and actual outcomes are both in the positive class.
2. False positive (FP): a case predicted to be in the positive class when the actual outcome belongs to the negative class.
3. False negative (FN): a case predicted to be in the negative class when the actual outcome belongs to the positive class.
4. True negative (TN): a case when the predicted and actual outcomes are both in the negative class.
5. Positive predictive value (PPV): the fraction of positive cases correctly predicted to be in the positive class out of all predicted positive cases, (TP)/(TP+FP). PPV is often referred to as precision, and represents the probability of a subject with a positive predictive value to truly belong to the positive class, $P(Y = 1|d = 1)$. It is the probability of

an applicant with a good predicted credit score to actually have a good credit score.

6. False discovery rate (FDR): the fraction of negative cases incorrectly predicted to be in the positive class out of all predicted positive cases, (FP)/(TP+FP). FDR represents the probability of false acceptance, P(Y = 0|d = 1), e.g., the probability of an applicant with a good predicted credit score to actually have a bad credit score.

7. False omission rate (FOR): the fraction of positive cases incorrectly predicted to be in the negative class out of all predicted negative cases, (FN)/(TN+FN). FOR represents the probability of a positive case to be incorrectly rejected, (P(Y = 1|d = 0)), e.g, the probability of an applicant with a bad predicted credit score to actually have a good score.

8. Negative predictive value (NPV): the fraction of negative cases correctly predicted to be in the negative class out of all predicted negative cases, (TN)/(TN+FN). NPV represents the probability of a subject with a negative prediction to truly belong to the negative class, P(Y = 0|d = 0), e.g., the probability of an applicant with a bad predicted credit score to actually have such score.

9. True positive rate (TPR): the fraction of positive cases correctly predicted to be in the positive class out of all actual positive cases, (TP)/(TP+FN). TPR is often referred to as sensitivity or recall; it represents the probability of the truly positive subject to be identified as such, P(d = 1|Y = 1). It is the probability of an applicant with a good credit score to be correctly assigned with such score.

10. False positive rate (FPR): the fraction of negative cases incorrectly predicted to be in the positive class out of all actual negative cases, (FP)/(FP+TN) . FPR represents the probability of false alarms – falsely accepting a negative case, P(d = 1|Y = 0), e.g., the probability of an applicant with a actual bad credit score to be incorrectly assigned with a good credit score.

11. False negative rate (FNR): the fraction of positive cases incorrectly predicted to be in the negative class out of all actual positive cases, (FN)/(TP+FN) . FNR represents the probability of a negative result given an actually positive subject, P(d = 0|Y = 1), e.g., the probability of an applicant with a good credit score to be incorrectly assigned with a bad credit score.

12. True negative rate (TNR): the fraction of negative cases correctly predicted to be in the negative class out of all actual negative cases, (TN)/(FP+TN). TNR represents the probability of a subject from the negative class to be assigned to the negative class, P(d = 0|Y = 0), e.g., the probability of an applicant with a bad credit score to be correctly assigned with such score.

5.1 Independence (or demographic parity, or statistical parity)

We say the random variables (R,A) satisfy independence if the sensitive characteristic A is statistically independent to the prediction R, and we write R⊥A. It requires the acceptance rate to be the same in all groups, where acceptance correspond to the event R=1. We can express this notion with the following formula:

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b) = ...$$

if A is binary, we can compute the Independence *unfairness* measure as:

$$U_I(a_1, a_2) = |P(R = 1 \mid A = a_1) - P(R = 1 \mid A = a_2)|$$

A lower value of this measure indicates more similar acceptance rates and therefore better fairness. Demographic parity (and disparate impact) ensure that the positive prediction is assigned to the two groups at a similar rate. One disadvantage of these two measures is that a fully accurate classifier may be considered unfair, when the base rates (i.e., the proportion of actual positive outcomes) of the various groups are significantly different. This definition ignores any possible correlation between Y and A, within groups recalls are not considered [2].

5.2 Separation (or equalized odds)

We say the random variables (R,A,Y) satisfy separation if the sensitive characteristics A are statistically independent to the prediction R given the target value Y, and we write R⊥A|Y. When R is binary (i.e., R=0 or R=1 and thus Y=0 or Y=1), requires the equivalence of true positive rate and false positive rate for each level of the protected attribute under analysis:

- TPR

$$P(R = 1 \mid Y = 1, A = a_1) = P(R = 1 \mid Y = 1, A = a_2) = ...$$

- FPR

$$P(R = 1 \mid Y = 0, A = a_1) = P(R = 1 \mid Y = 0, A = a_2) = ...$$

If A is binary (that is, $A = a_1$ or a_2), the we can compute two Separation *unfairness* measures (U) as:

$$U_{S\_TPR}(a_1, a_2) = |P(R = 1 \mid Y = 1, A = a_1) - P(R = 1 \mid Y = 1, A = a_2)|$$

$$U_{S\_FPR}(a_1, a_2) = |P(R = 1 \mid Y = 0, A = a_1) - P(R = 1 \mid Y = 0, A = a_2)|$$

This measure was designed to overcome the disadvantages of measures such as demographic parity. The measure computes the difference between the false positive rates (FPR), and the difference between the true positive rates (TPR) of the two groups. Smaller differences between groups indicate better fairness. In contrast to demographic parity and disparate impact measures, a fully accurate classifier will necessarily satisfy the two equalized odds constraints. Nevertheless, since equalized odds relies on the actual ground truth (i.e., Y), it assumes that the base rates of the two groups are representative and were not obtained in a biased manner. One use case that demonstrates the effectiveness of this measure investigated the COMPAS algorithm used in the United States criminal justice system. For predicting recidivism, although its accuracy was similar for both groups (African-Americans and Caucasians), it was discovered that the odds were different. It was discovered that the system had falsely predicted future criminality (FPR) among African-Americans at twice the rate predicted for white people [29]; importantly, the algorithm also induced the opposite error, significantly underestimating future crimes among Caucasians (FNR).

5.3 Sufficiency

We say the random variables (R,A,Y) satisfy sufficiency if the sensitive characteristics A are statistically independent to the target value Y given the prediction R, and we write Y⊥A|R. The sufficiency criterion, where R is binary (i.e., R=0 or R=1 and thus Y=0 or Y=1), requires the equality of PPV and NPV for each level of the protected attribute under analysis:

- PPV

$$P(Y = 1 \mid R = 1, A = a_1) = P(Y = 1 \mid R = 1, A = a_2) = ...$$

- NPV

$$P(Y = 1 \mid R = 0, A = a_1) = P(Y = 1 \mid R = 0, A = a_2) = ...$$

if A is binary, we can compute the Sufficiency *unfairness* measure as:

$$U_{S\_PPV}(a_1, a_2) = |P(Y = 1 \mid R = 1, A = a_1) - P(Y = 1 \mid R = 1, A = a_2)|$$

$$U_{S\_NPV}(a_1, a_2) = |P(Y = 1 \mid R = 0, A = a_1) - P(Y = 1 \mid R = 0, A = a_2)|$$

According to [30], sufficiency often comes for free (at least approximately) as a consequence of standard machine learning practices. The flip side is that imposing sufficiency as a constraint on a classification system may not be much of an intervention. In particular, it would not effect a substantial change in current practices. Therefore we could expect low values with respect to this measure.

5.4 Overall accuracy equality

A classifier satisfies this definition if the subject in the protected and unprotected groups have equal prediction accuracy, that is, the probability of a subject from one class to be assigned to it. It has to satisfies the following formula:

$$P(R = Y \mid A = a_1) = P(R = Y \mid A = a_2) = ...$$

if A is binary, we can compute the OAE *unfairness* measure as:

$$U_{OAE}(a_1, a_2) = |P(R = Y \mid A = a_1) - P(R = Y \mid A = a_2)|$$

The definition assumes that true negatives are as desirable as true positives.

All the definitions above can be easily extended to the case of non-binary attributes – i.e. $m > 2$ – by taking the mean of indexes computed considering all the possible pairs of levels in the attribute A:

$$U(a_1, ..., a_m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} U(a_i, a_j)$$

The unfairness measures range in the interval [0, 1]. They assume values equal to zero for a perfectly fair classification and higher values for unfair behavior.

5.5 Other unfairness measures

These measures are not considered in this study but are proposed as further investigations in future works.

5.5.1 Treatment Equality

This definition looks at the ratio of errors that the classifier makes rather than at its accuracy. A classifier satisfies this definition if the subjects in the protected and unprotected groups have an equal ratio of FN and FP, satisfying the formula:

$$\frac{FN_{A=a1}}{FP_{A=a1}} = \frac{FN_{A=a2}}{FP_{A=a2}} = ...$$

if A is binary, we can compute the Treatment Equality *unfairness* measure as:

$$U_{S\_GroupRatio}(a_1, a_2) = |\frac{FN_{A=a1}}{FP_{A=a1}} - \frac{FN_{A=a2}}{FP_{A=a2}})|$$

5.5.2 Group calibration

for any predicted probability score S, subjects in both protected and unprotected groups should not only have an equal probability to truly belong to the positive class, but this probability should be equal to S. That is, if the predicted probability score is s, the probability of both male and female applicants to truly belong to the positive class should be:

$$P(Y = 1|S = s, G = m) = P(Y = 1|S = s, G = f) = s$$

The intuition behind this definition is that if a classifier states that a set of applicants have a certain probability s of having a good credit score then approximately s percent of these applicants should indeed have a good credit score.

5.5.3 Balance for positive class

A classifier satisfies this definition if the subjects constituting the positive class from both protected and unprotected groups have equal average predicted probability score S. This means that the expected value of probability score for the protected and unprotected groups with positive actual outcome Y is the same, satisfying the formula:

$$E(S|Y = +, A = a) = E(S|Y = +, A = b)\ \forall a, b \in A$$

Violation of this balance means that, for example, one group of applicants with good credit score would consistently receive higher probability score than applicants with a good credit score from the other group.

5.5.4 Balance for negative class

This definition states that subjects constituting negative class from both protected and unprotected groups should also have equal average predicted probability score S. That is, the expected value of probability assigned by the classifier to male and female applicant with bad actual credit score should be same:

$$E(S|Y = -, A = a) = E(S|Y = -, A = b)\ \forall a, b \in A$$

Part III

Analysis

Chapter 6

Balance measures behaviours

6.1 Categorical attributes

How imbalance is penalized strictly depends on the index and attribute considered. Different attribute cardinalities are considered to better evaluate each index in different contexts. In particular m = 2,3,4,6,16 are reported, with m number of attribute's categories. Following considerations are merged with base paper ones adding also new indices insights.

All the plots are taken from run1 (thus a given seed of training set).

We start from the 'Sex' attribute of cardinality 2.
Among the first four base paper indices, IIR is the one which strongly penalizes distribution imbalance. By generalizing Shannon and Simpson we can be more strict, indeed Hill_100 and IIR ranges are near.

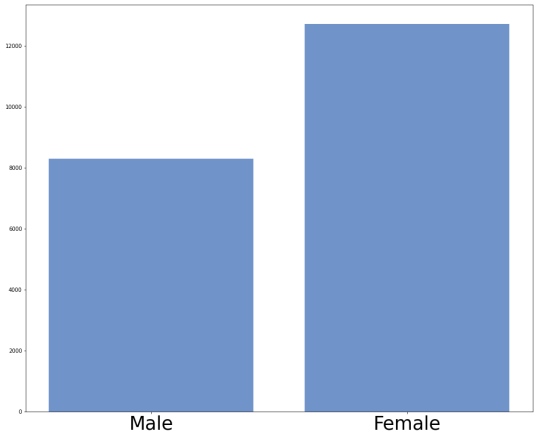


Figure 6.1. dataset:'default_uci', attribute:'SEX', m=2

INDEX	**MIN-MAX**
Gini	0.9555-0.9589
Shannon	0.9677-0.9701
Simpson	0.9149-0.921
IIR	0.6517-0.6628
Renyi_10	0.8022-0.8126
Renyi_100	0.7313-0.7411
Hill_11	0.7352-0.7477
Hill_100	0.6601-0.6714

Let's proceed with the 'age category' attribute with $m = 3$. As before, IIR is the lowest index and Hill_100 approach it (higher values of q can even be lower than IIR). At the same time for α=q=100, Hill<Renyi and it is confirmed in the following examples. In fact we generally see that Simpson Index has lower values than Shannon.

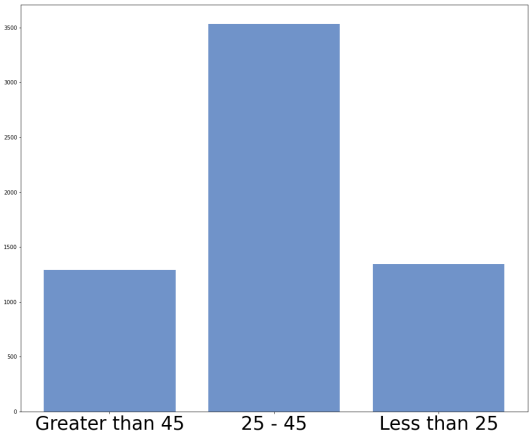


Figure 6.2. dataset:'compas', attribute:'age_cat', m=3

INDEX	VALUE
Gini	0.8715
Shannon	0.8912
Simpson	0.6933
IIR	0.3661
Renyi_10	0.5645
Renyi_100	0.5132
Hill_11	0.4239
Hill_100	0.3787

Proceeding to 'race' attribute with $m = 6$, there is a relevant disproportion, proved by general lower values of the indices. Now the difference between Hill_100 and IIR is higher. One could think that it depends on attribute cardinality, but it is refuted by the following example regarding the 'marriage' attribute with $m = 4$. It presents an high disproportion and the two ranges distance is higher, suggesting that in these cases for Hill (even more for Renyi) it is harder to approach IIR degree of penalty (need high value of the parameters).

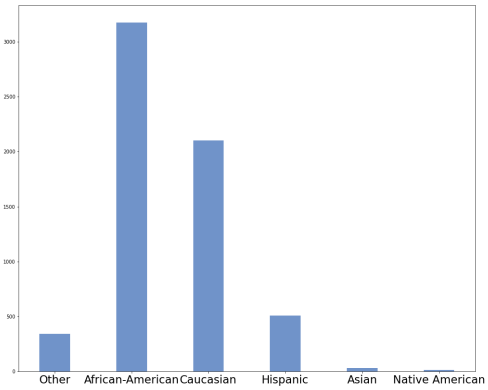


Figure 6.3. dataset:'compas', attribute:'race', m=6

INDEX	**VALUE**
Gini	0.7312
Shannon	0.6212
Simpson	0.312
IIR	0.0035
Renyi_10	0.4112
Renyi_100	0.3747
Hill_11	0.2151
Hill_100	0.1914

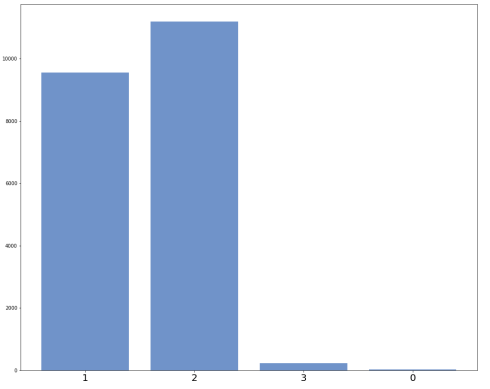


Figure 6.4. dataset:'dccc', attribute:'marriage', m=4

INDEX	MIN-MAX
Gini	0.6791-0.6804
Shannon	0.5438-0.5455
Simpson	0.346-0.3474
IIR	0.0032-0.0036
Renyi_10	0.4897-0.4939
Renyi_100	0.4586-0.4648
Hill_11	0.3223-0.3263
Hill_100	0.2962-0.3016

Finally, for attribute 'education' of cardinality 16, what said before is confirmed by an higher cardinality but lower disproportion and nearer ranges distance of Hill and IIR.

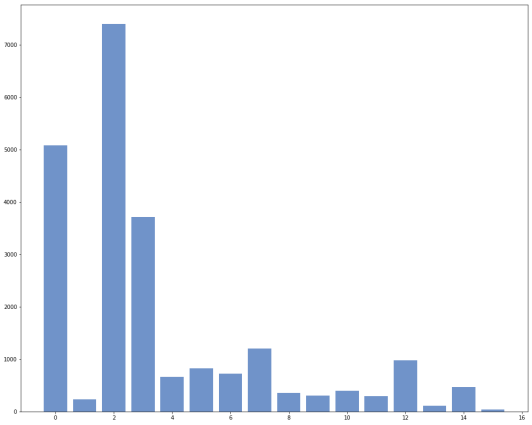


Figure 6.5. dataset:'census income', attribute:'education', m=16

INDEX	MIN-MAX
Gini	0.8633-0.8653
Shannon	0.7324-0.7357
Simpson	0.2829-0.2864
IIR	0.0043-0.0053
Renyi_10	0.4501-0.4554
Renyi_100	0.4101-0.415
Hill_11	0.1629-0.1663
Hill_100	0.1412-0.144

Chapter 7

Unfairness discussion

Since a lot of datasets and models have been evaluated, it could be interesting exploring unfairness indices values starting from a subset of datasets and the moving to an higher set to see if some aspects recur.

Moreover, for each model two versions are present: baseline and smote (the same algorithm trained on the oversampled training set).

First of all, only base paper datasets measures are shown, except for Compas and juvenile because they can obscure models differences (since these two datasets are trained with a black box algorithm).

Dataset -	Attribute	Diff Ind	Diff TPR	Diff FPR	Diff PPV	Diff NPV	Diff OAE
dccc -	Sex	0.0138	0.0117	0.0095	0.0285	0.0311	0.0324
dccc -	Education	0.0550	0.1649	0.0313	0.4086	0.0971	0.0955
statlog -	Status	0.1344	0.1828	0.0876	0.2572	0.1667	0.0988
statlog -	Sex	0.0634	0.0596	0.0237	0.0983	0.0442	0.0194
statlog -	Foreign worker	0.1687	0.4683	0.1165	0.3667	0.1844	0.1080
income -	Education	0.2385	0.2984	0.0986	0.3100	0.1457	0.0881
income -	Race	0.0901	0.1228	0.0394	0.1450	0.0528	0.0587
income -	Sex	0.1822	0.0889	0.0831	0.0124	0.1088	0.1227
income -	Native country	0.1614	0.4363	0.0965	0.4522	0.1186	0.1264
student_math -	Sex	0.0646	0.0970	0.0926	0.1211	0.1668	0.1250
student_math -	Age	0.2452	0.2461	0.2753	0.2643	0.3276	0.2075
student_math -	Mother's job	0.1218	0.1373	0.3102	0.1855	0.3380	0.1507
student_math -	Father's job	0.1377	0.1836	0.3037	0.1718	0.4247	0.1807
student_math -	Mother's education	0.2096	0.1718	0.3663	0.2097	0.3155	0.1751
student_math -	Father's education	0.0914	0.0655	0.1610	0.0966	0.1921	0.0760
student_port -	Sex	0.0621	0.0678	0.1757	0.0516	0.1285	0.0332
student_port -	Age	0.3502	0.3438	0.3137	0.3259	0.3301	0.2385
student_port -	Mother's job	0.1821	0.1525	0.4050	0.0501	0.3194	0.1213
student_port -	Father's job	0.2389	0.1666	0.4310	0.0763	0.4006	0.1328
student_port -	Mother's education	0.2349	0.2251	0.2946	0.1295	0.2497	0.1999
student_port -	Father's education	0.3016	0.2743	0.3514	0.2684	0.3266	0.2498

Table 7.1. Mean unfariness measures over five run for Logistic Regression (only base paper attributes)

For each pair unfairness measure (column) – model (row), the average fairness index, for each dataset, is compared between baseline model and smote one. In this sense we aim to see if there are some sistematic differences in term of fairness by applying smote or not. It is important to point out that the measures used, given a model, are the mean over the five runs, as illustrated as following for the LR model.

In each subplot, there are five bar pair comparisons (blu is baseline, orange smote). Each pair corresponds to a dataset, in order from left to right: dccc, income, statlog, student mat, student port. For each dataset the mean unfairness difference has been computed among its attributes. For example dccc has two attributes Sex and Education, which baseline Independence differences are respectively 0.0138 and 0.0550. So the dataset mean unfairness is the mean between these two values, thus 0.0344.

For how unfairness has been defined, the shorter a bar is the fairer the model is.

Starting by analyzing the grid by column (for each model), we describe how the algorithm behaves for different unfairness measures depending on whether baseline or smote is used.

- LR: Looking at Independence, True Positive Rate, False Positive Rate and Overall Accuracy Equality, for nearly all datasets smote leads to higher degree of unfairness (even for a little amount). It is something that happens still systematically by at its opposite for Sufficiency Criterion, in this case smote application mitigates inequalities in terms of Positive Predictive value and Negative Predictive value.

- SVM: For the first three unfairness indices (Independence, TPR, FPR) baseline is fairer than smote in a per-dataset fashion. This feature still remains proceeding to the two remaining algorithms. With respect to OAE, the differences between baseline and smote encountered in LR are not consistent in other algorithms, sometimes smote is fairer, other it is unfairer than base case. Conversely, Sufficiency indices are lower after smote application, achieving also big gaps depending on the dataset. It is confirmed also for KNN and RF.

- What said above, remains valid for KNN and RF.

Analysing the grid by row we aim to spot if, given a measure, there are differences between different algorithms. Given an unfairness index, distributions over datasets seem to be quite similar for all algorithm except for some slightly differences. Generally, differences between baseline and smote seems to be higher in SVM model.

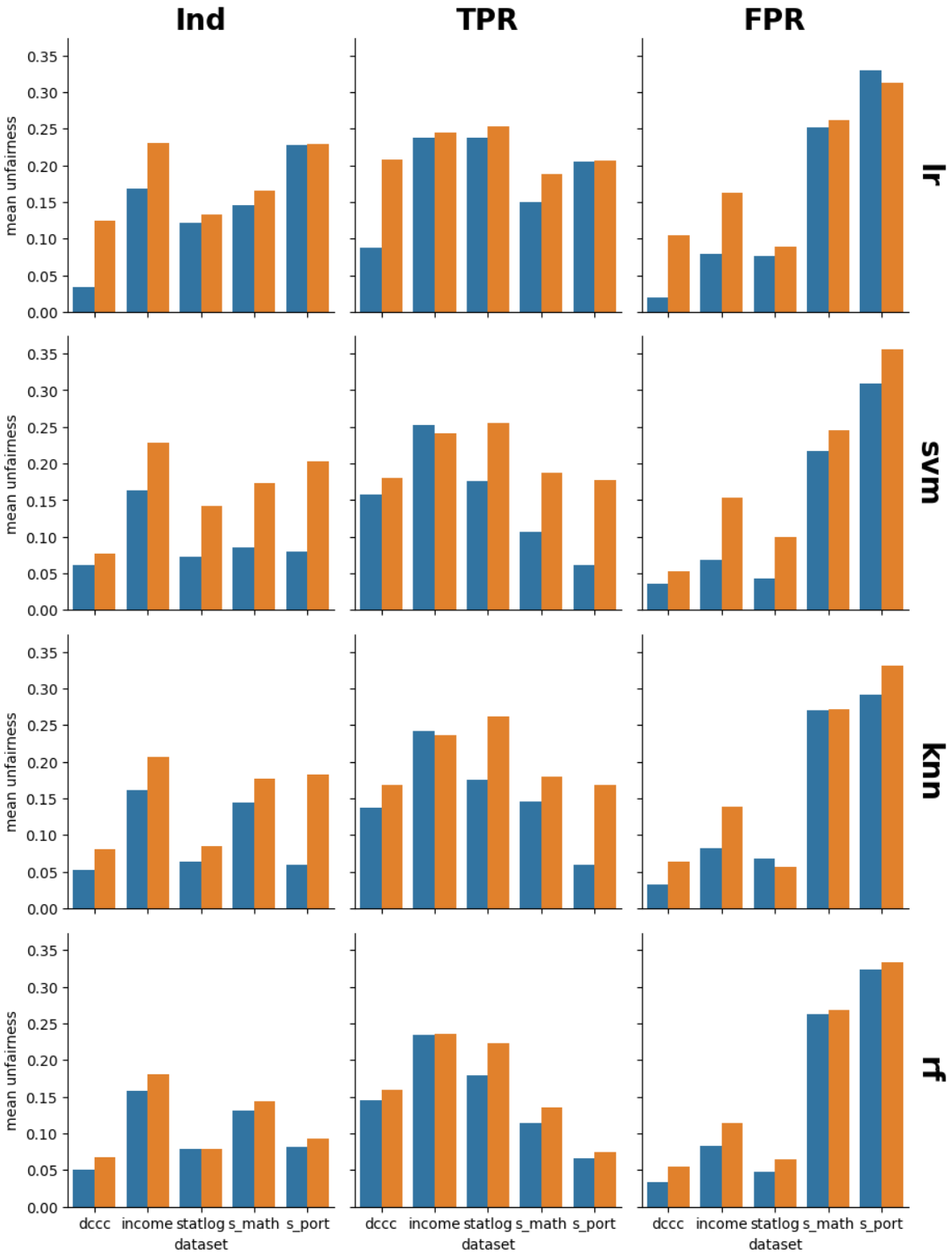


Figure 7.1. datasets mean Independence, Separation measures

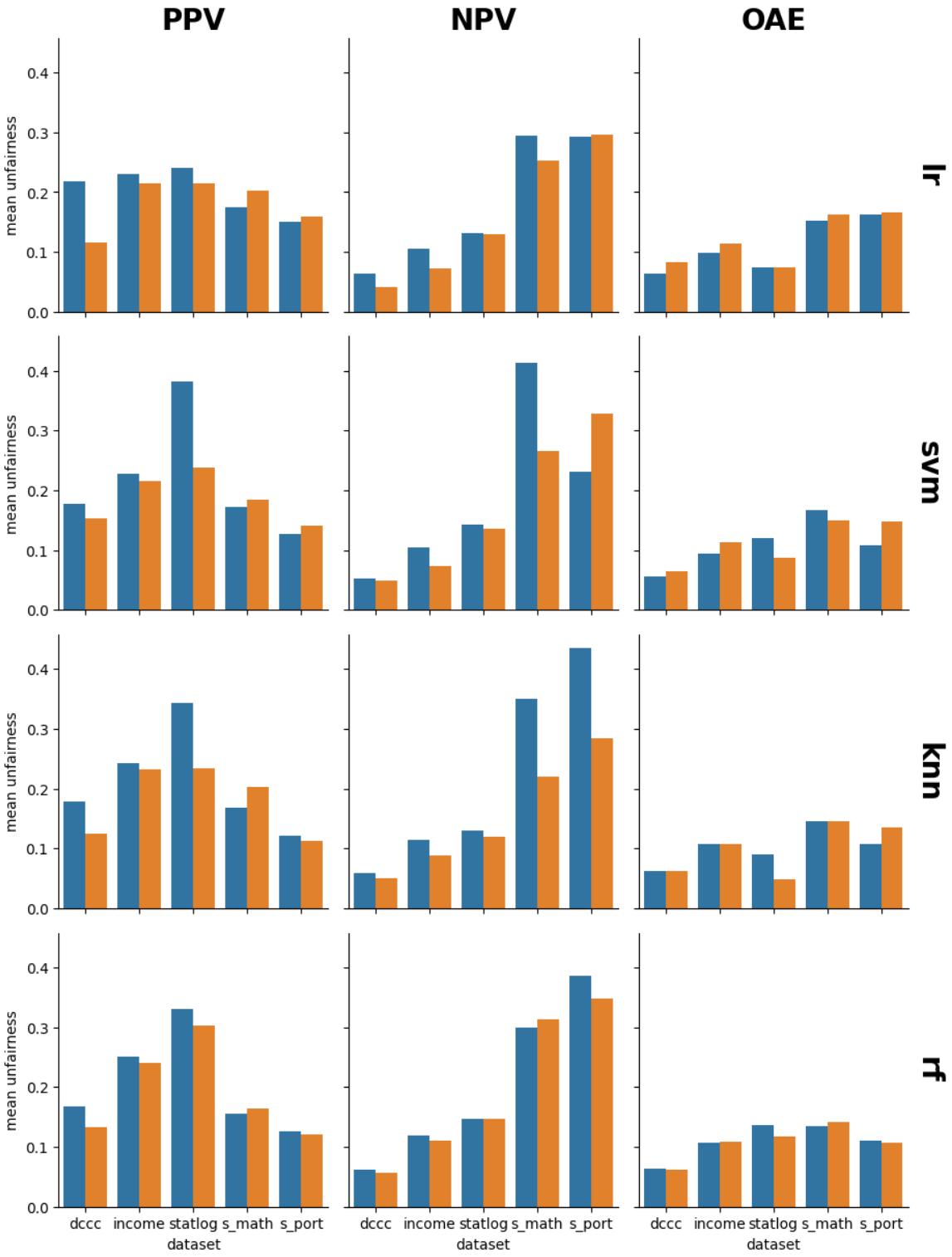


Figure 7.2. datasets mean Overall accuracy equality, Sufficiency measures

Afterwards we proceed to analyze the same pairs Fair index – model, in an aggregated fashion: given a model, take the mean over all datasets of that fairness index. We aim to see if previous observations are robusts to an aggregated evaluation of unfairness values.

We can clearly see that independently of the algorithm, baseline Independence, TPR, FPR are lower than smote ones. On the contrary smote PPV and NPV are lower than baseline. No conclusions can be made for AOE which comparisons look quite similar.

Other observations:

- RF seems to be the model which gaps between baseline and smote are smaller;
- With respect to Independence, TPR, FPR, SVM provides higher gaps as observed in the per-dataset analysis;
- Looking at Sufficency NPV higher differences are achieved by KNN, with a substantial mitigation of nearly 0.10;
- Independence, TPR: SVM, KNN, RF achieve, on average, more mitigated results(on baselines) with respect to LR;
- FPR, PPV and OAE are pretty similar among all four models;
- NPV seems to be penalized by KNN, with an average value of 0.28 when other models are below 0.25.

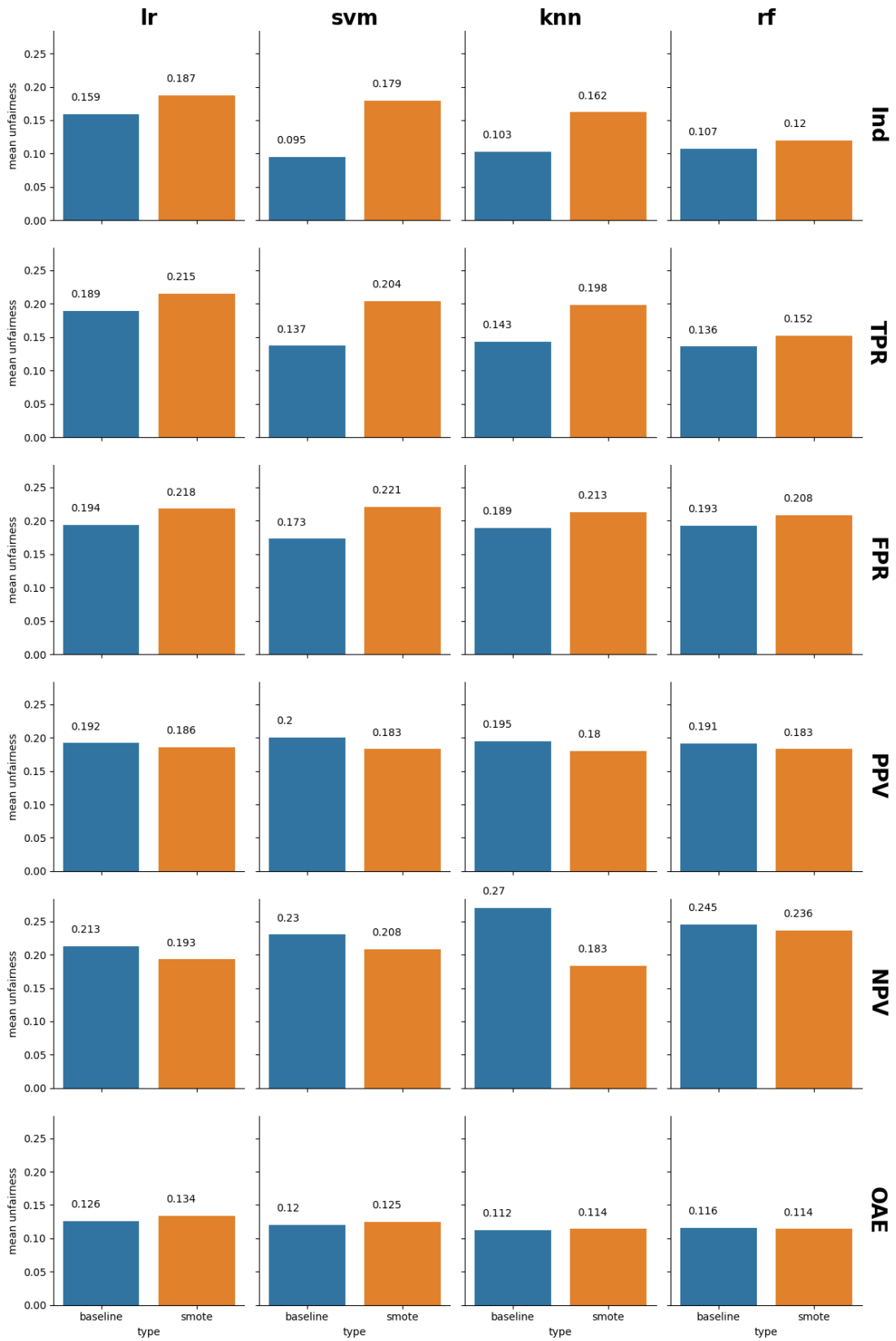


Figure 7.3. base paper aggregated fairness measures

To assess if previous findings are confirmed in a wider scenario with more data, two additional datasets are added: Term deposit and Titanic. Community has been excluded since smote version was not available.

Dataset -	Attribute	Diff Ind	Diff TPR	Diff FPR	Diff PPV	Diff NPV	Diff OAE
term deposit -	education	0.0892	0.1450	0.0704	0.1041	0.0368	0.0582
term deposit -	marriage	0.0529	0.1119	0.0410	0.0496	0.0266	0.0434
Titanic -	Sex	0.2132	0.0915	0.0584	0.5604	0.4440	0.0630

We directly explore aggregated results after to new datasets introduction.

Observations:

- Still in term of mean level of unfairness Independence and TPR of SVM, KNN, RF are lower than LR (the gap seems to be increased a little);
- Talking about NPV, it seems again penalized by KNN with a value near 0.4. Moreover, baseline unfairness is higher than smote one, confirming previous consideration;
- Again RF is the model with smallest differences between baseline and smote.

General mitigation behaviours are confirmed:

- baseline mitigates Independence, TPR, FPR over smote;
- smote mitigates mainly sufficiency NPV. Now PPV bars are pretty similar;
- Overall Accuracy equality is balanced between the two variants.

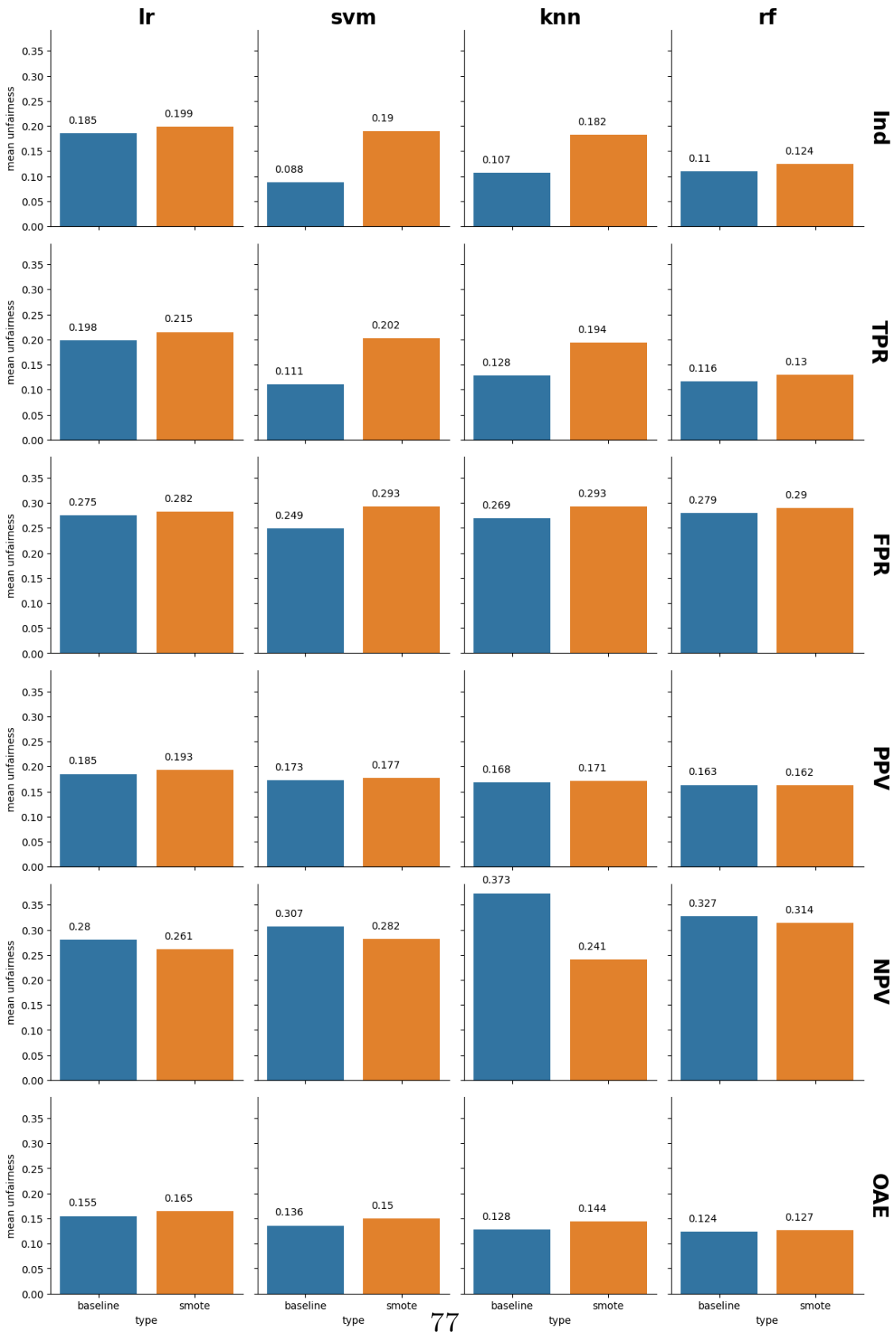


Figure 7.4. extensive aggregated fairness measures

7.1 Unfairness evaluation by attribute cardinality

A further investigation over unfairness values, for the different pairs fair measure-model, can be done by splitting out attributes by their cardinality. In particular $m = 2,5,8$ have been chosen for comparison. It is important to specify that the support (the number of attributes) of each cardinality is not the same. They are respectively 7,9,3. So in terms of comparison consistency, $m = 8$ lacks of a proper support with respect to other two cardinalities.
The datasets and attributes considered are the ones belonging to the extensive analysis.

Firstly we limit our observations to cardinalities 2 and 5 since they have comparable supports.

- With respect to Separation FPR and Sufficiency NPV on average, attributes with m=2 result in fairer measure compared to m=5.
- Looking at the grid as a whole, except for Independence and TPR gaps in unfairness between the two cardinalities does not depend on the model taken into consideration.
- For both cardinalities, higher Independence, TPR, are reached for Logistic Regression (less fair).
- Out of 24 plots, unfairness of $m = 5$ is higher than $m = 2$ in 15 cases. Moreover, often the gap is wider than the opposite (especially for FPR, NPV), if $m = 2$ is greater than $m = 5$ it is by a small quantity .

By adding $m = 8$ in the evaluation, it comes out to be the unfairest in 14/24 cases and bigger than $m = 2$ in 23/24 plots. This last proportion suggests higher unfairness measure for attribute of cardinality 8 than 2. However it is suggested to further investigate this speculation by adding datasets and attribute of cardinality 8 to make the two supports better comparable.

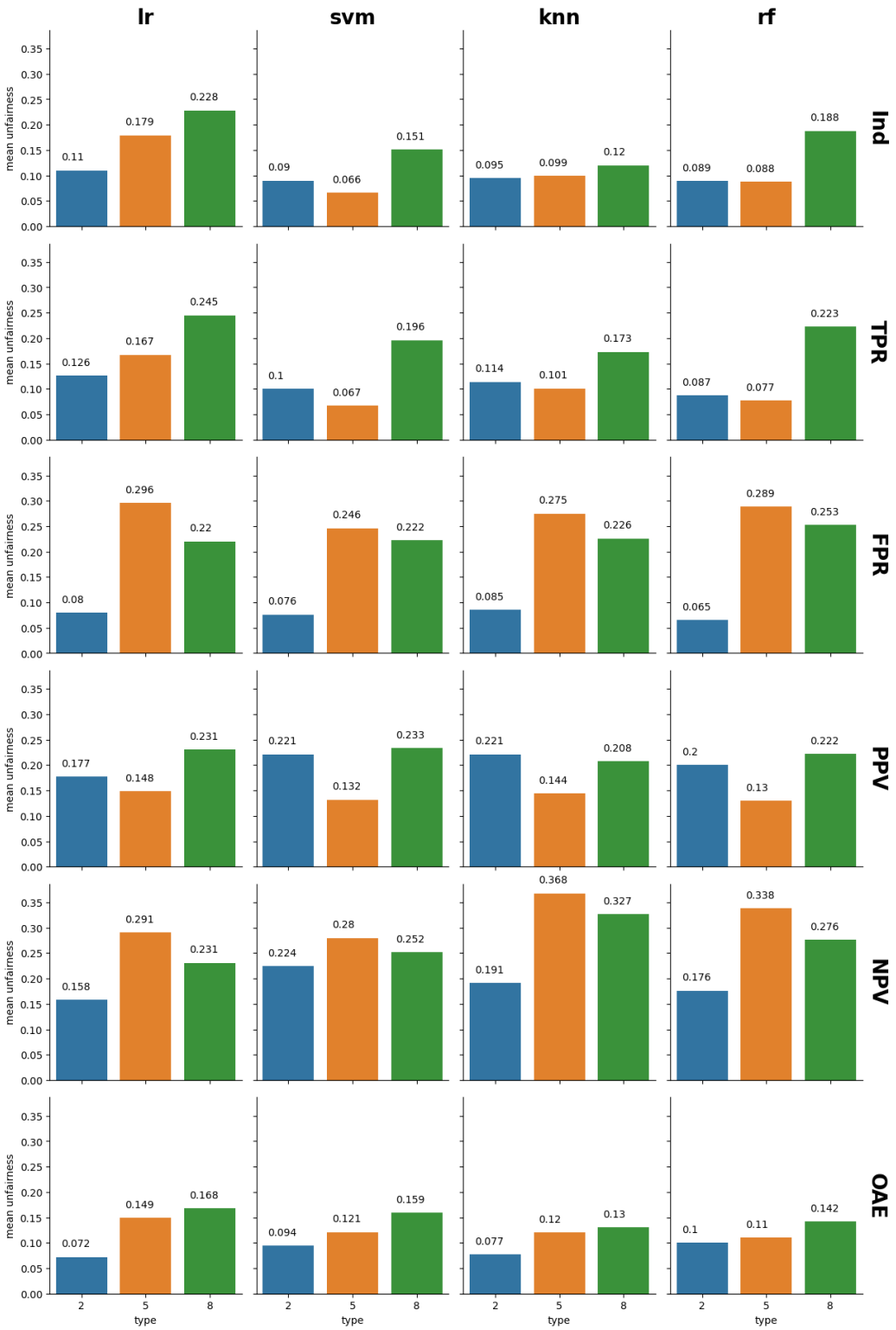


Figure 7.5. mean unfairness-model grid by attributes cardinality

Chapter 8

Indexes relationship

To evaluate the relationship between balance measures and unfairness ones, the same criteria of base paper have been applied: Box-plot distribution of unfairness for low/high risk attributes (based on a defined threshold) and Spearman correlation coefficient between balance and unfairness. As regards box-plots, for low risk attribute, we expect unfairness distribution shifted towards low values (fair behaviour). Conversely for for high risk attribute, we expect unfairness distribution shifted towards higher values. It can vary depending on the balance-unfairness pair at hand. Correlation matrix gives us insights about predictability of downstream unfairness thanks to upstream balance evaluation.

Remember the notation used:

- Balance measure in [0,1], high values correspond to balance;
- Unfairness measure in [0,1], high values correspond to unfairness.

We expect good predictive balance indices to be negatively correlated to unfairness.

8.1 Base paper

In analyzing output measures and their relations, we start from base paper datasets, attributes, inference algorithm and indices (balance and unfairness). The reason is to check if previous findings are confirmed over the five different runs performed in this study.

- Retained datasets: Compas, Juvenile, dccc, statlog, income, student math, student port;
- Excluded attributes: Marriage from dccc;
- Balance indices: Gini-Simpson, Shannon, Simpson, Inverse Imbalance Ratio;
- Unfairness indices: Independence, Separation (PPV), Separation (NPV). Then: Sufficiency (PPV, NPV) and Overall accuracy equality;
- Algorithm: Logistic Regression

Dataset -	Attribute	Gini	Shannon	Simpson	IIR	Diff Ind	Diff TPR
Compas -	Ethnicity	0.7312	0.6212	0.3120	0.0035	0.2480	0.2880
Compas -	Sex	0.6165	0.7023	0.4456	0.2351	0.0502	0.0250
juvenile -	Sex	0.4401	0.5460	0.2821	0.1440	0.0215	0.1169
juvenile -	Stranger	0.9489	0.9628	0.9027	0.6312	0.0314	0.0370
juvenile -	Country of origin	0.6135	0.4427	0.0434	0.0019	0.4136	0.4256
juvenile -	Area of origin	0.7092	0.6783	0.3279	0.0285	0.1300	0.0596
juvenile -	Age category	0.6626	0.5999	0.3957	0.0086	0.0605	0.4112
juvenile -	Age	0.8982	0.8334	0.6382	0.0167	0.0493	0.3110
dccc -	Sex	0.9565	0.9684	0.9166	0.6547	0.0181	0.0006
dccc -	Education	0.7351	0.5664	0.2839	0.0013	0.0558	0.1813
statlog -	Status	0.9236	0.8985	0.7514	0.1533	0.1299	0.1318
statlog -	Sex	0.8556	0.8932	0.7476	0.4493	0.0770	0.1325
statlog -	Foreign worker	0.1431	0.2290	0.0770	0.0386	0.1302	0.3750
income -	Education	0.8634	0.7324	0.2832	0.0043	0.2367	0.2873
income -	Race	0.3272	0.3453	0.0887	0.0097	0.0833	0.1555
income -	Sex	0.8825	0.9135	0.7897	0.4895	0.1882	0.1077
income -	Native country	0.2029	0.1765	0.0060	0.0000	0.1575	0.4339
student_math -	Sex	0.9974	0.9981	0.9949	0.9034	0.0040	0.0363
student_math -	Age	0.9144	0.8230	0.6041	0.0143	0.4123	0.4432
student_math -	Mother's job	0.9373	0.9239	0.7495	0.2551	0.1400	0.1306
student_math -	Father's job	0.7323	0.6952	0.3537	0.0692	0.0992	0.1026
student_math -	Mother's education	0.9205	0.8601	0.6985	0.0211	0.1352	0.1040
student_math -	Father's education	0.9362	0.8761	0.7459	0.0241	0.0500	0.0302
student_port -	Sex	0.9748	0.9818	0.9509	0.7262	0.1167	0.0973
student_port -	Age	0.9044	0.8122	0.5747	0.0163	0.3525	0.3763
student_port -	Mother's job	0.9287	0.9128	0.7228	0.1954	0.1708	0.1707
student_port -	Father's job	0.7551	0.7192	0.3815	0.0675	0.1506	0.1487
student_port -	Mother's education	0.9337	0.8640	0.7379	0.0076	0.1477	0.1547
student_port -	Father's education	0.9303	0.8787	0.7276	0.0400	0.3838	0.3686

Table 8.1. Base paper logistic regression from run2

To evaluate imbalance indices values for each measure inside a run, the mean over all measurements has been taken (i.e. take the mean over the

columns of the previous table). We aim to see which values on average they take. Same evidences come out from different runs: Gini and Shannon are the less strict imbalance measures, conversely the IIR is the indices enforcing more imbalance. Therefore it is proven that Gini and Shannon have the tendency to assume high values. Simpson index is on the middle in terms of penalization enforced, accordingly to previous evaluations.

RUNS/INDEX	Gini	Shannon	Simpson	IIR
RUN1	0.769	0.736	0.527	0.176
RUN2	0.771	0.739	0.535	0.179
RUN3	0.769	0.737	0.530	0.178
RUN4	0.769	0.736	0.530	0.172
RUN5	0.771	0.738	0.534	0.179

Table 8.2. Inside-run means over all attributes of imbalance indices

8.1.1 Unfairness measures vs. Balance classification

The following figures reports, for each pair of balance-unfairness measure, a box-plot that shows the distribution of unfairness measure values for higher risk vs lower risk attributes. The values used for the two type of indices are the run1 values of the model being considered. First of all, to compare distributions with base paper ones, baseline Logistic Regression has been choosen. We start with a threshold of 33%, corresponding to "imbalance" for higher risk, and "unknown" + "balanced" for lower risk. The more a boxplot leans to the right the more unfair the treatment of those attributes is. The more a box-plot is close to the left (zero) the more the relative attributes are threated fairly. When the two boxes (Red and Yellow) do not overlap, it means that the imbalance-based approach to risk identification is able to discriminate between fair and unfair classification.

Base paper results are replicated and confirmed:

- Gini: good discrimination ability for the true positive rates of the Separation criterion. No discrimination for other two unfairness measures.
- Imbalance Ratio: good discrimination ability for both the indicators of Separation and a limited ability for the Independence one.

- Shannon: good discrimination for the Independence, excellent for the TPR, bad for FPR.

- Simpson: the limited ability to discriminate Independence present in base paper here is not confirmed.

All indices, except Simpson, were able to detect TPR differences. No index, except IIR, was able to anticipate discrimination in terms of FPR differences.

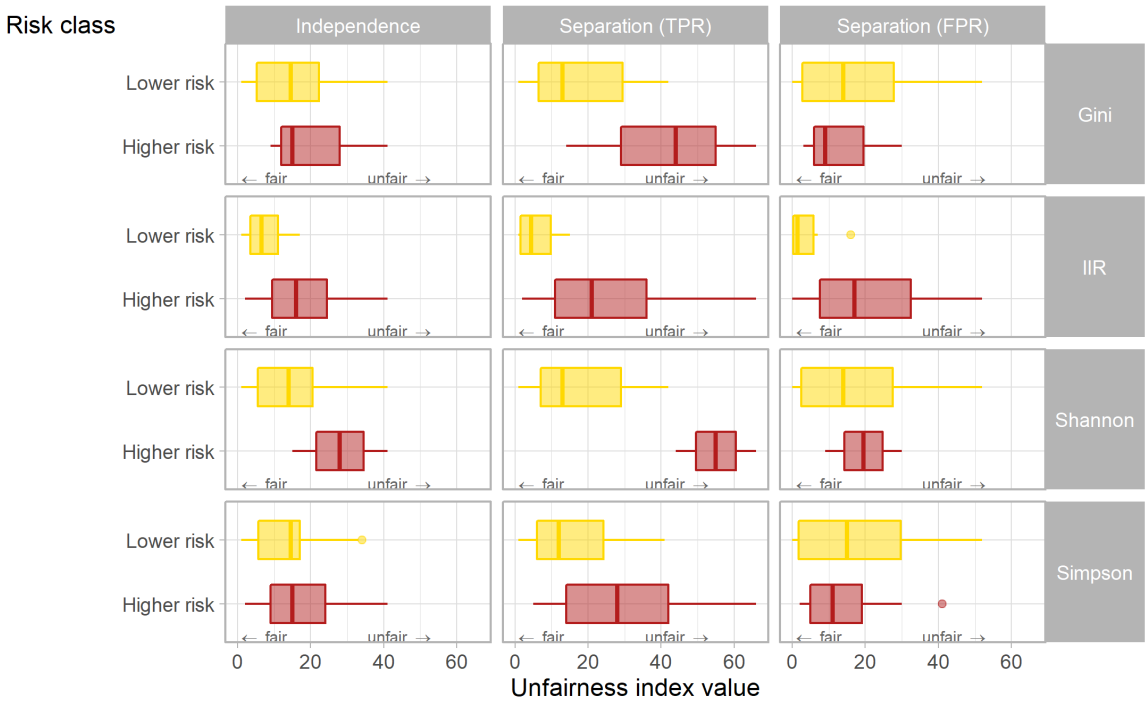


Figure 8.1. LR Ind, TPR, FPR distribution, threshold 33%

It's interesting to further investigate these distribution, in the same scenario (base paper), by introducing new unfairness measures: Sufficiency positive predictive values, Sufficiency negative predictive values, Overall Accuracy Equality.

- Gini: only on Sufficiency PPV we get good discrimination. Notably for NPV, there is a bad reversal of distribution where higher risk attributes have low unfairness values and lower risk ones are associated to unfairness (the opposite of what is desired).

- Imbalance Ratio: good discrimination for PPV and moderately good for NPV. No discrimination capabilities on OAE.
- Shannon: it behaves very well on Sufficiency PPV and good on Overall Accuracy Equality.
- Simpson: good discrimination for PPV. No discrimination for other two unfairness measures.

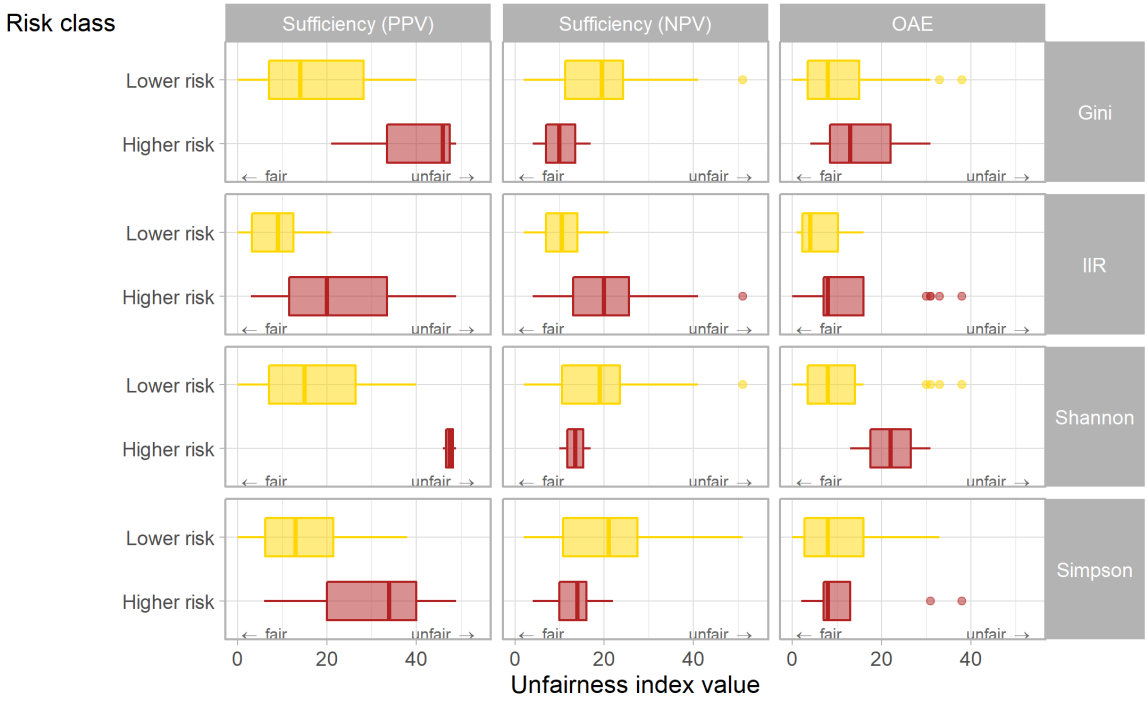


Figure 8.2. LR PPV, NPV, OAE distribution, threshold 33%

Considering all unfairness measures as a whole, IIR and Shannon are the two achieving more discrimination out of the six indices considered, with respectively 5/6 and 4/6.

For each unfairness index we order index in terms of discrimination capability:

1. Independence: Shannon, IIR;
2. TPR: Shannon, Gini, IIR;
3. FPR: IIR (others not capable);

4. PPV: Shannon, Gini, IIR;

5. NPV: IIR (others not capable);

6. OAE: Shannon (others not capable).

In two cases, for Separation and Sufficiency, only IIR is able to detect differences in unfairness distribution between the two category of risk at 33%.

By taking into consideration that our study positive classes are generally the minority ones, unfairness predictability among these results favoured by Shannon and Gini indices (however IIR is able to detect it). As regards negative classes (the majority ones) only IIR is good on discrimination (FPR, NPV).

8.1.2 Boxplots - Model comparison with threshold 33%

Models unfairness distributions can be compared between each other to see if model choice can influence box-plots configuration among the two levels of risk: high risk and low risk attributes. For each model, run1 unfairness measures are used. Compas and Juvenile have been excluded to not hinder models comparability. In the following grid we have over rows the four balance indexes considered. On the columns there are three unfairness measures: Independence, TPR and FPR. Each subplot shows for each of the four models, the unfairness measures vs. Balance classification over the y axis. therefore towards above there are high values of unfairness. Sub plots axis have been inverted to favour plot interpretability: unfairness values along the y axis and models over the x axis.

We start from the first three unfairness measures: Independence, TPR, FPR.

- Gini: only the Random Forest seems to have discrimination capabilities in terms of Independence (previously Logistic regression failed as now). All models are able to discriminate TPR with LR the most notable. On the contrary, no model performs well as regards FPR differences;

- Imbalance Ratio: all models are good on Independence. For TPR, only LR reaches satisfying discriminative distribution. On FPR, all models look similar. Generally IIR seems robusts over different models;

- Shannon: good on Independence excepting for SVM. Very good discrimination for TPR where LR gets has very polarized box-plots. Bad for FPR, no model acts discriminative in this sense;
- Simpson: only TPR seems satisfying, all models have discrimination capabilities.

Looking at columns(unfairness indices) we can make some further considerations:

- Independence: for this measure, Random Forest is the model getting best results on the four balance indices considered;
- True Positive Rate: here Logistic regression shows more polarized distributions among two risks categories;
- False Positive Rate: it's very difficult, independently on the model, to anticipate this measure with the notably exception of IIR.

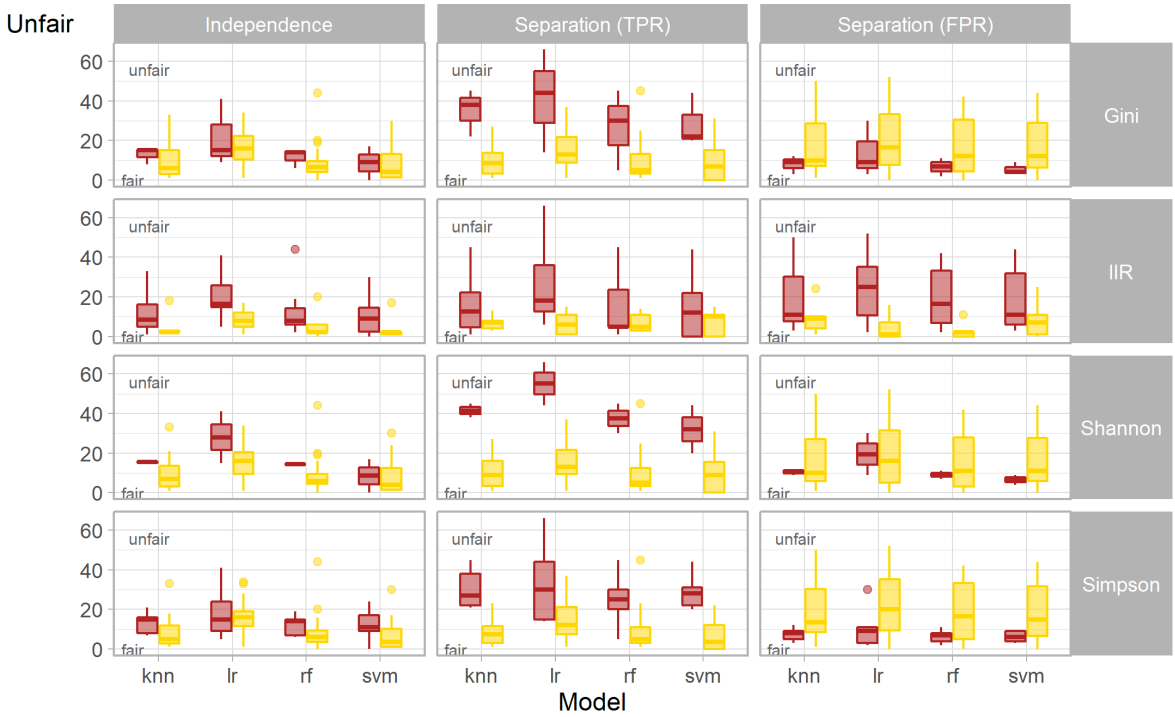


Figure 8.3. Models Ind, TPR, FPR distribution, threshold 33%

Same type of considerations can be extended to Sufficiency and Overall Accuracy Equality (columns of the grid).

- PPV: all models performs similar given the imbalance index. Among these, Shannon is the one which get higher differences for all considered models;
- NPV: very hard to discriminate unfairness differences for this measure. Only IIR models are able at some extents, especially LR and RF;
- OAE: Polarization is very limited and value tend to be concentrated in short ranges. Among indices, Shannon acts the best, especially for LR and RF.

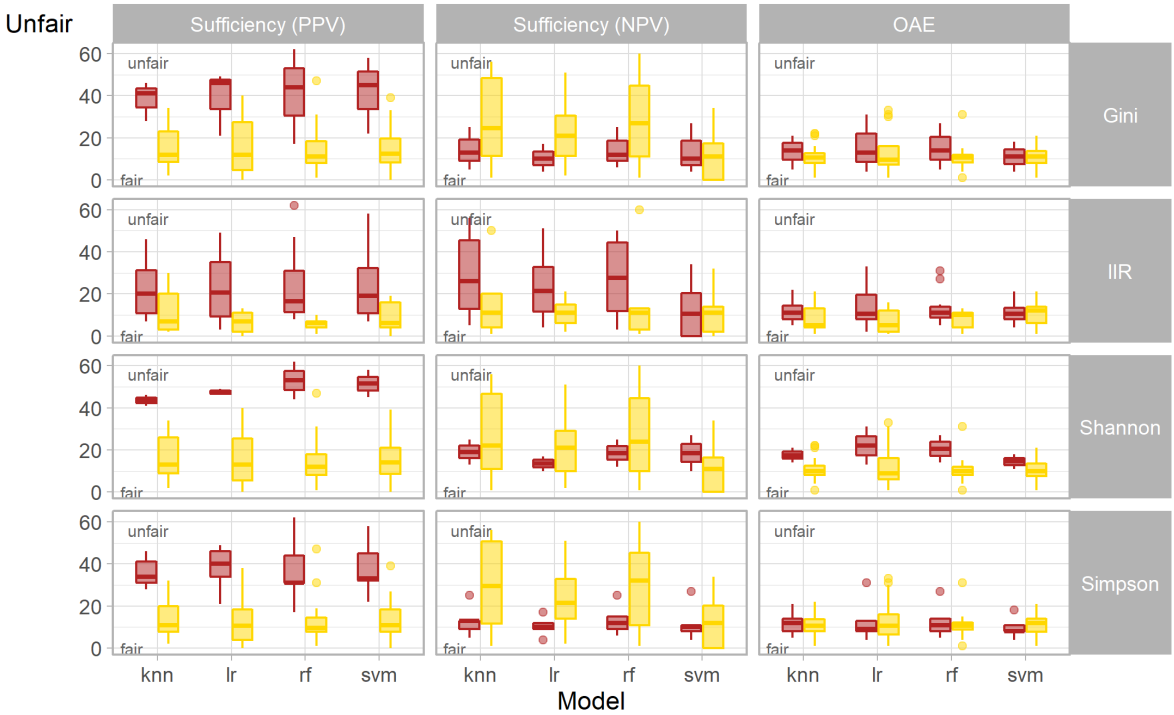


Figure 8.4. Models PPV, NPV, OAE distribution, threshold 33%

Generally when distribution discriminative capability is within the reach of all models, RF and LR are the ones getting higher performances. It's unlikely that hard problems (for which unfairness discrimination is difficult) are saved by a specific model (examples for RF: Independence-Gini, IIR-PPV, IIR-NPV).

Previous consideration on Shannon and IIR are robust among different models: The first performs well on TPR and PPV, the second on FPR and

NPV (it still remains the only one able to discriminate on these two unfairness measures). However. among NPV-IIR SVM performs worse than others.

The exclusivity of Shannon on OAE is confirmed over the four models.

8.1.3 Boxplots - Comparison between baseline and smote

In this sub section we aim to see if there are differences in unfairness distribution (between the two levels of risk) among baseline and smote models. The grid has all six Unfairness measures over the rows and the four balance indexes over the columns. Each subplot has the same structure as before, with the difference that along the x axis there are the two type of training data being used: baseline and oversampled (smote). Following we show only grids for Logistic Regression and Random Forest.

In this first grid no differences emerge between the two variants. There are not cases for which a baseline overlapping is a smote polarization or the other way around. Only for TPR-Simpson, smote seems to favour discrimination.

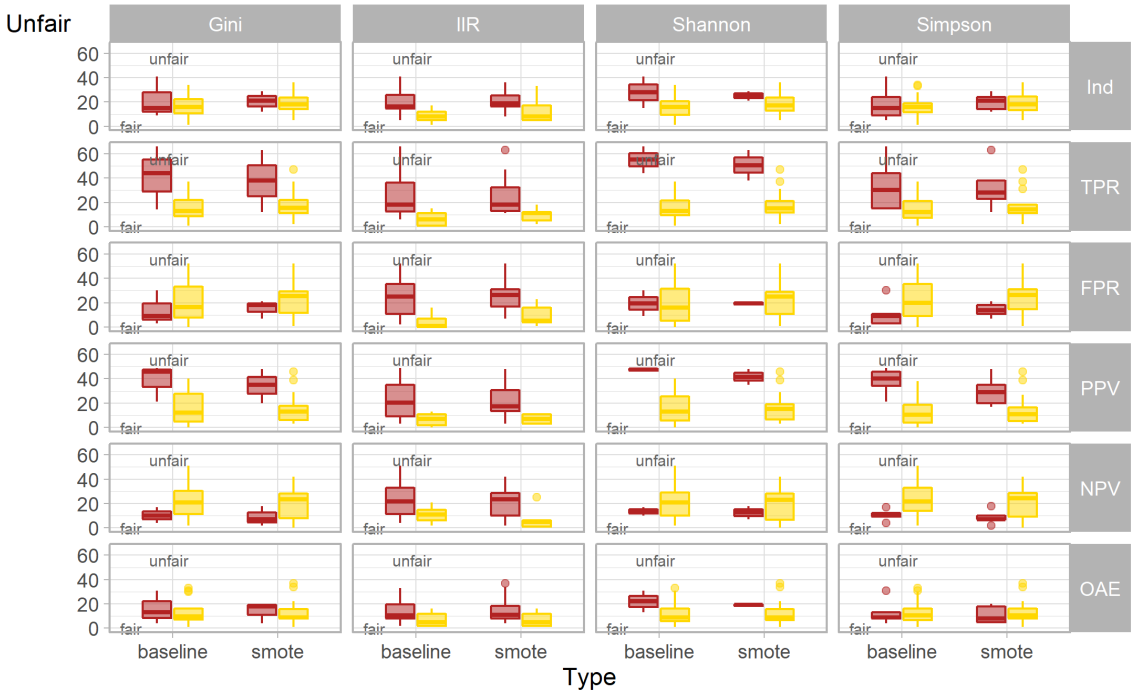


Figure 8.5. LR: baseline-smote comparison

For RF it is generally the same as for LR with the notable difference that

there are some cases for which smote weakens discrimination capabilities of the algorithms. Some examples are: Ind-Gini, OAE-Gini, FPR-IIR, NPV-IIR, Ind-Shannon.

There is more RF confidence in balance-fair relationship if smote is not applied.

Conversely there are some cases for SVM in which smote makes discrimination capabilities higher, leading to greater polarization.

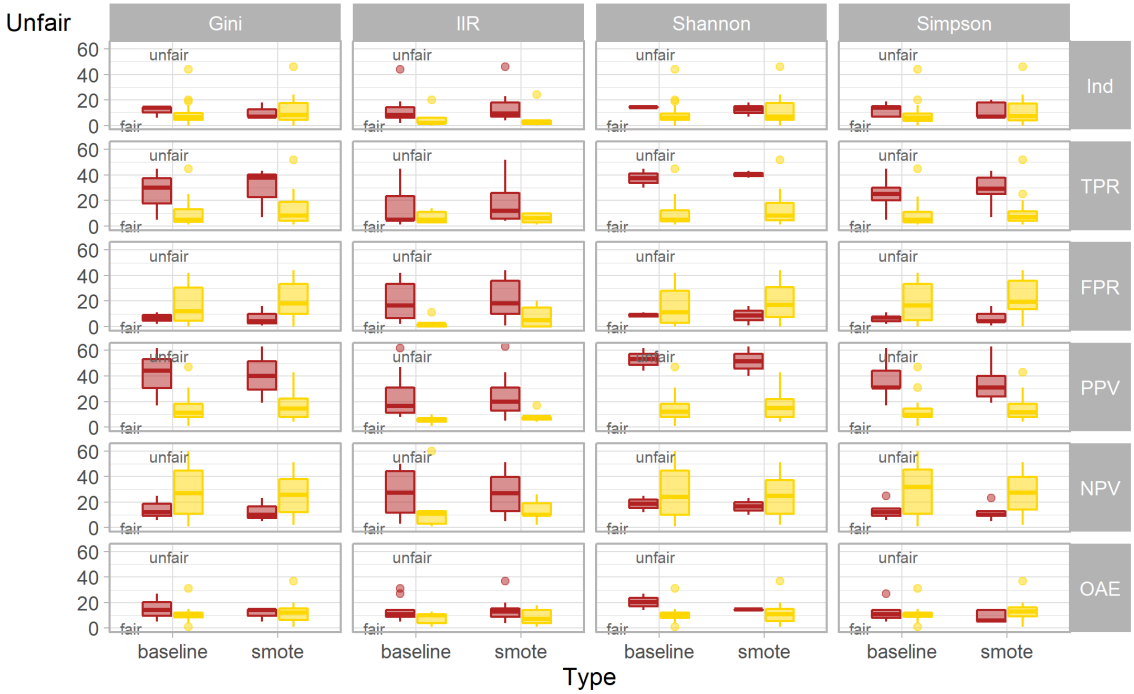


Figure 8.6. RF: baseline-smote comparison

8.1.4 Boxplots - Risks threshold comparison

Until now only the threshold 33% has been explored. Here we aim to see if by varying this value, the discrimination capabilities of some indices improve or not. The two following grids report distribution of SVM and RF models. As above on the y axis there is the unfairness value, over the x axis the four values of the threshold (%): 25,33,50,75.

Threshold comparison for SVM model:

- Gini: As assumed in the base paper, it is worth trying lower thresholds. In fact it seems to benefit, especially for TPR and PPV, of 25% value. It's interesting that also unfairness measures for which there were no discrimination capabilities now show polarized distribution. It is the case of NPV and OAE.
- IIR: There are no evident advantages on using a different threshold from 33%
- Shannon: As for Gini, no improvements on Independence and FPR. Same better discrimination on TPR and PPV, mitigated differences on NPV and OAE.
- Simpson: There seems to be no notable difference among the four thresholds.

In general we can say that no measure is favoured by high value of the threshold (75%). Only IIR in TPR and PPV, appears to gain some discrimination with a 50% threshold.

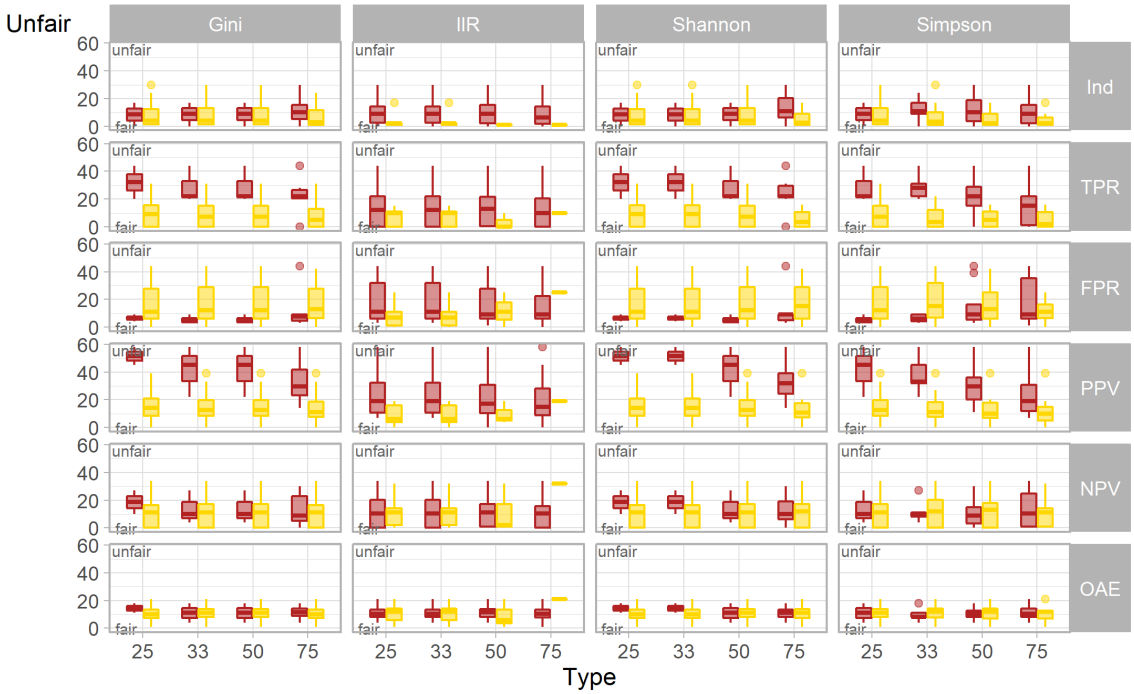


Figure 8.7. SVM: thresholds comparison

Same considerations can be made for RF model. Gini and Shannon exhibit higher polarization of unfairness distribution for lower value of the threshold

(25%), especially the first. We see again that improvements over the RF model seems to be higher than others. It will be confirmed in the 'Correlation' subsection.

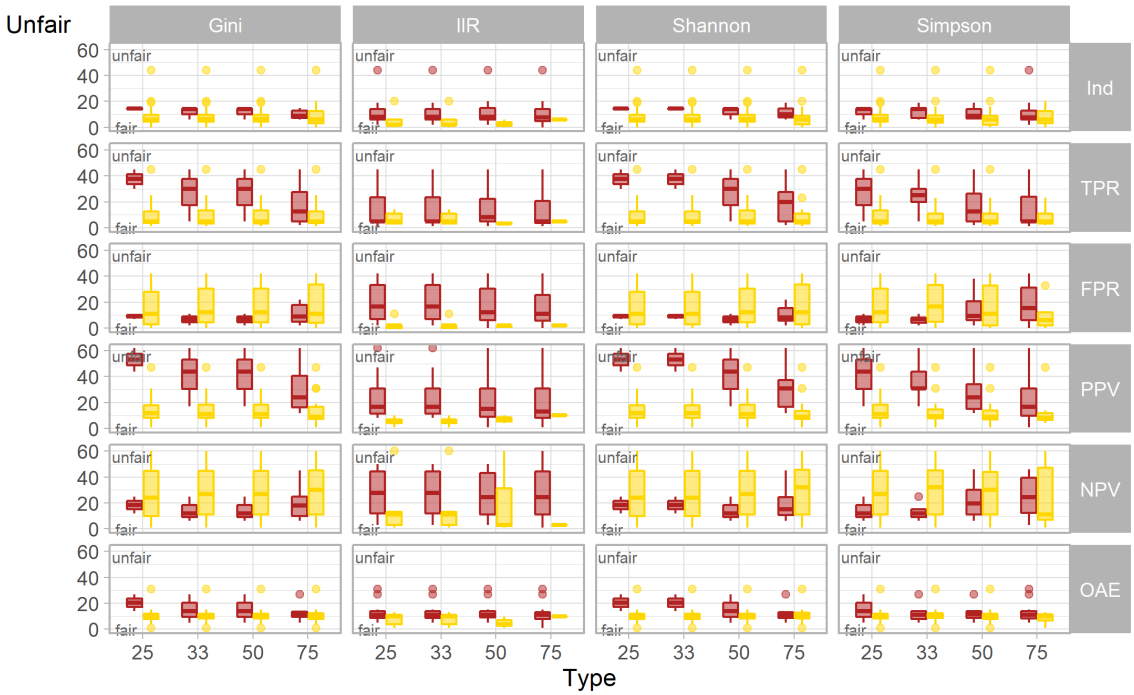


Figure 8.8. RF: thresholds comparison

It's interesting to notice that, as regards 25% and 33%, RF succeed in NPV-IIR but fails in NPV-Shannon while SVM the opposite. Further models (as the KNN and LR) and thresholds can be evaluated for a deeper exploration.

8.1.5 Correlation

According to [31], the Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of $+1$ or -1 occurs when each of the variables is a perfect monotone function of the other. Intuitively, the Spearman correlation between two variables will be high when observations have a similar (or identical for a correlation of 1) rank (i.e. relative position label of the observations within the variable: 1st, 2nd, 3rd,

etc.) between the two variables, and low when observations have a dissimilar (or fully opposed for a correlation of 1) rank between the two variables. Spearman's coefficient is appropriate for both continuous and discrete ordinal variables.

$$\rho_s = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2} \sqrt{\sum_i (r_i - \bar{r})^2}}$$

r_i and s_i are respectively the rank of the first and second variable of the i-th observation.

For each pair balance-unfairness, the average and standard deviation of the correlation over 5 runs is shown.

	Gini	Shannon	Simpson	IIR
Ind	-0.15(0.05)	-0.21(0.03)	-0.29(0.02)	-0.44(0.05)
TPR	-0.4(0.09)	-0.52(0.05)	-0.55(0.05)	-0.67(0.04)
FPR	0.16(0.08)	0.04(0.08)	-0.06(0.06)	-0.18(0.07)

Table 8.3. LR model: correlations in the format mean(std)

With respect to base paper results these are mitigated, there are negative correlations but with a little lower absolute values. Still, some aspects are in common: for all three unfairness measures, negative correlation gets bigger from Gini to IIR. So Inverse Imbalance Ratio still results being the index which better, on average, detect discrimination in real cases (being aware of its limitation if one class is empty).

Differences in estimates, is probably given by the different framework used to make Logistic regression classification: python Scikit-learn has been used instead of R. Hence, it resulted in a distinct hyperparameters optimization and final configuration.

models comparisons

Since other three models have been used as inference algorithms, it could be interesting evaluating how imbalance-unfairness relationship varies. Balance measures are applied ex-ante on trainset, so they do not depend on the model being used. Conversely, unfairness measures are computed on testset using model predictions. Therefore unfairness and correlation are likely to

be different among LR, SVM, KNN, RF.

In comparing models, Compas, Juveline datasets are excluded since their scores are pre-computed from a black box algorithm. As for now, the retained datasets are: dccc, statlog, income, student math, student port.

	Gini	Shannon	Simpson	IIR
Ind	-0.21(0.08)	-0.29(0.07)	-0.32(0.03)	-0.36(0.06)
TPR	-0.47(0.13)	-0.58(0.06)	-0.63(0.06)	-0.58(0.08)
FPR	0.14(0.11)	-0.02(0.11)	-0.08(0.09)	-0.06(0.11)

Table 8.4. LR model w/o Compas and Juvenile: correlations in the format mean(std)

By comparing LR correlations with/without Compas and Juvenile, we can notice that in the second case the standard deviation associated to the estimation is higher. It could be interesting to assess if this variability is given by the lower number of datasets used or because the two black box algorithm are associated to high confidence predictions and consequently more accurate fairness measures. However differences are not big, but sistematic.

For each bal-fair pair and model, mean spearman correlation over the five runs is plotted. Correlation values have been multiplied by -1 to facilitate graphs interpretability.

For each unfairness measure (row of the graph) we can identify the best algorithm, in term of correlation, based on the four analyzed indices.

- In the first row, as regards Independence measure, we can see that on average SVM (generally followed by RF) achieves higher level of correlation over the four balance indices. Moreover this model is associated to low variance along the same fair index.

- For True positive Rates, the four models are pretty balanced and no one outperforms others.

- False Positive Rates is the most difficult to predict given its low correlation with the four indices. Generally the one performing best is the RF model.

Concerning False Positive Rate differences, only IIR achieves negative correlation over all four models. This makes this imbalance measure robust (on avarage) to the use of more models (in fact for all three fair indices, is the only one always getting negative desired correlation).

Looking at models performances in terms of correlation, we show the number of top2 placements (first or second highest negative correlation among the four models and over the four balance indexes of a given Unfairness measure) of models for each of the Unfairness measures: Independence, Separation, Sufficiency, Overall Accuracy Equality. As concern Independence, SVM and Random Forest are the algorithm achieving a good number of top2. in particular, SVM places always first and RF second. In this sense LR and KNN have systematically lower correlation. In terms of Separation, KNN and RF are good. Conversely, SVM does not reach top2 position over none of the balance indexes.

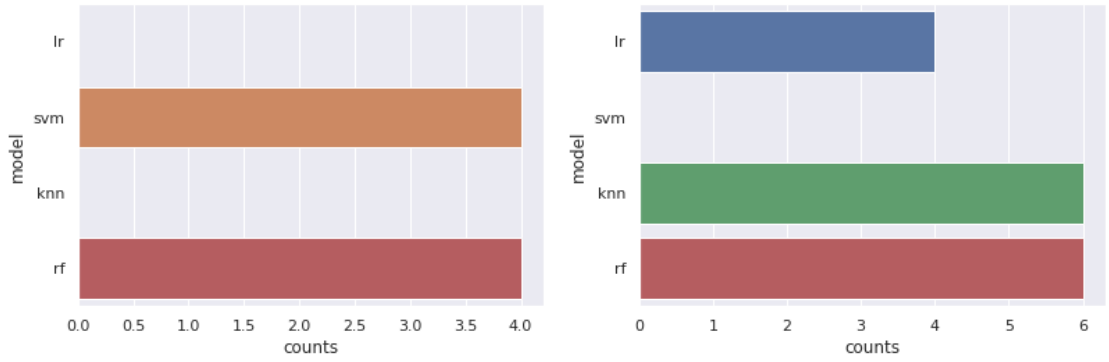
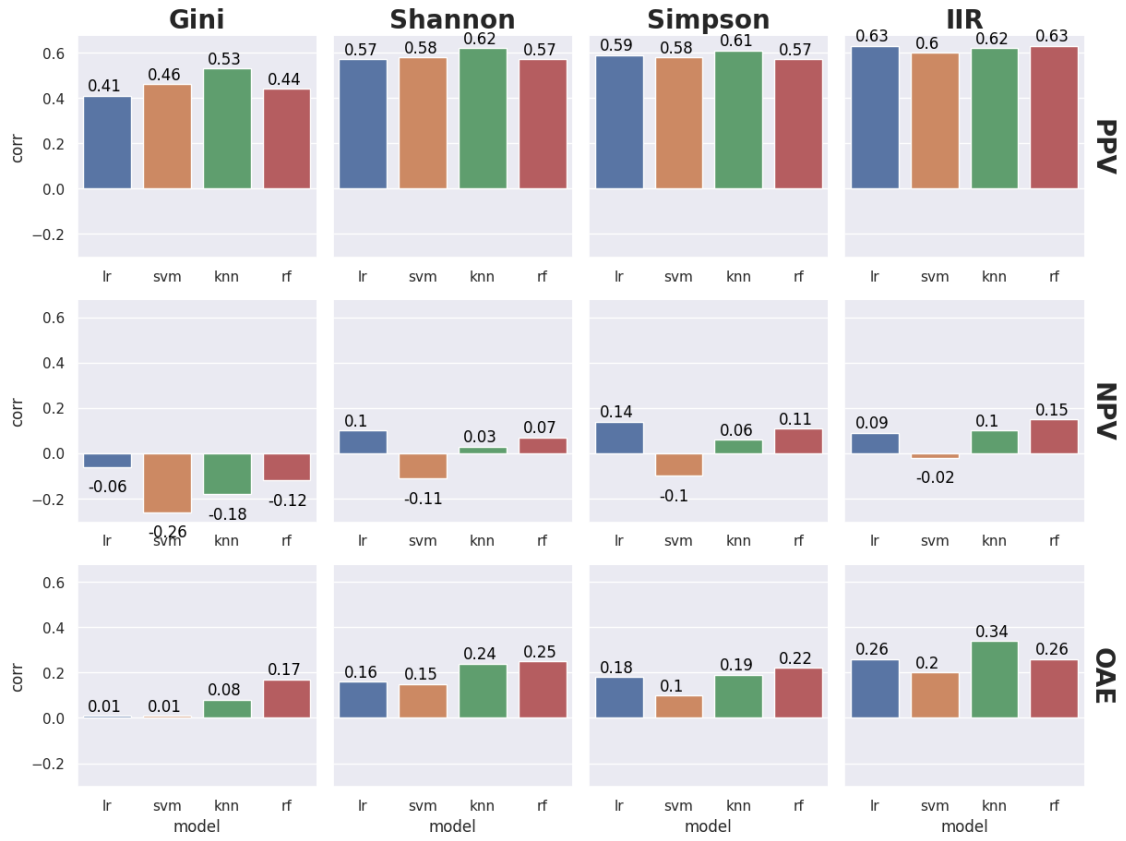


Figure 8.9. Left: Independence top2 placements. Right: Separation top2 placements.

We now compare imbalance-unfairness model correlations by taking into consideration the three unfairness criteria introduced: Sufficiency (PPV and NPV) and Overall Accuracy Equality. However the attributes being considered are still base paper ones.

Among the three new unfairness measures, PPV is the one whose values of correlation are satisfying for all models and indices. Correlation with OAE seems more easily achievable especially for KNN and RF. These two also seems better than others as concern NPV.

The general robustness to different models and the good behaviour of IIR is here confirmed with higher values compared to other indices



In terms of Sufficiency top2 placements, SVM is the model which get fewer victories. The same goes for OAE. In both cases RF and KNN get higher number of victories.

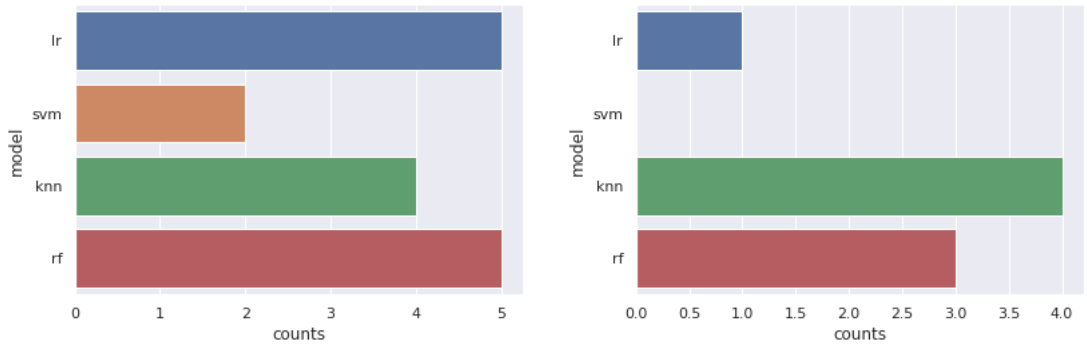


Figure 8.10. Left: Sufficiency top2 placements. Right: OAE top2 placements.

Taking into consideration all unfairness measures, KNN and RF are the ones which get more top2 victories, thus higher negative correlations.

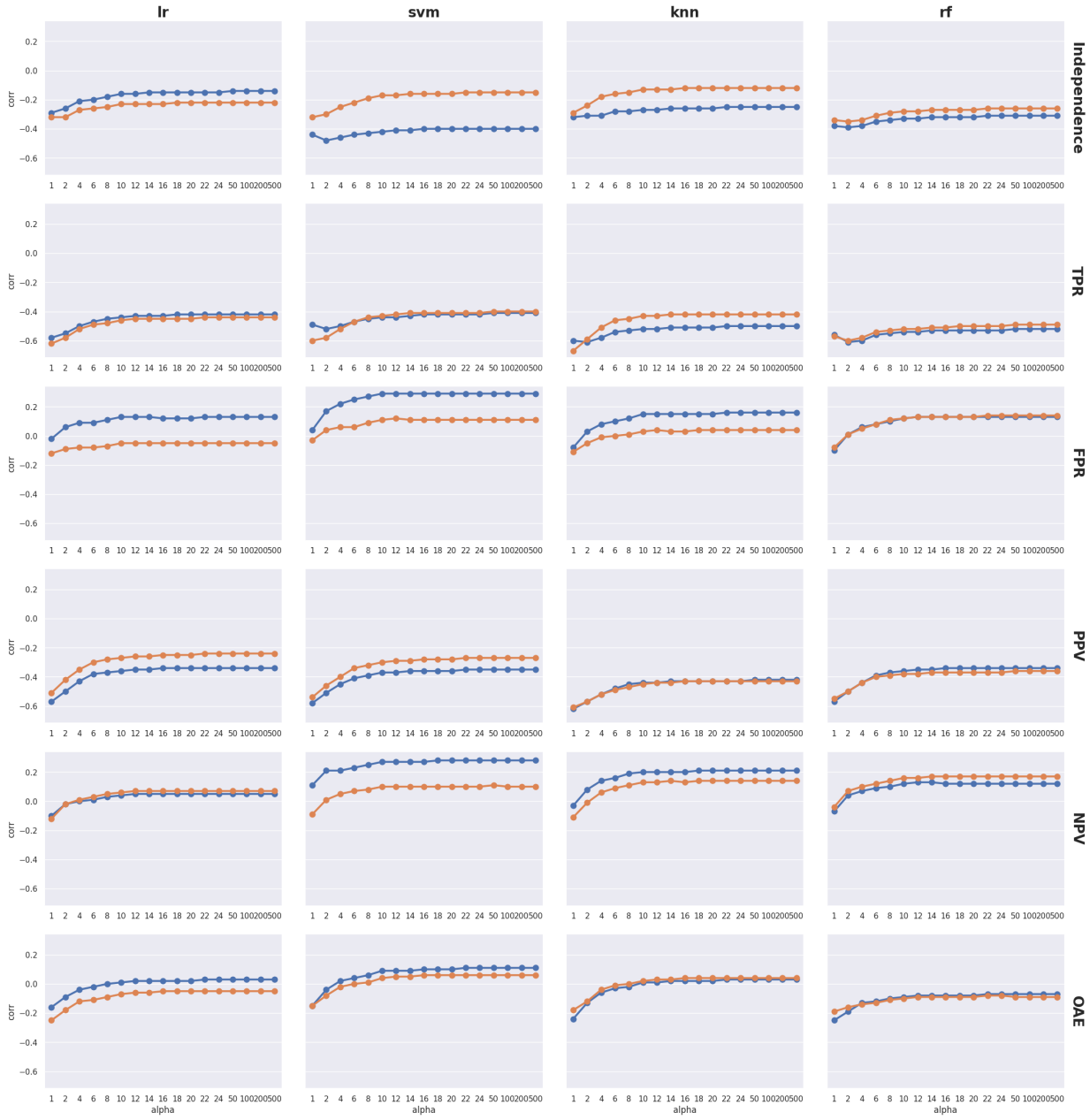
Renyi and Hill

In this paragraph we explore mean correlation of Renyi and Hill indices as a function of their parameters α and q. For each sub plots baseline(blue) and smote(orange) curves are compared.

The following picture report the plots for Renyi index. Same consideration can be made for Hill.

Generally we see correlation as an increasing monotone function of alpha. In some unusual case, for TPR and Independence, the correlation is lower for $\alpha = 2$. For all other measures it does not happen. Moreover smote curves are under the baseline's manly for False Positive Rate with 3 out of 4 more evident differences. RF curves differences are the lowest; generally they overlap.

It could be interesting to further study these indices in terms of previously discussed box-plots.



8.2 extensive analysis

The base paper analysis is here extended by adding further datasets and attributes. The purpose is to check if previous considerations apply here as well.
Additional datasets: Term-deposit, Titanic ;

Added attributes: Marriage of dccc.

Dataset -	Attribute	m	Gini	Shannon	Simpson	IIR
dccc -	Marriage	4	0.6804	0.5455	0.3474	0.0036
term deposit -	education	8	0.9226	0.8531	0.5983	0.0013
term deposit -	marriage	4	0.7218	0.6612	0.3934	0.0032
Titanic -	Sex	2	0.8945	0.9225	0.8092	0.5097

Table 8.5. Additional datasets and attributes

8.2.1 Boxplots - Models comparison

unfairness distribution for the two levels of risk (33%) are now compared in the extensive scenario.

- Independence: RF remains the best in terms of discrimination capabilities over the four indices;
- TPR: Also here LR seems pretty good;
- FPR: Discrimitation capabilities on this unfairness measure are still difficult to get. Moreover, for IIR index, only RF remained robust to dataset extension; other models seem to be penalized.

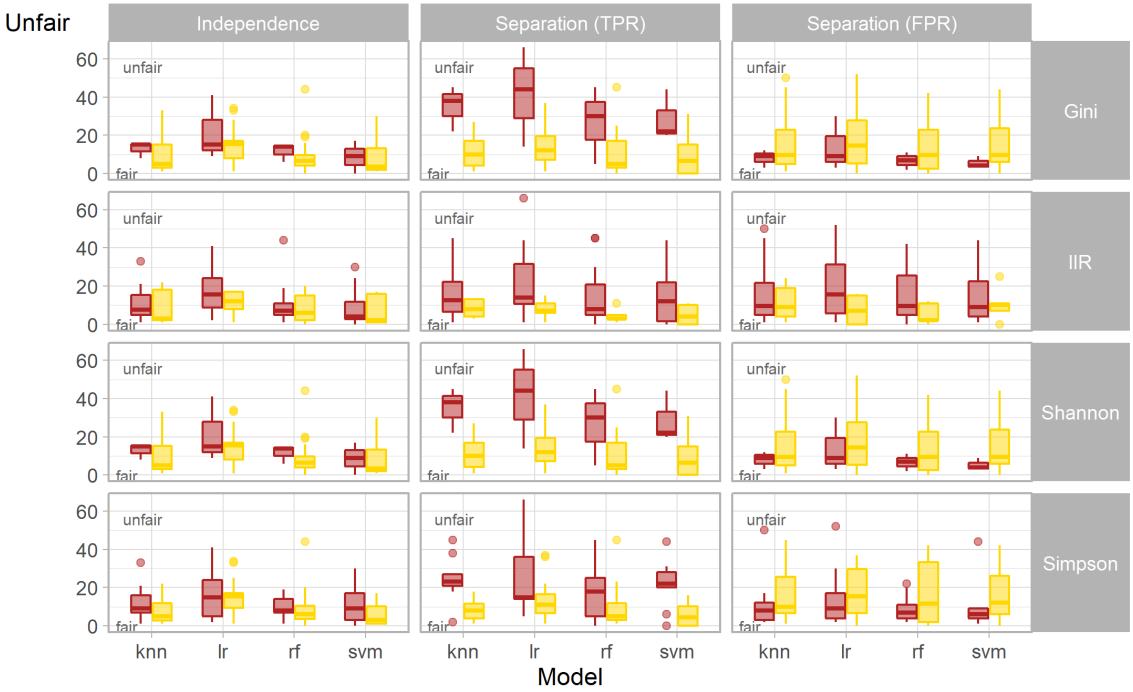


Figure 8.11. Extensive: models Ind, TPR, FPR distribution, threshold 33%

- PPV: all models still performs similar given the imbalance index. Shannon and Gini seem to perform best;
- NPV: very hard to discriminate unfairness differences for this measure. Discrimination capability of IIR is not confirmed for none of the models;
- OAE: Polarization has become more difficult.

Figure 8.12. Extensive: models PPV, NPV, OAE distribution, threshold 33%

8.2.2 Boxplots - Risks threshold comparison

As regards SVM, Gini and Shannon indices are again favoured by lower threshold (25%) especially for TPR and PPV unfairness measures. Still for IIR there are not notable differences in using different value of the risk levels choice. Simpson behaves better with the starting threshold of 33%. However, high values of the threshold do not lead to farther distribution among the two levels of risk.

By looking at RF, we can see that IIR is not able to enforce discrimination in terms of NPV distribution for none of the thresholds. But the same index remains able to discriminate among FPR. Other considerations for RF are still valid.

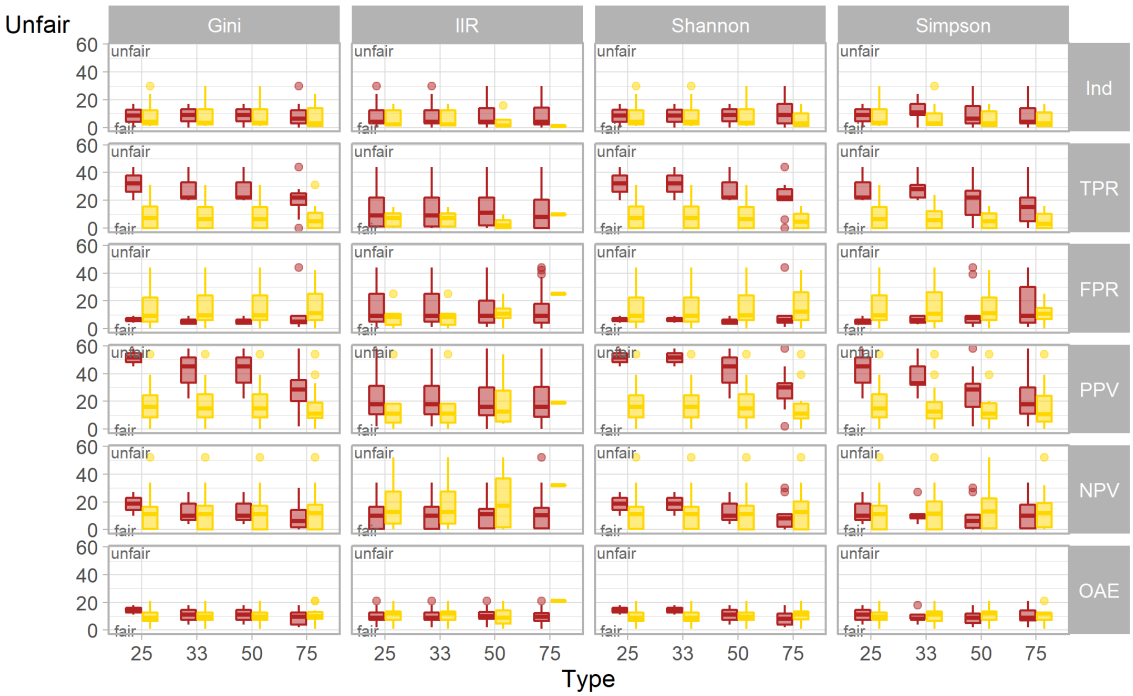


Figure 8.13. SVM: thresholds comparison

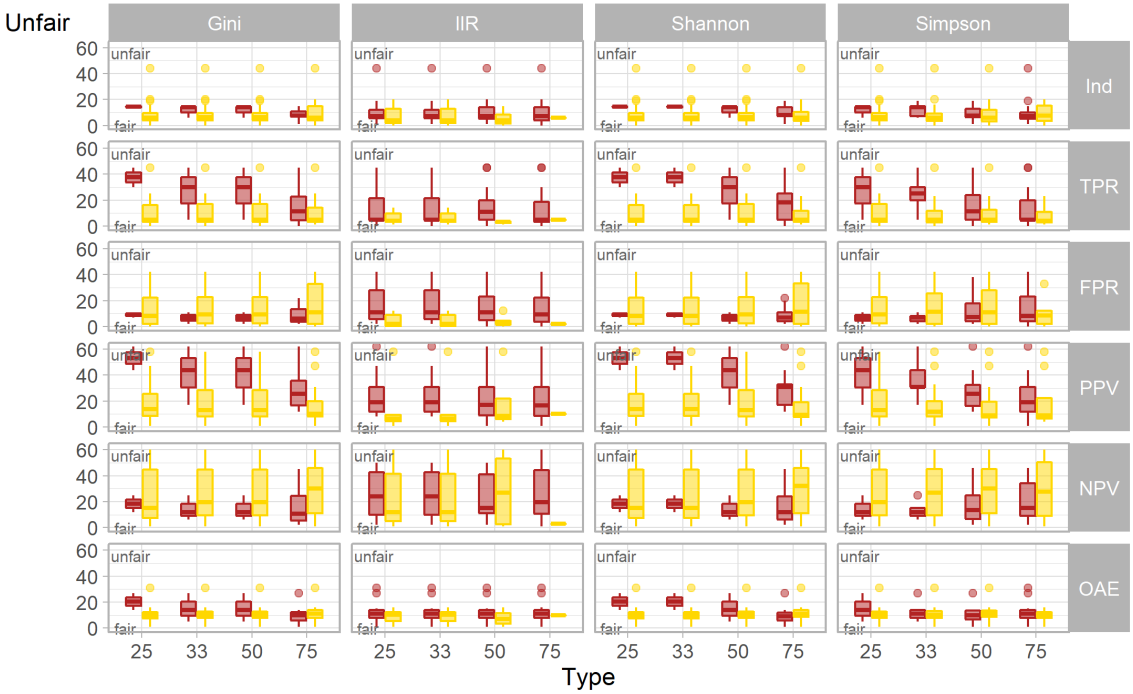


Figure 8.14. RF: thresholds comparison

8.2.3 Correlation

Same type of observations, in terms of mean correlation over the five runs, can be made in the extensive scenario.

Again, for each unfairness index (row) we analyze which aspects come out from the grid:

- Independence: All values of correlations are lower. The model which retained higher values is SVM. The more penalized, after other datasets introduction, is LR;
- TPR: It is more robust to Attributes introduction: values are still very high. In particular KNN is now the best correlated index over the imbalance ones;
- FPR: lower correlations than before. IIR does not retain all negative correlations over the models; only RF is able to do so (in small amount) among the Shannon, Simpson, IIR.

As regards independence, top2 placements distribution among model is the of the base paper one. Still SVM gets all first places, while RF second ones. With respect to Separation, again good performances of KNN and RF over the other 2 models.

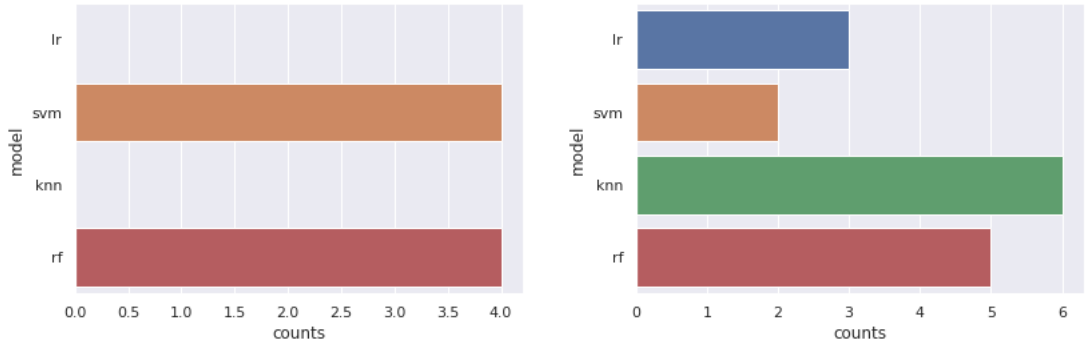
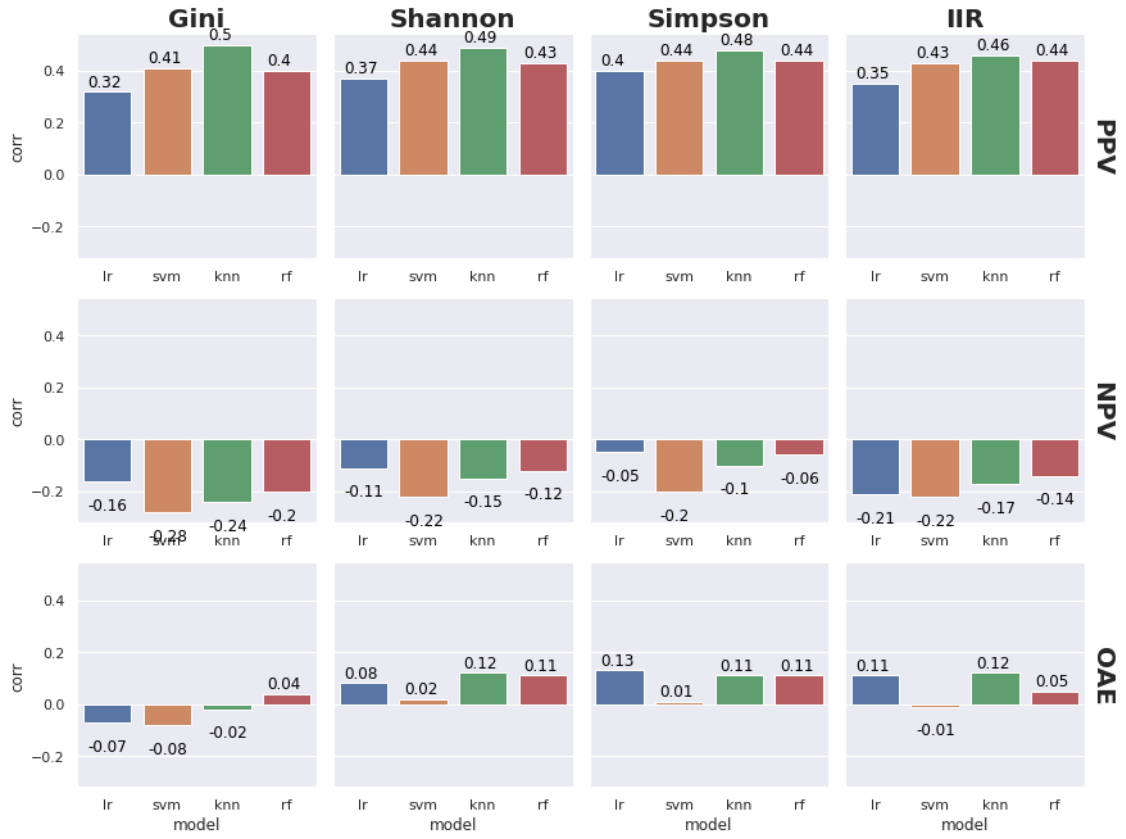


Figure 8.15. Left: Independence top2 placements. Right: Separation top2 placements.

Moving to the other three unfairness indices, General values of correlations are lower than base paper case.

- PPV: Values are still high. Now KNN is the model with higher results over all imbalance indices;
- NPV: No model or imbalance index is able to reach negative correlation. The model which mitigates more the phenomenon is RF. It is here confirmed that IIR is not able anymore of anticipating NPV discrimination given the positive value of correlation;
- OAE: Shannon, Simpson and IIR are the indices which general negative correlations are higher.



RF and KNN are confirmed to be the models better performing in terms of both Sufficiency and OAE over the four balance indexes considered. Again in terms of OAE top2 placements, SVM gets no first or second places, suggesting it is the model reaching lower performances (in OAE).

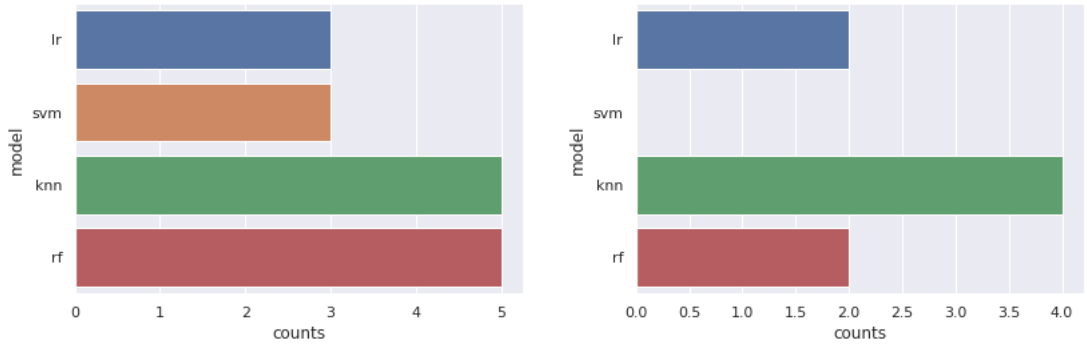


Figure 8.16. Left: Sufficiency top2 placements. Right: OAE top2 placements.

Chapter 9

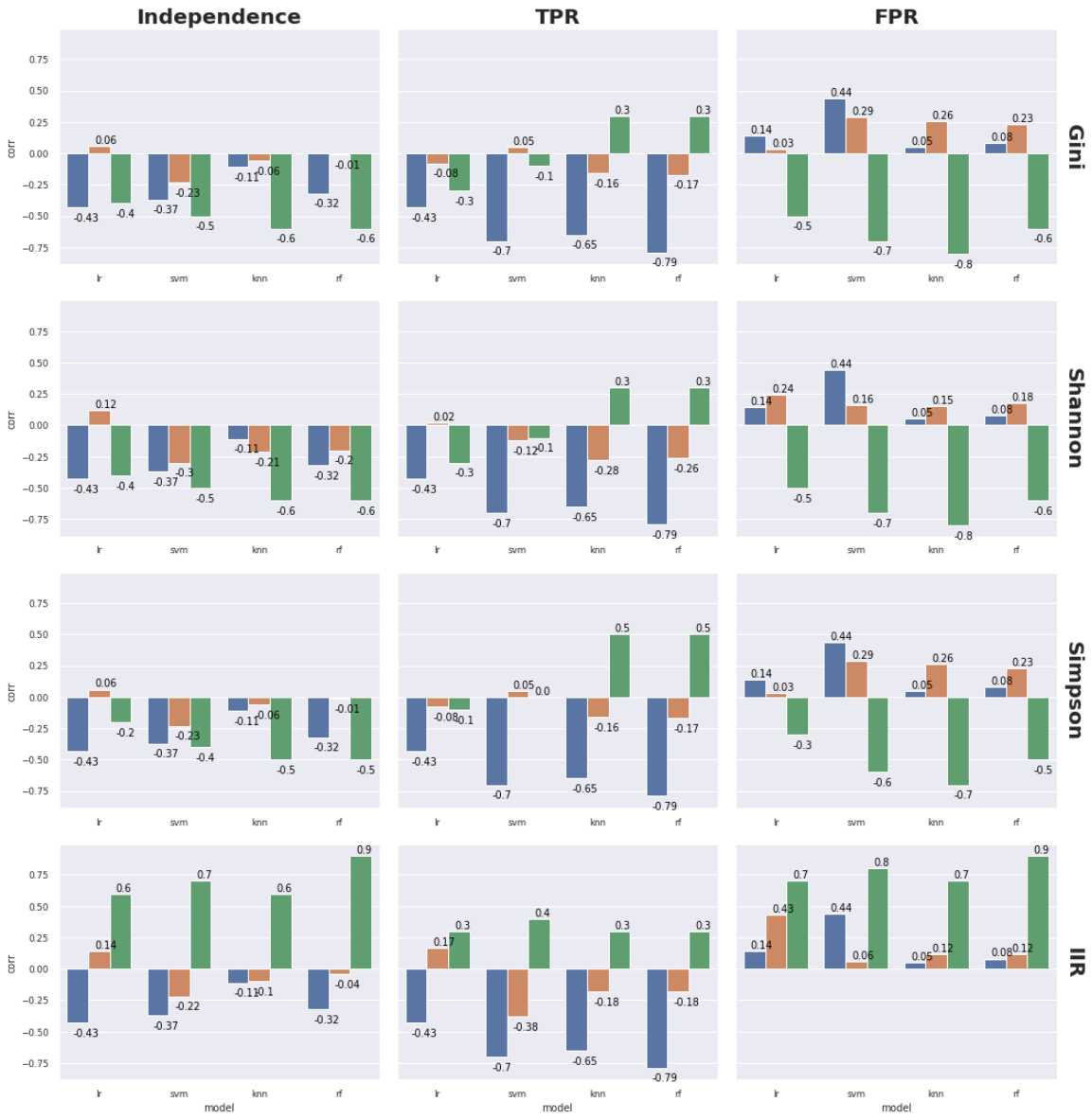
Cardinality comparison

Further investigation of correlation between imbalance and unfairness indices can be done by splitting attributes by their cardinality, as it was done for unfairness. Thus, the cardinalities taken into consideration are m = 2,5,8, still being aware of the different support sizes (7,9,3).

We start by looking at plots for the following imbalance Indices: Gini, Shannon, Simpson, IIR. Cardinality reflections have been splitted in two couples 2-5 and 5-8. Each plot in the grids stands for a combination of imbalance index and unfairness one among all the ones studied in this thesis. Unfairness measures have been divided in two groups of three measure:

- Independence, Separation;
- Sufficiency, OAE.

It's interesting to notice that, in case of a dichotomous attribute, Spearman correlation of an unfairness measure is constant over all the imbalance indices because of the rank computation behind correlation. In fact in case of linear correlation (Pearson), it changes depending on the pair unfairness - imbalance. Graphically it can be noticed by all blue bars in a column (for a given unfairness index) be the same.



pair 2-5

For independence, for all models, m=2 results in higher negative correlation with respect to m=5. The same happens for TPR where high level of correlation are reached for m=2 (-0.6,-0.8).

FPR is the measure whose correlation with analyzed indices is difficult to get: none of the algorithms and cardinality go below zero.

As concern Independence, LR is the only algorithm which, for m=5, does not go below zero for none of indices. In this sense, svm seems to work better. Generally the models which catch negative correlation over different imbalance indices are preferred. And talking about TPR, KNN and RF are

the one which reach always negative values for both levels of cardinality.

pair 5-8

Introducing m=8, it's interesting to see that correlations for this cardinality maintain similar values over first three indices with low fluctuations. Differently, values change in correspondence of IIR, for which there is an overturning of the bars. So for m=8 IIR fails in catching negative correlation with Independence, TPR, FPR.
For Independence and FPR, Gini, Shannon, Simpson, negative correlation of m=8 is higher than m=5.
Generally, for m=8 and all models, correlation between Gini, Shannon, Simpson and Independence and FPR is convincing. Until now m=8 is the only cardinality for which negative correlation on FPR is reached.

pair 2-5

m=2 does not reach levels of correlation as for Independence and TPR. For PPV it still maintains good values.

- PPV: m=2 negative correlation generally higher than m=5. Only IIR achieves negative correlations for all models;
- NPV: No indices or model is good for the two cardinalities considered;
- OAE: m=2 negative correlation generally higher than m=5.

cardinality 8

For m=8 high value of predictability are achieved. Still the same behaviour as before with the IIR is present: bars overturn to positive correlation for IIR. In this sense we can see that IIR, for m=8, does not reach negative value for none of the unfairness measures and models studied. However, for other imbalance indices, the correlation can also reach -0.8, with general satisfying results for all three unfairness indices.
Maybe it is temptative to prefer other indices instead of IIR when the cardinality of attributes is 8.

Part IV

Final Part

Chapter 10

Conclusions and Future works

Following we report the major observations retrieved.

10.1 Imbalance measures

Gini and Shannon generally bring to lower penalization, independently on attribute cardinality or the type of disproportion. They are followed by Simpson and IIR. Renyi and Hill are able to approach a tighter penalization (Hill(q) $<$ Renyi(α) when q=α); in particular Hill is able to reach IIR ranges. The extent to which the Hill and IIR ranges are near only depends on the attribute disproportion, disregarding its cardinality. We can say that, in case of high disproportion, Hill has an harder time in approaching IIR.

10.2 Unfairness measures

The following remarks concern the confirmations on both levels of study: Base paper and Extensive.

Independently on the algorithm used to train the model, on average, baseline independence and Separation are lower than smote ones. Smote, instead, mitigates Sufficiency unfairness. There are no notable differences in terms of Overall Accuracy Equality.

At both levels of study, we also see that Random Forest is the model which unfairness differences between baseline and smote are lower. As regards Independence and Separation, SVM is the model which baseline fairness gap from smote is wider. Logistic regression is the model which baseline Independence and TPR are higher than other three models. KNN strongly penalizes baseline NPV unfairness.

With respect to unfairness values by attribute cardinality we see that, except for Independence and TPR, gaps in unfairness between cardinalities (2,5,8) do not depends on the model being considered. As regards the two measures mentioned, we notice that what stated at aggregated level is confirmed: Logistic regression is more unfairer for all three cardinalities reported. With respect to the data and unfairness measures used, $m = 8$ results on unfairer measures than $m = 2$ (remembering the difference in support between the two).

10.3 Indexes relationship

As regards unfairness distribution (using Logistic Regression) for the two levels of risk (threshold=33%), base paper results on Independence and Separation are confirmed at section 8.1.1: IIR is the index resulting in better discrimination capabilities. By introducing the other three unfairness measures (Sufficiency and OAE), only IIR is able to detect NPV differences in distribution among the two levels. Instead Shannon is the only one capable of it for OAE. The only measure for which IIR fails is OAE. While the second 'best' measure, Shannon, fails in FPR and NPV (for which only IIR gets desired results).

Previous consideration on Shannon and IIR are robust among different models: The first performs well on TPR and PPV (confirmed in the extensive analysis), the second on FPR and NPV (it still remains the only one able to discriminate on these two unfairness measures). This second case is not robust in the extensive scenario, where only for FPR, the RF was capable of separating distribution. The exclusivity of Shannon on OAE is confirmed over the four models and the extensive assessment. For Independence, OAE and some particular pair balance-unfairness indexes, Random Forest seems to be the model providing better distribution polarization.

With respect to correlation between balance and unfairness indexes, Logistic Regression model in base paper scenario confirms that negative correlation gets bigger in the order Gini, Shannon, Simpson, IIR. So Inverse Imbalance Ratio still results being the index which better, on average, detect discrimination in real cases.

At both levels of study SVM is the model which correlations with independence is higher over the four balance indexes. Conversely the same model does not perform as well as the others for OAE unfairness. As regards TPR and PPV (especially the second), KNN is the model which gets higher correlations with respect to other four models. In general, Independence, TPR, PPV are the easiest to correlate with. As concern Independence and Separation, Random Forest is the model achieving the highest number of top2 placement in terms of negative correlations among the balance indexes. As regards Sufficiency and OAE, KNN (and then RF) achieves highest number of top2 placements.

Generally we can say that high values of negative correlation are reached for TPR, PPV and moderately high for Independence. On the opposite, correlating with FPR and NPV is very difficult and moderately difficult in the case of OAE.

10.3.1 Risk thresholds

At both levels of assessment and models considered (SVM and RF), Gini and Shannon benefit of a lower threshold in terms of Separation TPR, Sufficiency PPV, and Overall Accuracy Equality. No measure is favoured by high values of the thresholds.

10.3.2 Comparison with smote

Among the two models considered here, LR and RF, there are not notable differences in unfairness distributions, on the two risk levels, between baseline and smote. For RF there are some cases (around 6 out of 24) for which baseline shows better polarization than smote (But it is not something specific of a given unfairness index or balance one).

10.3.3 Renyi and Hill

Correlation is an increasing function of α and q. Except for some cases, baseline and smote trends have similar values; especially in correspondence of RF.

10.3.4 Correlation by attribute cardinality

For $m = 8$ negative correlation on Gini, Shannon and Simpson comes to be positive in correspondence of IIR (it happens for all unfairness indexes). Still $m = 8$ is the only cardinality which reaches negative correlation with Separation FPR and Sufficiency NPV (with the exception of IIR). Especially for Independence, TPR, PPV, $m = 2$ negative correlation is higher than $m = 5$.

10.4 Future works

The study carried out can be extended by adding other datasets (consequently attributes), algorithms, balance measures or unfairness measures among the proposed ones in section 5.5. 'Community and crimes' dataset was left out since no smote version of it was available; it can be used as additional dataset for further analysis. Where boxplot differences among smote and thresholds have been performed, only 2 models grids out of 4 have been shown to not load too images and get confused. It could be interesting to evaluate smote differences from baseline also in terms of Spearman correlation between indexes.

Renyi and Hill indexes have been investigated only in terms of correlation with unfairness measures as a function of their parameters. Further analysis can be done by extending unfairness distribution box plots also to Renyi and Hill for some values of their parameters to see if the polarization gets lower or bigger.

The chapter 9 'Cardinality comparison' can be extended by cardinalities being considered and type of investigation performed: insertion of box-plot unfairness distribution between levels of risk over the chosen cardinalities.

Continuous attributes indexes proposed in section 4.5 were not used since a problem emerged later: computing unfairness measures with respect to

continuous attributes. Hence, we were not able to compare unfairness distribution and correlation among categorical balance measures and continuous ones of specific attributes as age. However, the proposed balance indexes are still valid and can be used as source material for further investigations.

Bibliography

[1] *Vetrò, A., Torchiano, M., Mecati, M. (2021). A data quality approach to the identification of discrimination risk in automated decision making systems. Government Information Quarterly*, https://doi.org/10.1016/j.giq.2021.101619.

[2] DANA PESSACH, EREZ SHMUELI , https://arxiv.org/pdf/2001.09784.pdf.

[3] Wanwan Zheng e Mingzhe Jin *The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study.*

[4] Florian Privé , https://privefl.github.io/R-presentation/penalised-genetics.html#32, page 16/31.

[5] Stratified cross validation, https://stats.stackexchange.com/questions/49540/understanding-stratified-cross-validation.

[6] Nitesh V. Chawla, et al , https://arxiv.org/pdf/1106.1813.pdf.

[7] smote image, https://www.lorenzogovoni.com/algoritmo-smote/.

[8] lr image, https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/.

[9] svm image, https://link.springer.com/article/10.1007/s42979-019-0006-z.

[10] knn image, https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4.

[11] rf image, https://deepai.org/machine-learning-glossary-and-terms/random-forest.

[12] default credit card dataset, https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

[13] German Credit Data, https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data).

[14] Adult Data Set, https://archive.ics.uci.edu/ml/datasets/adult.

[15] Bank Marketing Data Set, https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

[16] Student Performance Data Set, https://archive.ics.uci.edu/ml/datasets/Student+Performance.
[17] TitanicSexism Data Set, https://www.kaggle.com/garethjns/titanicsexism-fairness-in-ml.
[18] Communities and Crime Data Set, https://archive.ics.uci.edu/ml/datasets/communities+and+crime.
[19] Compas Data Set, https://github.com/propublica/compasanalysis/blob/master/compas-scores-two-years.csv(2016).
[20] Juvenile Data Set, http://cejfe.gencat.cat/en/recerca/opendata/jjuvenil/reincidencia-justicia-menors/index.html(2016).
[21] Renyi entropy, https://en.wikipedia.org/wiki/R%C3%A9nyi_entropy.
[22] Hill index, https://en.wikipedia.org/wiki/Diversity_index.
[23] Generalized entropy index, https://en.wikipedia.org/wiki/Generalized_entropy_index.
[24] Theil index, https://en.wikipedia.org/wiki/Theil_index.
[25] Gini coefficient, https://en.wikipedia.org/wiki/Gini_coefficient.
[26] Pietra-Ricci coefficient, https://dipartimenti.unicatt.it/scienze-statistiche-SepDip8.pdf.
[27] Confusion matrix image, https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781838555078/6/ch06lvl1sec34/confusion-matrix.
[28] Confusion Matrix description, https://fairware.cs.umass.edu/papers/Verma.pdf.
[29] Julia Angwin, Jeff Larson et al, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[30] Classification chapter of fairmlbook, https://fairmlbook.org/classification.html.
[31] Spearman's rank correlation coefficient, https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.