7th ICML Workshop on Automated Machine Learning (2020)

# Analysis of Imbalance Strategies Recommendation using a Meta-Learning Approach

**Afonso José Costa** AJCOSTA@STUDENT.DEI.UC.PT
*CISUC, Department of Informatics Engineering, University of Coimbra, Portugal*

**Miriam Seoane Santos** MIRIAMS@DEI.UC.PT
*CISUC, Department of Informatics Engineering, University of Coimbra, Portugal*

**Carlos Soares** CSOARES@FE.UP.PT
*Fraunhofer AICOS and LIACC, Faculty of Engineering, University of Porto, Portugal*

**Pedro Henriques Abreu** PHA@DEI.UC.PT
*CISUC, Department of Informatics Engineering, University of Coimbra, Portugal*

## Abstract

Class imbalance is known to degrade the performance of classifiers. To address this issue, imbalance strategies, such as oversampling or undersampling, are often employed to balance the class distribution of datasets. However, although several strategies have been proposed throughout the years, there is no one-fits-all solution for imbalanced datasets. In this context, meta-learning arises a way to recommend proper strategies to handle class imbalance, based on data characteristics. Nonetheless, in related work, recommendation-based systems do not focus on how the recommendation process is conducted, thus lacking interpretable knowledge. In this paper, several meta-characteristics are identified in order to provide interpretable knowledge to guide the selection of imbalance strategies, using Exceptional Preferences Mining. The experimental setup considers 163 real-world imbalanced datasets, preprocessed with 9 well-known resampling algorithms. Our findings elaborate on the identification of meta-characteristics that demonstrate when using preprocessing algorithms is advantageous for the problem and when maintaining the original dataset benefits classification performance.

## 1. Introduction

Class imbalance is known to severely affect the data quality. Concerning imbalanced binary datasets, they comprise a majority (more represented) and a minority (underrepresented) class. The classifiers trained on asymmetrical distributions are often biased towards the majority class since the minority one is under-represented (López et al., 2013). Hence, the classifier is not able to generalise, leading to an increase of the misclassification rate of the minority class (López et al., 2013). Regarding this matter, several strategies, usually involving preprocessing algorithms, have been proposed in the literature to overcome this issue. In particular, data-level strategies are the most commonly used, due to its simplicity, efficiency and classifier-independence (Santos et al., 2018). These resampling techniques are characterised by altering the distribution of the training set, generating new synthetic instances of the minority class (oversampling) or removing samples from the majority class (undersampling). The oversampling strategies benefit from not discarding points from the majority class when compared with undersampling algorithms, which can lead to neglecting important concepts of the dataset (Santos et al., 2015). Therefore, we consider only over-

sampling algorithms, including the ones extended with data-cleansing techniques, such as Tomek-Links (TL) or Edited Nearest Neighbours (ENN). However, it has been reported that no imbalance strategy is suitable for all problems (Zhang et al., 2019). Instead of employing "brute-force" approaches (experimenting with all techniques) (De Morais et al., 2017), the recommendation of imbalanced strategies, based on the meta-characteristics of the dataset, is a research topic that aims for the automatic selection of preprocessing algorithms, using meta-learning approaches (Zhang et al., 2019; De Morais et al., 2017). However, these recommendations do not allow understanding the behaviour of such techniques, since the meta-learner only outputs the recommended imbalance strategy, without providing meaningful information concerning the meta-characteristics of the dataset. Accordingly, as recent literature has come to acknowledge, extracting knowledge from the produced recommendations becomes a fundamental aspect to fully understand the relationship between data characteristics and the success of preprocessing techniques (Loyola-González et al., 2016). Hence, the goal of this paper is to elaborate on a methodology to extract knowledge regarding the behaviour of oversampling algorithms. We investigate the relation between the classification performance of each resampling strategy and the characteristics of datasets, identifying scenarios where some strategies are more advantageous or where dismissing any preprocessing can be beneficial. To this end, two research questions were formulated: 1) ***What are the scenarios where not dealing with the imbalance of classes is beneficial?*** and 2) ***Which relations exist between dataset characteristics and the optimal preprocessing algorithm?***.

To answer these questions, Exceptional Preferences Mining (EPM) (de Sá et al., 2016) was employed to extract interpretable rules. EPM is a data mining framework that aims at finding interesting rules from subgroups of the dataset, where the "interest" is concerned with a target attribute. A subgroup is deemed "exceptional" if the label ranking of the subgroup is significantly different than the label ranking of the dataset (de Sá et al., 2018).

## 2. Related Work

In this section, an overview of related research on the topic of the recommendation of imbalance strategies is provided. Loyola-González et al. (2016) studied the effect of resampling strategies associated with different classifiers, on 95 real-world datasets, using Contrast Pattern Miners (CPM). In short, a *contrast pattern* is a descriptive expression, for instance, $[SepalWidth \leq 3.7]$, that appears frequently in a class and rarely in the remaining classes of the dataset (Loyola-González et al., 2016). The preprocessing strategies employed were both oversampling and undersampling algorithms. Backed by their findings, they proposed an empirical recommendation of resampling algorithms, based on 6-bins discretization of the Imbalance Ratio (IR). They concluded that SMOTE, Tomek Links and SMOTE-TL are the top performing approaches. Furthermore, the authors refer that a knowledge-seeking meta-analysis could bring new insights about the resampling algorithms' behaviour and it would be beneficial to aid researchers when selecting a resampling strategy, based on the meta-characteristics of the dataset.

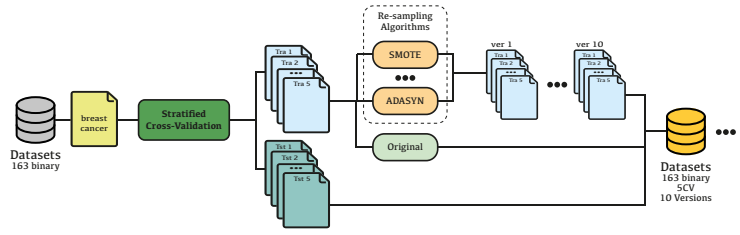
De Morais et al. (2017) and Zhang et al. (2019) propose recommendation systems based on a meta-learning approach, to provide the user with a preprocessing algorithm, along with its optimal hyperparameters. The recommendation is inferred from a meta-database, com-

posed by the training datasets' meta-features and the performance associated with imbalance strategies. For each new test dataset, the recommended algorithm is the one assigned to the closest training instance (each instance represents a dataset). The recommendation for this test instance is computed based on the similarity between the meta-characteristics of the test and training instances, using the $k$-Nearest Neighbours ($k$-NN) algorithm. The former work utilises meta-features from Simple and Statistical groups (Rivolli et al., 2018) and 7 under-sampling techniques, on 29 real-world datasets, whereas the latter uses meta-features from the Simple, Statistical, Complexity, Landmarking and Model-based groups and a more complete set of imbalance strategies, including algorithms from both data-level and algorithmic-level domains (Stefanowski, 2015), on 80 real-world datasets. In the end, the authors agree that there is no preprocessing algorithm that suits all scenarios.

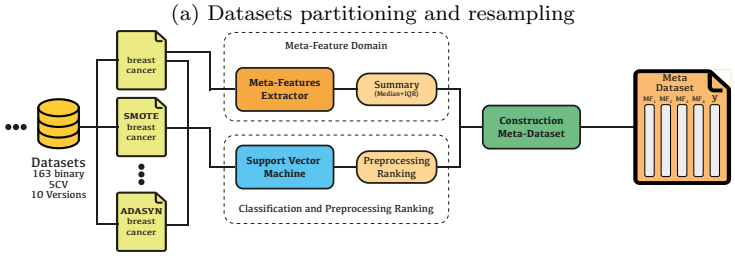
Overall, it is observable that related research handles the recommendation of imbalance strategies with meta-learning approaches, based on the meta-characteristics of the datasets. However, they do not provide any general knowledge about the scenarios of application of preprocessing algorithms, nor how the behaviour of imbalance strategies can be related to meta-characteristics. An analogy can be established with black-box models, due to the impossibility of understanding how the recommendation process is conducted. From the authors' knowledge, there are no other works that address these important questions.

## 3. Experimental Setup

In this work, a collection of 163 real-world binary datasets was retrieved from the UCI, Kaggle, OpenML and KEEL repositories, containing numerical and categorical attributes, where the latter was integer encoded from 0 to $m - 1$, where $m$ stands for the number of unique discrete values within each feature. The experimental setup can be divided into three phases: – 1) Partitioning and resampling (Figure 1a); – 2) Meta-features extraction and performance evaluation (Figure 1b) and – 3) Exceptional Preferences Mining.




(a) Datasets partitioning and resampling




(b) Performance evaluation and meta-feature extraction.

Figure 1: Experimental setup.

Concerning the first phase, the datasets were partitioned in 5 folds (stratified CV). The selected state-of-the-art oversampling techniques are: ROS, SMOTE, SafeLevel-SMOTE, Borderline-SMOTE, ADASYN, AHC, ADOMS, SMOTE-TL and SMOTE-ENN (Santos et al., 2018), which are implemented in the KEEL framework (Alcalá-Fdez et al., 2009). The resampling procedures were run 10 times (for each dataset), due to the stochastic processes associated with resampling techniques.

Regarding the second phase, the performance ($F1$-measure) of the Support Vectors Machine (SVM) classifier was evaluated on the 10 versions of each imbalance strategy, and the original dataset, considering the hyperparameters tuned for the original (non-resampled) dataset. Next, the performance on the 10 versions of each resampling algorithm was summarized using the median and interquartile range. The median $F1$-measure respecting each label (9 preprocessing algorithms and original dataset) was ranked, originating a ground-truth preference ranking of preprocessing algorithms (including the original dataset). For instance, the preference relation can be represented as (de Sá et al., 2016): SMOTE $\succ$ ADASYN $\succ \cdots \succ$ ORIGINAL.

Concerning meta-feature extraction (also included on the second phase), the open-source python *pymfe*[1] (Rivolli et al., 2018) library was chosen and the extraction took place only on the original (not resampled) datasets. All meta-features available on the library were extracted, plus the custom-implemented typology of minority instances (Napierala and Stefanowski, 2016). Afterwards, the meta-dataset is constructed from the mean meta-features of the 10 versions and the ground-truth rankings. An analogy can be established with conventional datasets, where the features are represented by the meta-features, the target attribute is a ranking of preprocessing algorithms and the patterns are the 163 datasets. To illustrate, $\mathbf{x}_{\text{meta}} = (MF_1, MF_2, \ldots, MF_Q, y)$, represents a meta-dataset pattern, where the attributes are the extracted meta-features, $Q$ is the number of meta-features and $y$ stands for the ranking of preprocessing algorithms, or a preference relation.

Finally, the EPM implementation of the Cortana Subgroup Discovery Tool[2] was utilised, with *on-the-fly* 8 bins discretization and a beam-search strategy (de Sá et al., 2016). The subgroups shown have a depth of 1 and undergone a Distribution of False Discoveries (DFD) validation (Duivesteijn and Knobbe, 2011), at a significance level $\alpha = 1\%$. The exceptional subgroups are deemed "exceptional" based on the *labelwise LWNorm* quality measure, which measures exceptional changes in the label ranking, from the perspective of individual labels (de Sá et al., 2018).

## 4. Results

Concerning the results of the first research question experiment (Section 4.1) the subgroups were extracted with the ranking of 10 labels (including the original dataset), whereas for second experiment results (Section 4.2) only the labels of the 9 preprocessing algorithms were utilised. The most relevant subgroups are shown on appendix Tables 1 and 3. Note that the complete set of exceptional subgroups is not shown due to its extent. We included only the most relevant ones for the descriptive analysis of the results.

1. *pymfe* library repository: `https://github.com/ealcobaca/pymfe`
2. Cortana website: `http://datamining.liacs.nl/cortana.html`

### 4.1 *What are the scenarios where not dealing with the imbalance of classes is beneficial?*

The motivation of this topic is to infer, in an imbalanced context, the meta-characteristics that indicate that it may not be necessary to preprocess such dataset. It is shown that **simpler classification tasks may not require preprocessing**, which is illustrated by the solid performance of landmarkers (simple and fast learning algorithms that characterise the dataset (Rivolli et al., 2018). In these cases, it was evident that the original imbalanced dataset scored the first rank. Furthermore, when the overall **complexity of the dataset is reduced** (complexity of the decision surface or dimensionality) using the original dataset is also the best option. The complexity meta-features (Lorena et al., 2019), scored low values for *L2* (error rate of a linear classifier), *N1* (fraction of borderline points), *N4* (non-linearity of NN classifier) and *T3* (average number of PCA dimensions per points).

Regarding the statistical meta-features, there is evidence that not performing resampling benefits classification performance, if the **data distribution has low variance**. This is corroborated by the low variance, covariance and first eigenvalue of the covariance matrix. Still concerning statistical properties, leptokurtic (positive kurtosis) and positive skewness are other distribution characteristics that favour maintaining the dataset imbalanced.

There are also some findings worth highlighting, concerning the typology of minority class instances. There is evidence that when a **high proportion of safe instances and a small amount of borderline instances** is present, it is also favourable to maintain the dataset imbalanced.

Conversely, there are some situations where the exceptional subgroups favoured the cases where resampling was employed. For instance, preprocessing is beneficial if the dataset is of **high dimensionality**, which is captured by the increase of *T2* (average number of features per dimension) and *T3* complexity measures. Also, it is observable that when the **number of borderline instances is elevated**, preprocessing needs to be performed, otherwise strong performance degradation is observable. These findings are summarized on Table 2.

### 4.2 *Which relations exist between dataset characteristics and the optimal preprocessing algorithm?*

The goal of this research question is to highlight the behaviour of the meta-features that evidence the use of a determined imbalance strategy. It is worth noting that some algorithms do not appear in any interesting subgroups if the ranking does not shift significantly from the average ranking or the subgroup's coverage (the number of patterns included on the subgroup) is reduced (de Sá et al., 2018).

**AHC** There is evidence that this algorithm is more suitable when presented with **less complex problems with reduced dimensionality**. Since one limitation of Hierarchical Clustering algorithms is that the performance is severely degraded in high dimensional feature spaces, it is expected that this algorithm would only be suitable for datasets with low dimensionality. This is corroborated by the values inferior than 0.1 of complexity measures *T3*, *L2*, *N1* and *N4* (except for $N1$ which indicates a value smaller than 0.1603), which depicts that both lower dimensionality of the problem and simpler decision boundaries favour this algorithm. Moreover, this strategy is also suitable when there is over 28% of rare points and/or over 73% of safe points. On the other hand, a low percentage (smaller

than 8.5%) of borderline instances has to be guaranteed otherwise, loss of performance is expectable.

**SMOTE-TL** It is the most suitable algorithm for **harder classification tasks and high dimensional datasets**. This is demonstrated by the fact that this algorithm scored the highest ranks when the landmarker meta-features scored low accuracies and higher *T2*. Furthermore, it is also applicable when there is a high amount of borderline instances (over 62%). This agrees with the Tomek Links cleaning procedure since it aims at removing the borderline samples, which are classified as Tomek Links, thus reducing the complexity of the decision surface, at the borderline regions (Batista et al., 2004).

**ADOMS** This algorithm was preferred when **the subgroup elements have low variance and small first principal component of the covariance matrix**. It consists of generating a new SMOTE-like instance along the line between the minority instance and the projection of the chosen neighbour, onto the first principal component (Tang and Chen, 2008). Even though the first principal component's direction is chosen, which explains the highest amount of variance of the dataset, it is observable that this algorithm seems to be only favourable when the overall variability of the training data is reduced.

**ROS** Random oversampling showed to be **more suitable when the attributes entropy is high**. The entropy is a measure of randomness in a variable (Castiello et al., 2005) and can be informative of the attributes capacity for class discrimination. For instance, if the attributes entropy is elevated, it indicates that the discriminatory power is significant (Rivolli et al., 2018). One possible explanation is that since there is higher redundancy on the data, algorithms that lack heuristics might be more suitable. Furthermore, since the discriminatory power is high, the remaining algorithms may degrade the performance since the generation of synthetic instances may diminish the discriminatory power (this is known as the problem of over-generalization for SMOTE-like approaches (Santos et al., 2018)). On the other hand, ROS randomly replicates minority class instances and no further information is added to the training data (Santos et al., 2018), therefore the discriminatory power is maintained.

## 5. Conclusion

In this paper, several meta-characteristics are identified which are suitable for guiding the selection of imbalance strategies, using the EPM framework. The results agree with related works, that have stated that simpler classification tasks may not require preprocessing, despite the imbalance degree (Jo and Japkowicz, 2004; Prati et al., 2004). Furthermore, it is observable that when preprocessing should be performed, AHC is not robust for complex learning tasks when compared with SMOTE-TL. The ADOMS algorithm works optimally when there is low variance and ROS is the best option when attributes entropy is high. These insights can be useful for the creation or enhancement of recommendation systems. Some directions for future work are integrating more preprocessing algorithms, such as the ones from the algorithmic-level, or provide a deeper understanding of classification differences among ranks. Rankings have the advantage of abstracting from the true values of performance but it may also be important to investigate the scenarios where steep variations on performance are observable and correlate these cases with the exceptional subgroups.

## Acknowledgments

This paper is a result of the project Safe Cities - Inovação para Construir Cidades Seguras, with the reference POCI-01-0247-FEDER-041435, co-funded by the European Regional Development Fund (ERDF), through the Operational Programme for Competitiveness and Internationalization (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement.

## References

J Alcalá-Fdez, L Sánchez, S García, M J del Jesus, S Ventura, J M Garrell, J Otero, C Romero, J Bacardit, V M Rivas, J C Fernández, and F Herrera. KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3): 307–318, 2009. ISSN 1433-7479. doi: 10.1007/s00500-008-0323-y.

Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20, 2004. ISSN 19310145. doi: 10.1145/1007730.1007735.

Ciro Castiello, Giovanna Castellano, and Anna Maria Fanelli. Meta-data: Characterization of input features for meta-learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3558 LNAI:457–468, 2005. ISSN 03029743. doi: 10.1007/11526018_45.

Romero F.A.B. De Morais, Pericles B.C. Miranda, and Ricardo M.A. Silva. A Meta-Learning Method to Select Under-Sampling Algorithms for Imbalanced Data Sets. *Proceedings - 2016 5th Brazilian Conference on Intelligent Systems, BRACIS 2016*, pages 385–390, 2017. doi: 10.1109/BRACIS.2016.076.

Cláudio de Sá, Wouter Duivesteijn, Carlos Soares, and Arno Knobbe. Exceptional Preferences Mining. In Toon Calders, Michelangelo Ceci, and Donato Malerba, editors, *Discovery Science*, pages 3–18, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46307-0.

Cláudio Rebelo de Sá, Wouter Duivesteijn, Paulo Azevedo, Alípio Mário Jorge, Carlos Soares, and Arno Knobbe. Discovering a taste for the unusual: exceptional models for preference mining. *Machine Learning*, 107(11):1775–1807, 2018. ISSN 15730565. doi: 10.1007/s10994-018-5743-z.

Wouter Duivesteijn and Arno Knobbe. Exploiting false discoveries - Statistical validation of patterns and quality measures in subgroup discovery. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 151–160, 2011. ISSN 15504786. doi: 10.1109/ICDM.2011.65.

Taeho Jo and Nathalie Japkowicz. Class Imbalances versus Small Disjuncts. *Sigkdd Explorations*, 6(1):40–49, 2004.

Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends

on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 2013. ISSN 00200255. doi: 10.1016/j.ins.2013.07.007.

Ana C. Lorena, Luís P.F. Garcia, Jens Lehmann, Marcilio C.P. Souto, and Tin K.A.M. Ho. How complex is your classification problem?: A survey on measuring classification complexity. *ACM Computing Surveys*, 52(5), 2019. ISSN 15577341. doi: 10.1145/3347711.

Octavio Loyola-González, José Fco Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, and Milton García-Borroto. Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing*, 175:935–947, 2016. ISSN 18728286. doi: 10.1016/j.neucom.2015.04.120.

Krystyna Napierala and Jerzy Stefanowski. Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, 46(3):563–597, 2016. ISSN 15737675. doi: 10.1007/s10844-015-0368-1.

Ronaldo C. Prati, Gustavo E.A.P.A. Batista, and Maria C. Monard. Class imbalances versus class overlapping: An analysis of a learning system behavior. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 2972:312–321, 2004. ISSN 03029743. doi: 10.1007/978-3-540-24694-7_32.

Adriano Rivolli, Luís P. F. Garcia, Carlos Soares, Joaquin Vanschoren, and André C. P. L. F. de Carvalho. Characterizing classification datasets: a study of meta-features for meta-learning. *CoRR*, 2018.

Miriam Seoane Santos, Pedro Henriques Abreu, Pedro J. García-Laencina, Adélia Simão, and Armando Carvalho. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of Biomedical Informatics*, 58: 49–59, 2015. ISSN 15320464. doi: 10.1016/j.jbi.2015.09.012.

Miriam Seoane Santos, Jastin Pompeu Soares, Pedro Henrigues Abreu, Helder Araújo, and João Santos. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches. *IEEE Computational Intelligence Magazine*, 13(4):59–76, 2018. ISSN 15566048. doi: 10.1109/MCI.2018.2866730.

Jerzy Stefanowski. Dealing with Data Difficulty Factors While Learning from Imbalanced Data. *Challenges in Computational Statistics and Data Mining*, 605:333–363, 2015. doi: 10.1007/978-3-319-18781-5.

Sheng Tang and Si Ping Chen. The generation mechanism of synthetic minority class examples. *5th Int. Conference on Information Technology and Applications in Biomedicine, ITAB 2008 in conjunction with 2nd Int. Symposium and Summer School on Biomedical and Health Engineering, IS3BHE 2008*, pages 444–447, 2008. doi: 10.1109/ITAB.2008.4570642.

Xueying Zhang, Ruixian Li, Bo Zhang, Yunxiang Yang, Jing Guo, and Xiang Ji. An instance-based learning recommendation algorithm of imbalance handling methods. *Applied Mathematics and Computation*, 351:204–218, 2019. ISSN 00963003. doi: 10.1016/j.amc.2018.12.020.

## Appendix A. Exceptional Preferences Mining Results

This appendix contains the exceptional subgroups computed by the EPM framework, using the *labelwise LWNorm* (de Sá et al., 2018) as the quality measure. Additionally, a summary of the cases when preprocessing should or should not be conducted is also provided, concerning the first experiment (Section 4.1). The column *coverage* on Tables 1 and 3 stands for the number of datasets included in the subgroup. The preprocessing algorithms were encoded with the letters *a*-*j* as follows:

- *a*: ADASYN
- *b*: ADOMS
- *c*: AHC
- *d*: Borderline-SMOTE
- *e*: ROS
- *f*: SMOTE
- *g*: SMOTE-ENN
- *h*: SMOTE-TL
- *i*: SafeLevel-SMOTE
- *j*: Original

| Coverage | LWNorm ($\times 10^{-2}$) | Ranking | Conditions |
|---|---|---|---|
| No preproc. | | | |
| 21 | 3.3992 | j>c>d>e>bf>i>a>h>g | *statistical_kurtosis* >=17.9168 |
| 21 | 2.8934 | j>b>d>e>f>c>i>a>g>h | *statistical_cov* <= 0.0234 |
| 21 | 2.7911 | j>b>d>c>f>e>a>i>g>h | *statistical_eigenvalues* <= 0.2581 |
| 21 | 2.7911 | j>b>d>c>f>e>a>i>g>h | *statistical_var* <= 0.2581 |
| 42 | 2.3839 | j>c>d>ef>b>i>h>a>g | *general_nr_inst* >= 376.0 |
| 21 | 2.8448 | j>c>e>d>f>b>h>i>g>a | *complexity_n1* <= 0.0675 |
| 41 | 2.8327 | j>c>e>d>b>f>h>i>g>a | *complexity_l2* <= 0.0421 |
| 24 | 2.6046 | j>c>d>e>f>b>i>h>a>g | *complexity_t3* <= 0.0031 |
| 41 | 2.2675 | j>c>e>d>b>h>f>gi>a | *complexity_n4* <= 0.0611 |
| 21 | 2.4585 | j>c>d>e>f>b>h>i>g>a | *taxonomy_border* <= 0.0858 |
| 41 | 2.2161 | j>c>d>b>e>f>h>i>g>a | *taxonomy_safe* >= 0.5334 |
| 41 | 2.9260 | j>c>e>d>b>f>h>i>g>a | *landmarking_linear_discr* >= 0.9225 |
| 21 | 2.7286 | j>e>c>d>b>f>h>i>g>a | *landmarking_nn* >= 0.9750 |
| 41 | 2.7001 | j>c>e>d>b>f>h>i>a>g | *landmarking_nn* >= 0.9052 |
| Do preproc. | | | |
| 21 | 3.0865 | h>c>d>f>ae>i>g>b>j | *statistical_kurtosis* <= -1.3063 |
| 21 | 2.2963 | h>a>ci>b>d>f>j>e>g | *statistical_sparsity* >= 0.4085 |
| 21 | 2.7049 | h>c>bg>i>d>f>a>e>j | *taxonomy_border* >= 0.6555 |
| 21 | 2.4990 | h>a>d>b>c>e>g>fi>j | *complexity_t3* >= 0.0668 |
| 22 | 2.2101 | h>a>g>c>f>e>d>i>b>j | *complexity_t2* >= 0.1250 |
| 41 | 2.1513 | c>h>f>b>a>d>e>i>j>g | *complexity_f3* >= 0.9831 |
| 41 | 2.1513 | c>h>f>b>a>d>e>i>j>g | *complexity_f4* >= 0.9831 |
| 21 | 2.3544 | h>c>b>a>f>d>ei>g>j | *landmarking_elite_nn* <= 0.5788 |
| 21 | 2.9550 | h>c>a>b>f>dg>i>e>j | *landmarking_best_node* <= 0.6557 |

Table 1: Exceptional subgroups reporting to the first experiment (Section 4.1).

| **Keep Dataset Imbalanced** | **Preprocessing** |
|---|---|
| • Low complexity of dataset shape<br>• Easy classification tasks<br>• Significant number of instances<br>• Low variance, leptokurtic and positively skewed distributions<br>• High ratio of safe instances<br>• Low ratio of borderline instances | • Dimensionality increases<br>• Classification difficulty increases<br>• Platykurtic distibution<br>• High fraction of borderline instances |

Table 2: Guidelines indicating when the dataset should be kept imbalanced versus employing preprocessing, concerning the first experiment (Section 4.1).

| Coverage | LWNorm ($\times 10^{-2}$) | Ranking | Conditions |
|---|---|---|---|
| AHC (c) | | | |
| 24 | 2.7773 | c>d>e>f>b>i>h>a>g | *complexity_t3* <= 0.0031 |
| 41 | 2.7519 | c>e>d>b>f>h>i>g>a | *complexity_l2* <= 0.0421 |
| 41 | 2.6012 | c>e>d>b>f>h>i>g>a | *complexity_n1* <= 0.1603 |
| 41 | 2.3660 | c>e>d>b>h>f>gi>a | *complexity_n4* <= 0.0611 |
| 41 | 2.7980 | c>e>d>b>f>h>i>g>a | *landmarking_linear_discr* >= 0.9225 |
| 22 | 2.4826 | c>d>e>b>f>h>gi>a | *landmarking_linear_discr* >= 0.9634 |
| 21 | 3.0606 | c>de>bf>i>a>h>g | *statistical_kurtosis* >= 17.9168 |
| 21 | 2.6686 | c>bf>d>e>i>a>h>g | *statistical_skewness* >= 2.2407 |
| 21 | 2.4160 | c>f>e>d>bh>i>g>a | *taxonomy_safe* >= 0.7375 |
| 21 | 2.5195 | c>d>e>f>b>h>i>g>a | *taxonomy_border* <= 0.0858 |
| SMOTE-TL (h) | | | |
| 21 | 3.2202 | h>c>d>f>ae>i>g>b | *statistical_kurtosis* <= -1.3063 |
| 21 | 2.5242 | h>c>a>b>f>g>d>i>e | *landmarking_best_node* <= 0.6557 |
| 22 | 2.3170 | h>a>g>c>f>e>d>i>b | *complexity_t2* >= 0.1250 |
| 21 | 2.3748 | h>a>d>b>c>e>g>fi | *complexity_t3* >= 0.0668 |
| 21 | 2.2878 | h>c>g>b>i>d>f>a>e | *taxonomy_border* >= 0.6555 |
| 41 | 2.2846 | h>f>c>a>d>i>e>b>g | *taxonomy_rare* >= 0.2062 |
| ADOMS (b) | | | |
| 21 | 3.0807 | b>d>e>f>c>i>a>g>h | *statistical_cov* <= 0.0234 |
| 21 | 2.9587 | b>d>c>f>e>a>i>g>h | *statistical_eigenvalues* <= 0.2581 |
| 21 | 2.9587 | b>d>c>f>e>a>i>g>h | *statistical_var* <= 0.2581 |
| 21 | 2.6592 | b>d>c>f>e>a>i>g>h | *statistical_sd* <= 0.4629 |
| 21 | 2.4649 | b>c>d>f>a>e>i>h>g | *statistical_mad* <= 0.1955 |
| ROS (e) | | | |
| 21 | 2.4550 | e>c>d>b>f>i>h>g>a | *info-theory_attr_ent* >= 2.5827 |
| 21 | 2.8046 | e>c>d>b>f>h>i>g>a | *landmarking_one_nn* >= 0.9000 |

Table 3: Exceptional subgroups reporting to the second experiment (Section 4.2).