**Davila-Pena, Laura, Saavedra-Nieves, Alejandro and Casas-Méndez, Balbina (2025) *On the influence of dependent features in classification problems: A game-theoreti perspective.* Expert Systems with Applications, 280 . ISSN 0957-4174.**

# On the influence of dependent features in classification problems: A game-theoretic perspective

Laura Davila-Pena, Alejandro Saavedra-Nieves,
Balbina Casas-Méndez

PII: S0957-4174(25)01068-1
DOI: https://doi.org/10.1016/j.eswa.2025.127446
Reference: ESWA 127446

To appear in: *Expert Systems With Applications*

Received date : 5 August 2024
Revised date : 14 February 2025
Accepted date : 25 March 2025

- We propose an influence measure for dependent features on classification problems
- Our measure is axiomatically characterized using cooperative game theory properties
- The introduced influence measure is a generalization of the Banzhaf-Owen value
- We provide a sampling method to approximate the measure in large-scale contexts
- We explore the importance of features in several datasets to validate our proposals

# On the influence of dependent features in classification problems: a game-theoretic perspective

Laura Davila-Pena[1,2], Alejandro Saavedra-Nieves[3], Balbina Casas-Méndez[4]

[1] *Corresponding author. Centre for Logistics and Sustainability Transportation (CeLSA), Department of Analytics, Operations and Systems, Kent Business School, University of Kent, CT2 7PE Canterbury, UK. ORCID: 0000-0003-2175-2546.* `l.davila-pena@kent.ac.uk`

[2] *MODESTYA Research Group, Department of Statistics, Mathematical Analysis and Optimization, Faculty of Mathematics, University of Santiago de Compostela, Campus Vida, 15782 Santiago de Compostela, Spain.* `lauradavila.pena@usc.es`

[3] *CITMAga, MODESTYA Research Group, Department of Statistics, Mathematical Analysis and Optimization, Faculty of Mathematics, University of Santiago de Compostela, Campus Vida, 15782 Santiago de Compostela, Spain. ORCID: 0000-0003-1251-6525.* `alejandro.saavedra.nieves@usc.es`

[4] *CITMAga, MODESTYA Research Group, Department of Statistics, Mathematical Analysis and Optimization, Faculty of Mathematics, University of Santiago de Compostela, Campus Vida, 15782 Santiago de Compostela, Spain. ORCID: 0000-0002-2826-218X.* `balbina.casas.mendez@usc.es`

**Abstract**

This paper deals with a new measure of the influence of each feature on the outcome in classification problems, accounting for potential dependencies among certain feature subsets. Within this framework, we consider a sample of individuals characterized by specific features, each feature encompassing a finite range of values, and classified based on a binary outcome. This measure turns out to be an influence measure explored in existing literature and related to cooperative game theory. We provide an axiomatic characterization of our proposed influence measure by tailoring properties from the cooperative game theory to our specific context. Furthermore, we demonstrate that our influence measure becomes a general characterization of the well-known Banzhaf-Owen value for games with a priori unions, from the perspective of classification problems. The definitions and results presented herein are illustrated through numerical examples and various applications, offering practical insights into our methodologies.

**Keywords:** Classification problems; Influence measure of features; Dependent features; Axiomatic characterization; Banzhaf-Owen value.

## 1 Introduction

Understanding how features influence a binary outcome becomes crucial in classification problems, with applications ranging from medical diagnosis to customer behavior analysis. In this paper, we start from a sample of individuals for which a set of features that take a finite number of values has been measured. These individuals are categorized using a classifier according to an outcome that can also take a finite set of values. Our objective is to provide a model-agnostic measure of the influence of each feature on predicting the outcome in cases where certain subsets of features are dependent. The motivation behind this research, along with its theoretical and practical significance, stems from the absence of a game theoretic-based influence measure in the literature that accounts for feature dependencies and is both experimentally validated and theoretically grounded. We fill this gap by proposing such a measure and providing an axiomatic characterization following the Banzhaf value (Banzhaf III, 1965) approach, a game-theoretical solution concept closely related to the well-known Shapley value (Shapley, 1953). The main difference between the two lies in the theoretical properties they fulfill, rendering one more suitable than the other depending on the nature of the problem under consideration (Feltkamp, 1995). The Shapley value has recently found widespread application in the context of the interpretability of machine learning models (Lundberg and Lee, 2017), and both the Banzhaf and Shapley values rely on players' contributions to certain coalitions. On the other hand, the Banzhaf value has been successfully employed in various contexts to determine rankings. Recent applications include its integration into a control framework, where it offers closed-form expressions as a function of the state and establishes steady-state conditions (Muros et al., 2017). Additionally, it has been used to define a measure of political power in a Spanish autonomous community (Arévalo-Iglesias and Álvarez-Mozos, 2020) and to analyze the risk of a terrorist attack by ranking terrorists within a network (Algaba et al., 2024).

In our case, the proposed measure of feature influence emerges as a generalization of the one studied in Datta et al. (2015) for scenarios with independent features, and is identified with the Banzhaf-Owen value (G. Owen, 1981) for classification problems where both dependent features and the outcome are binary. Although the problem addressed is not inherently novel, this is the first time, to the best of our knowledge, that an influence measure grounded in the Banzhaf-Owen value is proposed, constituting the primary contribution of our work. The Banzhaf-Owen value is a cooperative game theory solution that extends upon the Banzhaf value and is particularly relevant in situations where players exhibit affiliations due to political, economic, or other considerations. In addition, we offer an axiomatic characterization of the influence measure for cases where the outcome is binary, implement it using the software R, and validate its efficacy through a variety of examples and real-world applications.

It is worth emphasizing that solutions for cooperative games based on players' marginal contributions, such as the Banzhaf or the Shapley values and extensions thereof, are useful tools that can be adapted to measure the influence or contribution of a feature in classification and other machine learning contexts. This adaptation is feasible because such a problem closely resembles those addressed by cooperative games in disciplines like economics and political science. On the other hand, in the context of cooperative games, affinities or interactions among agents are modeled using a priori unions. Furthermore, as noted by Datta et al. (2015), a dataset of binary features where every possible combination of these feature values appears with frequency 1 can be identified with a cooperative game. Based on this, we have proposed a generalization of the Banzhaf-Owen value–one of the most widely recognized solutions for games with a priori unions–as an influence measure. The game theory methodology also enables the influence measures derived in this way to be endowed with specific axiomatic characterizations through desirable properties. These properties enhance the interpretation of the influence measure and allow for selecting the most appropriate measure according to the concrete problem at hand. We believe that the axiomatic characterization of our measure is one of the main and innovative contributions of this work. Although characterizations can also be found in Aas et al. (2021) or Davila-Pena et al. (2022) for Shapley values with dependence in linear and nonlinear models, or without dependence in classification, respectively, and in Datta et al. (2015) for the Banzhaf value without dependence in classification, our measure is precisely a generalization of Datta et al.'s approach that contributes to addressing datasets where dependencies exist among certain subsets of features. To achieve this, we incorporate a partition over the set of features in Datta et al.'s model, following the idea of cooperative games provided with a coalitional structure, that is, a structure of a priori unions. Notably, there are other solutions for cooperative games with a priori unions, such as the Owen value (G. Owen, 1977). Further exploration of this approach in machine learning is suggested by Li et al. (2024), though without detailed analysis or application in this field. While our paper is developed in the context of binary features and outcomes, we outline a procedure for extending it to more general cases involving multi-class or continuous features. Our method shows robustness when tested on datasets from various contexts, including clinical, road safety, and musical tastes. Furthermore, due to its model-agnostic nature, it is compatible with some of the most widely used classifiers in practice. Finally, although the computational complexity of the proposed influence measure may be a limitation, we provide a procedure for its estimation that is statistically justified, and we also experimentally validate the accuracy of our approximation through a small benchmarking study. To enhance the clarity and impact of our proposal, Table 1 explicitly outlines the key contributions of this paper in comparison with the closest prior related work.

| Closest literature | Gap | Main contributions |
|---|---|---|
| Datta et al. (2015) | No dependent features | A model-agnostic method to assess the influence of features in classification problems involving dependent features.<br>An axiomatic characterization of our measure by considering properties taken from cooperative game theory. |
| | High computational complexity | A sampling method to approximate the proposed influence measure in large-scale datasets. |
| | Limited validation | Validation of the proposed influence measure on several and varied datasets.<br>Test of different classifiers. |

Table 1: Summary of key contributions of the paper.

The organization of the paper is as follows. Section 2 presents a review of related literature and Section 3 offers the basic preliminaries concerning both influence measures and cooperative games. In Section 4, we provide the axiomatic characterization of our measure, an extension of such beyond binary classification, and an approximation technique for large-scale contexts. Section 5 relates our proposal with the Banzhaf-Owen value. Our methodology is validated and applied in Section 6 through various numerical experiments. The paper finishes with some concluding remarks in Section 7 and three appendices.

## 2 Related literature

A critical challenge for the future in modern data science and artificial intelligence is achieving adequate levels of explanation for the knowledge derived from diverse methods and techniques, ensuring their appropriate utilization (Carrizosa et al., 2021). Since 2018, the European Union has recognized the so-called "right to explanation" concerning algorithms designed for decision-making processes (European Commission, 2020). Burkart and Huber (2021) discuss the reasons behind the necessity for explanations of machine learning models and the various domains where such explanations are in demand, including healthcare, the automobile industry, and recommender systems, among others. A considerable body of literature is dedicated to the pursuit and enhancement of explanation in machine learning. A prevalent objective is to quantify the influence of the features on predictions, often facilitating the selection of relevant features and enabling the analysis of model behavior with a reduced set of predictor variables. Different methodologies have been explored to address the problem of explaining machine learning models. One of them involves considering permutations of the sample values (Altmann et al., 2010). Additionally, some approaches are tailored to specific methods. For instance, within the realm of classification, Ghaddar and Naoum-Sawaya (2018) propose an iterative procedure for feature selection when making use of support vector machines (SVMs). Xiao et al. (2024) allow transformers and capsule networks to better capture hierarchies in multivariate time series classification, enhancing feature representation and classification accuracy. Another suggested approach consists in using concepts from cooperative game theory, assessing each feature's contribution to prediction within any coalition of features (Carrizosa et al., 2021). All these references underscore the notable interest in the literature regarding explanations for classification models and, more broadly, for machine learning.

It is also noteworthy to mention the work of Cohen et al. (2007), who present a new feature selection algorithm on the basis of Shapley's contribution values of the features to classification precision. This method iteratively estimates the utility of features, facilitating subsequent forward selection or backward elimination procedures. Meanwhile, Štrumbelj and Kononenko (2010) provide a local measure of the influence of the features in classification problems, i.e., operating at the individual instance level, and make use of the Shapley value. This procedure is extended in Štrumbelj and Kononenko (2011) to explain regression models and their predictions for individual instances. A paper close to the preceding two studies is Datta et al. (2015). In this work, the authors explore the global influence of different features in a classification problem. Theoretically, they restrict their influence measure to scenarios where the outcome takes only two values, and it is for these specific cases that they discuss its relation with the Banzhaf value of so-called simple cooperative games. Another relevant reference within the context of interpretability in complex machine learning models leveraging game theory is due to Lundberg and Lee (2017). They introduce SHAP (SHapley Additive exPlanations), a unified framework for interpreting predictions. This framework is supported by theoretical foundations across several prediction models, with a focus on determining the importance of the features for each specific prediction. In Casalicchio et al. (2019), alongside introducing a local measure of feature importance for individual observations and two visualization tools, the authors describe a procedure based on the Shapley value. This approach equally distributes the overall model performance among features according to their marginal contributions, enabling comparison of the importance of features across different models. Smith and Alvarez (2021) employ the methodology proposed in Lundberg and Lee (2017) to analyze a COVID-19 database collected in the early stages of the pandemic in Wuhan, China. Meanwhile, Jothi et al. (2021) concentrate on implementing a novel feature selection and data mining classifier system. According to their method, the Shapley value is set to serve as the feature selector for the data mining classifier applied to mental health data. Davila-Pena et al. (2022) expand upon the solution proposed in Štrumbelj and Kononenko (2010) by providing a global measure of the influence of features in a classification problem, along with an axiomatic characterization. The authors apply this methodology to a sample of COVID-19 patients obtained in Spain during the first wave of the pandemic, the first four-month period of 2020. Recently, Nahiduzzaman et al. (2024) introduce a novel approach for accurately classifying three types of lung cancer together with normal lung tissue by using CT (Computed Tomography) images. The integration of SHAP into their framework enhances explanatory capabilities, offering valuable insights for decision-making and bolstering confidence in real-world lung cancer diagnoses. Table 2 summarizes the references where methods derived from cooperative game theory are proposed and applied to improve the interpretation of machine learning models.

| Context | Reference | Game theory value |
|---|---|---|
| Selection of features | Cohen et al. (2007) | Shapley value |
| Explanation of individual classifications | Štrumbelj and Kononenko (2010) | Shapley value |
| Explanation of regression predictions | Štrumbelj and Kononenko (2011) | Shapley value |
| Explanation of classification of a database | Datta et al. (2015) | Banzhaf value |
| Interpretation of predictions | Lundberg and Lee (2017) | Shapley value |
| Importance of features across models | Casalicchio et al. (2019) | Shapley value |
| Identifying mortality factors | Smith and Alvarez (2021) | Shapley value |
| Mental health data classification | Jothi et al. (2021) | Shapley value |
| Global explanation in classification | Davila-Pena et al. (2022) | Shapley value |
| Lung cancer classification | Nahiduzzaman et al. (2024) | Shapley value |

Table 2: Summary of recent references on game-theoretic methods for machine learning models.

According to Štrumbelj and Kononenko (2010), the main shortcoming of existing general explanation methods is that they do not consider all possible dependencies and interactions among feature values. Since then, researchers have worked to incorporate such dependencies in their proposals, although it should be noted that dependency has not been explicitly addressed from a game theory perspective. In linear regression problems, A. B. Owen and Prieur (2017) analyze the global explanation of the model using Shapley's contributions of the features and address the problem of dependent features using an ANOVA decomposition approach. Giudici and Raffinetti (2021) extend this framework using the so-called Lorenz Zonoid decomposition. Aas et al. (2021) address the problem of explaining individual predictions by treating dependent features using an extension of the Kernel SHAP method which is a computationally efficient approximation of Shapley values for the case of large problems. The method is illustrated with examples of linear and non-linear models.

| Context | Reference | Solution concept/approach |
|---|---|---|
| Linear regression model | A. B. Owen and Prieur (2017) | Shapley value/ANOVA |
| Linear regression model | Giudici and Raffinetti (2021) | Shapley value/Lorenz Zonoids |
| Linear and non linear models | Aas et al. (2021) | Extended Kernel SHAP |
| To estimate the conditional expectations | Olsen et al. (2022) | Local Shapley value |
| General overview | Li et al. (2024) | Shapley/Owen values |
| Classification | This paper | Banzhaf-Owen value |

Table 3: Summary of recent references on game-theoretic methods for machine learning models with dependence among features.

In contrast to papers that use methods from the statistical domain to model dependencies between features, Olsen et al. (2022) use machine learning methods to compute local Shapley values in a regression environment. Li et al. (2024) provide a recent overview of the Shapley value as one of the main approaches from artificial intelligence to explain machine learning models, including the consideration of possible complex dependencies between features. The work reveals the potential of the Owen value (G. Owen, 1977), which is a generalization of the Shapley value suited to the case where features can be grouped into a priori coalitions. An example is given of machine learning-based decision models in automatic driving in which the nature of traffic naturally leads to the appearance of a partitioning of features in which those associated with the vehicle would form an a priori coalition. Table 3 provides a summary of references that propose and apply methods from cooperative game theory to enhance the interpretation of machine learning models in cases where feature dependence is present.

## 3 Preliminaries

We start this section by introducing the fundamental concepts regarding datasets and influence measures necessary for the subsequent formal presentation of our model. Following this, we revisit key principles from cooperative game theory, given the close relationship between the influence measure proposed in this paper and a well-established solution concept within cooperative games.

### 3.1 Datasets and influence measures

Let $X = \{X_1, \dots, X_k\}$ be the set of features, with $K = \{1, \dots, k\}$ the set of indices of the features, and $Y$ an outcome. Also, let $\mathcal{A}_l$ denote the finite set of possible values or states that feature $X_l$, $l \in K$, can take, $\mathcal{A} = \prod_{l=1}^{k} \mathcal{A}_l$, and $\mathcal{B}$ the finite set of values that variable $Y$ can take. We will make use of datasets obtained from finite sets, denoted generally as $N$, consisting of $n$ individuals. That is, we have samples, in the form of $\mathcal{M} = \{(X^i, Y^i)\}_{i=1}^{n}$, where $X^i = (X_1^i, \dots, X_k^i)$ and $Y^i$, $i \in N$, are the observed values of the features and the prediction of the outcome, respectively, corresponding to individual $i$. These predictions are obtained by a classifier, $f$, trained on the same set of individuals from which the true outcome values, $Y_{\text{obs}}$, were observed. Note that $\cup_{i=1}^{n} X^i \subseteq \mathcal{A}$, so that the set of individuals can be identified with a set of different feature profiles.

Formally, a dataset is a three-tuple $(X, Y, \mathcal{M})$ where $(X, Y)$ is a $k$-dimensional features vector and an outcome and $\mathcal{M}$ is a sample of size $n$. Given a dataset $(X, Y, \mathcal{M})$ with a binary outcome, i.e., $|\mathcal{B}| = 2$, Datta et al. (2015) define a measure of influence of the features, for each $l \in K$, as follows:

$$\chi_l(X, Y, \mathcal{M}) = \sum_{(X^i, Y^i) \in \mathcal{M}} \sum_{\substack{\left((X^i_{-l}, a_l), b\right) \in \mathcal{M}: \\ a_l \in \mathcal{A}_l,\, b \in \mathcal{B}}} |Y^i - b|, \tag{1}$$

where $(X^i_{-l},\, a_l) = (X^i_1, \dots, X^i_{l-1}, a_l, X^i_{l+1}, \dots, X^i_k)$. Note that $\chi$ computes the number of times that a change in the state of feature $X_l$ causes a change in the outcome. Datta et al. (2015) examine the relationship between this influence measure and the Banzhaf value (Banzhaf III, 1965) for specific cooperative games. For this reason, we review some fundamental concepts from cooperative game theory below, which will be necessary for extending the influence measure to cases involving dependent features.

### 3.2 Cooperative games and values

A cooperative game with transferable utility, often abbreviated as TU game, is a pair $(G, v)$ where $G$ denotes a finite set of players and $v$, the characteristic function, satisfies $v(\emptyset) = 0$ and assigns a number, $v(R) \in \mathbb{R}$, to each subset $R \subseteq G$. For each coalition or subset of players $R$, $v(R)$ represents the benefit that the players in $R$ can guarantee to each other. A TU game $(G, v)$ is called a simple game if: i) $v(R) \leq v(W)$ whenever $R \subseteq W \subseteq G$ (i.e., the game is a monotone game), ii) $v(R) \in \{0, 1\}$ for all $R \subseteq G$, and iii) $v(G) = 1$. A simple game $(G, v)$ is a weighted majority game if there exists a vector of weights $w = (w_1, \ldots, w_g)$ for the players, where $g = |G|$, with $w_l \geq 0$ for all $l \in \{1, \ldots, g\}$, and a positive real number $a \in \mathbb{R}^+$, referred to as the quota, such that $v(R) = 1$, $R \subseteq G$, if and only if $\sum_{l \in R} w_l \geq a$.

One of the focal aspects in cooperative game theory revolves around the definition of values. Some of these values serve as procedures for distributing the worth associated to the cooperation among players, as the Shapley value and the Owen value, while others can be used as ranking indices, as the Banzhaf value and the Banzhaf-Owen value. Given a TU game $(G, v)$ and $l \in G$, the Banzhaf value (Banzhaf III, 1965) is given by

$$B_l(G, v) = \sum_{R \subseteq G \setminus \{l\}} \frac{1}{2^{g-1}} \cdot \Big( v(R \cup \{l\}) - v(R) \Big). \tag{2}$$

G. Owen (1981) extends the Banzhaf value to the class of TU games with a priori unions. A TU game with a priori unions (G. Owen, 1977) is a three-tuple $(G, v, P)$, where $(G, v)$ is a TU game and $P = \{P_1, \ldots, P_m\}$ is a partition of $G$ representing affinities among players, which might stem from familiar, political, or economic motives, among other factors. We denote by $M = \{1, \ldots, m\}$ the set of indices of the unions in $P$, and we usually identify a union with its index. Given a TU game with a priori unions $(G, v, P)$, $t \in M$, and $l \in P_t$, the Banzhaf-Owen value (G. Owen, 1981) is given by

$$BO_l(G, v, P) = \sum_{S \subseteq M \setminus \{t\}} \sum_{R \subseteq P_t \setminus \{l\}} \frac{1}{2^{m-1}} \cdot \frac{1}{2^{|P_t|-1}} \cdot \Big( v(\cup_{u \in S} P_u \cup R \cup \{l\}) - v(\cup_{u \in S} P_u \cup R) \Big).$$

Note that both the Banzhaf and Banzhaf-Owen values assign to each player the average contribution they make to the various coalitions they can join. This model, or its generalization, is therefore well suited for quantifying the influence of a feature within a database consisting of a set of individuals in which the values of certain features and the result of their classification by an outcome are known. A thorough analysis of the properties and full characterizations of the Banzhaf-Owen value can be found in Amer et al. (2002), Alonso-Meijide et al. (2007) or Lorenzo-Freire (2017), among others.

## 4 Main results

The aim of this section is to extend the influence measure proposed by Datta et al. (2015) to the context of dependent features. Since the authors related their influence measure to the Banzhaf value, our proposal is inspired by the Banzhaf-Owen value for TU games with a priori unions (G. Owen, 1981). This approach enables us to capture potential interactions between features, providing a more nuanced understanding of feature influence when independence cannot be assumed.

## 4.1 Datasets and influence measures under dependency

In this context, let $P = \{P_1, \ldots, P_m\}$ represent a partition of $K$ that reflects possible dependencies or interactions between certain subsets of features. A trivial partition is $P^k = \{\{1\}, \ldots, \{k\}\}$, where each subset of the partition consists of a singleton.

Now, a dataset with interactions is a four-tuple $(X, Y, P, \mathcal{M})$ where $(X, Y)$ is a $k$-dimensional features vector and an outcome, $P$ is a partition of $K$, and $\mathcal{M}$ is a sample of size $n$; in other words, $(X, Y, \mathcal{M})$ is a dataset and $P$ is a partition of $K$, the set of indices of the features. $D(X, Y)$ denotes the family of all datasets with interactions where $(X, Y)$ are the features vector and the outcome. Our goal is to make use of techniques used in classification problems to define a measure for studying the influence of features on the predicted value of the outcome under the assumption of possible dependencies between subsets of these features. First, let us state the formal definition of such an influence measure within this context.

**Definition 4.1.** *An influence measure for $D(X, Y)$ is a map, $I$, that assigns to every dataset with interactions, $(X, Y, P, \mathcal{M}) \in D(X, Y)$, a vector $I(X, Y, P, \mathcal{M}) \in \mathbb{R}^k$. The measure $I_l(X, Y, P, \mathcal{M}) \in \mathbb{R}$, $l \in K$, is a metric of the importance of $X_l$ in determining the predicted values of $Y$ over $\{(X^i)\}_{i=1}^n$.*

The main objective is to show that there is a unique measure of influence that satisfies a set of natural axioms, which we introduce and describe in the following subsection. In what follows, we assume that $|\mathcal{B}| = 2$, i.e., the outcome $Y$ can only take two distinct values.

## 4.2 Axioms

A feature $X_l$, with $l \in K$, is said to be *non-influential* in the dataset with interactions $(X, Y, P, \mathcal{M})$ if $Y^i = Y^j$ for all $i, j \in N$ such that $X^i_{-l} = X^j_{-l}$, where $X^i_{-l}$, with $i \in N$ and $l \in K$, denotes the vector $X^i$ after removing the $l$-th coordinate. First, we introduce the dummy property for influence measures.

**(DP)** **Dummy property**. An influence measure $I$ satisfies the *dummy property* if, for every $(X, Y, P, \mathcal{M}) \in D(X, Y)$ and every non-influential feature in the dataset $(X, Y, P, \mathcal{M})$, $X_l$, with $l \in K$, it holds that $I_l(X, Y, P, \mathcal{M}) = 0$.

Given a dataset with interactions $(X, Y, P, \mathcal{M})$ and a bijective mapping $\sigma$ from $K$ to itself, we define $\sigma(X, Y, P, \mathcal{M}) = (\sigma(X), Y, \sigma(P), \sigma(\mathcal{M}))$ in the natural way, consisting in relabelling the features according to $\sigma$, i.e., making the index of $l$, with $l \in K$, in the initial dataset become now $\sigma(l)$. We write $\sigma(\mathcal{M}) = \{(\sigma(X^i), Y^i)\}_{i=1}^n$. Given a bijective mapping $\tau$ from $\mathcal{A}_l$, $l \in K$, to itself, we define $\tau(X, Y, P, \mathcal{M}) = (X, Y, P, \tau(\mathcal{M}))$ in a similar manner, consisting of relabelling the values of $\mathcal{A}_l$ according to $\tau$, i.e., making the value of $a_l$, with $a_l \in \mathcal{A}_l$, in the initial dataset becoming now $\tau(a_l)$. We write $\tau(\mathcal{M}) = \{(\tau(X^i), Y^i)\}_{i=1}^n$.

These concepts enable us to introduce various notions of symmetry that measures of influence, like those examined here, should satisfy.

**(FSY)** **Feature symmetry**. An influence measure $I$ satisfies *feature symmetry property* if, for every $(X, Y, P, \mathcal{M}) \in D(X, Y)$ such that $P = P^k$ and a bijective mapping $\sigma$ from $K$ to itself, it holds that $I_l(X, Y, P, \mathcal{M}) = I_{\sigma(l)}(\sigma(X, Y, P, \mathcal{M}))$ for all $l \in K$.

**(SSY)** **State symmetry**. An influence measure $I$ satisfies *state symmetry property* if, for every $(X, Y, P, \mathcal{M}) \in D(X, Y)$ such that $P = P^k$ and a bijective mapping $\tau$ from $A_l$, $l \in K$, to itself, it holds that $I_q(X, Y, P, \mathcal{M}) = I_q(\tau(X, Y, P, \mathcal{M}))$ for all $q \in K$.

**(SY)** **Symmetry**. An influence measure $I$ satisfies *symmetry property* if it satisfies both feature symmetry (FSY) and state symmetry (SSY).

Below we present some common properties that an influence measure should satisfy when considering a partition structure. Since $|\mathcal{B}| = 2$, we can assume that $\mathcal{B} = \{0, 1\}$. Consequently, we define $W(\mathcal{M}) = \{i \in N : Y^i = 1\}$ and $L(\mathcal{M}) = \{i \in N : Y^i = 0\}$ as the sets of sample profiles where the outcome takes the values 1 and 0, respectively. Thus, in general, given a set of individuals $N$, with $n \leq |\mathcal{A}|$, $W, L \subseteq N$, and $W \cap L = \emptyset$, we can identify a sample with $(W, L)$ and a dataset with interactions with $(X, Y, P, (W, L))$.

**(DU)** **Disjoint union**. An influence measure $I$ satisfies *disjoint union property* if for every $(X, Y, P, (Q, R \cup R')), (X, Y, P, (R \cup R', Q)) \in D(X, Y)$ where $X$ is a set of features and $Q, R$, and $R'$ are pairwise disjoint sets satisfying $|Q \cup R \cup R'| \leq |\mathcal{A}|$ for all $l \in K$, it holds that

$$I_l(X, Y, P, (Q, R)) + I_l(X, Y, P, (Q, R')) = I_l(X, Y, P, (Q, R \cup R'))$$

and

$$I_l(X, Y, P, (R, Q)) + I_l(X, Y, P, (R', Q)) = I_l(X, Y, P, (R \cup R', Q)).$$

**(II)** **Indifference to interactions**. An influence measure $I$ satisfies *indifference to interactions property* if for all $(X, Y, P, \mathcal{M}) \in D(X, Y)$, for all $t \in M$, with $l, q \in P_t$ and $l \neq q$, it holds that $I_l(X, Y, P, \mathcal{M}) = I_l(X, Y, P_{\_q}, \mathcal{M})$, where $P_{\_q}$ is the partition that results from removing $q$ from its original subset to create a unitary subset, i.e.,

$$P_{\_q} = \{P_1, \ldots, P_{t-1}, P_t \backslash \{q\}, P_{t+1}, \ldots, P_m, \{q\}\}.$$

**(RP)** **Relevance of dependent feature profiles**. An influence measure $I$ satisfies *relevance of dependent feature profiles property* if for all $(X, Y, P, \mathcal{M}) \in D(X, Y)$ and for all $t \in M$ such that $|P_t| = 1$, with $l \in P_t$, it holds that $I_l(X, Y, P, \mathcal{M}) = I_l(X, Y, P^k, \mathcal{M}^t)$, where

$$\mathcal{M}^t = \{(X^i, Y^i) \in \mathcal{M}, i \in N : \text{if } u \in M \backslash \{t\}\text{, then } X^i_q \approx X^i_v \text{ for all } q, v \in P_u\}^1 \quad (3)$$

is the subsample of $\mathcal{M}$ where features within the same union, except for $P_t$, precisely adhere

[1] Note that this expression holds even when $|P_t| \neq 1$.

to a pre-adjusted dependency model. In a union consisting of two binary features, the acceptable values for both features will be either identical or opposite, depending on whether the dependency is positive or negative.

Note that each of these properties extends specific axioms from the cooperative game theory literature to the context of influence measures. For instance, (DP) is an extension of the null player property utilized in the axiomatization of solutions for TU games, as evidenced in works such as Feltkamp (1995) for the Banzhaf value. Datta et al. (2015) also employ this property from a binary classification perspective. Additionally, (FSY), (SSY), and (SY) share similarities with properties outlined in Datta et al. (2015). These properties, in turn, expand upon those used in the axiomatization of the Banzhaf-Owen value for a TU game with a priori unions, among others, as seen in works like Alonso-Meijide et al. (2007). Furthermore, (DU) has the same essence as the union-intersection axiom employed in Lehrer (1988) to characterize the Banzhaf value for TU games. It also serves as a generalization of the axiom with the same name used in Datta et al. (2015). Lastly, (II) and (RP) extend properties of indifference in unions and the quotient game for single-player unions introduced in Alonso-Meijide et al. (2007) to axiomatize the Banzhaf-Owen value.

The axioms presented above will be used in Subsection 4.3 to introduce and axiomatically characterize an influence measure assuming potential dependencies among features.

### 4.3 Axiomatic characterization

The following proposition extends the result of Datta et al. (2015) to the case of a set of binary outcomes, when there is no partition structure on the affinities of the features.

**Proposition 4.2.** *An influence measure for $D(X,Y)$, $I$, satisfies (DP), (SY), and (DU) if and only if there exists a constant $C$ such that for every dataset with interactions $(X,Y,P^k,\mathcal{M})$ and every feature $l \in K$,*

$$I_l(X,Y,P^k,\mathcal{M}) = C \cdot \sum_{(X^i,Y^i)\in\mathcal{M}} \; \sum_{\substack{((X^i_{-l},a_l),b)\in\mathcal{M}:\\ a_l\in\mathcal{A}_l,\ b\in\mathcal{B}}} |Y^i - b|. \tag{4}$$

*Moreover, it holds that $I_l(X,Y,P^k,\mathcal{M}) = C \cdot \chi_l(X,Y,\mathcal{M})$.*

*Proof.* The proof follows the same principles as the (non-trivial) proof in Datta et al. (2015). The only difference between Datta et al.'s statement and ours is their consideration of datasets and a binary outcome, while we are focused on datasets with interactions. The main connection lies in the inclusion of the trivial partition $P^k$ in our axioms (DP), (SY), and (DU). □

Now, we present an axiomatic characterization of our influence measure in cases where a partition exists over the set of features, reflecting their dependencies and interactions.

*Remark* 4.1. It is important to note the analogy of this result with that obtained by Alonso-Meijide et al. (2007) for the Banzhaf-Owen value of TU games with a priori unions. ◇

**Theorem 4.3** (Existence and uniqueness)**.** *An influence measure for $D(X,Y)$, $I$, satisfies (DP), (SY), (DU), (II), and (RP) if and only if there exists a constant $C$ such that for every dataset with interactions*

*$(X, Y, P, \mathcal{M})$, $t \in M$, $l \in P_t$, it holds that*

$$I_l(X, Y, P, \mathcal{M}) = C \cdot \Psi_l(X, Y, P, \mathcal{M}), \tag{5}$$

*where*

$$\Psi_l(X, Y, P, \mathcal{M}) = \sum_{(X^i, Y^i) \in \mathcal{M}^t} \sum_{\substack{((X^i_{\_l}, a_l), b) \in \mathcal{M}^t : \\ a_l \in \mathcal{A}_l,\ b \in \mathcal{B}}} |Y^i - b|. \tag{6}$$

*Proof.* We prove (a) the existence and (b) the uniqueness of an influence measure for $D(X, Y)$ satisfying (DP), (SY), (DU), (II), and (RP).

(a) (Existence). First we prove the existence of an influence measure satisfying the required axioms.

1. The influence measure $\Psi$ satisfies (DP), (SY), and (DU):
   According to Proposition 4.2, it suffices to check Equation (4). Let $(X, Y, P, \mathcal{M}) \in D(X, Y)$, $t \in M$, and $l \in P_t$. As we will deal with $P = P^k$, we have $M = K$ and $P_t = \{l\}$, meaning $t = l$ and $\mathcal{M}^t = \mathcal{M}^l = \mathcal{M}$ when applying Equation (5). Thus, $\Psi_l(X, Y, P, \mathcal{M})$ reduces to

   $$\Psi_l(X, Y, P^k, \mathcal{M}) = \sum_{(X^i, Y^i) \in \mathcal{M}} \sum_{\substack{((X^i_{\_l}, a_l), b) \in \mathcal{M} : \\ a_l \in \mathcal{A}_l,\ b \in \mathcal{B}}} |Y^i - b|.$$

2. The influence measure $\Psi$ satisfies (II):
   Let $(X, Y, P, \mathcal{M}) \in D(X, Y)$, $P_t \in P$, and $l, q \in P_t$ be distinct features. Let also $P_{\_q} = \{P'_1, \ldots, P'_{m+1}\}$, where $P'_u = P_u$ for every $u \in M \backslash \{t\}$, $P'_t = P_t \backslash \{q\}$, and $P'_{m+1} = \{q\}$, and let $M' = \{1, 2, \ldots, m, m+1\}$. Then, $m' = m + 1$, $|P'_t| = |P_t| - 1$, and $\left|P'_{m+1}\right| = 1$. Therefore,

   $$\Psi_l(X, Y, P_{\_q}, \mathcal{M}) = \sum_{(X^i, Y^i) \in \mathcal{M}'^t} \sum_{\substack{((X^i_{\_l}, a_l), b) \in \mathcal{M}'^t : \\ a_l \in \mathcal{A}_l,\ b \in \mathcal{B}}} |Y^i - b|,$$

   where $\mathcal{M}'^t = \{(X^i, Y^i) \in \mathcal{M}, i \in N : \text{if } u \in M' \backslash \{t\}, \text{then } X^i_q \approx X^i_v \text{ for all } q, v \in P_u\}$. Note that $\mathcal{M}'^t = \mathcal{M}^t$ because $M' \backslash \{t\} = \{1, \ldots, t-1, t+1, \ldots, m, m+1\}$ and $\left|P'_{m+1}\right| = 1$. Thus, $\Psi_l(X, Y, P_{\_q}, \mathcal{M}) = \Psi_l(X, Y, P, \mathcal{M})$.

3. The influence measure $\Psi$ satisfies (RP):
   Let $(X, Y, P, \mathcal{M}) \in D(X, Y)$, $P_t \in P$ such that $|P_t| = 1$, and take $l \in P_t$ the only feature in $P_t$. Thus, we consider $\mathcal{M}^t$ as in Equation (3). Then,

   $$\begin{aligned} \Psi_l(X, Y, P^k, \mathcal{M}^t) &= \sum_{(X^i, Y^i) \in \mathcal{M}^t} \sum_{\substack{((X^i_{\_l}, a_l), b) \in \mathcal{M}^t : \\ a_l \in \mathcal{A}_l,\ b \in \mathcal{B}}} |Y^i - b| \\ &= \Psi_l(X, Y, P, \mathcal{M}). \end{aligned}$$

(b) (Uniqueness). Now, we prove its uniqueness. Let us assume that two influence measures $I_1$ and $I_2$ satisfy (DP), (SY), (DU), (II), and (RP). Then we can find a dataset with interactions $(X, Y, P, \mathcal{M}) \in D(X, Y)$, with a partition $P$ on $K$ with the maximum number of unions such that $I_1(X, Y, P, \mathcal{M}) \neq I_2(X, Y, P, \mathcal{M})$, i.e., $(I_1)_l(X, Y, P, \mathcal{M}) \neq (I_2)_l(X, Y, P, \mathcal{M})$ for some $l \in K$. From Proposition 4.2, it suffices to consider the case where $|P| < k$. Let us take $P_t \in P$ such that $l \in P_t$. We can distinguish two cases:

- *Case 1.* $|P_t| = 1$. Then, we have $P_t = \{l\}$. Since both influence measures satisfy (RP), then

  $$(I_1)_l(X, Y, P, \mathcal{M}) = (I_1)_l(X, Y, P^k, \mathcal{M}^t)$$

  and

  $$(I_2)_l(X, Y, P, \mathcal{M}) = (I_2)_l(X, Y, P^k, \mathcal{M}^t),$$

  where $(I_1)_l(X, Y, P^k, \mathcal{M}^t)$ and $(I_2)_l(X, Y, P^k, \mathcal{M}^t)$ reduce to the influence measure for $l$ considered in Datta et al. (2015), whose uniqueness is ensured by (DP), (SY), and (DU).

- *Case 2.* $|P_t| > 1$. Then, there is some $q \in P_t$ such that $l \neq q$. By (II), it holds that

  $$(I_1)_l(X, Y, P, \mathcal{M}) = (I_1)_l(X, Y, P_{-q}, \mathcal{M})$$

  and

  $$(I_2)_l(X, Y, P, \mathcal{M}) = (I_2)_l(X, Y, P_{-q}, \mathcal{M}).$$

  By the maximality of partition $P$, it follows that

  $$(I_1)_l(X, Y, P_{-q}, \mathcal{M}) = (I_2)_l(X, Y, P_{-q}, \mathcal{M})$$

  and, hence, this leads to $(I_1)_l(X, Y, P, \mathcal{M}) = (I_2)_l(X, Y, P, \mathcal{M})$, also obtaining a contradiction.

This concludes the proof. □

Example 4.4 below presents a simple case study that demonstrates how to infer dependencies between features directly from the dataset and highlights the practical impact of these inferred dependencies on the results. It also illustrates the direct computation of the influence measure, including detailed calculations for determining various influence scores and their interpretation.

**Example 4.4.** *We analyze a dataset of 20 patients diagnosed with COVID-19 in Galicia, Spain, during the first wave of the pandemic, between March and April 2020. Five key features are considered. $X_1$ represents patient age, with a value of 0 if the patient is under 60 years old and 1 for those 60 years old and older. $X_2$, $X_3$, $X_4$, and $X_5$ indicate the presence or absence of dementia, respiratory pathologies, cancer, and diabetes, respectively, where the value 1 denotes the presence of the corresponding condition and 0 its absence. Finally, the outcome, $Y$, takes a value of 1 if the patient required hospitalization, admission to the intensive care unit, or passed away, and 0 if none of these three events occurred. The data is presented in columns 1, 2, 4, and 5 of Table 4.*

*First, we calculate the influence measure ignoring possible dependencies or interactions between features, resulting in $\Psi(X, Y, P = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}, \mathcal{M}) = (4, 4, 14, 2, 6)$. Next, trying to identify possible dependencies between features, we construct Table 5 summarizing the data for features $X_1$ and*

| Feature profile<br>($X$) | Outcome<br>($Y$) | Changes<br>($X_5$) | Feature profile<br>($X$) | Outcome<br>($Y$) | Changes<br>($X_5$) |
|---|---|---|---|---|---|
| **00000** | **0** | **0** | 10100 | 0 | 1 |
| **00001** | **0** | **0** | **11000** | **0** | **1** |
| **00010** | **0** | **0** | **11001** | **1** | **1** |
| **00011** | **0** | **0** | **11010** | **0** | **0** |
| **00100** | **1** | **0** | **11011** | **0** | **0** |
| **00101** | **1** | **0** | **11100** | **1** | **0** |
| **00110** | **1** | **0** | **11111** | **1** | **0** |
| **00111** | **1** | **0** | 01100 | 0 | 1 |
| 10101 | 1 | 1 | 01101 | 1 | 1 |
| **11101** | **1** | **0** | **11110** | **1** | **0** |

Table 4: Dataset of COVID-19 patients and computation of changes to measure the influence of diabetes.

*$X_2$. For convenience and clarity in explaining the methodology, we use a relatively small sample, which is nonetheless sufficient to ensure that none of the expected cell frequencies in the matrix fall below 5, making the test adequate. Applying Pearson's chi-squared test of independence with Yates' continuity correction, we obtain that the test statistic takes a value of 5 (p-value $= 0.02535$). For a significance level of 0.01, we would accept the hypothesis of independence, $H_0$, for these two features, so that the calculation of the influence measure corresponds to the original proposal of Datta et al.. However, with a significance level of 0.05, we reject $H_0$, indicating a significant difference in dementia incidence between those under and over 60 years old. It is easy to check, using Fisher's exact test for small samples, that for the remaining pairs of features, we cannot reject the hypothesis of independence because we obtain in all cases a p-value between 0.275 and 0.350. However, for $X_1$ and $X_2$, applying Fisher's exact test, we obtain p-value $= 0.0109604$. Hence, by the results of both tests, we calculate the influence measure assuming a positive dependence between features $X_1$ and $X_2$, yielding $\Psi(X, Y, P = \{\{1,2\},\{3\},\{4\},\{5\}\}, \mathcal{M}) = (4,4,14,2,2)$.*

| Dementia<br>Age | $0 \equiv$ No | $1 \equiv$ Yes | Totals |
|---|---|---|---|
| $0 \equiv < 60$ | 8 | 2 | 10 |
| $1 \equiv \geq 60$ | 2 | 8 | 10 |
| Totals | 10 | 10 | 20 |

Table 5: Contingency table of age vs. dementia.

*In both scenarios, respiratory pathologies and cancer emerge as the most and least influential features, respectively. When dependency between features $X_1$ and $X_2$ is not considered, diabetes ranks as the second most influential feature. However, when this dependency is accounted for, age and dementia, tied, become the next most important features after respiratory pathologies, with diabetes ranking last, tied with cancer.*

*This small example also allows us to illustrate the direct calculation of our influence measure. To do this, we focus on the feature $X_5$ (diabetes). First, if we ignore the dependency between age and dementia, based on Equation (1), we determine for each sample profile whether a change in the value of $X_5$ results in a different outcome, while holding all other feature values constant. For instance, consider the feature profile 00000, which has an outcome of 0. Changing the value of $X_5$ results in the feature profile 00001, which also has an outcome of 0. Therefore, no changes occur in the outcome, as indicated by the 0 in column 3 of Table 4. Now, take the feature profile 10101, which has an outcome of 1. Changing the value of $X_5$ produces the feature profile 10100, with an outcome of 0. As a result, the number of changes recorded in column 3 for the feature profile 10101 is 1, and the number of changes recorded in column 6 for the feature profile 10100 is also naturally 1. By proceeding in this way for all feature profiles, columns 3 and 6 of Table 4* *record these changes, obtaining a total of 6 changes. As mentioned above, this total quantifies the influence of this condition, diabetes. The influence scores for the remaining features are computed analogously. Next, after statistically verifying the dependency between age and dementia, we group these two features into a union and, using Equation* (6)*, we repeat the calculation. However, this time, we restrict the analysis to the subsample of profiles that exhibit the positive dependence between these two features, which are those highlighted in bold in Table 4. As can be observed, this reduces the total number of changes to 2, which corresponds to the influence score for feature* $X_5$ *under this dependency scenario. As a result, the influence of diabetes is three times lower assuming the strong association between age and dementia. This variation in the influence score of diabetes, from considering age and dementia as independent to accounting for their dependency, could indicate that, when analyzed as a group, age and dementia might collectively capture more of the outcome variation, reducing the influence of diabetes.* △

To deepen the interpretability of the proposed influence measure, it is also worth considering its practical application. In this regard, its role in supporting decision-making in real-life classification tasks is especially remarkable. Focusing on health-related problems such as the one considered in Example 4.4, influence scores serve, in particular, to alert medical professionals to the importance of certain patient features, such as age or pre-existing conditions, including respiratory pathologies, as opposed to the lesser importance of others. It should be noted that these features potentially constitute an additional difficulty for patients with a certain disease, which should be taken into account in both their care and resource planning.

A review of the existing literature in this line of research further underscores the importance of having a correct measure of feature influence in real-life classification problems. In the same medical context as above, Smith and Alvarez (2021) also recognize the importance of identifying the most influential features–such as age or length of hospitalization—to predict COVID-19 mortality, thereby enabling medical professionals to provide special care to COVID-19 patients with these risk factors. Similarly, Jothi et al. (2021) highlight the interest in identifying key features such as loss of pleasure, self-disgust, or suicide attempts, to aid clinicians in diagnosing specific mental illnesses. Finally, Ghaddar and Naoum-Sawaya (2018) underline the importance for physicians to identify the most relevant features in the diagnosis of different classes of tumors based on gene expressions.

### 4.4 Extension beyond binary classification

Our proposal can be extended to handle multi-class classifications and continuous outcomes, broadening its applicability to more complex scenarios. In these cases, the influence measure can assess outcome changes using its original formulation in (6), which also allows for quantifying such changes. In general, this definition can be alternatively reformulated to keep the spirit that motivates our proposal of influence measure, which is based on counting the number of changes in the outcome, as stated in Remark 4.2.

*Remark* 4.2. Let $(X, Y, P, \mathcal{M})$ be a dataset with interactions, where $\mathcal{M} = \{(X^i, Y^i)\}_{i=1}^n$, and take

$t \in M$ and $l \in P_t$. Thus, the influence measure in (5) can be rewritten as

$$I_l(X,Y,P,\mathcal{M}) = C \cdot \Psi_l(X,Y,P,\mathcal{M}) = C \cdot \sum_{(X^i,Y^i)\in\mathcal{M}^t} \sum_{\substack{((X^i_{-l},a_l),b)\in\mathcal{M}^t:\\ a_l\in\mathcal{A}_l,\; b\in\mathcal{B}}} (\mathbb{1}_{\{i\}})_b, \tag{7}$$

where

$$(\mathbb{1}_{\{i\}})_b = \begin{cases} 1, & \text{if } Y^i \neq b, \\ 0, & \text{otherwise.} \end{cases}$$ ◇

It is worth noting that a generalization of Theorem 4.3 to the case of a non-binary outcome is not trivial. While most of the axioms used in the result naturally extend to scenarios with multi-class or even continuous outcomes, this is not the case for the disjoint union (DU), for which the binary character of the outcome is crucial. Reformulating this axiom along the lines of Lehrer (1988)'s union-intersection axiom could be a future task of interest. Moreover, the non-binary nature of the outcome introduces multiple ways to account for changes in its various values, suggesting the potential need for an additional axiom to ensure the uniqueness of any axiomatic characterization of the proposed influence measure. Example 4.5 below shows the extension of the method of calculating the influence measure on a case with a non-binary outcome.

**Example 4.5.** *We modify Example 4.4 by transforming the initially binary outcome $Y$ into a multi-class outcome. Now, $Y$ takes the value 1 if the patient required hospitalization, 2 if the patient was admitted to the intensive care unit, 3 if the patient was deceased, and 0 if none of these events occurred. The new data are presented in columns 1, 2, 4, and 5 of Table 6.*

| Feature profile ($X$) | Outcome ($Y$) | Changes ($X_5$) | Feature profile ($X$) | Outcome ($Y$) | Changes ($X_5$) |
|---|---|---|---|---|---|
| 00000 | 0 | 0 | 10100 | 0 | 1 |
| 00001 | 0 | 0 | 11000 | 0 | 1 |
| 00010 | 0 | 0 | 11001 | 2 | 1 |
| 00011 | 0 | 0 | 11010 | 0 | 0 |
| 00100 | 1 | 0 | 11011 | 0 | 0 |
| 00101 | 1 | 0 | 11100 | 3 | 0 |
| 00110 | 1 | 1 | 11111 | 3 | 1 |
| 00111 | 2 | 1 | 01100 | 0 | 1 |
| 10101 | 3 | 1 | 01101 | 2 | 1 |
| 11101 | 3 | 0 | 11110 | 2 | 1 |

Table 6: Dataset of COVID-19 patients with non-binary outcome.

*For illustrative purposes, the influence measure is calculated in this case for the fifth feature, $X_5$, ignoring potential dependencies between features. For each sample profile, it is again assessed whether a change in the value of $X_5$ leads to a different outcome value while keeping all other features constant. These changes are recorded in columns 3 and 6 of Table 6, resulting in a total of 10 changes. This differs with the total of 6 changes achieved in Example 4.4, where the outcome was treated as binary.* △

As mentioned, the extension of the proposed influence measure in (7) can be applicable in situations where the (single) outcome is non-binary. However, adapting the method to more complex classification settings, such as multi-label problems, introduces additional challenges. The notion of "outcome changes" in such a context could be interpreted in several ways, potentially depending on how the relationships between the components of the multi-dimensional outcome are modeled. This aspect requires further investigation, and extending the results presented in this paper to such scenarios is not straightforward.

The obtaining of the influence measure $I(X, Y, P, \mathcal{M})$ (5) entails high computational complexity, as the computation of $\Psi(X, Y, P, \mathcal{M})$ (6) requires the evaluation of all elements in the considered subsample $\mathcal{M}^t$. Thus, we present below an alternative to its exact calculation.

### 4.5 A brief note on the computational complexity

It seems intuitive that our influence measure can be approximated by statistical sampling in large-scale settings, in line with what Saavedra-Nieves and Fiestras-Janeiro (2021) do for the Banzhaf-Owen value. To mitigate the high computational complexity in large-scale datasets with interactions, we propose a specific sampling method (Cochran, 2007) to approximate $\Psi_l(X, Y, P, \mathcal{M})$ for every $(X, Y, P, \mathcal{M})$ and $l \in K$.

The steps of the procedure are described as follows:

1. The sampling population is $\mathcal{M}^t$, the subsample of $\mathcal{M}$ in (3).
2. The parameter to be estimated is $\Psi_l(X, Y, P, \mathcal{M})$ for each feature $l$.
3. The sampling unit to be studied is

$$u(X^i, Y^i)_l = \sum_{\substack{((X^i_{-l}, a_l), b) \in \mathcal{M}^t: \\ a_l \in \mathcal{A}_l,\ b \in \mathcal{B}}} |Y^i - b|, \text{ for each } (X^i, Y^i) \in \mathcal{M}^t.$$

4. We take without replacement a sample of $\ell$ observed features and their associated outcomes in $\mathcal{M}^t$, i.e., $\{(X^1, Y^1), \ldots, (X^\ell, Y^\ell)\}$.
5. The estimation of $\Psi_l(X, Y, P, \mathcal{M})$ is obtained as $\overline{\Psi}_l = |\mathcal{M}^t| \cdot \left( \frac{1}{\ell} \sum_{j=1}^{\ell} u(X^j, Y^j)_l \right)$, where $\ell$ denotes the sampling size.

After applying this procedure to each feature, $\overline{\Psi} = (\overline{\Psi}_1, \ldots, \overline{\Psi}_k)$ represents the estimated $\Psi(X, Y, P, \mathcal{M})$. Let $l \in K$ be an arbitrary feature. From a statistical point of view (Cochran, 2007), the estimator $\overline{\Psi}_l$ is unbiased since $\mathbb{E}(\overline{\Psi}_l) = \Psi_l(X, Y, P, \mathcal{M})$, where $\mathbb{E}(\cdot)$ denotes the mean operator over the set of features in $\mathcal{M}^t$. Besides, it readily follows that $\text{MSE}(\overline{\Psi}_l) = \frac{1}{\ell}\text{Var}(u(X^i, Y^i)_l)$, where MSE denotes the mean square error. This fact ensures the consistency of $\overline{\Psi}$ since $\lim_{\ell \to |\mathcal{M}^t|} \text{MSE}(\overline{\Psi}_l) = 0$ for each $l \in K$.

Serfling's inequality for the sum of non-independent random variables (Serfling, 1974) can be used for bounding the error in estimating $\Psi_l(X, Y, P, \mathcal{M})$ for a fixed $l \in K$. Take $\varepsilon > 0$ and $\alpha \in (0, 1)$, consider the dataset with interactions $(X, Y, P, \mathcal{M})$, and denote

$$w_l = \max_{(X^i, Y^i), (X^j, Y^j) \in \mathcal{M}^t} (u(X^i, Y^i)_l - u(X^j, Y^j)_l).$$

Thus, if

$$\ell \geq \frac{ln(2/\alpha) w_l^2 |\mathcal{M}^t| (|\mathcal{M}^t| + 1)}{ln(2/\alpha) w_l^2 |\mathcal{M}^t| + 2\varepsilon^2},$$

then

$$\mathbb{P}(|\overline{\Psi}_l - \Psi_l| \geq \varepsilon) \leq \alpha.$$

We refer the reader to Proposition A.1 in Appendix A of the Online Resource Section (ORS) to follow the proof of such result.

Finally, in Step 4 of the procedure for approximating the influence measure, we consider sampling without replacement to select elements for the sampling population. This choice is justified by the nature of the problem, where each feature profile is uniquely identified with a different individual and, thus, the number of changes should be evaluated only once. As stated in Cochran (2007), from a statistical point of view, the non-replacement hypothesis in sampling theoretically ensures a lower variance of the resulting estimators with respect to those obtained under replacement in sampling. However, it is known that for sufficiently large populations, simple random sampling with replacement would give similar results in terms of accuracy, although alternative and specific bounds of error should be provided.

In summary, this section introduced a novel influence measure that accounts for dependent features, supported by an axiomatic characterization grounded in game theory principles. We also explored potential extensions for handling non-binary classifications, discussed the computational complexity of our proposal, and presented an approximation method. Figure 1 displays a flowchart summarizing the steps for determining the influence scores of features in classification problems under dependency.

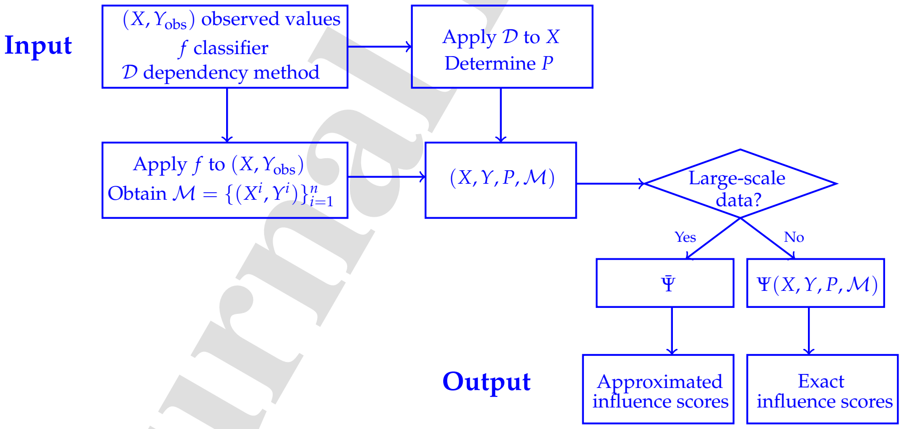
Input
$(X, Y_{\text{obs}})$ observed values
$f$ classifier
$\mathcal{D}$ dependency method
Apply $\mathcal{D}$ to $X$
Determine $P$
Apply $f$ to $(X, Y_{\text{obs}})$
Obtain $\mathcal{M} = \{(X^i, Y^i)\}_{i=1}^n$
$(X, Y, P, \mathcal{M})$
Large-scale data?
Yes
No
$\bar{\Psi}$
$\Psi(X, Y, P, \mathcal{M})$
Output
Approximated influence scores
Exact influence scores

Figure 1: Flowchart for the influence score computation.

The choice of the dependency method, $\mathcal{D}$, is naturally influenced by the characteristics of the dataset under study. For instance, in Example 4.4, we applied Pearson's chi-squared test of independence and Fisher's exact test, as the dataset contained no missing values. However, in the context of missing data, alternative statistical techniques for testing independence exist (Tian and Li, 2017). One such example is the score test statistic, which is similar to Pearson's squared test when data is complete (Lipsitz and Fitzmaurice, 1996).

## 5 The influence measure and games with a priori unions

In this section, we study the proposed measure of influence from a game-theoretical perspective. For this purpose, we mainly follow the ideas in Datta et al. (2015).

First, we will consider a TU game associated to any sample. Let $\mathcal{M} = \{(X^i, Y^i)\}_{i=1}^n$ be a sample such that $\mathcal{A} = \cup_{i=1}^n X^i$ and $|\mathcal{A}_l| = 2$, for all $l \in K$. Thus, the sample $\mathcal{M}$ corresponds to the TU game $(K, v^{\mathcal{M}})$ defined, for all $R \subseteq K$, by

$$v^{\mathcal{M}}(R) = Y^i \Longleftrightarrow \exists\, i \in N \text{ such that } X_l^i = \begin{cases} 1 & \text{if } l \in R, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

If, moreover, we particularly consider the case where $|\mathcal{B}| = 2$, we can assume that the characteristic function of the game fulfills $v^{\mathcal{M}}(R) \in \{0, 1\}$, for all $R \subseteq K$.

Using this proposal of TU game, Datta et al. (2015) innovatively relate their introduced influence measure to the Banzhaf value of the TU game $(K, v^{\mathcal{M}})$. More precisely, when the sample $\mathcal{M}$ corresponds to the TU game $(K, v^{\mathcal{M}})$ and $|\mathcal{B}| = 2$, it follows that

$$B(K, v^{\mathcal{M}}) = \frac{\chi(X, Y, \mathcal{M})}{|\mathcal{A}|}, \tag{9}$$

indicating that the influence measure coincides with the raw Banzhaf value, which is the Banzhaf value without the constant that normalizes it by considering the total number of possible coalitions.[2]

However, as previously justified, it is natural to assume the presence of a partition of the features describing potential dependencies among them. Consequently, a dataset with interactions $(X, Y, P, \mathcal{M})$, with $\mathcal{M}$ corresponding to a TU game $(K, v^{\mathcal{M}})$ and $P$ a partition of $K$, can be identified with a TU game with a priori unions $(K, v^{\mathcal{M}}, P)$.

The following result extends Equation (9) to the case of dependent features and shows that the influence measure for potential dependent features introduced in this paper (see Equation (5)) is a generalization of the Banzhaf-Owen value for simple games with a priori unions.

**Proposition 5.1.** *Let $(X, Y, P, \mathcal{M})$ be a dataset with interactions where the associations between the features within each union are positive. Suppose that $\mathcal{A} = \cup_{i=1}^n X^i$, $|\mathcal{A}_l| = 2$ for all $l \in K$, and $|\mathcal{B}| = 2$. If $(K, v^{\mathcal{M}})$ is the TU game corresponding to $\mathcal{M}$, then it holds that, for every $t \in M$, $l \in P_t$,*

$$BO_l(K, v^{\mathcal{M}}, P) = \frac{\Psi_l(X, Y, P, \mathcal{M})}{|\mathcal{A}^t|},$$

*where $\mathcal{A}^t = \{a = (a_1, \ldots, a_k) \in \mathcal{A} \,:\, \text{if } u \in M \backslash \{t\} \text{ then } a_q = a_v \text{ for all } q, v \in P_u\}$.*

*Proof.* Take $(X, Y, P, \mathcal{M})$ a dataset with positive interactions such that $\mathcal{A} = \cup_{i=1}^n X^i$, with $|\mathcal{A}_l| = 2$ for all $l \in K$, and $|\mathcal{B}| = 2$. We can identify the TU game $(K, v^{\mathcal{M}})$ corresponding to $\mathcal{M}$ according

[2] Given a TU game $(G, v)$ and $l \in G$, the raw Banzhaf value is given by $2^{g-1} \cdot B_l(G, v)$ (see Equation (2)).

to Equation (8). Then, for every feature $l \in K$, it holds that

$$\begin{aligned}
\Psi_l(X,Y,P,\mathcal{M}) &= \sum_{(X^i,Y^i)\in\mathcal{M}^t} \ \sum_{\substack{((X^i_{-l},a_l),b)\in\mathcal{M}^t:\\ a_l\in\mathcal{A}_l,\ b\in\mathcal{B}}} |Y^i - b| \\
&= 2\cdot \sum_{\substack{(X^i,Y^i)\in\mathcal{M}^t:\\ X^i_l=1}} \ \sum_{\substack{((X^i_{-l},0),b)\in\mathcal{M}^t:\\ b\in\mathcal{B}}} |Y^i - b| \\
&= 2\cdot \sum_{S\subseteq M\setminus\{t\}} \ \sum_{R\subseteq P_t\setminus\{l\}} \Big( v^{\mathcal{M}}(\cup_{u\in S}P_u \cup R \cup \{l\}) - v^{\mathcal{M}}(\cup_{u\in S}P_u \cup R) \Big) \\
&= 2\cdot BO_l(K,v^{\mathcal{M}},P)\cdot 2^{m-1}\cdot 2^{|P_t|-1} \\
&= BO_l(K,v^{\mathcal{M}},P)\cdot 2^{m-1+|P_t|} \\
&= BO_l(K,v^{\mathcal{M}},P)\cdot |\mathcal{A}^t|,
\end{aligned}$$

where the second and third equalities hold because, with $|\mathcal{B}| = |\mathcal{A}_l| = 2$, we can assume $\mathcal{B} = \mathcal{A}_l = \{0,1\}$. □

Below, we consider an example that illustrates the previous result.

**Example 5.2.** *Let us take the dataset with interactions $(X,Y,P,\mathcal{M})$ where $X = \{X_1, X_2, X_3, X_4\}$, that is, $k = 4$, and such that $\mathcal{A}_l = \mathcal{B} = \{0,1\}$ for all $l \in K$. Suppose that $\mathcal{A} = \cup_{i=1}^{n} X^i$ and $n = 16$. In addition, for $i \in N$, we have that*

$$Y^i = \begin{cases} 1 & \text{if } X^i \in \{0011, 1110, 1011, 0111, 1111\}, \\ 0 & \text{otherwise.} \end{cases}$$

*The dataset with interactions $(X,Y,P,\mathcal{M})$ can be identified with a weighted majority game with a priori unions $(K, v^{\mathcal{M}}, P)$ where, for example, the vector of weights is $w = (1,\ 1,\ 4,\ 3)$ and the quota is $a = 6$. Table 7 displays the influence measure introduced in this paper for the above-specified dataset, considering all possible partitions of the feature set. By Proposition 5.1, this influence measure coincides with the Banzhaf-Owen value. All results have been computed using the R library powerindexR*[3].

| Scenario | Feature set partitions ($P$) | Influence measures of the features ($BO$) |
|---|---|---|
| 1 | $\{\{1\},\{2\},\{3\},\{4\}\}$ | (0.125, 0.125, 0.625, 0.375) |
| 2 | $\{\{1\},\{2\},\{3,4\}\}$ | (0.000, 0.000, 0,625, 0.375) |
| 3 | $\{\{1\},\{3\},\{2,4\}\}$ | (0.000, 0.125, 0.500, 0.375) |
| 4 | $\{\{1\},\{4\},\{2,3\}\}$ | (0.250, 0.125, 0.625, 0.250) |
| 5 | $\{\{2\},\{3\},\{1,4\}\}$ | (0.125, 0.000, 0.500, 0.375) |
| 6 | $\{\{2\},\{4\},\{1,3\}\}$ | (0.125, 0.250, 0.625, 0.250) |
| 7 | $\{\{3\},\{4\},\{1,2\}\}$ | (0.125, 0.125, 0.750, 0.250) |
| 8 | $\{\{1,2\},\{3,4\}\}$ | (0.000, 0.000, 0.750, 0.250) |
| 9 | $\{\{1,3\},\{2,4\}\}$ | (0.000, 0.250, 0.500, 0.250) |
| 10 | $\{\{1,4\},\{2,3\}\}$ | (0.250, 0.000, 0.500, 0.250) |
| 11 | $\{\{1\},\{2,3,4\}\}$ | (0.000, 0.125, 0.625, 0.375) |
| 12 | $\{\{2\},\{1,3,4\}\}$ | (0.125, 0.000, 0.625, 0.375) |
| 13 | $\{\{3\},\{1,2,4\}\}$ | (0.125, 0.125, 0.500, 0.375) |
| 14 | $\{\{4\},\{1,2,3\}\}$ | (0.125, 0.125, 0.625, 0.000) |
| 15 | $\{\{1,2,3,4\}\}$ | (0.125, 0.125, 0.625, 0.375) |

Table 7: Numerical results for our influence measure in the 15 possible scenarios of Example 5.2.

[3] More information is available at `https://cran.r-project.org/web/packages/powerindexR/index.html`.

*From Table 7, we can clearly observe the behavior of various properties of our influence measure. Notably, in scenarios 1 and 2, features 3 and 4 exhibit identical influence, reflecting the property of indifference to interactions. This property is further confirmed by looking at the influence of features 3 and 4 in scenarios 2 and 12. Additionally, in scenario 1, where all feature unions have size 1, every feature is influential. However, when dependencies between groups of two or three features are considered, some features become non-influential, as demonstrated in scenarios 9 and 10.* △

*Remark* 5.1. Proposition 5.1 is not satisfied if $|\mathcal{B}| < \infty$, $|\mathcal{B}| \neq 2$ and $\Psi$ is naturally rewritten as suggested in Remark 4.2. Now, if instead of this expression, one considers the original form given in Equation (6), and the TU game corresponding to the sample is monotone, it is easy to see that Proposition 5.1 still holds. This remains true even though the game is no longer simple when the outcome is non-binary. ◇

To summarize, this section has shown the conditions under which a dataset with interactions can be identified with a game with a priori unions, and we have demonstrated that the influence measure introduced in this paper coincides with the Banzhaf-Owen value of such a game. An example illustrates the computation of the influence measure through the Banzhaf-Owen value and provides clarification on the interpretation of some of the properties that this measure satisfies. Additionally, Remark 5.1 underlines the existence of various possibilities for extending the methodology proposed in this study to the case of a non-binary outcome.

# 6 Numerical results

This section analyzes the performance of our influence measure on three different datasets that will be further described below. The proposed influence measure has been implemented in `R` 4.3.3 and a set of numerical experiments were run on a quad-core Intel i7-8665U CPU with 16 GB RAM.

The following subsections report the computational study. We first present the experimental details in Subsection 6.1, where we describe the scenarios under which we will compute our influence measure. These will depend on both the predictive model used (namely, the classifier) and the chosen coalitional structure. Next, Subsection 6.2 explores the influence of certain factors on the severity of a car crash, while Subsection 6.3 examines the influence of listening to various music groups or singers on the likelihood of listening to other artists. We will also discuss how the chosen classifiers and partitions impact feature rankings. Subsection 6.4 applies the estimation methodology presented in this paper to a large-scale banking context. All the datasets used are available at `https://github.com/LauraDavilaPena/GT-based_IM`.

## 6.1 Experimental details

We compute our influence measure for each dataset in several scenarios. On the one hand, we consider three different predictive models from the families of random forests (RFs), support vector machines (SVMs), and logistic regression (LR), chosen for their well-documented performance as classifier types (Fernández-Delgado et al., 2014). We use the base implementations of these models provided by the `RWeka`[4] library in `R` software, specifically the Breiman's random

[4] More information is available at `https://cran.r-project.org/web/packages/RWeka/RWeka.pdf`.

forest classifier (Breiman, 2001), the Platt's sequential minimal optimization (SMO) algorithm for training a support vector machine classifier (Platt, 1998), and a modified le Cessie and van Houwelingen's multinomial logistic regression model with a ridge estimator (le Cessie and van Houwelingen, 1992), respectively. On the other hand, we consider three distinct coalitional structures: 1) a singleton-based partition (SBP), 2) a constructed-by-design partition (CDP), and 3) a hierarchical clustering partition (HCP). For this latter case, the number of clusters is selected to match the number of coalitions in the constructed-by-design partition. Furthermore, given the binary nature of our datasets, we use the Jaccard distance (Remark 6.1) as a dissimilarity measure to obtain these clusters. Table 8 provides a summary of the nine scenarios considered in this study.

| | | Coalitional structure | | |
|---|---|---|---|---|
| | | SBP | CDP | HCP |
| Predictive model | RF | RF_SBP | RF_CDP | RF_HCP |
| | SVM | SVM_SBP | SVM_CDP | SVM_HCP |
| | LR | LR_SBP | LR_CDP | LR_HCP |

Table 8: Summary of scenarios based on the predictive model and the coalitional structure considered.

*Remark* 6.1. Let $(X, Y, \mathcal{M})$ be a binary dataset, where $\mathcal{M} = \{(X^i, Y^i)\}_{i=1}^n$, and consider two features $X_l$ and $X_{l'}$. The Jaccard distance between $X_l$ and $X_{l'}$, $d_J(X_l, X_{l'})$, is calculated as follows:

$$d_J(X_l, X_{l'}) = \frac{\sum_{i=1}^n \mathbb{1}_{\{X_l^i=1 \text{ or } X_{l'}^i=1\}} - \sum_{i=1}^n \mathbb{1}_{\{X_l^i=1 \text{ and } X_{l'}^i=1\}}}{\sum_{i=1}^n \mathbb{1}_{\{X_l^i=1 \text{ or } X_{l'}^i=1\}}},$$

where the numerator represents the number of individuals for which features $X_l$ and $X_{l'}$ differ (i.e., one of them is 1 while the other is 0), and the denominator counts the number of individuals for whom at least one of these features takes the value 1. This distance ranges from 0 to 1, where values closer to 0 indicate very similar features and values closer to 1 indicate highly dissimilar features, respectively. ◇

## 6.2 On the analysis of car crash fatalities

In this section, we apply our proposed influence measure to the relevant context of occupant safety in car crashes. The severity of an accident can vary significantly based on the type of collision. For instance, a head-on collision differs substantially from a side-impact collision or a vehicle rollover. Furthermore, the risk of fatality in a car older than ten years old is twice as high as in a newer vehicle. We will specifically look at the variables that influence the likelihood of fatalities in vehicle accidents. To achieve this, we examine the influence of various factors that help us describe the nature of a road accident.

### 6.2.1 Description of the data

We consider the variables included in the `nassCDS` dataset from the `DAAG R` package (Maindonald and Braun, 2021). This dataset contains information on car crashes in the US in the period 1997-2002, reported by the police, in which there is an injury (to person or property) and at least one vehicle is towed. The data is limited to front-seat occupants and includes only a subset of the recorded variables, with additional restrictions. For our application, we have conveniently adapted this database to focus on binary variables and to address a classification problem. The selected outcome corresponds to a feature in the original dataset. Table 9 lists the characteristics of both the original and the transformed datasets, referred to as `cars_original` and `cars_binary`, respectively. This latter one consists of a sample of 17,565 observations for the 10 binary features listed in Table 10.

| | # of observations | # of features | Nature of features |
|---|---|---|---|
| `cars_original` | 17,565 | 15 | multi-class, continuous |
| `cars_binary` | 17,565 | 10 + outcome | binary |

Table 9: Characteristics of the original and transformed datasets for the analysis of car crash fatalities.

| Outcome | | Description |
|---|---|---|
| | `deceased` | Binary variable indicating whether the person involved in the car accident is deceased (1) or not (0). |
| **Feature** | | **Description** |
| 1 | `dvcat` | Binary variable indicating whether the vehicle, at the moment of the accident, was traveling at a speed higher than 55 km/h (1) or not (0). |
| 2 | `airbag` | Binary variable indicating whether the vehicle had an airbag system (1) or not (0). |
| 3 | `seatbelt` | Binary variable indicating whether the person involved was wearing a seat belt (1) or not (0). |
| 4 | `frontal` | Binary variable indicating whether the vehicle crash was frontal (1) or non-frontal (0). |
| 5 | `sex` | Binary variable indicating the sex of the person involved: 1 for male and 0 for female. |
| 6 | `ageOFocc` | Binary variable indicating whether the person was 30 years old or less (1) or over 30 years old (0). |
| 7 | `abcat` | Binary variable indicating airbag activation: 1 if one or more airbags in the vehicle were activated (even if not deployed) and 0 if none were deployed (either due to malfunction or being disabled). |
| 8 | `occRole` | Binary variable indicating whether the person involved was the driver (1) or a passenger (0) of the vehicle. |
| 9 | `deploy` | Binary variable indicating whether the airbag functioned correctly (1) or was unavailable or not functioning (0). |
| 10 | `age` | Binary variable indicating whether the vehicle was 10 years old or more (1) or less than 10 years old (0). |

Table 10: Summary of the considered features in the analysis of car crash fatalities.

### 6.2.2 Performance of our influence measure

From the sample, we obtain the influence measure both with and without a coalitional structure of the features across various scenarios. Initially, we consider the case of the singletons-based partition (SBP), meaning we do not account for potential affinity relations between the features; this corresponds to the original case presented by Datta et al. (2015). Next, we examine a constructed-by-design partition (CDP) where features 1, 4, and 6 are grouped in one block, 2, 7, and 9 in another block, and features 3, 5, 8, and 10 act individually. The choice for the first group is motivated by the common belief that younger drivers tend to drive faster, and high-speed driving is often associated with frontal accidents. For the second group, we consider all features related to airbags. Finally, we adopt a coalitional structure specified by hierarchical clustering, fixing the number of clusters at six. This hierarchical clustering partition (HCP) joins features 2, 7, and 9 in the same block, features 3, 4, and 8 in another block, and leaves 1, 5, 6, and 10 alone. For each case, we compute the resulting influence measure using the above-specified random forest (RF), support vector machine (SVM), and logistic regression (LR) classifiers as predictive models. Table 11 displays the numerical results.

| Feature | RF_SBP | SVM_SBP | LR_SBP | RF_CDP | SVM_CDP | LR_CDP | RF_HCP | SVM_HCP | LR_HCP |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.4280 | 0 | 0.1952 | 0.5680 | 0 | 0.2590 | 0.0319 | 0 | 0.0383 |
| 2 | 0.0087 | 0 | 0.0016 | 0.0037 | 0 | 0 | 0.0080 | 0 | 0.0006 |
| 3 | 0.0285 | 0 | 0.0829 | 0 | 0 | 0 | 0.0193 | 0 | 0.0349 |
| 4 | 0.1197 | 0 | 0.3473 | 0.1354 | 0 | 0.1312 | 0.1354 | 0 | 0.1312 |
| 5 | 0.0345 | 0 | 0.0170 | 0 | 0 | 0 | 0.0011 | 0 | 0 |
| 6 | 0.0351 | 0 | 0.0065 | 0.0289 | 0 | 0.0006 | 0.0032 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0.0363 | 0 | 0.0033 | 0 | 0 | 0 | 0.0298 | 0 | 0 |
| 9 | 0.0281 | 0 | 0.0048 | 0.0174 | 0 | 0 | 0.0216 | 0 | 0.0018 |
| 10 | 0.0175 | 0 | 0.0014 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 11: Numerical results in the analysis of car crash fatalities.

Based on these findings, several conclusions can be drawn. Features 1 and 4 consistently hold the top two positions in the ranking of the most influential features. Specifically, they refer to the vehicle's speed and whether the collision is frontal. Factors such as the individual's role (passenger or driver), represented by feature 8, seat belt usage (feature 3), or the person's age (feature 6) rank third in the different rankings obtained, as can be seen in Table 12. Notably, regardless of the scenario examined, the airbag activation, represented by feature 7, appears to have no influence.

| Position | RF_SBP | SVM_SBP | LR_SBP | RF_CDP | SVM_CDP | LR_CDP | RF_HCP | SVM_HCP | LR_HCP |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | - | 4 | 1 | - | 1 | 4 | - | 4 |
| 2 | 4 | - | 1 | 4 | - | 4 | 1 | - | 1 |
| 3 | 8 | - | 3 | 6 | - | 6 | 8 | - | 3 |

Table 12: Top three features in the analysis of car crash fatalities.

### Assessing the accuracy of the estimation

Next, we address the problem of measuring the influence of such features from the perspective of its approximation. It is worth noting that, despite having 17,565 observations, we can still compute the measure directly. This is important, as it allows us to evaluate the performance of the procedure outlined in Section 4.5 in practice.

For this purpose, we consider a sampling size of $\ell = 295$ observations, which corresponds to 10% of the total in $\mathcal{M}^t$. Table 13 displays the associated results.

In view of the results obtained, the conclusions regarding the most influential features and those for which our measure gives a value of 0 remain consistent. Table 14 completes this study by including the results of estimating the influence measure using simple random sampling with

| Feature | RF_SBP | SVM_SBP | LR_SBP | RF_CDP | SVM_CDP | LR_CDP | RF_HCP | SVM_HCP | LR_HCP |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.7551 | 0 | 0.2080 | 0.5292 | 0 | 0.2520 | 0.0296 | 0 | 0.0365 |
| 2 | 0.0062 | 0 | 0.0006 | 0.0077 | 0 | 0 | 0.0061 | 0 | 0.0003 |
| 3 | 0.0403 | 0 | 0.1006 | 0 | 0 | 0 | 0.0186 | 0 | 0.0335 |
| 4 | 0.1358 | 0 | 0.3153 | 0.1241 | 0 | 0.1304 | 0.1241 | 0 | 0.1304 |
| 5 | 0.0540 | 0 | 0.0114 | 0 | 0 | 0 | 0.0009 | 0 | 0 |
| 6 | 0.0233 | 0 | 0.0057 | 0.0274 | 0 | 0.0007 | 0.0021 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0.0210 | 0 | 0.0006 | 0 | 0 | 0 | 0.0144 | 0 | 0 |
| 9 | 0.0426 | 0 | 0.0051 | 0.0322 | 0 | 0 | 0.0176 | 0 | 0.0015 |
| 10 | 0.0091 | 0 | 0.0011 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 13: Estimation of the influence measure using sampling without replacement in the analysis of car crash fatalities by using 10% of the sample.

| Feature | RF_SBP | SVM_SBP | LR_SBP | RF_CDP | SVM_CDP | LR_CDP | RF_HCP | SVM_HCP | LR_HCP |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.4874 | 0 | 0.1965 | 0.5263 | 0 | 0.24686 | 0.0293 | 0 | 0.0362 |
| 2 | 0.0075 | 0 | 0.0014 | 0.0077 | 0 | 0 | 0.0061 | 0 | 0.0004 |
| 3 | 0.0293 | 0 | 0.0828 | 0 | 0 | 0 | 0.0194 | 0 | 0.0342 |
| 4 | 0.1370 | 0 | 0.3375 | 0.1231 | 0 | 0.1318 | 0.1231 | 0 | 0.1318 |
| 5 | 0.0388 | 0 | 0.0161 | 0 | 0 | 0 | 0.0009 | 0 | 0 |
| 6 | 0.0351 | 0 | 0.0057 | 0.0261 | 0 | 0.0008 | 0.0021 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |  |
| 8 | 0.0355 | 0 | 0.0034 | 0 | 0 | 0 | 0.0146 | 0 | 0 |
| 9 | 0.0296 | 0 | 0.0043 | 0.0330 | 0 | 0 | 0.0171 | 0 | 0.0022 |
| 10 | 0.0134 | 0 | 0.0015 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 14: Estimation of the influence measure using sampling with replacement ($wR$) in the analysis of car crash fatalities by using 10% of the sample.

replacement ($wR$). While it seems obvious from their construction that the numerical results may differ from the exact ones, their accuracy is supported by the study of correlations for each pair of non-null vectors of the influence measure, as shown in Table 15.

| Correlations | RF_SBP | LR_SBP | RF_CDP | LR_CDP | RF_HCP | LR_HCP |
|---|---|---|---|---|---|---|
| Real vs. estimation | 0.9945 | 0.9947 | 0.9995 | 0.9999 | 0.9941 | 0.9999 |
| Real vs. estimation ($wR$) | 0.9998 | 0.9998 | 0.9994 | 0.9997 | 0.9941 | 0.9998 |

Table 15: Correlations between exact and approximate numerical results in the analysis of car crash fatalities.

### Assessing the impact of the imbalanced distribution in the outcome

It is also worth noting that the SVM classifier does not yield conclusive results when using measures based on Datta et al.'s methodology. In such cases, influence measures always return a value of 0. As discussed in the literature, the imbalanced distribution between 0s and 1s in the outcome can lead to influence measure outcomes that may not accurately reflect reality. As shown in Table 16, SVM does not predict any occurrences of the value 1 for the outcome in this particular case study.

| Value | Original outcome | RF | SVM | LR |
|---|---|---|---|---|
| 0 | 16,788 | 17,496 | 17,565 | 17,464 |
| 1 | 777 | 69 | 0 | 101 |

Table 16: Summary of the outcomes from both the original database and those predicted by the classifiers.

To check this conjecture, we draw a subsample from the original database consisting of 777 instances with an outcome of 1 and 777 instances with an outcome of 0 (balanced in the outcome). Table 17 presents the numerical results for this ad hoc subsample. Interestingly, in this case, we observe that the influence measures based on Datta et al.'s methodology are not zero for SVM's scenarios.

| Feature | RF_SBP | SVM_SBP | LR_SBP | RF_CDP | SVM_CDP | LR_CDP | RF_HCP | SVM_HCP | LR_HCP |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.9485 | 2.0193 | 2.2368 | 2.7856 | 2.8868 | 3.1978 | 4.0967 | 4.0967 | 4.9517 |
| 2 | 0.2278 | 0.1750 | 1.2432 | 0.1542 | 0.0398 | 3.8010 | 0.3413 | 0.0220 | 0.4551 |
| 3 | 8.1519 | 3.6049 | 7.1145 | 7.5443 | 1.9662 | 0 | 15.9695 | 6.5242 | 14.9720 |
| 4 | 3.3333 | 6.3977 | 5.1030 | 4.5925 | 8.9384 | 7.1334 | 4.5925 | 8.9384 | 7.1334 |
| 5 | 0.7902 | 1.1004 | 0 | 0.2194 | 2.6329 | 0 | 0.4857 | 1.5731 | 0 |
| 6 | 3.7877 | 2.5071 | 4.7799 | 5.1481 | 2.9292 | 6.3882 | 9.4809 | 4.0458 | 10.2239 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 1.5714 | 2.0438 | 0 | 0.9451 | 1.1477 | 0 | 1.1941 | 1.6707 | 0 |
| 9 | 0.0026 | 0 | 0.0103 | 0.0100 | 0 | 0.0398 | 0 | 0 | 0.0220 |
| 10 | 0.7143 | 0.2471 | 0 | 0 | 1.6203 | 0 | 1.5216 | 0 | 0 |

Table 17: Numerical results in the analysis of car crash fatalities over the considered subsample.

A consistent finding shared with the original case is the lack of influence of feature 7 (the airbag activation) across all studied scenarios. Furthermore, Table 18 presents additional insights, showcasing the top three features.

| Position | RF_SBP | SVM_SBP | LR_SBP | RF_CDP | SVM_CDP | LR_CDP | RF_HCP | SVM_HCP | LR_HCP |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 4 | 3 | 3 | 4 | 4 | 3 | 4 | 3 |
| 2 | 6 | 3 | 4 | 6 | 6 | 6 | 6 | 3 | 6 |
| 3 | 4 | 6 | 6 | 4 | 1 | 2 | 4 | 1 | 4 |

Table 18: Top three features in the analysis of car crash fatalities over the considered subsample.

Features 3, 4, and 6, which relate to seat belt usage, crash type, and the person's age, respectively, are usually the most influential factors (with slight variations in order) for the influence measure. It is reasonable to assume that the three mentioned features play a pivotal role in the consequences of a car crash fatality. Feature 3 almost always ranks first in the three scenarios considered for RF and LR, with the only exception of the partition constructed by mere feature observation for LR. In the other cases, feature 4 ranks first. In the hierarchical clustering scenario with SVM, feature 6 is replaced by feature 1 (the vehicle's speed). More disparities are found when considering the partition constructed by mere feature observation. Feature 3 is replaced by feature 1 when SVM is selected and by feature 2 (the presence of an airbag) when LR is used.

The proposed influence measure has the potential to handle any dataset as an input, regardless of its imbalanced nature. Thus, a key advantage of our method is its broad applicability, as readers could check, even when dealing with imbalanced data. Table 17 displays the performance of our approach on a balanced dataset derived from the original one. However, modifying the dataset to achieve balance inevitably alters the problem's original nature, which is an important issue to consider. Since real-world datasets vary significantly and are often inherently imbalanced, our method's ability to handle such cases enhances its practical relevance and adaptability.

### Assessing the impact of missing feature values

We have already observed how the characteristics of the dataset under study influence the obtained results. Although our influence measure is designed to be applicable to any complete dataset with interactions, we aimed to assess the impact of missing feature values on the inferred partition and, consequently, on the influence scores.

For the `cars_binary` dataset (Table 9) we randomly removed 5% of the values from each feature. The missing values were then imputed using the `mice` package in `R` software (van Buuren and Groothuis-Oudshoorn, 2011), employing a Random Forest-based imputation method, which is well-suited for categorical data. Using the imputed dataset, we applied hierarchical clustering with the Jaccard distance to obtain the hierarchical clustering partition (HCP). The resulting HCP matched the one obtained with the complete `cars_binary` dataset. Table 19 presents the numerical results applying our influence measure under the `RF_HCP`, `SVM_HCP`, and `LR_HCP` scenarios to this imputed dataset with interactions. The Spearman correlations between these influence scores and those in Table 11 are 0.9058 and 0.7658 for `RF_HCP` and `LR_HCP`, respectively. These high correlations suggest that, when the proportion of missing data is not too large and an appropriate imputation method is chosen, the influence scores can largely be preserved.

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| `RF_HCP` | 0.0407 | 0.0068 | 0.0475 | 0.2189 | 0.0018 | 0.0071 | 0 | 0.0516 | 0.0138 | 0.0050 |
| `SVM_HCP` | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| `LR_HCP` | 0.0492 | 0.0003 | 0.0141 | 0.0618 | 0.0018 | 0 | 0 | 0.0121 | 0.0015 | 0 |

Table 19: Numerical results in the analysis of car crash fatalities over a dataset with missing feature values.

We also examined more extreme cases where the percentage of missing data exceeded 50%. In these situations, hierarchical clustering based on the Jaccard distance, when applied to the imputed dataset, did not always return the same partition. Naturally, this led to more substantial differences in the influence scores.

As a final thought, the numerical results presented here seem to be strongly dependent on the selected sample and the methodology employed in each case. Nevertheless, it is notable that the ultimate discrepancies are minimal in terms of rankings, with only a few variations in the positions. The incorporation of information regarding feature affinities through partitions may justify this fact. To provide deeper insights into the generalizability of our model-agnostic influence measure, we expand its experimental validation by analyzing additional scenarios. In particular, we consider two additional classifiers implemented via `RWeka`: a multilayer perceptron (MLP), based on a neural network model for complex pattern recognition, and a bagging (B) classifier, an ensemble meta-estimator. These classifiers are evaluated across the three coalitional structures described earlier. Table 20 presents the numerical results obtained by applying our influence measure to all scenarios—both the original nine from Table 8 and six additional ones, following the same terminology—in the `cars_binary` dataset. Furthermore, Table 21 highlights the top three features for each of these 15 scenarios. A brief comparative analysis of our method's performance across these different predictive models, along with a discussion on the impact of classifier selection and coalitional structures, is provided in Appendix B of the Online Resource

Section (ORS).

| Feature | RF_SBP | SVM_SBP | LR_SBP | MLP_SBP | B_SBP | RF_CDP | SVM_CDP | LR_CDP | MLP_CDP | B_CDP | RF_HCP | SVM_HCP | LR_HCP | MLP_HCP | B_HCP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.4280 | 0 | 0.1952 | 0.0740 | 0.0630 | 0.5680 | 0 | 0.2590 | 0.0982 | 0.0836 | 0.0319 | 0 | 0.0383 | 0.0319 | 0.0053 |
| 2 | 0.0087 | 0 | 0.0016 | 0.0008 | 0.0023 | 0.0037 | 0 | 0 | 0 | 0 | 0.0080 | 0 | 0.0006 | 0.0012 | 0.0059 |
| 3 | 0.0285 | 0 | 0.0829 | 0.0092 | 0.0076 | 0 | 0 | 0 | 0 | 0 | 0.0193 | 0 | 0.0349 | 0.0104 | 0.0095 |
| 4 | 0.1197 | 0 | 0.3473 | 0.0463 | 0.0369 | 0.1354 | 0 | 0.1312 | 0.0529 | 0.0481 | 0.1354 | 0 | 0.1312 | 0.0529 | 0.0481 |
| 5 | 0.0345 | 0 | 0.0170 | 0.0016 | 0.0017 | 0 | 0 | 0 | 0 | 0 | 0.0011 | 0 | 0 | 0.0011 | 0.0011 |
| 6 | 0.0351 | 0 | 0.0065 | 0.0208 | 0.0177 | 0.0289 | 0 | 0.0006 | 0.0248 | 0.0233 | 0.0032 | 0 | 0 | 0.0032 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0.0363 | 0 | 0.0033 | 0.0117 | 0.0065 | 0 | 0 | 0 | 0 | 0 | 0.0298 | 0 | 0 | 0.0142 | 0.0054 |
| 9 | 0.0281 | 0 | 0.0048 | 0.0022 | 0.0056 | 0.0174 | 0 | 0 | 0 | 0 | 0.0216 | 0 | 0.0018 | 0.0021 | 0.0118 |
| 10 | 0.0175 | 0 | 0.0014 | 0.0050 | 0.0005 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 20: Numerical results in the analysis of car crash fatalities using additional classifiers.

| Position | RF_SBP | SVM_SBP | LR_SBP | MLP_SBP | B_SBP | RF_CDP | SVM_CDP | LR_CDP | MLP_CDP | B_CDP | RF_HCP | SVM_HCP | LR_HCP | MLP_HCP | B_HCP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | - | 4 | 1 | 1 | 1 | - | 1 | 1 | 1 | 4 | - | 4 | 4 | 4 |
| 2 | 4 | - | 1 | 4 | 4 | 4 | - | 4 | 4 | 4 | 1 | - | 1 | 1 | 9 |
| 3 | 8 | - | 3 | 6 | 6 | 9 | - | 6 | 6 | 6 | 8 | - | 3 | 8 | 3 |

Table 21: Top three features in the analysis of car crash fatalities using additional classifiers.

### 6.3 On the analysis of musical taste in Spotify

Next, we use our proposed influence measure to analyze musical tastes from a Spotify database. Our goal is to identify the most influential singers or music groups.

#### 6.3.1 Description of the data

The original dataset (Köhler, 2017) contains a total of 285 artists and 1226 users and indicates, for each of the users, whether they have listened to each artist or not, making it inherently a binary dataset. For our study, we have decided to consider only the 75 most listened-to artists, all with an audience share of no less than 5%, to ensure results in a reasonable computational time and to avoid examining those artists who are already of little interest. Additionally, to apply our methodology, we need to frame a classification problem and define our outcome. We create 10 independent classification problems by designating 10 selected artists as the outcome, aiming to examine whether listening to the remaining 74 artists affects each specified outcome.

Table 22 presents the 75 selected artists, showcasing the percentage of listeners and the decade in which their band was created. Artists chosen as outcomes for the classification problems are highlighted in bold.

#### 6.3.2 Performance of our influence measure

In this example, we compare the results obtained using two different coalitional structures. First, we consider the case where the partition is determined by the decade of the band's foundation, as shown in Table 22, which will serve as our CDP. Table 23 summarizes the number of artists in each group. Then, as in Subsection 6.2, we consider hierarchical clustering techniques to obtain a

| Artist | Decade | % of listeners | Artist | Decade | % of listeners |
|---|---|---|---|---|---|
| linkin.park | 1990 | 16.2 | johnny.cash | 1950 | 7.3 |
| **coldplay** | 1990 | 16.1 | kings.of.leon | 2000 | 7.3 |
| red.hot.chili.peppers | 1980 | 15.1 | **amy.winehouse** | 2000 | 7.2 |
| rammstein | 1990 | 15 | depeche.mode | 1980 | 7.2 |
| system.of.a.down | 1990 | 13.2 | bullet.for.my.valentine | 1990 | 7 |
| **metallica** | 1980 | 12.2 | in.extremo | 1990 | 7 |
| die.toten.hosen | 1980 | 11.8 | blink.182 | 1990 | 6.9 |
| billy.talent | 2000 | 10.8 | slipknot | 1990 | 6.7 |
| **the.killers** | 2000 | 10.8 | death.cab.for.cutie | 1990 | 6.5 |
| **the.beatles** | 1960 | 10.6 | daft.punk | 1990 | 6.4 |
| jack.johnson | 2000 | 10 | limp.bizkit | 1990 | 6.3 |
| **muse** | 1990 | 10 | sum.41 | 2000 | 6.2 |
| beatsteaks | 1990 | 9.5 | fall.out.boy | 2000 | 6.1 |
| foo.fighters | 1990 | 9.5 | schandmaul | 1990 | 6.1 |
| nirvana | 1980 | 9.5 | kanye.west | 2000 | 6 |
| radiohead | 1990 | 9.4 | seeed | 1990 | 6 |
| arctic.monkeys | 2000 | 9.1 | tenacious.d | 1990 | 6 |
| placebo | 1990 | 9.1 | **rihanna** | 2000 | 6 |
| bloc.party | 2000 | 8.8 | papa.roach | 1990 | 5.8 |
| evanescence | 1990 | 8.8 | portishead | 1990 | 5.8 |
| rise.against | 2000 | 8.7 | rage | 1990 | 5.8 |
| the.kooks | 2000 | 8.7 | franz.ferdinand | 2000 | 5.7 |
| mando.diao | 2000 | 8.6 | marilyn.manson | 1980 | 5.6 |
| the.white.stripes | 1990 | 8.4 | nelly.furtado | 2000 | 5.6 |
| deichkind | 1990 | 8.3 | **queen** | 1970 | 5.5 |
| incubus | 1990 | 8.3 | **bob.marley** | 1960 | 5.5 |
| farin.urlaub | 1990 | 8.2 | feist | 1990 | 5.5 |
| in.flames | 1990 | 8.2 | massive.attack | 1980 | 5.4 |
| clueso | 2000 | 8.1 | queens.of.the.stone.age | 1990 | 5.4 |
| peter.fox | 2000 | 8 | iron.maiden | 1970 | 5.2 |
| the.offspring | 1980 | 8 | avril.lavigne | 2000 | 5.1 |
| air | 1990 | 7.8 | amon.amarth | 1990 | 5.1 |
| subway.to.sally | 1990 | 7.7 | apocalyptica | 1990 | 5.1 |
| nightwish | 1990 | 7.7 | gorillaz | 1990 | 5.1 |
| the.prodigy | 1990 | 7.4 | nine.inch.nails | 1980 | 5.1 |
| disturbed | 1990 | 7.3 | oasis | 1990 | 5.1 |
| **ac.dc** | 1970 | 7.3 | children.of.bodom | 1990 | 5 |
| green.day | 1980 | 7.3 | | | |

Table 22: List of the 75 most listened-to music bands in the database, with the decade of the band's creation and percentage of users who listen to it. In bold, those artists selected as outcomes for the classification problems.

partition of the bands for each of the considered outcomes, that is, the HCP. Note that the large number of bands involved makes it computationally infeasible to calculate exactly the influence measure for the SBP, that is, Datta et al.'s influence measure (4).

| Decade | 1950 | 1960 | 1970 | 1980 | 1990 | 2000 |
|---|---|---|---|---|---|---|
| # of artists | 1 | 2 | 3 | 10 | 40 | 19 |

Table 23: Distribution of artists per decade.

Appendix C in the Online Resource Section (ORS) includes the numerical results for the overall lists of bands in both cases under study. In particular, Table C.4 in the ORS presents the results obtained with our influence measure when bands are grouped according to the CDP, and Table C.5 in the ORS does the same for the case of considering the HCP. In view of these results, only for those classification problems where the outcomes are listening to or not listening to Rihanna or Metallica does our influence measure have some non-zero component when using SVM or LR. In particular, when studying the potential influences of other bands on the likelihood of listening to Metallica, we find that Iron Maiden and Nightwish have an influence when using a

HCP in combination with LR.

| | | Position 1 | Position 2 | Position 3 | | Position 1 | Position 2 | Position 3 |
|---|---|---|---|---|---|---|---|---|
| coldplay | RF_CDP | avril.lavigne | muse | air | RF_HCP | avril.lavigne | muse | air |
| | SVM_CDP | - | - | - | SVM_HCP | - | - | - |
| | LR_CDP | - | - | - | LR_HCP | - | - | - |
| metallica | RF_CDP | nightwish | rage | evanescence | RF_HCP | nightwish | iron.maiden | amon.amarth |
| | SVM_CDP | - | - | - | SVM_HCP | - | - | - |
| | LR_CDP | - | - | - | LR_HCP | nightwish | iron.maiden | - |
| the.killers | RF_CDP | green.day | the.white.stripes | kings.of.leon | RF_HCP | green.day | franz.ferdinand | evanescence |
| | SVM_CDP | - | - | - | SVM_HCP | - | - | - |
| | LR_CDP | - | - | - | LR_HCP | - | - | - |
| the.beatles | RF_CDP | kings.of.leon | air | franz.ferdinand | RF_HCP | kings.of.leon | muse | apocalyptica |
| | SVM_CDP | - | - | - | SVM_HCP | - | - | - |
| | LR_CDP | - | - | - | LR_HCP | - | - | - |
| muse | RF_CDP | avril.lavigne | evanescence | placebo | RF_HCP | evanescence | avril.lavigne | - |
| | SVM_CDP | - | - | - | SVM_HCP | - | - | - |
| | LR_CDP | - | - | - | LR_HCP | - | - | - |
| ac.dc | RF_CDP | rage | air | the.prodigy | RF_HCP | rage | evanescence | air |
| | SVM_CDP | - | - | - | SVM_HCP | - | - | - |
| | LR_CDP | - | - | - | LR_HCP | - | - | - |
| amy.winehouse | RF_CDP | kings.of.leon | nelly.furtado | peter.fox | RF_HCP | kings.of.leon | bob.marley | the.beatles |
| | SVM_CDP | - | - | - | SVM_HCP | - | - | - |
| | LR_CDP | - | - | - | LR_HCP | - | - | - |
| rihanna | RF_CDP | limp.bizkit | the.prodigy | amy.winehouse | RF_HCP | evanescence | seeed | nelly.furtado |
| | SVM_CDP | nelly.furtado | kanye.west | - | SVM_HCP | seeed | nelly.furtado | kanye.west |
| | LR_CDP | nelly.furtado | kanye.west | - | LR_HCP | seeed | nelly.furtado | kanye.west |
| queen | RF_CDP | amon.amarth | the.prodigy | muse | RF_HCP | feist | muse | the.beatles |
| | SVM_CDP | - | - | - | SVM_HCP | - | - | - |
| | LR_CDP | - | - | - | LR_HCP | - | - | - |
| bob.marley | RF_CDP | seeed | jack.johnson | the.killers | RF_HCP | seeed | the.killers | peter.fox |
| | SVM_CDP | - | - | - | SVM_HCP | - | - | - |
| | LR_CDP | - | - | - | LR_HCP | - | - | - |

Table 24: Top three bands in the analysis of musical taste in Spotify.

Table 24 summarizes the three most influential bands for each scenario. The numerical results show some similarities, which are detailed below. For instance, according to RF, Avril Lavigne, Muse, and Air are the three most influential bands for listening to Coldplay. In the case of The Killers, Green Day is the most influential band; for The Beatles, Kings of Leon; for Muse, Avril Lavigne and Evanescence swap positions when using the CDP and the HCP; for AC/DC, Rage and Air are both in the top three; for Amy Winehouse, the most influential band is Kings of Leon; and for Queen, the only coincidence is that Muse is in position 3 when using the CDP and in position 2 when using the HCP. Finally, Seeed is the most influential band for listening to Bob Marley, although The Killers are in positions 3 and 2 for the CDP and the HCP, respectively. In all these cases, no other bands are influential for SVM and LR. However, for Metallica, Nightwish is the most influential band according to RF and both partition structures, as well as LR with the HCP. Also in this latter case, Iron Maiden is in the second position under RF and LR. When analyzing the case of Rihanna, Nelly Furtado ranks first for the CDP with SVM and LR, second for the HCP with SVM or LR, and third for the HCP with RF. Seeed is the most influential when using the HCP with SVM and LR, and ranks second with RF. We also highlight the case of Kanye West, in positions 2 and 3 for both SVM and LR with CDP and HCP, respectively.

### The case of the 15 most listened-to music bands

As the reader can check in the previous section, the SVM and LR classifiers also did not yield conclusive results in the analysis of musical tastes. Therefore, we have narrowed our study to focus only on the 15 most listened-to bands in our original dataset. According to Table 22, these bands have been listened to by more than 9.5% of the users included in the database.

The outline of this section follows our previous approach. We once again apply RF, SVM, and LR as classifiers on the dataset under consideration, and we obtain the influence measure using the two aforementioned partitions: CDP and HCP. Despite focusing on the 15 most listened-to bands, we encountered inconclusive results with SVM when assessing influences. For this reason, such results are not made explicit in order to reduce the length of the paper.

First, we analyze the case of using the CDP. The corresponding numerical results are shown in Table 25. For each classification problem (by columns), we highlight the most influential band in bold. Using RF, the following conclusions can be drawn. Linkin Park is the most influential band for listening to System of a Down, Metallica and Die Toten Hosen; Rammstein, to Billy Talent; System of a Down, to Rammstein; Die Toten Hosen, to Nirvana; Billy Talent, to Linkin Park and Muse; Muse, to Beatsteaks; Beatsteaks, to Red Hot Chilli Peppers, The Killers, and Jack Johnson; Foo Fighters, to Coldplay; and Nirvana, to The Beatles. When considering LR, Coldplay and Foo Fighters are both the most influential bands on Beatsteaks; Rammstein, on Metallica and Die Toten Hosen; Metallica, on The Beatles; Muse, on Rammstein; Beatsteaks, on The Killers and Foo Fighters; and Foo Fighters, on Coldplay, Billy Talent, Muse, and Beatsteaks. Only Foo Fighters and Beatsteaks share a bilateral relation where each band is mutually the most influential on the other.

| RF_CDP | linkin.park | coldplay | red.hot.chili.peppers | rammstein | system.of.a.down | metallica | die.toten.hosen | billy.talent | the.killers | the.beatles | jack.johnson | muse | beatsteaks | foo.fighters | nirvana |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| linkin.park | - | 0.0000 | 0.0028 | 0.5128 | **1.0256** | **0.1114** | **0.8350** | 0.0282 | 0.0000 | 0.0000 | 0.0028 | 0.0090 | 0.0271 | 0.0000 | 0.0000 |
| coldplay | 0.0000 | - | 0.0693 | 0.0241 | 0.0271 | 0.0000 | 0.0226 | 0.0367 | 0.0905 | 0.0000 | 0.0056 | 0.0090 | 0.0211 | 0.0754 | 0.0000 |
| red.hot.chili.peppers | 0.0000 | 0.0000 | - | 0.3110 | 0.0000 | 0.0000 | 0.0078 | 0.0038 | 0.0000 | 0.0117 | 0.0109 | 0.0113 | 0.0000 | 0.0000 | 0.0313 |
| rammstein | 0.0000 | 0.0030 | 0.0166 | - | 0.7903 | 0.0167 | 0.6996 | **0.0423** | 0.0082 | 0.0000 | 0.0056 | 0.0000 | 0.0030 | 0.0030 | 0.0057 |
| system.of.a.down | 0.0000 | 0.0121 | 0.0111 | **0.7149** | - | 0.0251 | 0.1128 | 0.0000 | 0.0137 | 0.0000 | 0.0084 | 0.0000 | 0.0121 | 0.0060 | 0.0115 |
| metallica | 0.0000 | 0.0000 | 0.0156 | 0.4205 | 0.0000 | - | 0.0078 | 0.0000 | 0.0000 | 0.0117 | 0.0219 | 0.0226 | 0.0000 | 0.0000 | 0.0234 |
| die.toten.hosen | 0.0000 | 0.0000 | 0.0078 | 0.5512 | 0.0000 | 0.0000 | - | 0.0075 | 0.0000 | 0.0117 | 0.0073 | 0.0038 | 0.0000 | 0.0000 | **0.0469** |
| billy.talent | **19.2817** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0988 | - | 0.0000 | 0.0000 | 0.0000 | **0.1107** | 0.0000 | 0.0000 | 0.0040 |
| the.killers | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0277 | 0.0000 | - | 0.0000 | 0.0000 | 0.0277 | 0.0000 | 0.0000 | 0.0000 |
| the.beatles | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| jack.johnson | 0.0454 | 0.0000 | 0.0000 | 0.0038 | 0.0000 | 0.0039 | 0.0040 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| muse | 0.0090 | 0.0090 | 0.0804 | 0.0000 | 0.0000 | 0.0474 | 0.0310 | 0.0226 | 0.0768 | 0.0000 | 0.0253 | - | **0.0483** | 0.0090 | 0.0345 |
| beatsteaks | 0.0181 | 0.0151 | **0.1304** | 0.0030 | 0.0000 | 0.0362 | 0.0395 | 0.0339 | **0.2936** | 0.0000 | **0.2331** | 0.0362 | - | **0.1388** | 0.0000 |
| foo.fighters | 0.0000 | **0.0392** | 0.0832 | 0.0030 | 0.0060 | 0.0084 | 0.0000 | 0.0169 | 0.0165 | 0.0000 | 0.0281 | 0.0090 | 0.0121 | - | 0.0029 |
| nirvana | 0.0000 | 0.0000 | 0.0352 | 0.2014 | 0.0000 | 0.0000 | 0.0469 | 0.0113 | 0.0000 | **0.0156** | 0.0036 | 0.0038 | 0.0000 | 0.0000 | - |

| LR_CDP | linkin.park | coldplay | red.hot.chili.peppers | rammstein | system.of.a.down | metallica | die.toten.hosen | billy.talent | the.killers | the.beatles | jack.johnson | muse | beatsteaks | foo.fighters | nirvana |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| linkin.park | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| coldplay | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | **0.0030** | 0.0000 | 0.0000 |
| red.hot.chili.peppers | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0078 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| rammstein | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | **0.0056** | **0.0056** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| system.of.a.down | 0.0000 | 0.0000 | 0.0000 | 0.0965 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| metallica | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | **0.0117** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| die.toten.hosen | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| billy.talent | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| the.killers | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| the.beatles | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| jack.johnson | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| muse | 0.0000 | 0.0000 | 0.0000 | **0.1448** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 |
| beatsteaks | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | **0.0027** | 0.0000 | 0.0000 | 0.0000 | - | **0.0030** | 0.0000 |
| foo.fighters | 0.0000 | **0.0030** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | **0.0028** | 0.0000 | 0.0000 | 0.0000 | **0.0030** | **0.0030** | - | 0.0000 |
| nirvana | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0039 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - |

Table 25: Influence measure (5) using the constructed-by-design partition in the analysis of musical taste in Spotify.

In view of Table 25, the most influential bands do not seem to coincide in the rankings resulting from using RF and LR as classifiers. However, we will use Pearson correlations on the numerical results as a measure of overall comparison of the rankings obtained (see Table 26). The greatest similarities between rankings are found, in this order, when we study the influence of listening to The Killers, Coldplay, Foo Fighters, The Beatles and Die Toten Hosen. All of these bands have a correlation of around or above 0.6.

| linkin.park | coldplay | red.hot.chili.peppers | rammstein | system.of.a.down | metallica | die.toten.hosen | billy.talent | the.killers | the.beatles | jack.johnson | muse | beatsteaks | foo.fighters | nirvana |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| - | 0.8824 | - | 0.1451 | - | -0.0099 | 0.5984 | 0.0448 | 0.9290 | 0.7143 | - | -0.0822 | 0.2272 | 0.8704 | - |

Table 26: Correlations between the rankings obtained for RF and LR with CDP.

| RF_HCP | linkin.park | coldplay | red.hot.chili.peppers | rammstein | system.of.a.down | metallica | die.toten.hosen | billy.talent | the.killers | the.beatles | jack.johnson | muse | beatsteaks | foo.fighters | nirvana |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| linkin.park | - | 0.3989 | 0.0000 | 1.2427 | 1.1437 | 0.0031 | 0.0269 | 0.0062 | 0.0078 | 0.0000 | 0.0082 | 0.0138 | 0.0126 | 0.0000 | 0.0211 |
| coldplay | 0.0000 | - | 0.5779 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| red.hot.chili.peppers | 0.0093 | **0.8825** | - | 0.3369 | 0.1669 | **0.4822** | 0.0000 | 0.0309 | 0.0117 | 0.0000 | 0.0408 | 0.0055 | 0.0067 | 0.0038 | 0.0070 |
| rammstein | 0.0556 | 0.0328 | 0.0335 | - | **1.4807** | 0.1669 | 0.0000 | 0.0340 | 0.0000 | 0.0000 | 0.0082 | 0.0083 | 0.0000 | 0.0000 | 0.0000 |
| system.of.a.down | 0.0216 | 0.0383 | 0.1673 | **1.4683** | - | 0.0433 | 0.0000 | 0.1453 | 0.0000 | 0.0000 | 0.0163 | 0.0083 | 0.0000 | 0.0000 | 0.0000 |
| metallica | 0.0247 | 0.0301 | 0.0000 | 1.2674 | 1.1376 | - | 0.0000 | 0.0309 | 0.0000 | 0.0000 | 0.0245 | 0.0083 | 0.0034 | 0.0038 | 0.0000 |
| die.toten.hosen | 0.0309 | 0.0055 | 0.1673 | 1.2798 | 0.2844 | 0.0185 | - | **0.1484** | 0.0156 | 0.0000 | 0.0109 | 0.0166 | 0.0000 | 0.0000 | **0.0352** |
| billy.talent | **21.3682** | 0.0820 | 0.0000 | 0.1329 | 0.2658 | 0.0464 | **0.0960** | - | 0.0000 | 0.0000 | 0.0082 | **0.0775** | 0.0502 | 0.0000 | 0.0282 |
| the.killers | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0307 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0045 | 0.0000 |
| the.beatles | 0.0000 | 0.0277 | 0.0608 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| jack.johnson | 0.0000 | 0.0000 | **1.3080** | 0.0037 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | **0.0115** | 0.0000 |
| muse | 0.0042 | 0.0000 | 0.0078 | 0.0000 | 0.0000 | 0.0000 | 0.0192 | 0.0000 | **0.0165** | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 |
| beatsteaks | 0.0126 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0045 | 0.0000 | 0.0000 | **0.3152** | 0.0000 | - | 0.0000 | 0.0070 |
| foo.fighters | 0.1004 | 0.0285 | 0.0044 | 0.0000 | 0.0000 | 0.0043 | 0.0000 | 0.0000 | 0.0041 | 0.0000 | 0.0162 | 0.0000 | **0.0753** | - | 0.0070 |
| nirvana | 0.1235 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0531 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - |

| LR_HCP | linkin.park | coldplay | red.hot.chili.peppers | rammstein | system.of.a.down | metallica | die.toten.hosen | billy.talent | the.killers | the.beatles | jack.johnson | muse | beatsteaks | foo.fighters | nirvana |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| linkin.park | - | 0.0000 | 0.0000 | 0.0185 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| coldplay | 0.0000 | - | 0.0532 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| red.hot.chili.peppers | 0.0000 | 0.0219 | - | 0.0278 | 0.0031 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| rammstein | **0.0031** | 0.0000 | 0.0000 | - | 0.0124 | **0.0093** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| system.of.a.down | 0.0000 | 0.0000 | 0.0000 | **0.7913** | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| metallica | 0.0000 | 0.0027 | 0.0000 | 0.4884 | 0.0031 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| die.toten.hosen | 0.0000 | 0.0000 | 0.0000 | 0.3369 | 0.0031 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| billy.talent | 0.0000 | 0.0000 | 0.0000 | 0.0309 | **0.0402** | 0.0062 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| the.killers | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| the.beatles | 0.0000 | 0.0000 | **0.0608** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| jack.johnson | 0.0000 | 0.0000 | 0.0456 | 0.0000 | 0.0000 | 0.0000 | **0.0036** | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| muse | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 | 0.0000 |
| beatsteaks | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 | 0.0000 |
| foo.fighters | 0.0000 | **0.0285** | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | - | 0.0000 |
| nirvana | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | **0.0067** | 0.0000 | - |

Table 27: Influence measure (5) using the hierarchical clustering partition in the analysis of musical taste in Spotify.

Second, we do a similar study from the results obtained with the partitions prescribed by the hierarchical clustering (see Table 27). In this case, using RF as a classifier, Red Hot Chili Peppers is the most influential on listening to Coldplay and Metallica according to our influence measure; Rammstein, on System of a Down (and vice versa); Die Toten Hosen, to Billy Talent and Nirvana; Billy Talent, to Linking Park, Die Toten Hosen and Muse; Jack Johnson, to Red Hot Chilly Peppers and Foo Fighters; Muse, to The Killers; Beatsteaks, to Jack Johnson; and Foo Fighters, to Beatsteaks. Regarding the results based on LR, we mention the following issues. Our influence measure shows that Rammstein is, in this case, the most influential band in listening to Linkin Park and Metallica; System of a Down, to Rammsten; Billy Talent, to Systems of a Down; The Beatles, to Red Hot Chilly Peppers; Jack Johnson, to Die Toten Hosen; Foo Fighters, to Coldplay, and finally, Nirvana, to Beatsteaks.

Table 28 shows the Pearson correlations between the influence measures obtained for each band when using RF and LR as classifiers. In this case, the greatest similarities between the rankings are found, in this order, when we examine the influence of listening to Rammstein, Red Hot Chilli Peppers, and Coldplay.

| linkin.park | coldplay | red.hot.chili.peppers | rammstein | system.of.a.down | metallica | die.toten.hosen | billy.talent | the.killers | the.beatles | jack.johnson | muse | beatsteaks | foo.fighters | nirvana |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| -0.0756 | 0.4911 | 0.6255 | 0.8122 | 0.1923 | 0.2026 | -0.1640 | - | - | - | - | - | -0.1328 | - | - |

Table 28: Correlations between the rankings obtained for RF and LR with HCP.

Finally, one aspect to be considered is the possible effect of partitioning on the resulting ranking. Although some similarities can apparently be observed between the rankings obtained under CDP and HCP, we determine the corresponding Pearson correlations in Table 29 between the rankings obtained when using RF as classifier as well as the case of considering LR.

| | linkin.park | coldplay | red.hot.chili.peppers | rammstein | system.of.a.down | metallica | die.toten.hosen | billy.talent | the.killers | the.beatles | jack.johnson | muse | beatsteaks | foo.fighters | nirvana |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RF | 0.9999 | -0.1964 | -0.1546 | 0.9484 | 0.7776 | -0.1768 | 0.0469 | -0.2691 | -0.1001 | - | 0.9859 | 0.8619 | -0.0137 | -0.2259 | 0.3248 |
| LR | - | 0.7746 | - | 0.3316 | - | 0.8190 | -0.0769 | - | - | - | - | - | -0.1132 | - | - |

Table 29: Correlations between the rankings obtained under CDP and HCP.

When using RF, the highest similarities in the rankings are found for Linkin Park, Jack Johnson, Rammstein, Muse and System of a Down, with correlations above 0.75. In these cases, the use of different partitions on the features causes small numerical variations in the rankings. This is not common for the rest of the bands, where even negative correlations are obtained. Nor is it when using LR as a classifier, where only Metallica and Coldplay exceed the 0.75 correlation threshold.

### 6.4 On the analysis of bank term deposit subscriptions

Given the good performance of the sampling method proposed in Subsection 4.5, as demonstrated in Subsection 6.2, this section applies the methodology to estimate our influence measure in the context of large-scale banking data. By applying our approximation technique, we can determine the features that influence the likelihood of clients subscribing to a term bank deposit.

### 6.4.1 Description of the data

We use the dataset described in Moro et al. (2011), which captures direct marketing campaigns conducted by a Portuguese banking institution between May 2008 and November 2010. These campaigns involved phone calls, often requiring multiple contacts with the same client to determine whether they would subscribe to the bank's term deposit product. As noted by the authors, identifying the key features influencing the subscription success can optimize resource allocation (e.g. human effort, call frequency and time, etc.) and aid in selecting a cost-effective and high-quality pool of potential customers.

The original dataset comprises 45,211 clients contacted during the marketing campaign and includes 16 features, capturing information about the clients or their interactions in the current or previous campaigns. The dataset also contains a binary outcome indicating whether a client subscribed to a term deposit during the current campaign. For our study, we modified this dataset by converting the features into binary variables, selecting 14 out of the original 16 features, and retaining the binary outcome. Table 30 summarizes the characteristics of both the original dataset (`bank_original`) and the transformed dataset (`bank_binary`). The transformed dataset comprises 43,193 clients and the 14 binary features detailed in Table 31. The reduction in the number of instances compared to the original dataset results from excluding clients with unknown information on job or education level.

| | # of observations | # of features | Nature of features |
|---|---|---|---|
| `bank_original` | 45,211 | 16 + outcome | binary, multi-class, continuous |
| `bank_binary` | 43,193 | 14 + outcome | binary |

Table 30: Characteristics of the original and transformed datasets for the analysis of banking data.

### 6.4.2 Performance of our influence measure's approximation technique

We estimate the influence measure using two different coalitional structures of the features informed by the insights in Moro et al. (2011). First, we consider a constructed-by-design partition (CDP), referred to as CDP_1, in which we group features related to clients (features 1–8), features related to the contact during the current campaign (features 9–12), and features related to contacts from previous campaigns (features 13–14).

Table 32 presents the results obtained by applying different classifiers to determine the outcome. Due to the large size of the dataset, the proposed sampling technique was employed. Specifically, we selected 10 subsamples, each containing 10% of the total observations. The influence measure was estimated for each subsample, and the results reported for each classifier represent the average of these 10 estimates.

The findings reveal that the SVM classifier fails to provide conclusive results for this new dataset, consistent with previous analyses. In contrast, the other two classifiers indicate that only

| Outcome | Description |
|---|---|
| y | Binary variable indicating whether the client has subscribed to a bank term deposit during the current marketing campaign (1) or not (0). |
| **Feature** | **Description** |
| 1 age | Binary variable indicating whether the client was over 35 years old at the contact date (1) or 35 years old or less (0). |
| 2 job | Binary variable indicating whether the client was employed (1) or unemployed, student, or retired (0). |
| 3 marital | Binary variable indicating whether the client was married (1) or not (0). |
| 4 education | Binary variable indicating whether the client had tertiary education (1) or primary or secondary education (0). |
| 5 default | Binary variable indicating whether the client had credit in default (1) or not (0). |
| 6 balance | Binary variable indicating whether the client's average yearly balance, in euros, was positive (1) or not (0). |
| 7 housing | Binary variable indicating whether the client had a housing loan (1) or not (0). |
| 8 loan | Binary variable indicating whether the client had a personal loan (1) or not (0). |
| 9 contact | Binary variable indicating whether the client was last contacted by telephone or cellular (1) or the communication type is unknown (0). |
| 10 day | Binary variable indicating whether the client was last contacted after the first 7 days of the month (1) or during the first 7 days of the month (0). |
| 11 duration | Binary variable indicating whether the last contact duration with the client was 5 minutes or more (1) or less than 5 minutes (0). |
| 12 campaign | Binary variable indicating whether the client was contacted multiple times during the current campaign (1) or only once (0). |
| 13 previous | Binary variable indicating whether the client was contacted before the current campaign (1) or not (0). |
| 14 poutcome | Binary variable indicating whether the client subscribed to a bank term deposit in the previous marketing campaign (1) or not (0). |

Table 31: Summary of the considered features in the analysis of banking data.

client-related features influence the outcome, while those related to the current and previous campaigns appear irrelevant. When using the RF classifier, the three most influential features in subscribing to a bank term deposit are, in order, education, housing, and job. With the LR classifier, education remains the most influential feature, but marital and housing occupy the second and third positions, respectively. Although differences are already noticeable in the first three positions, it is easy to see discrepancies in the overall rankings of the most influential features. The main difference is that feature default becomes inconclusive when the LR classifier is used.

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RF_CDP_1 | 2.0980 | 2.5983 | 2.1834 | 3.9346 | 0.0824 | 0.2193 | 3.7356 | 0.6515 | 0 | 0 | 0 | 0 | 0 | 0 |
| Position | 5 | 3 | 4 | 1 | 8 | 7 | 2 | 6 | 9-14 | 9-14 | 9-14 | 9-14 | 9-14 | 9-14 |
| SVM_CDP_1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Position | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| LR_CDP_1 | 0.1180 | 2.5159 | 4.7047 | 4.8854 | 0 | 0.1936 | 4.0875 | 0.7092 | 0 | 0 | 0 | 0 | 0 | 0 |
| Position | 7 | 4 | 2 | 1 | 8-14 | 6 | 3 | 5 | 8-14 | 8-14 | 8-14 | 8-14 | 8-14 | 8-14 |

Table 32: Estimation, using sampling without replacement, of the influence measure and rankings in the analysis of banking data using the first partition informed by Moro et al. (2011), CDP_1.

Some disparities are observed when simple random sampling with replacement is considered for the approximation of the influence measure. In this case, features related to the contact during the current campaign gain influence when using the RF classifier. Additionally, features 1 (client's age) and 8 (presence of a loan) become non-influential when using the LR classifier. Table 33 displays such results. However, despite these discrepancies between the two estimated results, the Spearman correlations between the influence measures in these two cases remain high, at approximately 0.82 for `RF_CDP_1` and 0.87 for `LR_CDP_1`.

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| `RF_CDP_1` | 0.6728 | 3.4268 | 1.6721 | 0.5558 | 0.1249 | 0.4825 | 0.3349 | 0.1200 | 0 | 0.1000 | 0.2333 | 0.0667 | 0 | 0 |
| Position | 3 | 1 | 2 | 4 | 8 | 5 | 6 | 9 | 12-14 | 10 | 7 | 11 | 12-14 | 12-14 |
| `SVM_CDP_1` | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Position | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| `LR_CDP_1` | 0 | 1.7315 | 0.0952 | 0.3968 | 0 | 0.0084 | 0.0909 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Position | 6-14 | 1 | 3 | 2 | 6-14 | 5 | 4 | 6-14 | 6-14 | 6-14 | 6-14 | 6-14 | 6-14 | 6-14 |

Table 33: Estimation, using sampling with replacement, of the influence measure and rankings in the analysis of banking data using the first partition informed by Moro et al. (2011), CDP_1.

Building on the same approach as in the preceding analysis, we have made a slight modification to the partition used earlier. Specifically, we introduce a second constructed-by-design partition (CDP), referred to as CDP_2, in which we group features related to personal data of clients (features 1–4), features related to financial data of clients (features 5–8), and the remaining features are grouped in the same way as before: those related to the contact of the current campaign (features 9–12) and those related to contacts made on previous campaigns (features 13–14). The results, in Table 34, show significant differences compared to the previous analysis. When considering the RF classifier, the three most influential features are previous contact with the client, education, and marital. However, the previous subscription to a bank term in the previous campaign appears inconclusive. Under the LR classifier, the duration of the call (feature duration) emerges as the single influential feature for subscribing to a bank deposit. Identical conclusions on the influence of features are drawn when the influence measure is estimated using simple random sampling with replacement, as shown in Table 35.

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| `RF_CDP_2` | 0.9458 | 0.025 | 2.2542 | 2.3500 | 0.0704 | 0.0019 | 0.5389 | 0.1093 | 0.9000 | 0.0286 | 0.7750 | 0.0536 | 4.2500 | 0 |
| Position | 4 | 12 | 3 | 2 | 9 | 13 | 7 | 8 | 5 | 11 | 6 | 10 | 1 | 14 |
| `SVM_CDP_2` | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Position | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| `LR_CDP_2` | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.017 | 0 | 0 | 0 |
| Position | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 1 | 2-14 | 2-14 | 2-14 |

Table 34: Estimation, using sampling without replacement, of the influence measure and rankings in the analysis of banking data using the second partition informed by Moro et al. (2011), CDP_2.

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| `RF_CDP_2` | 0.9875 | 0.0083 | 2.2917 | 2.3917 | 0.0685 | 0.0037 | 0.5833 | 0.1185 | 0.8571 | 0.0214 | 0.7107 | 0.0607 | 5 | 0 |
| Position | 4 | 12 | 3 | 2 | 9 | 13 | 7 | 8 | 5 | 11 | 6 | 10 | 1 | 14 |
| `SVM_CDP_2` | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Position | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| `LR_CDP_2` | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.036 | 0 | 0 | 0 |
| Position | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 1 | 2-14 | 2-14 | 2-14 |

Table 35: Estimation, using sampling with replacement, of the influence measure and rankings in the analysis of banking data using the second partition informed by Moro et al. (2011), CDP_2.

### Assessing the impact of the imbalanced distribution in the outcome

From the numerical results obtained, we can visualize that the SVM classifier once again fails to produce conclusive results, as it returns a value of 0. As shown in Table 36, SVM predicts the lowest number of occurrences of the outcome value 1 in this particular case study.

| Value | Original outcome | RF | SVM | LR |
|---|---|---|---|---|
| 0 | 38,172 | 41,310 | 41,769 | 41,588 |
| 1 | 5,021 | 1,883 | 1,424 | 1,605 |

Table 36: Summary of the outcomes from both the original database and those predicted by the classifiers.

To assess the potential effect of an imbalanced distribution of 0s and 1s in the outcome, we draw a subsample from the original database consisting of 5,021 instances with an outcome of 1 and 5,021 instances with an outcome of 0 (balanced in the outcome). Table 37 and Table 38 present the results of estimating the influence measure using sampling without and with replacement, respectively, for this ad hoc subsample. From this analysis, we observe that the influence measures based on Datta et al.'s methodology are not entirely zero in SVM scenarios.

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RF_CDP_1 | 0.0021 | 0 | 0 | 0.0042 | 0 | 0 | 0.0063 | 0 | 0.4000 | 0 | 0 | 0 | 0 | 0 |
| Position | 4 | 5-14 | 5-14 | 3 | 5-14 | 5-14 | 2 | 5-14 | 1 | 5-14 | 5-14 | 5-14 | 5-14 | 5-14 |
| SVM_CDP_1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.4000 | 0 | 0 | 0 | 0 | 0 |
| Position | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 1 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 |
| LR_CDP_1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.4000 | 0 | 0 | 0 | 0 | 0 |
| Position | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 1 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 |

Table 37: Estimation, using sampling without replacement, of the influence measure and rankings over the considered subsample using the first partition informed by Moro et al. (2011), CDP_1.

| Feature | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RF_CDP_2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.4500 | 0 | 1.3250 | 0 | 0 | 0 |
| Position | 3-14 | 3-14 | 3-14 | 3-14 | 3-14 | 3-14 | 3-14 | 3-14 | 2 | 3-14 | 1 | 3-14 | 3-14 | 3-14 |
| SVM_CDP_2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.6500 | 0 | 0 | 0 | 0 | 0 |
| Position | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 1 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 |
| LR_CDP_2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.6500 | 0 | 0 | 0 | 0 | 0 |
| Position | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 | 1 | 2-14 | 2-14 | 2-14 | 2-14 | 2-14 |

Table 38: Estimation, using sampling without replacement, of the influence measure and rankings over the considered subsample using the first partition informed by Moro et al. (2011), CDP_2.

The results indicate that feature 9 (related to the method of contacting the customer) is the only influential feature for both SVM and L, regardless of the partition used. The main differences arise in the influence measures obtained using RF. When considering CDP_1, four features influence the decision to subscribe to a bank deposit: the method of contacting the client (telephone or not), housing, education, and age. However, for the CDP_2 partition, only call duration and contact method are influential.

## 7 Concluding remarks

In this paper, we have introduced a method to analyze the influence of specific features in a classification problem where dependencies might exist among them. Our influence measure extends

Datta et al.'s approach to accommodate the existence of a coalitional structure on such features. Given its game-theoretic nature, we provide both an axiomatic characterization and a computational complexity analysis of this influence measure, as is typical in cooperative game theory. Furthermore, we demonstrate that the Banzhaf-Owen value serves as a special case of our measure under conditions of binary databases and no repeated profiles. The methodology presented here is subsequently applied to investigate influential variables in diverse domains, including health-related conditions, traffic accidents, musical taste on Spotify, and the likelihood of subscribing to a bank term deposit.

In this context, several lines of research stemming from our proposal remain open and need to be addressed in the future. From a methodological point of view, one promising direction is extending the main axiomatic characterization theorem obtained in this study to accommodate non-binary outcomes. A key challenge is the reformulation of the disjoint union axiom along with the possible incorporation of some additional axiom that would guarantee the uniqueness of a general measure of influence. From a practical perspective, a broader analysis incorporating a wider variety of datasets would strengthen the validation of our method. Future studies could further investigate how the influence measure performs across diverse datasets. Furthermore, we believe that our influence measure has the potential to be used in a wide range of real-world applications. One example is the booming sector of streaming platforms such as Amazon Prime, Disney+, Max, or Netflix. It would be particularly interesting to conduct a similar analysis to what we performed for the Spotify database, aiming to identify which audiovisual content has the greatest (or least) impact on the viewing of others. The inherent similarities among some of these media resources make our influence measure with dependencies applicable in this domain. This would enable us to propose personalized recommendations to users, which is of major relevance to platforms' stakeholders (Bourreau and Gaudin, 2022).

However, applying our methodology to real-world problems still presents various challenges. In particular, our method assumes that the existing dependencies among the features are known, which may not always be the case. As evidenced in Section 6, the influence measure is affected by the coalitional structure considered. Therefore, performing further dependency analyses before selecting the partition would further strengthen the conclusions drawn from the associated measure of influence. Techniques such as Pearson's chi-squared test of independence or Fisher's exact test, as demonstrated in Example 4.4, or calculating Cramèr's V or the Jaccard index between pairs of binary features are useful tools to infer feature dependencies directly from the dataset. Additionally, alternative dependency methods tailored to the specific characteristics of the datasets could be explored. These methods allow for the quantification of partition robustness by filtering out weaker interactions through appropriate significance levels and thresholds, thereby reducing the risk of capturing noise or grouping features with weak interactions instead of meaningful patterns. Furthermore, these criteria are context-dependent and can be adjusted by decision-makers, enabling our model to focus on more robust and relevant dependencies.

We also observed the strong dependence of the chosen databases and the outcome determined by a certain classifier. We thus highlight the potential value of conducting a more comprehensive analysis on the performance of our influence measure using other classifiers in the existing literature. To the best of the authors' knowledge, random forests, support vector machines, and logistic regression-type classifiers are among the most widely used methods for binary data, which motivated their inclusion in our study. We also present preliminary experimental results for two additional classifiers based on neural networks and ensemble methods, taking base implementations from the RWeka library. Calibrating different parameters of these classifiers, as well as testing other classification methods on a large collection of datasets, might yield more conclusive results for some of the analyzed scenarios.

## Acknowledgments

This work is part of the R+D+I project PID2021-124030NB-C32, granted by MICIU/AEI/ 10.13039/ 501100011033/ and by "ERDF A way of making Europe"/EU. This research was also funded by *Grupos de Referencia Competitiva* ED431C 2021/24 from the *Consellería de Cultura, Educación e Universidades, Xunta de Galicia*. The authors would like to thank the computational resources of the *Centro de Supercomputación de Galicia* (CESGA). They also extend their thanks to the Editor-in-Chief, the Associate Editor, and the three anonymous referees for their valuable comments, which helped improve an earlier version of this paper. The authors are grateful to A. Pérez González for her useful insights.

### Declaration of interest

The authors declare that there is no conflict of interest.

## References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*, 103502.

Algaba, E., Prieto, A., and Saavedra-Nieves, A. (2024). Risk analysis sampling methods in terrorist networks based on the Banzhaf value. *Risk Analysis*, *44*(2), 477–492.

Alonso-Meijide, J., Carreras, F., Fiestras-Janeiro, G., and Owen, G. (2007). A comparative axiomatic characterization of the Banzhaf-Owen coalitional value. *Decision Support Systems*, *43*(3), 701–712.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

Amer, R., Carreras, F., and Giménez, J. (2002). The modified Banzhaf value for games with coalition structure: an axiomatic characterization. *Mathematical Social Sciences*, *43*(1), 45–54.

Arévalo-Iglesias, G., and Álvarez-Mozos, M. (2020). Power distribution in the Basque Parliament using games with externalities. *Theory and Decision*, *89*(2), 157–178.

Banzhaf III, J. (1965). Weighted voting does not work: A mathematical analysis. *Rutgers Law Review*, *19*(2), 317-343.

Bourreau, M., and Gaudin, G. (2022). Streaming platform and strategic recommendation bias. *Journal of Economics & Management Strategy*, *31*(1), 25–47.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Burkart, N., and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245–317.

Carrizosa, E., Molero-Río, C., and Morales, D. R. (2021). Mathematical optimization in classification and regression trees. *TOP*, *29*, 5–33.

Casalicchio, G., Molnar, C., and Bischl, B. (2019). Visualizing the feature importance for black box models. In M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Ifrim (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 11051, pp. 655–670). Springer, Cham.

Cochran, W. G. (2007). *Sampling Techniques* (3rd ed.). John Wiley & Sons Inc.

Cohen, S., Dror, G., and Ruppin, E. (2007). Feature selection via coalitional game theory. *Neural Computation*, *19*(7), 1939–1961.

Datta, A., Datta, A., Procaccia, A., and Zick, Y. (2015). Influence in classification via cooperative game theory. *Proceedings of the Twenty–fourth International Joint Conference on Artificial Intelligence*, 511–517.

Davila-Pena, L., García-Jurado, I., and Casas-Méndez, B. (2022). Assessment of the influence of features on a classification problem: An application to COVID-19 patients. *European Journal of Operational Research*, *299*(2), 631–641.

European Commission. (2020). *White Paper on Artificial Intelligence: a European approach to excellence and trust.* Retrieved from `https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en` (Accessed: February 14, 2025)

Feltkamp, V. (1995). Alternative axiomatic characterization of the Shapley and Banzhaf values. *International Journal of Game Theory*, *24*(2), 179–186.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Ghaddar, B., and Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*, *265*(3), 993–1004.

Giudici, P., and Raffinetti, E. (2021). Shapley-Lorenz explainable artificial intelligence. *Expert Systems with Applications*, *167*, 114104.

Jothi, N., Husain, W., and Rashid, N. (2021). Predicting generalized anxiety disorder among women using Shapley value. *Journal of Infection and Public Health*, *14*, 103–108.

Köhler, V. (2017). *lastfm dataset.* Github Gist: `https://gist.github.com/victorkohler/0931d181ef126e0740d8aac6933f13f4`. (Accessed: February 14, 2025)

le Cessie, S., and van Houwelingen, J. (1992). Ridge estimators in logistic regression. *Applied Statistics*, *41*(1), 191-201.

Lehrer, E. (1988). An axiomatization of the Banzhaf value. *International Journal of Game Theory*, *17*(2), 89–99.

Li, M., Sun, H., Huang, Y., and Chen, H. (2024). Shapley value: from cooperative game to explainable artificial intelligence. *Autonomous Intelligent Systems*, *4*, 2.

Lipsitz, S. R., and Fitzmaurice, G. M. (1996). The score test for independence in R x C contingency tables with missing data. *Biometrics*, *52*(2), 751–762.

Lorenzo-Freire, S. (2017). New characterizations of the owen and banzhaf–owen values using the intracoalitional balanced contributions property. *TOP*, *25*, 579–600.

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

Maindonald, J. H., and Braun, W. J. (2021). *Data Analysis and Graphics using R. An Example-based Approach* (3rd ed.). Cambridge University Press. (The `DAAG` package was created to support this text. `https://CRAN.R-project.org/package=DAAG` [`R` package version 1.25.6.])

Moro, S., Laureano, R., and Cortez, P. (2011). Using Data Mining for bank direct marketing: An application of the CRISP-DM methodology. In P. Novais, J. Machado, C. Analide, and A. Abelha (Eds.), *Proceedings of the European Simulation and Modelling Conference - ESM'2011* (pp. 117–121). Guimarães, Portugal: EUROSIS.

Muros, F., Algaba, E., Maestre, J. M., and Camacho, E. F. (2017). Banzhaf value as a design tool in coalitional control. *Systems & Control Letters*, *104*, 21–30.

Nahiduzzaman, M., Faisal Abdulrazak, L., Arselene Ayari, M., Khandakar, A., and Islam, S. R. (2024). A novel framework for lung cancer classification using lightweight convolutional neural networks and ridge extreme learning machine model with SHapley Additive exPlanations (SHAP). *Expert Systems with Applications*, *248*, 123392.

Olsen, L. H. B., Glad, I. K., Jullum, M., and Aas, K. (2022). Using Shapley values and variational autoencoders to explain predictive models with dependent mixed features. *Journal of Machine Learning Research*, *23*(213), 1–51.

Owen, A. B., and Prieur, C. (2017). On Shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, *5*(1), 986–1002.

Owen, G. (1977). Values of games with a priori unions. In R. Henn and O. Moeschlin (Eds.), *Mathematical Economics and Game Theory* (Vol. 141, pp. 76–88). Springer, Berlin, Heidelberg.

Owen, G. (1981). Modification of the Banzhaf-Coleman index for games with a priori unions. In M. J. Holler (Ed.), *Power, Voting, and Voting Power* (pp. 232–238). Physica, Heidelberg.

Platt, J. C. (1998). Fast training of support vector machines using sequential minimal optimization. In C. J. C. Burges, B. Schölkopf, and A. J. Smola (Eds.), *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

Saavedra-Nieves, A., and Fiestras-Janeiro, M. G. (2021). Sampling methods to estimate the Banzhaf–Owen value. *Annals of Operations Research*, *301*(1), 199–223.

Serfling, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, *2*(1), 39–48.

Shapley, L. (1953). A value for n-person games. *Annals of Mathematics Studies*, *28*(7), 307–317.

Smith, M., and Alvarez, F. (2021). Identifying mortality factors from Machine Learning using Shapley values – a case of COVID19. *Expert Systems with Applications*, *176*, 114832.

Štrumbelj, E., and Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, *11*(1), 1–18.

Štrumbelj, E., and Kononenko, I. (2011). A general method for visualizing and explaining black-box regression models. In A. Dobnikar, U. Lotrič, and B. Šter (Eds.), *Adaptive and Natural Computing Algorithms* (Vol. 6594, pp. 21–30). Springer, Berlin, Heidelberg.

Tian, G.-L., and Li, H.-Q. (2017). A new framework of statistical inferences based on the valid joint sampling distribution of the observed counts in an incomplete contingency table. *Statistical Methods in Medical Research*, *26*(4), 1712–1736.

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

Xiao, Z., Xu, X., Xing, H., Zhao, B., Wang, X., Song, F., . . . Feng, L. (2024). DTCM: Deep Transformer Capsule Mutual Distillation for Multivariate Time Series Classification. *IEEE Transactions on Cognitive and Developmental Systems*, *16*(4), 1445–1461.

**CRediT authorship contribution statement**

**L. Davila-Pena**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing.

**A. Saavedra-Nieves**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing.

**B. Casas-Méndez**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing.

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: