Received: 27 May 2022 | Revised: 19 April 2023 | Accepted: 2 May 2023

DOI: 10.1002/sam.11624

WILEY

RESEARCH ARTICLE

# Categorical classifiers in multiclass classification with imbalanced datasets

**Maurizio Carpita** | **Silvia Golia**

Department of Economics and Management, University of Brescia, Brescia, Italy

**Correspondence**
Silvia Golia, C.da S.Chiara, 50 - 25122 Brescia, Italy.
Email: silvia.golia@unibs.it

**Abstract**

This paper discusses, in a multiclass classification setting, the issue of the choice of the so-called categorical classifier, which is the procedure or criterion that transforms the probabilities produced by a probabilistic classifier into a single category or class. The standard choice is the Bayes Classifier (BC), but it has some limits with rare classes. This paper studies the classification performance of the BC versus two alternatives, that are the Max Difference Classifier (MDC) and Max Ratio Classifier (MRC), through an extensive simulation and some case studies. The results show that both MDC and MRC are preferable to BC in a multiclass setting with imbalanced data.

KEYWORDS

Bayes Classifier, categorical classifier, imbalanced data, multiclass classification

## 1 | INTRODUCTION

The classification task is one of the most important issues in real applications. In this paper, we focus on multiclass target variables; that is, variables that admit $k$ nonoverlapping classes or categories and the units are to be classified into one and only one of them. We can distinguish two steps in the classification procedure. The first step is to identify the so-called *probabilistic classifier*, which is a suitable method that assigns a probability to all of the categories that can be assumed by the target variable. The second step is to identify the so-called *categorical classifier*, which is a procedure or criterion that transforms the probabilities produced by the probabilistic classifier into a single category or class.

There is a broad literature that discusses the problem of how to find the best probabilistic classifier in both the dichotomous and polytomous contexts. However, less attention has been given to the choice of the criterion to be used to pass from the probabilistic classifier to the predicted class (i.e., to the choice of the categorical classifier).

The standard method is the Bayes Classifier (BC), which assigns a unit to the most likely class based on the probabilistic classifier. This categorical classifier has the advantage that it minimizes (on average) the test error rate [18]. Consequently, it is the optimal criterion to use if one is interested in the accuracy of the classification. BC favors the prevalent class. However, this characteristic can be a limit when the prevalent class is not the one of interest, or the data are not balanced. There are many real-world applications that are characterized by class imbalance (i.e., there are one or more classes that are under-represented in the dataset and most of the time, these classes are the ones of interest). In addition, there is a broad literature that has discussed the binary case [14], where the class imbalance problem was addressed through the following approaches:

data level, algorithm level, cost-sensitive learning and ensemble-based. The data level approach balances the data by resampling methods; for example, under- and/or over-sampling of one or more categories, such as the well-known SMOTE [6]. The algorithm level approach modifies the existing classifier learning algorithms to bias the learning toward the minority class, whereas the cost-sensitive learning approach incorporates approaches at the data level, at the algorithmic level, or at both levels jointly, considering higher misclassification costs for the units belonging to the minority class. The last approach is based on ensemble techniques (e.g., bagging or boosting) and usually combines an ensemble learning algorithm and one of the other approaches, specifically data level and cost-sensitive approaches. Nevertheless, most of the techniques that have been developed for the binary case are not directly applicable to the multiple classes case, which turns out to be a challenging issue. Some studies have faced the class imbalance problem in the multiclass setting by reducing the multiclass problem by banalization techniques, such as One-vs-One (OVO) or One-vs-All (OVA) schemes, or by following the four approaches developed for the binary case (see, e.g., [13, 15, 25, 26]). The methods that have been developed to answer the class imbalance problem work as pre-processing techniques or at the probabilistic classifier level, without taking into account the role that the choice of the categorical classifier could have in the final classification. In this paper, we work on the categorical classifier and we show that a suitable choice can improve the classification.

In previous papers (see, e.g., [16, 17]), we have investigated the performances of different categorical classifiers (some of them have also taken the ordinal nature of the target variable), and the Maximum Difference Classifier (MDC) has been found to be promising. In this paper, we use a simulation study and real applications to examine the classification performance of BC, MDC, and a new proposal, which we have denoted as the Maximum Ratio Classifier (MRC). The last two classifiers are based on a comparison between the predicted probabilities and the sample frequencies. This paper shows that these sample frequencies can be seen as the output of the null model. So, the resulting two classifiers come from the comparison of the "full" model (the probabilistic classifier) with the null model. In this sense, we can say that they are based on an index of performance following Cramer's approach [12] for the dichotomous case. However, this reasoning cannot be applied to BC, which does not consider the null model. The conclusions of this study will show that MDC and MRC represent better alternatives to the BC when the target variable has rare classes.

The rest of this paper is organized as follows. Section 2 introduces the definition of the three classifiers under study. Section 3 recalls the measures used in the paper to evaluate the classification performance of a classifier. Meanwhile, Section 4 reports the description of a simulation study. Section 5 shows the results on four real datasets. Conclusions follow in Section 6.

## 2 | THREE CATEGORICAL CLASSIFIERS

In this section, we define the three classifiers for multiclass classification problems, which are the object of our study. Let the target variable $Y$ be a categorical random variable with $k$ categories, $Y \in \{1, 2, \dots, k\}$, and let us assume that there is the following relationship between the probabilities of $Y$ and a set of $m$ predictors $\boldsymbol{X}^T = [X_1, X_2, \dots, X_m]$:

$$P_j = P(Y = j|\boldsymbol{X}) \quad j = 1, 2, \dots, k.$$

These conditional probabilities $P_j$, estimated with a model using a random sample of $n$ observations $\{(y_i, \boldsymbol{x}_i)\,; i = 1, 2, \dots, n\}$, can be collected into $n$ vectors of the type

$$\widehat{\boldsymbol{P}}_i^T = \left[\widehat{P}_{1i}, \widehat{P}_{2i}, \dots, \widehat{P}_{ki}\right] \quad i = 1, 2, \dots, n$$

with $\widehat{P}_{ji} \geq 0$ for all $j$ and $\boldsymbol{\iota}^T\widehat{\boldsymbol{P}}_i = \sum_{j=1}^{k} \widehat{P}_{ji} = 1$, where $\boldsymbol{\iota}$ is the all-ones vector.

The estimated conditional probabilities of $Y$ can be expressed, for each observation $y_i$, with the one-hot encoding sample vector $\boldsymbol{w}_i$ as follows:

$$\Pr(y_i) = \boldsymbol{w}_i^T\widehat{\boldsymbol{P}}_i = \sum_{j=1}^{k} w_{ji}\widehat{P}_{ji} = \prod_{j=1}^{k} \widehat{P}_{ji}^{w_{ji}} \quad i = 1, 2, \dots, n \tag{1}$$

where $\boldsymbol{w}_i^T = [w_{1i}, w_{2i}, \dots, w_{ki}]$ is, for $y_i = j$, a vector of $k-1$ zeros and a 1 in the $j$th position (see [2], Section 4.3.4).

If a generic $\boldsymbol{w}$ is considered, then instead of the observed $\boldsymbol{w}_i$ we can use the $\Pr(y_i)$ in (1) as a probabilistic classifier. The rule that maximizes (1) for each given $\widehat{\boldsymbol{P}}_i$ is the well-known *Bayes Classifier* (BC):

$$\begin{aligned}\text{BC} : \widehat{\boldsymbol{w}}_i &= \arg\max_{\boldsymbol{w}} \Pr(y_i) \\ &= \arg\max_{\boldsymbol{w}} \boldsymbol{w}^T\widehat{\boldsymbol{P}}_i \quad i = 1, 2, \dots, n.\end{aligned} \tag{2}$$

Note that, for each $\widehat{\boldsymbol{w}}_i$, there is one in-sample predicted category $\widehat{y}_i$. Moreover, in the case of binary classification ($k = 2$), the two estimated probabilities are $\widehat{P}_{1i}$ and $\widehat{P}_{2i} = 1 - \widehat{P}_{1i}$, so that (2) is the standard criterion "if $\widehat{P}_{1i} > 0.5$ then $\widehat{y}_i = 1$ else $\widehat{y}_i = 0$".

A useful interpretation of the BC rule (2) can be obtained when parametric models are adopted. In this case, the estimated conditional probabilities $P_j(\hat{\boldsymbol{\theta}}) = P(Y = j|\boldsymbol{x};\hat{\boldsymbol{\theta}})$ depend on the vector of parameters $\boldsymbol{\theta}$ that are estimated with an optimization method. For statistical models, the maximum likelihood (ML) method finds the maximum of the Log-Likelihood, which can be written using (1), as follows:

$$LL(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log\left[\Pr\left(y_i|\hat{\boldsymbol{\theta}}\right)\right] = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ji}\log\left[P_{ji}(\hat{\boldsymbol{\theta}})\right]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ji}\log\left[P\left(Y_i = j|\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}\right)\right] \qquad (3)$$

Equation (3) is an extension of the Log-Likelihood defined for the binary case by Cramer [12] to the multiclass case.

Let $LL(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ be the maximized Log-Likelihood (3) defined as function of the vector $\boldsymbol{y}^T = [y_1, y_2, \dots, y_n]$; applying the BC we choose $\boldsymbol{y} = \hat{\boldsymbol{y}}$, obtaining:

$$LL(\hat{\boldsymbol{y}}|\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{y}} LL(\boldsymbol{y}|\hat{\boldsymbol{\theta}}). \qquad (4)$$

In other terms, with the (2) rule, the in-sample predicted categories are those with higher estimated conditional probabilities, so that the BC is *optimal* in the sense that, for given $P_{ji}(\hat{\boldsymbol{\theta}})$, it maximizes $LL(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$, and hence the fit $\hat{\boldsymbol{y}}$ to the given $P_{ji}(\hat{\boldsymbol{\theta}})$. Note that the Log-Likelihood function has two uses: in the estimation step, $LL(\boldsymbol{\theta})$ is maximized with respect to probabilities (parameters) conditionally to sample data; whereas in the classification step, $LL(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ is maximized with respect to data conditionally to the estimated probabilities.

The other two useful simple classification rules can be obtained by comparing the estimated conditional probabilities (1) with their benchmarks, which are the in-sample relative frequencies observed for the $k$ categories, $\boldsymbol{f}^T = [f_1, f_2, \dots, f_k]$. These observed relative frequencies correspond to the estimated unconditional probabilities of $Y$ obtained under the so-called *null model*. To this aim, we define two *indices of performance* of the model for the $n$ observations:

$$\mathrm{Pf_R}\,(y_i) = \Pr(y_i)\,/\boldsymbol{w}_i^T\boldsymbol{f} = \boldsymbol{w}_i^T\left[\hat{\boldsymbol{P}}_i/\boldsymbol{f}\right] = \sum_{j=1}^{k} w_{ji}\left(\frac{\hat{P}_{ji}}{f_j}\right)$$
$$= \prod_{j=1}^{k}\left(\frac{\hat{P}_{ji}}{f_j}\right)^{w_{ji}} \qquad i = 1, 2, \dots, n \qquad (5)$$

$$\mathrm{Pf_D}\,(y_i) = \Pr(y_i) - \boldsymbol{w}_i^T\boldsymbol{f} = \boldsymbol{w}_i^T\left[\hat{\boldsymbol{P}}_i - \boldsymbol{f}\right] = \sum_{j=1}^{k} w_{ji}\left(\hat{P}_{ji} - f_j\right)$$
$$= \prod_{j=1}^{k}\left(\hat{P}_{ji} - f_j\right)^{w_{ji}} \qquad i = 1, 2, \dots, n \qquad (6)$$

where the ratio of two vectors in (5) is defined term-by-term. It is worth noting here that $\mathrm{Pf_R}\,(y_i)$ is an extension to the multiclass case of the index of performance defined by Cramer [12] for the binary case. If a generic $\boldsymbol{w}$ is used in (5) and (6) instead of the observed $\boldsymbol{w}_i$, then we can use $\mathrm{Pf_R}\,(y_i)$ and $\mathrm{Pf_D}\,(y_i)$ as probabilistic classifiers, respectively.

The rule that, for each given $\hat{\boldsymbol{P}}_i$, maximizes (5), is named the *Maximum Ratio Classifier* (MRC):

$$\mathrm{MRC}: \hat{\boldsymbol{w}}_i = \arg\max_{\boldsymbol{w}} \mathrm{Pf_R}\,(y_i)$$
$$= \arg\max_{\boldsymbol{w}} = \boldsymbol{w}^T\left[\hat{\boldsymbol{P}}_i/\boldsymbol{f}\right] \qquad i = 1, 2, \dots, n \qquad (7)$$

whereas the rule that, for each given $\hat{\boldsymbol{P}}_i$, maximizes (6), is named the *Maximum Difference Classifier* (MDC):

$$\mathrm{MDC}: \hat{\boldsymbol{w}}_i = \arg\max_{\boldsymbol{w}} \mathrm{Pf_D}\,(y_i)$$
$$= \arg\max_{\boldsymbol{w}} = \boldsymbol{w}^T\left[\hat{\boldsymbol{P}}_i - \boldsymbol{f}\right] \qquad i = 1, 2, \dots, n \qquad (8)$$

Note that, for each $\hat{\boldsymbol{w}}_i$, there is one in-sample predicted category $\hat{y}_i$ for both MRC and MDC, which are not necessarily the same. For the binary case, the classification rule (7) was proposed by Cramer [12], so MRC represents its extension to the multiclass case. It is easy to verify that, for binary classification, MRC and MDC always give the same predicted category, so (7) and (8) are the Cramer criterion "if $\hat{P}_{1i} > f_1$ then $\hat{y}_i = 1$ else $\hat{y}_i = 0$". This is useful in the case of unbalanced samples (see [12], Section 5.1). However, MRC and MDC do not give the same results when $k > 2$. For example, for $k = 3$ and sample frequencies $\boldsymbol{f}^T = [0.30, 0.20, 0.50]$, if the estimated probabilities are $\hat{\boldsymbol{P}}^T = [0.15, 0.25, 0.60]$, then MDC gives $\hat{y} = 3$ (for this example the same predicted category of BC), whereas MRC gives $\hat{y} = 2$.

Both MRC and MDC, as BC, have a useful interpretation in the ML framework. If $\hat{P}_{ji} = P_{ji}(\hat{\theta})$ are the ML estimates of the conditional probabilities and $\boldsymbol{y}$ is the observed sample of $Y$, then by using (5) we obtain:

$$LR(\hat{\boldsymbol{\theta}}) = 2\sum_{i=1}^{n}\log\left[\mathrm{Pf_R}\left(y_i|\hat{\boldsymbol{\theta}}\right)\right] = 2\left[LL(\hat{\boldsymbol{\theta}}) - LL_0\right] \qquad (9)$$

which is the classical *Likelihood Ratio* statistic to test the significance of the model. $LL(\hat{\boldsymbol{\theta}})$ is the Log-Likelihood of the *full* model (with $\boldsymbol{x}$) and $LL_0$ the Log-Likelihood of the *null* model (without $\boldsymbol{x}$) (see [12] for the dichotomous case).

Let $LR(\mathbf{y}|\hat{\boldsymbol{\theta}})$ be the maximized Likelihood Ratio (9) defined as function of the vector $\boldsymbol{y}$; applying the MRC we choose $\boldsymbol{y} = \hat{\boldsymbol{y}}$, obtaining:

$$LR(\hat{\boldsymbol{y}}|\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{y}} LR(\boldsymbol{y}|\hat{\boldsymbol{\theta}}). \tag{10}$$

In other terms, with the (7) rule, the in-sample predicted categories are those with higher estimated conditional probabilities in term of *relative difference* with respect to the sample frequencies, so that the MRC is *optimal* in the sense that, for given $P_{ji}(\hat{\boldsymbol{\theta}})$, it maximizes $LR(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$, and hence the fit $\hat{\boldsymbol{y}}$ to the given $P_{ji}(\hat{\boldsymbol{\theta}})$. Also note in this case the two uses of the Likelihood Ratio function: first, in the estimation step $LR(\boldsymbol{\theta})$ is maximized with respect to probabilities (parameters) conditionally to sample data; and second, in the classification step $LR(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ is maximized with respect to data conditionally to the estimated probabilities.

Finally, using (6) we define:

$$LD(\hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} \mathrm{Pf}_{\mathrm{D}}\left(y_i|\hat{\boldsymbol{\theta}}\right) = L(\hat{\boldsymbol{\theta}}) - L_0 \tag{11}$$

with $L(\hat{\boldsymbol{\theta}}) = \exp[LL(\hat{\boldsymbol{\theta}})]$ and $L_0 = \exp(LL_0)$; (11) is the *Likelihood Difference* statistic between the full and the null models. Let $LD(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ be the maximized Likelihood Difference (11) defined as function of the vector $\boldsymbol{y}$; applying the MDC we choose $\boldsymbol{y} = \hat{\boldsymbol{y}}$, obtaining:

$$LD(\hat{\boldsymbol{y}}|\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{y}} LD(\boldsymbol{y}|\hat{\boldsymbol{\theta}}). \tag{12}$$

In other terms, with the (8) rule, the in-sample predicted categories are those with higher estimated conditional probabilities in term of *absolute difference* with respect to the sample frequencies, so that the MDC is *optimal* in the sense that, for given $P_{ji}(\hat{\boldsymbol{\theta}})$, it maximizes $LD(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$, and hence the fit $\hat{\boldsymbol{y}}$ to the given $P_{ji}(\hat{\boldsymbol{\theta}})$. Again, note the two uses of the Likelihood Difference function: first, in the estimation step, $LD(\boldsymbol{\theta})$ is maximized with respect to probabilities (parameters) conditionally to sample data; and second, in the classification step, $LD(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ is maximized with respect to data conditionally to the estimated probabilities.

## 3 | MEASURES TO EVALUATE PERFORMANCE OF A CLASSIFIER

When the task is to evaluate the predictive performance of a classifier, it is generally a good rule to consider several indices that give a wide comprehension of the behavior of the classifier. For example, the simple percentage of correct predictions gives a limited view of the predictive ability of a classifier, especially when the classes are not balanced (i.e., when one or more classes are under-represented in the dataset). Following [22], in addition to the overall accuracy, in this paper we will use the per-class precision, recall and F1-score, and their macroaverage version, used in the multiclass setting.

In a multiclass classification problem with $k > 2$ classes, the $k \times k$ confusion matrix can be reduced to $k$ $2 \times 2$ confusion matrices, one for each class label $j = 1, 2, \ldots, k$.

The per-class Precision ($Pre_j$), Recall ($Rec_j$), and F1-score ($F1_j$) can be calculated as follows:

$$Pre_j = \frac{tp_j}{tp_j + fp_j}$$
$$Rec_j = \frac{tp_j}{tp_j + fn_j}$$
$$F1_j = 2 \times \frac{Pre_j \times \mathrm{Rec}_j}{Pre_j + \mathrm{Rec}_j},$$

where $tp_j$, $fp_j$, $fn_j$, and $tn_j$ are the number of true positives, false positives, false negatives, and true negatives for the $j$th class. Precision quantifies the class agreement of the data labels with the positive labels given by the classifier, whereas recall quantifies the effectiveness of the classifier in identifying positive labels. The F1-score is used to integrate recall and precision into a single metric by means of their harmonic mean. Two strategies can be applied to summarize the $k$ values of these indices: macroaveraging, which is obtained by taking the arithmetic mean of the per-class indices; and micro-averaging, which is obtained by summing the counts to get cumulative $tp$, $fn$, $tn$, and $fp$ and then calculating a performance measure. Macroaveraging treats all classes equally, while micro-averaging favors larger classes [22]. In this paper, we will use macroaverage precision (Macro Pre), recall (Macro Rec) and F1-score (Macro F1) because we do not want to discriminate less frequent classes. For Macro F1, we stress that the used formula is the arithmetic mean over individual F1-scores and not the harmonic mean of Macro Pre and Macro Rec, as introduced in [22]. There is some evidence to show that the second formula overly favors heavily biased classifiers and can yield misleadingly high evaluation scores [20].

The overall accuracy (OvAc) is the rate of correct classification, which is defined as:

$$OvAc = \frac{1}{n} \sum_{j=1}^{k} tp_j$$

where $n$ is the total number of cases. OvAc is equal to Micro Precision, Micro Recall and Micro F1 measures and it is maximized by BC.

It can also be of interest to compare the per-class indices of the classes computing their difference in absolute value. A useful indicator of these differences is their maximum, which is denoted as Maximum Distance Between Indices (MDB Ind), where $Ind = Pre$, $Rec$ and $F1$, and is defined as:

$$MDB\ Ind = \max_{j \neq s} \mid Ind_j - Ind_s \mid .$$

The lower the MDB Ind, the better the classification.

## 4 | SIMULATION STUDY

We used the Dirichlet random variable (r.v.) to simulate the probability distribution of a nominal variable with $k$ categories. This r.v. is parameterized by a vector $\boldsymbol{\alpha}$ of $k$ positive real numbers and is a multivariate generalization of the Beta r.v. The appealing characteristic of the Dirichlet r.v. $\boldsymbol{D} = \{D_1, D_2, \dots, D_k\}$ for the present context is that a single realization is composed of $k$ values $d_j$ such that each $d_j \in (0, 1)$ and $\sum_{j=1}^{k} d_j = 1$, so it can be seen as the probability mass function of a $k$-variate discrete r.v., which is the so-called "target variable." Moreover, there is a link between the alpha parameters and the expected value of the marginals; that is, $E\left(D_j\right) = \alpha_j / \sum_{j=1}^{k} \alpha_j$. This allows us to control the form of the probability mass function of this target variable controlling the expected value of the marginals.

In our simulation, each realization of a given Dirichlet r.v. had a double use. First, it was seen as the probability mass function of a $k$-variate discrete r.v. and used to randomly extract a realization from it that represented the actual (observed) class of the target variable. At the same time, it was considered as the output of a probabilistic classifier for the target variable and was used in producing the classification following the three categorical classifiers defined in Section 2. This simulation scheme avoids the specification of a statistical model for data generation. Consequently, the results will be general and not linked to a specific model. Nevertheless, it prevents out-of-sample prediction, so the predictive performances will be in-sample.

The structure of the simulation setting is as follows. For each set of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_k)$, we extracted 5000 realizations of the corresponding Dirichlet r.v. The choice of this sample size refers to the fact that a high sample size allows us to investigate the performance of the analyzed categorical classifiers in a less problematic framework, given that they are working in the context of large samples. The observed class was generated from the 5000 realizations. Then, the BC, MDC, and MRC were applied to produce the predicted class, and the performance measures described in Section 3 were calculated. This scheme was repeated 1000 times and the mean values of the indicators and their standard deviations were recorded.

To consider a variety of situations and manage the imbalance complexity, we left the task to control the choice of the alphas $\boldsymbol{\alpha}$ to the imbalance ratio (IR) [21], remembering that each $\alpha_j$ is obtained as $\alpha_j = E\left(D_j\right) \times \sum_{j=1}^{k} \alpha_j$, where, in our simulation, $E\left(D_j\right) = f_j$, the frequency of class $j$. The IR is defined as the ratio of the frequencies of the majority and the minority class, that is

$$IR = \frac{f_{max}}{f_{min}}$$

and it can be used to quantify the degree of imbalance [15].

Moreover, given that, with the same IR, there can be different configurations of the probability mass function of the target variable, we have considered the mean imbalance ratio (MIR), which is obtained as

$$MIR = \frac{1}{k} \sum_{j=1}^{k} \frac{f_{max}}{f_j}.$$

The higher the MIR, the more $c_{max}$, the class with $f_{max}$ as frequency, absorbs units and the dataset is imbalanced.

The sequence of numbers from 1.5 to 13.5 with the increment of 0.5 composes the set of selected IRs. For each IR, we considered five different and increasing values of MIR, labeled as MIR 1, …, MIR 5; Table 1 reports the range of these five MIR with respect to the 25 IRs.

Moreover, we fixed the sum of the $\alpha_j$ equal to 20. This choice does not have an impact on the simulation results. The choice of particular values for the alphas is not a sensitive issue. A preliminary analysis that applied a linear transformation to the set of alphas has shown that the characteristics of the probability mass function of the target variable simulated with the multivariate Dirichlet before and after the transformation remain almost unchanged.

Operationally, the simulations were made using the R open source software and the R package rminer [9] was used to calculate most of the predictive performance indices.

### 4.1 | The balanced case

First, we have investigated if the three classifiers defined in Section 2 have the same performance in the balanced case, which were obtained by imposing the same value to all of the $\alpha_j$; that is, $\alpha_j = 6.667$, $\alpha_j = 1.5$ and $\alpha_j = 5$ for the case of the three, four and five classes respectively.

**TABLE 1** Range of MIR (mean imbalance ratio) by number of classes.

| | MIR 1 | MIR 2 | MIR 3 | MIR 4 | MIR 5 |
|---|---|---|---|---|---|
| 3 classes | 1.235–4.667 | 1.261–5.213 | 1.287–5.268 | 1.314–5.331 | 1.340–5.399 |
| 4 classes | 1.130–4.468 | 1.201–4.515 | 1.229–4.917 | 1.294–5.640 | 1.300–6.711 |
| 5 classes | 1.176–3.895 | 1.193–4.570 | 1.223–5.249 | 1.241–5.627 | 1.284–6.563 |

**TABLE 2** Macroaveraged and OvAc indices for the balanced cases.

| | 3 classes | | | 4 classes | | | 5 classes | | |
|---|---|---|---|---|---|---|---|---|---|
| | BC | MDC | MRC | BC | MDC | MRC | BC | MDC | MRC |
| Macro Pre | 0.445 | 0.445 | 0.445 | 0.473 | 0.473 | 0.473 | 0.311 | 0.312 | 0.312 |
| Macro Rec | 0.445 | 0.444 | 0.444 | 0.473 | 0.473 | 0.473 | 0.311 | 0.310 | 0.310 |
| Macro F1 | 0.445 | 0.442 | 0.441 | 0.473 | 0.473 | 0.472 | 0.311 | 0.308 | 0.306 |
| OvAc | 0.445 | 0.442 | 0.442 | 0.473 | 0.473 | 0.473 | 0.311 | 0.308 | 0.307 |

Table 2 reports the mean value, over 1000 replications, of the macroaveraged indices and OvAc defined in Section 3.

When all of the classes are equally represented, the three categorical classifiers perform in the same way.

## 4.2 | Imbalanced case: three classes

The first simulation for the imbalanced case concerns the three classes case. The three values $\alpha_j$ were chosen as follows. For a given IR, we generated the frequency of the first class $f_1$ from a continuous uniform distribution $U(0.4,0.6)$. This class is the one with the highest frequency ($f_{max}$). The frequency of the second class, $f_2$, was obtained as $f_2 = f_1/IR$ and this is the class with the lowest frequency ($f_{min}$). The frequency of the third class was obtained as $f_3 = 1 - (f_1 + f_2)$. The resulting triplet was considered only if $f_2 < f_3$. The MIR was then calculated. This procedure was repeated a large number of times, obtaining a set of feasible triplets with their MIR. Afterward, we selected the five final triplets according to their MIR equal to its minimum, first, second and third quartile and maximum. Then, the $\alpha_j$ were computed as $\alpha_j = f_j \times 20$, with $20 = \sum_{j=1}^{k} \alpha_j$. Table 1 reports the range of these five MIR with respect to the 25 IRs.

Figure 1 shows the mean value, over 1000 replications, of the OvAc and the macroaveraged indices of the three classifiers by MIRs and IRs; ○ identifies BC, $\Delta$ MDC and + MRC. The plots of the corresponding standard deviations are reported in the appendix (Figure A1).

We can observe the following behavior for all of the combinations of IR and MIR. OvAc for BC is always higher than the corresponding values for MDC and MRC, even if its standard deviations are higher. This behavior does not surprise us because BC favors the class with the highest frequency, which is better predicted. The BC's Macro Pre also appears to be higher, nevertheless the associate standard deviations show increasing values. This can be explained by the presence, for most of the combinations of IR and MIR, of a multimodal distribution of the index along the 1000 replications. In contrast, the competitor classifiers MDC and MRC have higher values for Macro Rec and Macro F1 because, even if they lose power in predicting the most frequent class, they gain more power in predicting the rare class and these two indices are sensible to the predictive ability of all of the classes. Now, let us consider the behavior of BC's Macro F1 for low values of MIR (MIR 1 and 2). It can be observed that there is a decrease in the index for small values of IR and a recovery for higher values of IR. This is related to the shape of the triplets of the frequencies, which depends on the simulation setting; as IR increases, $f_{min}$ obviously decreases and for small MIR, this causes the remaining two frequencies to be almost the same (e.g., MIR 1: IR = 5 with $\boldsymbol{f}^T = [0.455,0.091,0.454]$, IR = 11.50 with $\boldsymbol{f}^T = [0.479,0.042,0.479]$; MIR 2: IR = 5 with $\boldsymbol{f}^T = [0.488,0.098,0.415]$, IR = 11.50 with $\boldsymbol{f}^T = [0.509,0.044,0.446]$).

By comparing the performances of MDC and MRC, it is possible to observe that MDC has in general higher values for OvAc and Macro F1, whereas the two classifiers have a comparable Macro Pre and Macro Rec for low values of IR. When IR increases, Macro Rec tends to be higher for MRC, whereas MDC's Macro Pre is slightly higher than that of MRC.

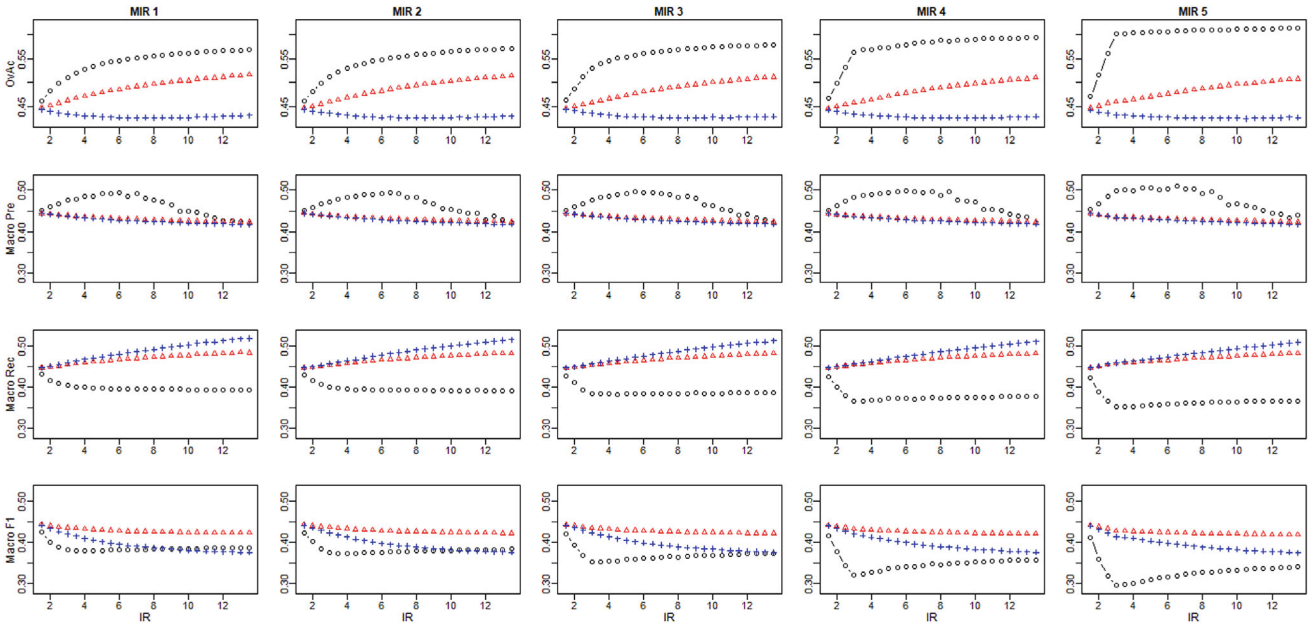

FIGURE 1 Simulation results for three classes: mean values of OvAc and Macroaveraged Pre, Rec and F1 by IR and MIR for the three classifiers (BC = ○; MDC = *Δ*; MRC = +).

### 4.3 | Imbalanced case: four and five classes

The simulation procedure with $k$ equal to 4 and 5 is similar but more complex than in the previous case with $k = 3$. The five values $\alpha_j$ of a Dirichlet with five components were chosen as follows. We generated five numbers from as many Beta distributions; that is, $Beta(1,5)$, $Beta(1,5)$, $Beta(2,5)$, $Beta(3,5)$ and $Beta(5,5)$. Figure 2 shows their probability density function. The median of these r.v. is equal, respectively, to 0.125, 0.263, 0.364 and 0.500. As the first parameter increases, the distribution becomes less skewed.

Given that the record of these five numbers, transformed so that they sum to one, represents the probability mass function of a five-class categorical variable, and given that we want to simulate imbalanced data, we need to have two small probabilities, one high probability and the remaining two probabilities in between. The choice of the particular values of the Beta parameters used in the simulation is related to this simulation scheme. The IR and MIR were then calculated. This procedure was repeated a large number of times. Afterward, for each of the 25 IR considered in the study, we selected the five final sets of frequencies according to their MIR equal to its minimum, first, second and third quartile and maximum. Then, the $\alpha_j$ were computed as $\alpha_j = f_j \times 20$, with $20 = \sum_{j=1}^{k} \alpha_j$. Table 1 reports the range of these five MIR with respect to the 25 IRs.

The four values $\alpha_j$ of a Dirichlet for the case with $k = 4$ classes were chosen as just explained, while removing one of the two $Beta(1,5)$.

Figures 3 and 4 show the mean value, over 1000 replications, of the OvAc and the macroaveraged indices of the three classifiers by MIRs and IRs; ○ identifies BC, *Δ* MDC and + MRC. The plots of the corresponding standard deviations are reported in the appendix (Figures A2 and A3).

When analyzing these two figures, we can observe that the relations between the three classifiers, previously highlighted in Section 4.2, are valid even if the number of the classes is four or five. Again, the higher values of Macro Pre for BC are associated with high standard deviations, which for most of the combinations of IR and MIR can be explained by the presence of a multimodal distribution of the index along the 1000 replications.

The sawtooth behavior of the indices for high values of MIR and IR, which is clearly evident for BC, is related to the similarity of the results in presence of similar configurations of the set of frequencies (e.g., four classes and MIR 4: Configuration 1) IR = 9.50 with $\boldsymbol{f}^T = [0.048, 0.077, 0.423, 0.452]$ and IR = 10.50 with $\boldsymbol{f}^T = [0.043, 0.063, 0.437, 0.456]$ versus Configuration 2 IR = 10 with $\boldsymbol{f}^T = [0.059, 0.154, 0.199, 0.588]$ and IR = 11 with $\boldsymbol{f}^T = [0.055, 0.122, 0.214, 0.609]$). It is interesting to observe that low values of BC's Macro F1 correspond to high values of BC's OvAc. This occurs when there is one frequency that absorbs more of the 50% of the units, as in Configuration 2 in the four classes example.

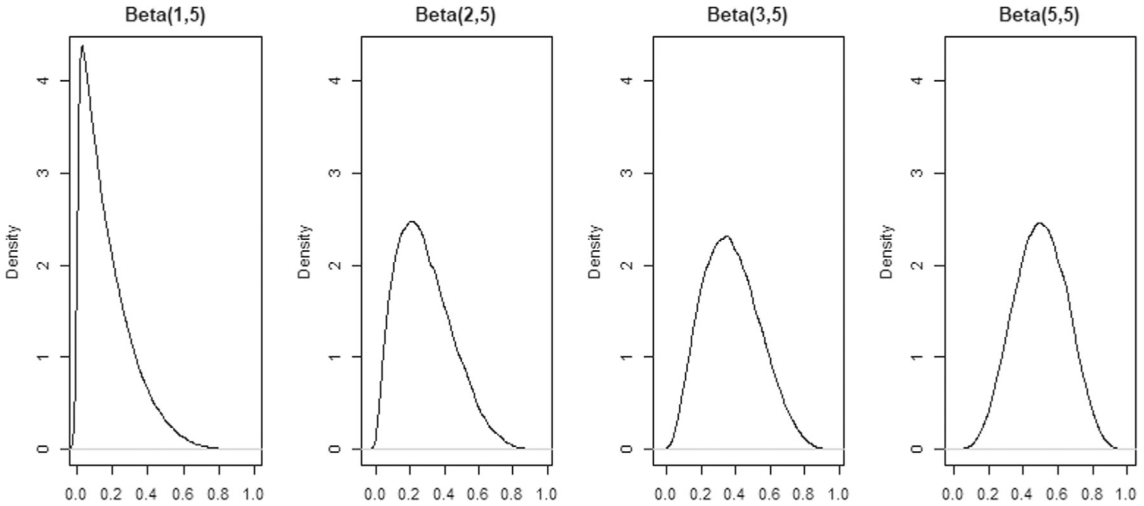

FIGURE 2 Probability density function of the Beta r.v. used in the simulation.

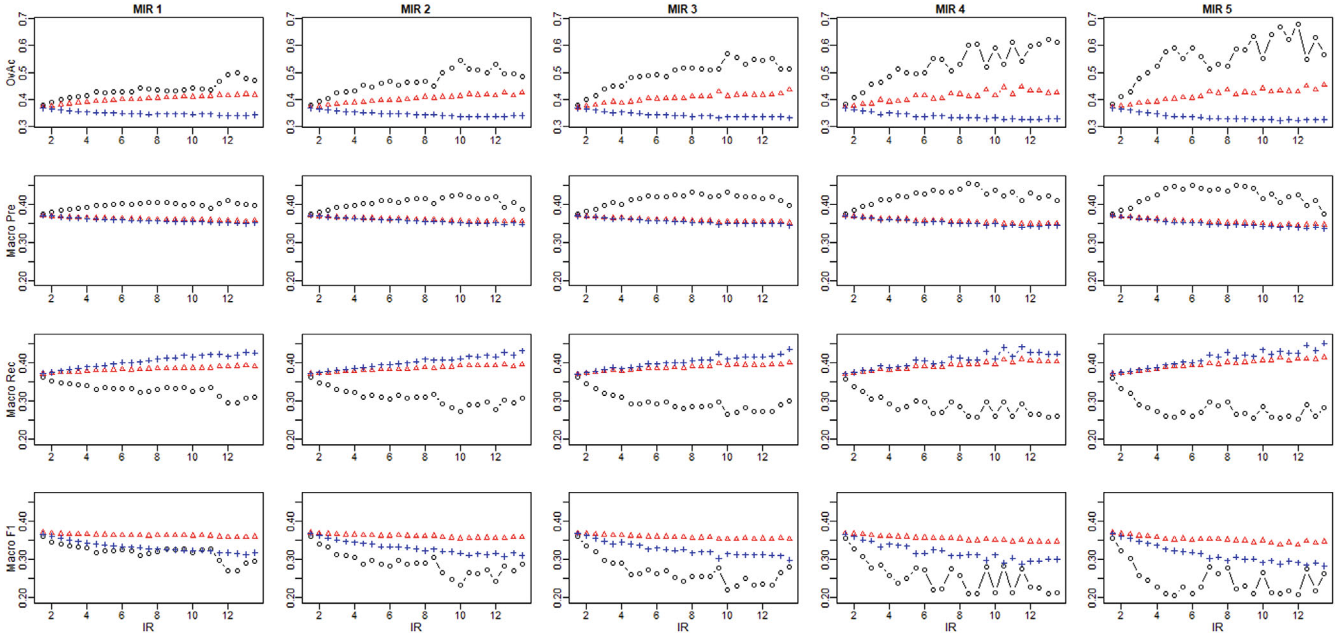

FIGURE 3 Simulation results for four classes: mean values of OvAc and Macroaveraged Pre, Rec and F1 by IR and MIR for the three classifiers (BC = ○; MDC = Δ; MRC = +).

Moreover, we can observe that the performance indices of MCD and MRC, in contrast to those of BC, are more stable as the imbalance between the classes increases.

## 5 | CASE STUDIES

In this section, we will discuss the results of the application of the three classifiers, BC, MDC, and MRC, to some case studies. When the data needed a probabilistic classifier first, and the categories were ordered, we used the *Cumulative Logit Model* (CLM) [1] to study and predict the occurrence probabilities of each category. Although this choice for the probabilistic classifier might not be the best for the analyzed cases, our attention is focused on the performance of different categorical classifiers more than on the performance of the probabilistic classifier. Only for the dataset with the largest sample size we will consider the XGBoost [7] as an alternative probabilistic classifier. This probabilistic classifier belongs to the ensemble-based approach to dealing with the class imbalance problem and has proven itself to be a good choice between the available boosting ensemble algorithms [25].

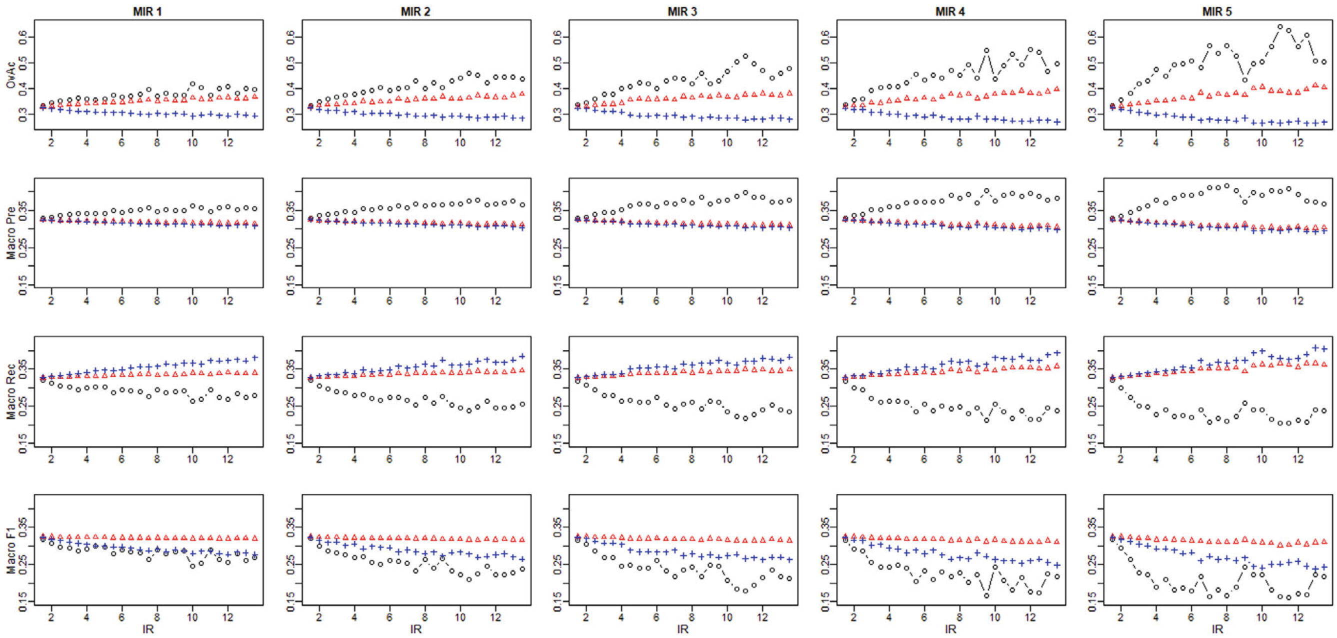

FIGURE 4 Simulation results for five classes: mean values of OvAc and Macroaveraged Pre, Rec and F1 by IR and MIR for the three classifiers (BC = ○; MDC = Δ; MRC = +).

CLM is defined as follows: let $Y$ be a categorical target variable with $k$ ordinal categories $\{1, 2, \ldots, k\}$, and let $\{X_1, \ldots, X_m\}$ be a set of $m$ explanatory variables; for the statistical unit $i$, the CLM has the following form:

$$\begin{aligned} \text{logit}\,[P(Y_i \leq j)] &= \log \frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)} \\ &= \beta_{0j} + \sum_{s=1}^{m} \beta_s x_{is}, \quad \text{for} \quad j = 1, 2, \ldots, k-1. \end{aligned}$$

Once the parameters have been estimated, it is possible to use the model for predictive purposes, so the CLM gives the $k$ predicted probabilities that are passed to the categorical classifier.

To evaluate the out-of-sample predictive performance of the three classifiers, we conducted a stratified five-fold cross-validation. The stratified $k$-fold cross-validation method [3, 27] builds each fold in such a way that the class proportion of the feature of interest inside the fold is approximately the same as in the original dataset, so each fold is a good representative of the entire original dataset. This implies that there is almost the same class proportion of the original dataset in both the training and test sets, and this is recommended in the presence of imbalanced data.

The first real dataset under study regards the prediction of the result of a soccer match. In this context, the target variable is the result obtained by the home team and it has three categories: loss (Class 1), draw (Class 2) and win (Class 3) of the home team. This variable has an ordinal nature (loss $\prec$ draw $\prec$ win) and therefore can be seen as an ordinal variable, even if the three categorical classifiers considered do not require an ordering in the categories. The dataset comes from the Kaggle European Soccer database (KES) [4] and comprises the decimal betting odds on the matches, provided by 10 betting companies, and the results of the corresponding matches. These 10 odds were averaged and transformed into probabilities of loss, draw and win [23, 24]. The focus was on the matches played in the Italian League Serie A during the seasons from 2008/2009 to 2015/2016.

The second dataset concerns the sensorial quality of the white and red variants of the Portuguese "Vinho Verde" wine [10]. The data are available at the UCI Machine Learning Repository [11]. The dataset comprises 11 of the most common physicochemical variables and a sensory preference variable that measures the sensorial quality of the wine, which is the target variable. This sensory preference variable was obtained from the evaluations of experienced judges who scored the wines, using a 0–10 scale, with 0 meaning very bad and 10 excellent, although not all of the possible scores were used by the judges (red wines: 3–8; white wines: 3–9). Moreover, the lowest and highest sensory preferences had very low frequency (red wines: 0.63% and 1.13%; white wines: 0.41% and 0.10%), causing extremely high IR (red wines: 68.1; white wines: 439.6) and MIR (red wines: 20.71; white wines: 82.93). Consequently, we decided to merge scores 3–4 and 7–8 for red wines and scores 3–4 and 8–9 for white wines, obtaining new target variables on a four-category (red wines) and five-category (white wines) ordinal scale.

**TABLE 3** Observed frequencies of the four target variables, sample sizes ($n$), number of explanatory variables ($m$), IR and MIR indices.

| | Class | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Data set** | **1** | **2** | **3** | **4** | **5** | $n$ | $m$ | **IR** | **MIR** |
| Soccer results | 0.270 | 0.260 | 0.470 | - | - | 3014 | - | 1.808 | 1.516 |
| Red wines | 0.039 | 0.426 | 0.399 | 0.135 | - | 1599 | 11 | 10.923 | 4.037 |
| White wines | 0.037 | 0.297 | 0.449 | 0.180 | 0.037 | 4898 | 11 | 12.135 | 5.855 |
| Heart disease | 0.593 | 0.182 | 0.118 | 0.118 | 0.044 | 297 | 13 | 13.477 | 5.557 |

**TABLE 4** Classifier performance indices (macroaveraged Pre, Rec and F1, OvAc and W.Kappa) for the real case studies.

| | Soccer results | | | Red wines | | | White wines | | | Heart disease | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **BC** | **MDC** | **MRC** | **BC** | **MDC** | **MRC** | **BC** | **MDC** | **MRC** | **BC** | **MDC** | **MRC** |
| Macro Pre | 0.526 | 0.461 | 0.465 | 0.446 | 0.475 | 0.401 | 0.411 | 0.371 | 0.304 | 0.423 | 0.429 | 0.401 |
| Macro Rec | 0.454 | 0.471 | 0.474 | 0.412 | 0.444 | 0.477 | 0.298 | 0.337 | 0.384 | 0.382 | 0.412 | 0.420 |
| Macro F1 | 0.399 | 0.460 | 0.462 | 0.417 | 0.442 | 0.387 | 0.298 | 0.323 | 0.296 | 0.380 | 0.410 | 0.396 |
| OvAc | 0.535 | 0.484 | 0.481 | 0.592 | 0.574 | 0.420 | 0.518 | 0.487 | 0.349 | 0.626 | 0.598 | 0.560 |
| W.Kappa | 0.309 | 0.326 | 0.331 | 0.510 | 0.528 | 0.516 | 0.405 | 0.468 | 0.473 | 0.706 | 0.730 | 0.713 |

The third dataset, denoted as the Cleveland Dataset, is a part of the wide Heart Disease Dataset, which is available at the UCI Machine Learning Repository [19] and refers to the data obtained from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. The aim is to study and predict the presence of heart disease patients. The target variable is the diagnosis of heart disease (angiographic disease status), which assumes five categories from 0 (no presence) to 4. The Heart Disease Dataset contains 76 attributes. In this study, we used a subset of 13 that were used in published papers on the subject.

Table 3 reports, for the four target variables, the observed class frequencies, the sample sizes $n$, the number of explanatory variables $m$, the IR and the MIR.

The soccer dataset has a class (win) that absorbs the 47% of the results, whereas the other two equally subdivide the remaining 53%; here there is no rare class but the data are still imbalanced. The other datasets have one (red wines and heart disease) or two (white wines) rare classes with a frequency lower than 5%, and this explains the high value of IR.

Table 4 reports the value of the macroaveraged indices and OvAc computed on the base of the stratified cross-validation prediction of the target variables of all of the available units (matches, wines and patients). The per-class Precision, Recall and F1-score plus the MDB indices are reported in the appendix (Table A1). Moreover, given that all of the target variables are on an ordinal scale, we have also computed the Cohen's Weighted Kappa index (W.Kappa) [8] to take into account a different cost in the miss-classification. The weights that we used are quadratic.

Let us consider the soccer dataset. It is well-known that the draw is the most difficult result to predict [5]. This also happens in this dataset: the draw (Class 2) has the lowest recall between the three results. Nevertheless, even if BC has almost no power to correctly predict it ($Rec_2 = 1.9\%$), the other two classifiers gain a certain power. In fact, $Rec_2$ for MDC and MRC is 15 and 16 times higher than the $Rec_2$ of BC. OvAc and Macro Prec are higher for BC but BC is outperformed by both MDC and MRC in terms of Macro Rec, Macro F1 and W.Kappa. Apart from OvAc, which is slightly lower, MRC outperforms MDC. We can conclude from this that overall MRC seems to be the best categorical classifier for this dataset.

Regarding the red wines and heart disease datasets, we can observe that MRC has a relatively good power in correctly predicting the rare classes (Rec between 0.415 and 0.444), unlike BC, which has no predictive power at all or low predictive power, and MDC, which has low predictive power but higher than the one of BC. When there is only one rare class (red wines and heart disease), MDC has the highest Macro Pre, Macro F1 and W.Kappa, OvAc slightly lower than the BC's OvAc and Macro Rec lower than the MRC's Macro Rec but higher than the BC's Macro Rec. We can conclude from this that MDC seems to be overall

**TABLE 5** Classifier performance indices (macroaveraged Pre, Rec and F1, OvAc and W.Kappa) for the White Wines estimated using XGBoost.

| | BC | MDC | MRC |
|---|---|---|---|
| Macro Pre | 0.395 | 0.370 | 0.363 |
| Macro Rec | 0.331 | 0.352 | 0.389 |
| Macro F1 | 0.343 | 0.358 | 0.371 |
| OvAc | 0.507 | 0.494 | 0.474 |
| W.Kappa | 0.464 | 0.473 | 0.486 |

the best categorical classifier for both red wines and heart disease datasets.

Let us consider the white wines dataset, for which the XGBoost was used as an alternative probabilistic classifier. Table 5 reports the value of the Macroaveraged indices, OvAc and W.Kappa.

By comparing the values of the indices reported in Tables 4 and 5, it can be observed that all of the values are higher for both the MDC and MRC when the XGBoost is considered, whereas for BC Macro Pre and OvAc are lower than those obtained by CLM. Without going into the question of the choice of the best model, we can observe that the comments concerning the comparison between BC and its two competitors that arise from Table 4 continue to be valid if Table 5 is instead evaluated. In fact, apart from OvAc and Macro Pre, the other indices are higher when computed by applying MDC or MRC instead of BC, so both are preferable. The use of XGBoost seems to have an effect on the performance of MRC when compared with MDC. The reduction in OvAc using MRC instead of MDC passes from 28.34% for CLM to 4.05% for XGBoost and MRC's Macro F1 becomes higher than MDC's Macro F1. Overall, considering CLM, MDC seems to be the best categorical classifier for the white wines dataset; whereas, considering XGBoost, MRC seems the preferable.

## 6 | CONCLUSIONS

This paper has examined the issue of choice of the so-called categorical classifier in a multiclass classification setting, the procedure or criterion that transforms the probabilities produced by a probabilistic classifier into a single category or class. Although BC is the standard choice, it has some limits with rare classes. Therefore, two alternative classifiers, MDC and MRC, based on comparing the predicted probabilities and the sample frequencies, were proposed. They have interesting characteristics, as they do not need the specification of a misclassification cost function to be minimized, or a balancing of the rare classes. Moreover, if a parametric model is used, they have a useful interpretation in the maximum likelihood framework.

First, we performed a broad simulation study involving target variables with three, four and five classes and a high sample size (5000) to investigate the performance of the analyzed categorical classifiers in a less problematic framework. The main findings from the simulation study are as follows. When all the categories are equally represented, as in the balanced case, the three categorical classifiers perform in the same way. In the imbalance cases, regardless of the number of classes of the target variable and for all combinations of IR and MIR, OvAc for BC is always higher than the corresponding values for MDC and MRC, whereas MDC and MRC have higher values for Macro Rec and Macro F1. BC's Macro Pre also appears to be higher. Nevertheless, the associated standard deviations show increasing values, which for most of the combinations of IR and MIR can be explained by the presence of a multimodal distribution of the index along the 1000 replications.

When the number of classes are four or five, the performance indices of MCD and MRC (in contrast to those of BC) are more stable as the imbalance between the classes increases. Finally, when comparing the performance of MDC and MRC, MDC has in general higher values for OvAc and Macro F1, whereas the two classifiers have a comparable Macro Pre and Macro Rec for low values of IR. When IR increases, Macro Rec tends to be higher for MRC, whereas MDC's Macro Pre is slightly higher than that of MRC.

In summary, the simulation studies reveal that both MDC and MRC are preferable to BC in a multiclass setting with imbalanced data. As pointed out in Section 4, the simulation scheme leads us to the evaluation of the performance of in-sample predictions that could be too optimistic. Nevertheless, we are interested in comparing the three categorical classifiers, rather than evaluating the performance of different models. Moreover, the results concerning the four case studies are based on out-of-sample predictions and confirm what emerged from the simulation.

From the analysis of the white wine dataset, for which we have used two alternative probabilistic classifiers, it seems that the choice of the probabilistic classifier can mostly affect the performance of MRC and MDC, improving all the indices. This could suggest that the choice of a probabilistic classifier could be an issue which matters. Although it is clear that our observations are limited to one real dataset and two models, this finding deserves further and closer analysis. This will be part of a future development of this paper.

A second issue of interest is the role played by the sample size on the simulation results. Now, given the type of simulation scheme, which does not involve the estimation of a model first to get the estimated probability mass function of the target variable, we expect that the sample size will not have a great impact on the performance of the categorical classifiers. Some preliminary simulations with smaller sample sizes and the results from the case studies, which have a sample size lower than or equal to the one used in the simulation ($n = 297, 1599, 3014, 4898$), seem to support this expectation. A future wider simulation study will investigate this aspect.

Finally, we would like to discuss our choice of the values of IR, limiting them to the 1.5–13.5 range and excluding larger values. This paper contains the proposal of two new categorical classifiers, therefore we thought it would be worth testing them under realistic, but not extreme conditions. As a matter of fact, conditions in which the ratio of the smallest to the largest class reaches 1 to 100 or even more, are extreme conditions, which appear in particular and specific contexts such as, for example, fraud detection. This issue surely deserves further analysis which will be part of a future development of this paper.

## ACKNOWLEDGMENTS

The authors wish to thank the two reviewers for their constructive and valuable comments, which have enhanced the quality of the paper.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in UCI Machine Learning Repository at https://archive.ics.uci.edu/ml/index.php. These data were derived from the following resources available in the public domain: - Wine Quality Data Set, https://archive.ics.uci.edu/ml/datasets/wine+quality - Heart Disease Data Set, https://archive.ics.uci.edu/ml/datasets/heart+disease

## ORCID

*Maurizio Carpita* https://orcid.org/0000-0001-7998-5102
*Silvia Golia* https://orcid.org/0000-0003-0015-8126

## REFERENCES

1. A. Agresti, *Analysis of ordinal categorical data*, 2nd ed., John Wiley & Sons, Hoboken, New Jersey, 2010.
2. C. M. Bishop, *Pattern recognition and machine learning*, Springer, New York, 2006.
3. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, CRC Press LLC, Boca Raton, 1984.
4. M. Carpita, E. Ciavolino, and P. Pasca, *Exploring and modelling team performances of the kaggle european soccer database*, Stat. Model. 19 (2019), 74–101.
5. M. Carpita and S. Golia, *Discovering associations between players' performance indicators and matches' results in the european soccer leagues*, J. Appl. Stat. 48 (2021), no. 9, 1696–1711.
6. N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, *Smote: Synthetic minority over-sampling technique*, J. Artif. Intell. Res. 16 (2002), 321–357.
7. T. Chen and G. Carlos, "*Xgboost: A scalable tree boosting system*," *KDD '16: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, Association for Computing Machinery, New York, NY, 2016, pp. 785–794.
8. J. Cohen, *Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit*, Psychol. Bull. 70 (1968), no. 4, 213–220.
9. P. Cortez, "*Data mining with neural networks and support vector machines using the R/rminer tool*," Advances in data mining—Applications and theoretical aspects, 10th industrial conference on data mining. *LNAI*, Vol 6171, P. Perner (ed.), Springer, Berlin, Germany, 2010, pp. 572–583.
10. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, *Modeling wine preferences by data mining from physicochemical properties*, Decis. Support. Syst. 47 (2009), 547–553.
11. P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, *Wine quality*, UCI Machine Learning Repository, 2009. https://archive.ics.uci.edu/ml/index.php.
12. J. S. Cramer, *Predictive performance of the binary logit model in unbalanced samples*, Statistician 48 (1999), no. 1, 85–94.
13. M. Deng, Y. Guo, C. Wang, and F. Wu, *An oversampling method for multi-class imbalanced data based on composite weights*, PLoS One 16 (2021), e0259227.
14. A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*, Springer Nature, Switzerland AG, 2018.
15. A. Fernández, V. López, M. Galar, M. del Jesus, and F. Herrera, *Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches*, Knowl.-Based Syst. 42 (2013), 97–110.
16. S. Golia and M. Carpita, "*On classifiers to predict soccer match results*," *ASMOD 2018 proceedings of the international conference on advances in statistical modelling of ordinal data*, S. Capecchi, F. D. Iorio, and R. Simone (eds.), Federico II Open Access University Press, Napoli, 2018, pp. 125–132.
17. S. Golia and M. Carpita, "*Comparing classifiers for ordinal variables*," *Book of short papers SIS 2020*, A. Pollice, N. Salvati, and F. S. Spagnolo (eds.), Pearson Publishing, London, 2020, pp. 1160–1165.
18. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning with applications in R*, Springer, New York, 2013.
19. A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, *Heart disease*, UCI Machine Learning Repository, 1988. https://archive.ics.uci.edu/ml/index.php.
20. J. Opitz and S. Burst, Macro F1 and macro F1, 2019. http://arxiv.org/abs/1911.03347.
21. A. Orriols-Puig and E. Bernadó-Mansilla, *Evolutionary rule-based systems for imbalanced data sets*, Soft. Comput. 13 (2009), no. 3, 213–225.
22. M. Sokolova and G. Lapalme, *A systematic analysis of performance measures for classification tasks*, Inf. Process. Manag. 45 (2009), no. 4, 427–437.

23. E. Strumbelj, *On determining probability forecasts from betting odds*, Int. J. Forecast. 30 (2014), 934–943.
24. E. Strumbelj and M. Sikonja, *Online bookmakers' odds as forecasts: The case of european soccer leagues*, Int. J. Forecast. 26 (2010), 482–488.
25. J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, *Boosting methods for multi-class imbalanced data classification: An experimental review*, J. Big Data 7 (2020), 1–47.
26. S. Wang and X. Yao, *Multiclass imbalance problems: Analysis and potential solutions*, IEEE Trans. Syst. Man Cybern.. Part B Cybern. 42 (2012), 1119–1130.
27. X. Zeng and T. R. Martinez, *Distribution-balanced stratified cross-validation for accuracy estimation*, J. Exp. Theor. Artif. Intell. 12 (2000), no. 1, 1–12.

**How to cite this article:** M. Carpita and S. Golia, *Categorical classifiers in multiclass classification with imbalanced datasets*, Stat. Anal. Data Min.: ASA Data Sci. J. (2023), 1–15. https://doi.org/10.1002/sam.11624

## APPENDIX A

See Figures A1–A3 and Table A1.

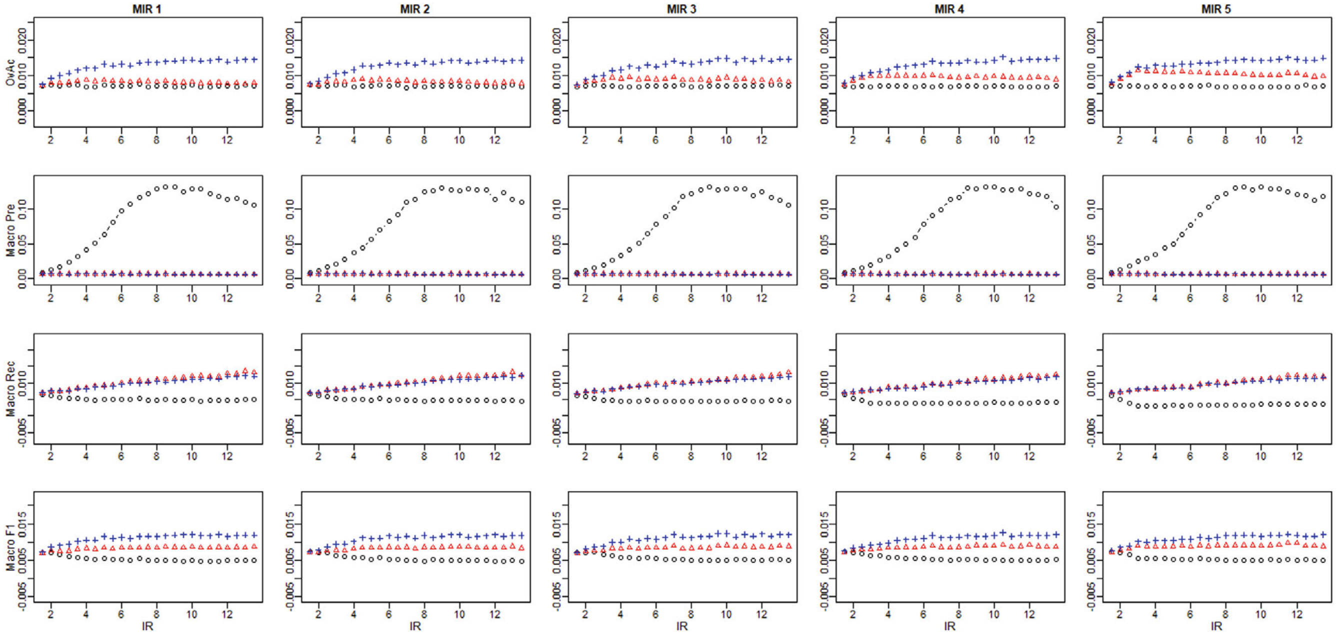

**FIGURE A1** Simulation results for three classes: standard deviation for OvAc and Macroaveraged Pre, Rec and F1 by IR and MIR for the three classifiers (BC = ○; MDC = Δ; MRC = +).

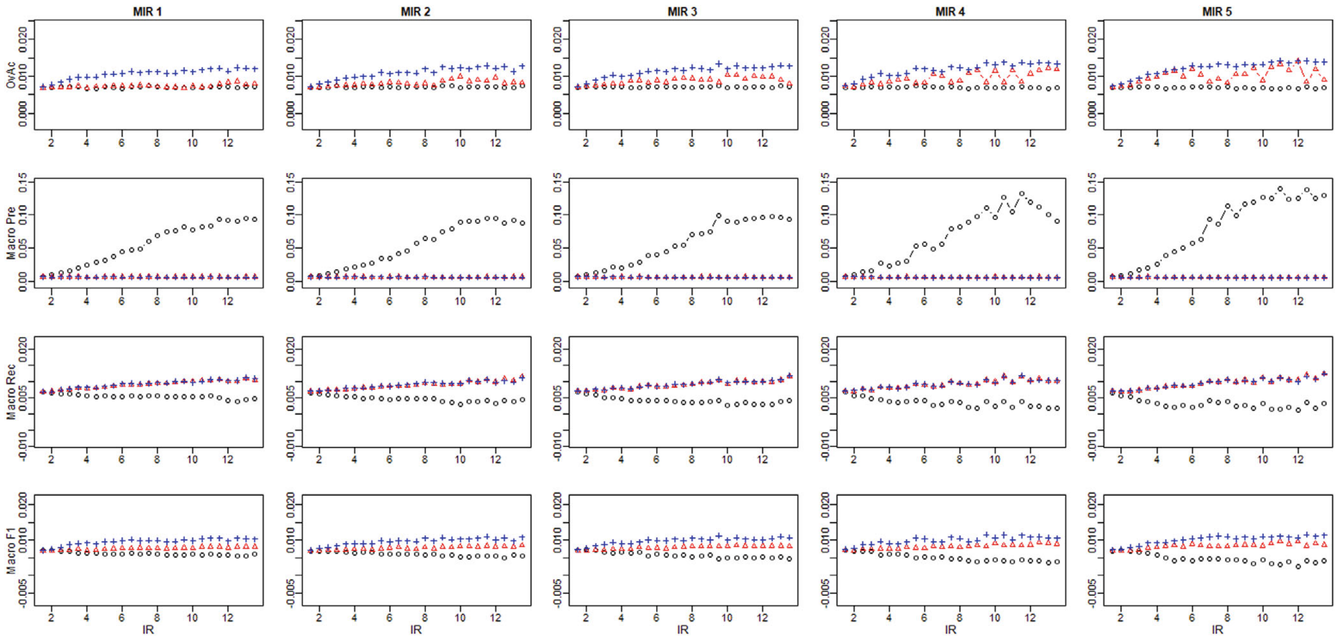

**FIGURE A2** Simulation results for four classes: standard deviation for OvAc and Macroaveraged Pre, Rec and F1 by IR and MIR for the three classifiers (BC = ○; MDC = Δ; MRC = +).

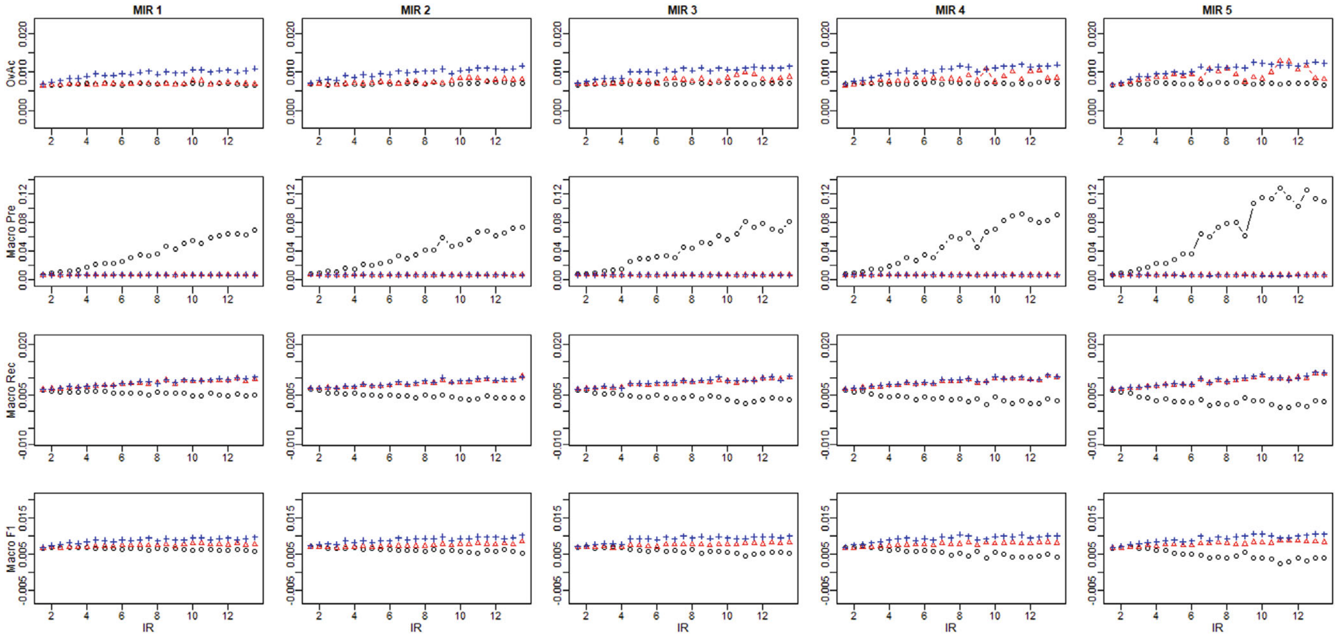

**FIGURE A3** Simulation results for five classes: standard deviation for OvAc and Macroaveraged Pre, Rec and F1 by IR and MIR for the three classifiers (BC = ○; MDC = Δ; MRC = +).

**TABLE A1** Classifier performance indices (per-class Pre, Rec and F1, and MDB) for the real case studies.

| | Soccer results | | | Red wines | | | White wines | | | Heart disease | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **BC** | **MDC** | **MRC** | **BC** | **MDC** | **MRC** | **BC** | **MDC** | **MRC** | **BC** | **MDC** | **MRC** |
| $Pre_1$ | 0.489 | 0.439 | 0.439 | 0.000 | 0.250 | 0.083 | 0.545 | 0.317 | 0.116 | 0.813 | 0.891 | 0.906 |
| $Pre_2$ | 0.536 | 0.312 | 0.311 | 0.648 | 0.652 | 0.595 | 0.545 | 0.513 | 0.441 | 0.322 | 0.332 | 0.318 |
| $Pre_3$ | 0.552 | 0.632 | 0.646 | 0.532 | 0.528 | 0.518 | 0.516 | 0.543 | 0.549 | 0.254 | 0.224 | 0.224 |
| $Pre_4$ | - | - | - | 0.603 | 0.471 | 0.409 | 0.446 | 0.372 | 0.304 | 0.336 | 0.338 | 0.363 |
| $Pre_5$ | - | - | - | - | - | - | 0.000 | 0.111 | 0.111 | 0.391 | 0.361 | 0.197 |
| $Rec_1$ | 0.491 | 0.604 | 0.604 | 0.000 | 0.048 | 0.444 | 0.033 | 0.071 | 0.470 | 0.934 | 0.805 | 0.744 |
| $Rec_2$ | 0.019 | 0.282 | 0.309 | 0.724 | 0.700 | 0.351 | 0.499 | 0.671 | 0.391 | 0.289 | 0.485 | 0.478 |
| $Rec_3$ | 0.852 | 0.528 | 0.507 | 0.600 | 0.511 | 0.386 | 0.727 | 0.431 | 0.328 | 0.091 | 0.137 | 0.206 |
| $Rec_4$ | - | - | - | 0.323 | 0.516 | 0.728 | 0.232 | 0.506 | 0.285 | 0.457 | 0.434 | 0.257 |
| $Rec_5$ | - | - | - | - | - | - | 0.000 | 0.006 | 0.444 | 0.138 | 0.200 | 0.415 |
| $F1_1$ | 0.490 | 0.508 | 0.508 | 0.000 | 0.080 | 0.140 | 0.062 | 0.116 | 0.186 | 0.869 | 0.846 | 0.817 |
| $F1_2$ | 0.036 | 0.296 | 0.310 | 0.684 | 0.675 | 0.441 | 0.521 | 0.581 | 0.415 | 0.305 | 0.395 | 0.382 |
| $F1_3$ | 0.670 | 0.575 | 0.568 | 0.564 | 0.520 | 0.442 | 0.604 | 0.480 | 0.411 | 0.134 | 0.170 | 0.214 |
| $F1_4$ | - | - | - | 0.420 | 0.492 | 0.524 | 0.305 | 0.429 | 0.294 | 0.387 | 0.380 | 0.301 |
| $F1_5$ | - | - | - | - | - | - | 0.000 | 0.011 | 0.177 | 0.205 | 0.257 | 0.267 |
| MDB Pre | 0.063 | 0.320 | 0.335 | 0.648 | 0.402 | 0.511 | 0.545 | 0.432 | 0.438 | 0.559 | 0.666 | 0.709 |
| MDB Rec | 0.833 | 0.323 | 0.295 | 0.724 | 0.653 | 0.377 | 0.727 | 0.666 | 0.185 | 0.842 | 0.668 | 0.538 |
| MDB F1 | 0.634 | 0.279 | 0.258 | 0.684 | 0.595 | 0.384 | 0.604 | 0.571 | 0.237 | 0.735 | 0.675 | 0.602 |