European Journal of Operational Research 299 (2022) 631–641

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Analytics, Computational Intelligence and Information Management

Assessment of the influence of features on a classification problem: An application to COVID-19 patients

Laura Davila-Pena[a,*], Ignacio García-Jurado[b], Balbina Casas-Méndez[a]

[a] MODESTYA Research Group, Department of Statistics, Mathematical Analysis and Optimisation and IMAT, Faculty of Mathematics, University of Santiago de Compostela, Campus Vida, Santiago de Compostela 15782, Spain

[b] MODES Research Group, Department of Mathematics and CITIC, Faculty of Computer Science, University of A Coruña, Campus de Elviña, A Coruña 15071, Spain

ARTICLE INFO

Article history:
Received 31 July 2020
Accepted 18 September 2021
Available online 24 September 2021

Keywords:
Machine learning
Classification
Influence of features
Shapley value
COVID-19

ABSTRACT

This paper deals with an important subject in classification problems addressed by machine learning techniques: the evaluation of the influence of each of the features on the classification of individuals. Specifically, a measure of that influence is introduced using the Shapley value of cooperative games. In addition, an axiomatic characterisation of the proposed measure is provided based on properties of efficiency and balanced contributions. Furthermore, some experiments have been designed in order to validate the appropriate performance of such measure. Finally, the methodology introduced is applied to a sample of COVID-19 patients to study the influence of certain demographic or risk factors on various events of interest related to the evolution of the disease.

1. Introduction

A classification problem consists of predicting the value of a qualitative response variable for one or more individuals, making use of the values we know of certain variables (features) of such individuals. Those predictions are based on the knowledge obtained through a training sample of individuals whose values of the features and of the response variable are known. Classification problems can be addressed by using machine learning techniques. Numerous classifiers have been proposed and analysed in the machine learning literature (see, for example, Fernández-Delgado, Cernadas, Barro, & Amorim, 2014).

In this article we make use of some classification techniques imported from machine learning to develop a methodological tool for the exploratory analysis of a training sample of the type described above. Specifically, our objective is to define a sensible measure to evaluate the influence of the features on the value of the response variable.

The relevance of our objective can be illustrated with a real problem of applied research that we recently faced. During the first wave of COVID-19 in Spain we had access to a database of 10,454 patients from Galicia (a region in the northwest of Spain) infected with COVID-19 from March 6, 2020 to May 7, 2020. Knowing the characteristics of individuals that significantly increase their probability of needing access to certain health infrastructures is highly useful for health authorities to make the right decisions. Therefore, we set out to use these data to find out which were the values of the features that most influenced the worsening of an infected patient's condition, so that such patient had to be hospitalised, had to be admitted to the ICU or even died. In Section 4 we apply the methodology we introduced and analysed in Sections 2 and 3 to explore this database.

The machine learning literature has been more focused on classification than on analysing the influence of features, although the latter is a problem of growing interest (see, for example, Section 4 of Carrizosa, Molero-Río, & Morales, 2021). In the emerging literature on this topic (cf. Burkart & Huber, 2021), we have identified the following shortcoming: we have not found a global measure of the influence of features on the response variable in a classification problem that is both theoretically well-founded and empirically contrasted. We thus fill the gap by introducing and analysing both theoretically and empirically an influence measure based on the Shapley value of cooperative games (Shapley, 1953). In the following paragraphs we present an overview of the current research background by discussing some of the works that we consider to be closest to our study.

* Corresponding author.
E-mail addresses: lauradavila.pena@usc.es (L. Davila-Pena), ignacio.garcia.jurado@udc.es (I. García-Jurado), balbina.casas.mendez@usc.es (B. Casas-Méndez).

https://doi.org/10.1016/j.ejor.2021.09.027
0377-2217/

When analysing a database, it is very likely to come across features that only cause noise and are not really influential for the response. For this reason, the selection of features is a prior problem to the study of influences we discuss here, since it is convenient to start from an already selected set of features and then comparatively study their influences. The problem of feature selection has given rise to a remarkable literature. For instance, Ghaddar & Naoum-Sawaya (2018) introduce an iterative approach to address feature selection in classification using support vector machines and apply it to a case of medical tumours diagnosis. Jothi, Husain, & Rashid (2021) attempt to identify individuals who may be suffering from mental illness by implementing the Shapley value as a feature selection of a data mining classifier system.

In the context of classification, Štrumbelj & Kononenko (2010) introduce a general procedure to assess the importance that the various features have had in the classification of a particular individual. Our approach is closely related to that paper but it is essentially different mainly because it is not locally oriented: we do not attempt to evaluate the influence of each feature on the classification of a particular individual, but rather to evaluate the influence of each feature value on the response variable. The Štrumbelj & Kononenko's procedure has been extended to regression models in Štrumbelj & Kononenko (2011). Although the evaluation of features contributions to model predictions is of great relevance, other papers have been devoted to determining the importance of features with respect to model performance. According to the latter approach, Casalicchio, Molnar, & Bischl (2019) describe a procedure, based on the Shapley value, that allows to fairly distribute the overall performance of a model among the features involved. Regarding our contribution, we propose a global measure in the context of classification and analyse it in some depth, both from a theoretical and empirical point of view. Furthermore, instead of focusing on how features affect the performance of a model trained on a dataset, our method is aimed at measuring the influence of the features on whether the response variable takes a certain value in a classification problem; to this end, it considers the knowledge provided by a dataset.

Probably the nearest paper to the subject of our research is Datta, Datta, Procaccia, & Zick (2015). In that paper, the authors also study how influential the various features are in a classification problem. They theoretically base their measure of influence in the binary case, that is, when both the features and the response variable take only two possible values. However, their measure of influence can also be used in the general non-binary case. Another difference with our approach is that they start from a set of observed cases of the feature vectors and an already fixed classifier, and study the influence of each feature for that classifier. In our approach we start from a training sample of individuals for whom we have observed their values of the features and of the response variable; we intend to know the influence of the feature values on the response in the population from which the training sample has been drawn. It is certainly possible to use the approach of Datta et al. (2015) to address our problem: train a classifier with the training sample, and then apply Datta et al.'s measure of influence. In fact, in Section 3 we compared the latter approach with our own and show that overall our procedure presents a better performance. In particular, it is seen that our measure, unlike that of Datta et al. (2015), distinguishes between positive and negative influences, thus improving the interpretability of the results. We refer the reader to Section 3 for a further discussion on this topic.

A common point of Štrumbelj & Kononenko (2010), Datta et al. (2015) and our work is that all three make extensive use of cooperative game theory tools, specially the Shapley value. The Shapley value is a rule for distributing the profits generated by a collection of cooperating agents and it has many applications in a wide range of fields. Just to give a few instances, Liu, Ji, Tang, & Li (2020) use the Shapley value for water resource allocation in multinational river basins, Saavedra-Nieves & Saavedra-Nieves (2020) propose a new quota system for the milk market that is based again on the Shapley value, Li & Chen (2020) make use of the Shapley value in their study of alliance formation in an assembly system where several upstream complementary suppliers produce components and sell them to a downstream manufacturer. Algaba, Fragnelli, & Sánchez-Soriano (2019) is a recent review of the Shapley value, its variants, and its applications. The objective of fairly distributing the benefits generated by the cooperation of a set of agents is essentially the same as fairly evaluating the contribution of the variables involved in a classification problem. In the field of game theory, the Shapley value has proven to be an exceptionally valuable tool for distributing such benefits. On the one hand, the Shapley value has been defined to always provide fair allocations. Furthermore, it usually has specific properties in the particular problems to which it is applied, which are often interpretable in a very insightful way. On the other hand, the Shapley value has an explicit formula that allows it to be calculated or approximated in a computationally affordable way, and it has been widely studied. In our particular case, the fact that the Shapley value depends on all possible subgroups of features means that somehow it accounts for all potential interactions between features when calculating our influence measure. On this way, this tool allows us to better understand the importance that the several features have on a classification problem. Thus, the last few years there has been an explosion of papers using the Shapley value to improve the interpretability of complex machine learning models.

A particularly relevant reference on this latter context is Lundberg & Lee (2017), which introduces SHAP (SHapley Additive exPlanations), a unified framework for interpreting predictions and providing theoretical foundation in several prediction models. This methodology is used, for example, in Smith & Alvarez (2021), who analyse a COVID-19 database collected at the early stages of the pandemic in Wuhan (China). The research conducted in our work does not fit within this framework because SHAP focuses on determining the importance of features for each particular prediction. As already mentioned, we instead study the features' importance not for each specific individual but for the whole dataset.

To summarise, the contribution of our paper is framed within a topic of great importance in machine learning today, which is the explainability and interpretability of models. Specifically, we propose a model-agnostic measure of the influence of the features involved in a classification problem. To do so, we make use of the Shapley value, just as it is increasingly done in the artificial intelligence literature. For our influence measure, we first provide a theoretical justification based on the properties that characterise it, and we then show its good performance with a computational study. Finally, we illustrate the versatility of our measure through the analysis of a COVID-19 database in Galicia (Spain).

The organisation of this paper is as follows. Section 2 presents the influence measure and discusses its theoretical basis, including an axiomatic characterisation. In Section 3 various experiments are carried out to validate in practice the behaviour of our measure, which is also compared with another approach from the literature. Section 4 uses the measure to explore data from a sample of COVID-19 patients to detect features that affect mortality, ICU admission, and patient hospitalisation, and to evaluate the influence of such features. Finally, Section 5 summarises the main conclusions of this work.

2. Assessing influence in classification

We start this section by formally establishing what we mean by *classification problem*. In one such problem we have a vector of features $X = (X_1, \dots, X_k)$ and a response variable Y. $K = \{1, \dots, k\}$ denotes the set of indices of the features. Each feature $X_j,\ j \in K$, takes values in a finite set $\mathcal{A}_j$ and Y takes values in a finite set $\mathcal{B}$. We also have a training sample $\mathcal{M} = \{(X^i, Y^i)\}_{i=1}^n$, where $X^i = (X_1^i, \ldots, X_k^i)$ and Y^i are the observed values of the features and the response variable corresponding to individual i. A classification problem is thus characterised by a triplet $(X, Y, \mathcal{M})$.

A *classifier* trained with sample $\mathcal{M}$ is a map $f^{\mathcal{M}}$ that assigns to every $a \in \mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_k$ (an observation of X) a probability distribution over $\mathcal{B}$, i.e., $f^{\mathcal{M}}(a) = (f_b^{\mathcal{M}}(a))_{b\in\mathcal{B}}$ with $f_b^{\mathcal{M}}(a) \geq 0$, for all $b \in \mathcal{B}$, and $\sum_{b\in\mathcal{B}} f_b^{\mathcal{M}}(a) = 1$. Each $f_b^{\mathcal{M}}(a)$ is the estimated probability that an individual whose observed values of the features are given by a belongs to group b of the response variable Y. From now on, $\mathcal{A}_V$, a_V, X_V, and X_V^i will denote the restrictions of $\mathcal{A}$, a, X, and X^i to the variables of V, respectively (for all $V \subseteq K$).

Our goal in this section is to use classification techniques to define a measure that allows us to study the influence of the features on the response variable. The formal definition of an influence measure is the one included below.

Definition 1. An *influence measure* for $(X, Y, \mathcal{M})$ is a map I that assigns to every $a_R \in \mathcal{A}_R$ $(R \subseteq K)$, $b \in \mathcal{B}$, and $T \subseteq K$ $(T \neq \emptyset)$ a vector $I(a_R, b, T) = (I_l(a_R, b, T))_{l\in T} \in \mathbb{R}^T$. The vector $I(a_R, b, T)$ provides an evaluation of the influence that each feature X_l $(l \in T)$ has on whether the response is worth b when X_R is worth a_R and we only take into account the features $\{X_l\}_{l\in T}$.

Remark 2. Note that R represents the subset of indices corresponding to those features whose values are to be fixed and, therefore, we will only take into account those observations that present these values. On the other hand, T is the non-empty subset of K corresponding to those features whose influences are to be studied. The appropriate selection of these subsets allows this influence measure to be used with considerable versatility, as illustrated in Section 4.

Throughout this section we present an influence measure based on the Shapley value of cooperative games. To strengthen the theoretical support for our measure, we introduce it axiomatically, i.e., we provide a collection of desirable properties and then show that there exists a unique measure that satisfies them. As we show below, two properties are sufficient to characterise our measure.

To begin with, in order to facilitate the reader's understanding, we include the definition of the Shapley value. First, recall that a cooperative game is a par (N, v), where N is the finite set of players, and $v : 2^N \to \mathbb{R}$ is the characteristic function of the game, which satisfies $v(\emptyset) = 0$. We usually interpret $v(S)$ as the gain that coalition $S \subseteq N$ can obtain. Also, $G(N)$ represents the set of all cooperative games with set of players N. In general, we identify (N, v) with its characteristic function, v. An extensively addressed problem in cooperative games is to allocate $v(N)$ among the cooperating agents. One of the most important allocation rules is the Shapley value, $\Phi : G(N) \to \mathbb{R}^N$, which represents a fair compromise for the players and it is defined by the following expression:

$$\Phi_i(v) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!} (v(S\cup\{i\}) - v(S)),$$

for all $v \in G(N)$ and $i \in N$. For more details on cooperative games see, for instance, González-Díaz, García-Jurado, & Fiestras-Janeiro (2010).

Regarding the problem that concerns us, our objective is to know how much certain values of some variables contribute to the prediction of a given set of individuals. In particular, for a single individual, $a = (a_1, \ldots, a_k)$, we consider the difference between the average prediction of our classifier when only the feature values corresponding to T are known, and the average prediction when no feature value is known. Thus, we take:

$$\triangle_{f_b^{\mathcal{M}}}(a, T) = \frac{1}{|\mathcal{A}_{K\setminus T}|} \sum_{a'_{K\setminus T}\in\mathcal{A}_{K\setminus T}} f_b^{\mathcal{M}}(a_T, a'_{K\setminus T}) - \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} f_b^{\mathcal{M}}(a'), \quad (1)$$

where $(a_T, a'_{K\setminus T})$ denotes the k-dimensional feature vector whose j-th component is a_j if $j \in T$ or a'_j if $j \in K\setminus T$.

The first desirable property we consider establishes that a measure of influence distributes among the features corresponding to T their total influence on the response variable (when X_R equals a_R). One way to estimate that total influence using the classifier $f^{\mathcal{M}}$ is given by the following expression:

$$\frac{1}{n_{a_R}^b} \sum_{(X^i,Y^i)\in\mathcal{M}_{a_R}^b} \triangle_{f_b^{\mathcal{M}}}(X^i, T), \quad (2)$$

where $\mathcal{M}_{a_R}^b$ denotes the subsample of $\mathcal{M}$ formed by the observations (X^i, Y^i) with $X_R^i = a_R$ and $Y^i = b$, and $n_{a_R}^b$ denotes the size of the subsample $\mathcal{M}_{a_R}^b$.

Notice that expression (2) can be interpreted as an estimation of the variability of the response variable due to the T features (using $f^{\mathcal{M}}$). Therefore, the first property we ask for an influence measure is the $f^{\mathcal{M}}$-Efficiency below.

$f^{\mathcal{M}}$-Efficiency. An influence measure I satisfies $f^{\mathcal{M}}$-Efficiency if, for every $(X, Y, \mathcal{M})$, every $a_R \in \mathcal{A}_R$ $(R \subseteq K)$, $b \in \mathcal{B}$, and $T \subseteq K$ $(T \neq \emptyset)$, it holds that $\sum_{l\in T} I_l(a_R, b, T)$ is equal to the amount in expression (2).

The second property that we consider is a fairness property that treats all features in a balanced way. Informally, it states that given two of these features, the effect of ignoring one to the measure of the influence of the other is identical for both features. Note that the marginal loss or gain of influence that the inclusion or exclusion of one feature causes to another feature is due to the dependency that exists between the two. The fact that the dependence between features is symmetrical, makes advisable the property of balanced contributions.

Balanced contributions. An influence measure satisfies Balanced Contributions if, for every $(X, Y, \mathcal{M})$, every $a_R \in \mathcal{A}_R$ $(R \subseteq K)$, $b \in \mathcal{B}$, $T \subseteq K$ $(T \neq \emptyset)$, and $l, m \in T$ with $l \neq m$,

$$I_l(a_R, b, T) - I_l(a_R, b, T\setminus\{m\}) = I_m(a_R, b, T) - I_m(a_R, b, T\setminus\{l\}).$$

Now we state and prove the main mathematical result of this section. It provides a characterisation and a formal expression of an influence measure that satisfies all the properties introduced above.

Theorem 3. *There exists a unique influence measure for $(X, Y, \mathcal{M})$ which satisfies the properties of $f^{\mathcal{M}}$-Efficiency and Balanced Contributions. For all $a_R \in \mathcal{A}_R$ $(R \subseteq K)$, $b \in \mathcal{B}$, $T \subseteq K$ $(T \neq \emptyset)$ and $l \in T$, this measure (that we denote by I^Φ) is given by*

$$I_l^\Phi(a_R, b, T) = \frac{1}{n_{a_R}^b} \sum_{(X^i,Y^i)\in\mathcal{M}_{a_R}^b} \Phi_l(v_{X^i}^b|_T), \quad (3)$$

where Φ denotes the Shapley value, $v_{X^i}^b$ denotes the game with set of players K given by

$$v_{X^i}^b(S) = \frac{1}{|\mathcal{A}_{K\setminus S}|} \sum_{a'_{K\setminus S}\in\mathcal{A}_{K\setminus S}} f_b^{\mathcal{M}}(X_S^i, a'_{K\setminus S}) - \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} f_b^{\mathcal{M}}(a'), \quad (4)$$

for all $S \subseteq K$, and $v_{X^i}^b|_T$ denotes the restriction of the game $v_{X^i}^b$ to the subsets of T.[1]

[1] The game in (4) results to be the same as the one used in Štrumbelj & Kononenko (2010) to assess the importance of the various features in the classification of a particular individual in a classification problem.

Proof. *Existence.* To show that I^{Φ} satisfies $f^{\mathcal{M}}$-Efficiency, take $a_R \in \mathcal{A}_R$ $(R \subseteq K)$, $b \in \mathcal{B}$, and $T \subseteq K$ $(T \neq \emptyset)$. Shapley (1953) proves that the Shapley value of cooperative games satisfies an efficiency property. In our case, this property implies that

$$\sum_{l \in T} \Phi_l(v^b_{X^i}|_T) = v^b_{X^i}(T).$$

Applying this result we obtain that:

$$\begin{aligned}
&\sum_{l \in T} I^{\Phi}_l(a_R, b, T) \\
&= \sum_{l \in T} \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \Phi_l(v^b_{X^i}|_T) \\
&= \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \sum_{l \in T} \Phi_l(v^b_{X^i}|_T) \\
&= \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} v^b_{X^i}(T) \\
&= \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \left(\frac{1}{|\mathcal{A}_{K\setminus T}|} \sum_{a'_{K\setminus T}\in \mathcal{A}_{K\setminus T}} f^{\mathcal{M}}_b(X^i_T, a'_{K\setminus T}) - \frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} f^{\mathcal{M}}_b(a') \right) \\
&= \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \triangle_{f^{\mathcal{M}}_b}(X^i, T),
\end{aligned}$$

which equals expression (2).

To show that I^{Φ} satisfies Balanced Contributions, let $a_R \in \mathcal{A}_R$ $(R \subseteq K)$, $b \in \mathcal{B}$, $T \subseteq K$ $(T \neq \emptyset)$, and $l, m \in T$ with $l \neq m$. Myerson (1980) proves that the Shapley value of cooperative games satisfies a property of balanced contributions. In our case, this property implies that

$$\Phi_l(v^b_{X^i}|_T) - \Phi_l(v^b_{X^i}|_{T\setminus\{m\}}) = \Phi_m(v^b_{X^i}|_T) - \Phi_m(v^b_{X^i}|_{T\setminus\{l\}}).$$

Applying this result we obtain that:

$$\begin{aligned}
&I^{\Phi}_l(a_R, b, T) - I^{\Phi}_l(a_R, b, T\setminus\{m\}) \\
&= \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \Phi_l(v^b_{X^i}|_T) - \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \Phi_l(v^b_{X^i}|_{T\setminus\{m\}}) \\
&= \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \left(\Phi_l(v^b_{X^i}|_T) - \Phi_l(v^b_{X^i}|_{T\setminus\{m\}})\right) \\
&= \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \left(\Phi_m(v^b_{X^i}|_T) - \Phi_m(v^b_{X^i}|_{T\setminus\{l\}})\right) \\
&= \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \Phi_m(v^b_{X^i}|_T) - \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} \Phi_m(v^b_{X^i}|_{T\setminus\{l\}}) \\
&= I^{\Phi}_m(a_R, b, T) - I^{\Phi}_m(a_R, b, T\setminus\{l\}).
\end{aligned}$$

Uniqueness. We show uniqueness by induction on the size of T. Suppose that I^1 and I^2 are two influence measures satisfying $f^{\mathcal{M}}$-Efficiency and Balanced Contributions. If $|T| = 1$, by $f^{\mathcal{M}}$-Efficiency,

$$I^1(a_R, b, T) = \frac{1}{n^b_{a_R}} \sum_{(X^i,Y^i)\in \mathcal{M}^b_{a_R}} v^b_{X^i}(T) = I^2(a_R, b, T).$$

Assume now that $I^1(a_R, b, S) = I^2(a_R, b, S)$ for all $S \subseteq T$ with $1 \leq |S| < |T|$. Then, by Balanced Contributions, for all $l, m \in T$, $l \neq m$,

$$I^1_l(a_R, b, T) - I^1_m(a_R, b, T) = I^2_l(a_R, b, T) - I^2_m(a_R, b, T). \tag{5}$$

Using $f^{\mathcal{M}}$-Efficiency,

$$\sum_{l \in T} I^1_l(a_R, b, T) = \sum_{l \in T} I^2_l(a_R, b, T). \tag{6}$$

By (5) and (6) it is obtained that:

$I^1_l(a_R, b, T) = I^2_l(a_R, b, T)$ for all $l \in T$.

This last expression gives the uniqueness. □

3. Empirical results

In this section we show the performance of the proposed influence measure (3) by means of a computational study. Three different experiments have been carried out using the software R. The objective of such simulations is to corroborate that the results obtained by the methodology introduced in the current work are in accordance with the expected ones. Furthermore, these results are compared with those obtained by the influence measure introduced in Datta et al. (2015), which counts the number of times that a modification in a feature results in a different classification. We provide the formal definition of such an influence measure below.

Definition 4. Given a training set $\mathcal{M} = \{(X^i, Y^i)\}_{i=1}^n$ and a classifier $f^{\mathcal{M}}$, the influence of the j-th feature is

$$\chi_j(f^{\mathcal{M}}) = \sum_{a' \in \{X^i\}} \sum_{\substack{a_j \in \mathcal{A}_j: \\ (a'_{-j}, a_j) \in \{X^i\}}} \min\left\{ \left| \arg\max_{b \in \mathcal{B}} f^{\mathcal{M}}_b(a'_{-j}, a_j) - \arg\max_{b \in \mathcal{B}} f^{\mathcal{M}}_b(a') \right|, 1 \right\},$$

where (a'_{-j}, a_j) denotes the vector $(a'_1, \ldots, a'_{j-1}, a_j, a'_{j+1}, \ldots, a_k)$, $\{X^i\}$ denotes $\{(X^i_1, \ldots, X^i_k)\}_{i=1}^n$, and $\mathcal{B} \subset \mathbb{N}$.

The classifier used in this paper is Breiman's random forest classifier (Breiman, 2001), implemented in Weka[2] and used through `RWeka`[3]. This choice is motivated by the excellent result of the random forest type classifiers (see, for example, Fernández-Delgado et al., 2014). The code was run on a quad-core Intel i7-8665U CPU with 16 GB RAM.

The procedure adopted in the experiments is as follows. We start from a sample of individuals from which their attributes and response are known, $\mathcal{M} = \{(X^i, Y^i)\}_{i=1}^n$. Right after, such sample is used to train a previously chosen classifier, obtaining $f^{\mathcal{M}}$. To evaluate the influence of feature X_j on the response Y taking the value b, the quantities $I^{\Phi}_j(a_j, b, K)$ and $\sum_{l \in K} I^{\Phi}_l(a_j, b, K)$ are computed and analysed for all $a_j \in \mathcal{A}_j$.

For the first experiment, a sample of 1000 instances with four binary features $\{X_1, X_2, X_3, X_4\}$ was generated. Such attributes take the values 0 and 1 with probability 0.5 (hence, $a_j \in \mathcal{A}_j = \{0, 1\}$, $j \in K$). In half of the instances, the value of Y coincides with the value of X_1, while in the remaining instances the value of Y coincides with the value of X_2; note thus that $b \in \mathcal{B} = \{0, 1\}$. The following step is to select those observations whose assigned class was $b = 1$. Afterwards, for each attribute X_j, $j \in K$, and each of its possible values, we study the influence that such feature had on the response when it took such value. Since the procedure by which the class has been generated is known, it is evident that the influence of attributes X_3 and X_4 should be independent of their values. Furthermore, the value 1 for features X_1 and X_2 should have a stronger influence in the classification than the value 0. Table 1 and Fig. 1 present the results obtained for this simulation, which took a runtime of 9.3 minutes.

Indeed, it can be observed that for attributes X_1 and X_2 the value $I^{\Phi}_j(a_j, b, K)$ is positive when $a_j = 1$ and negative when $a_j = 0$, which means that features X_1 and X_2 taking the value 1 works in favour of the response resulting in 1, unlike what happens if these features are worth 0. Note also that $\sum_{l \in K} I^{\Phi}_l(a_j, b, K)$ is the total influence of the four features on the response being 1 when

[2] http://www.cs.waikato.ac.nz/ml/weka.

[3] https://cran.r-project.org/web/packages/RWeka/index.html.

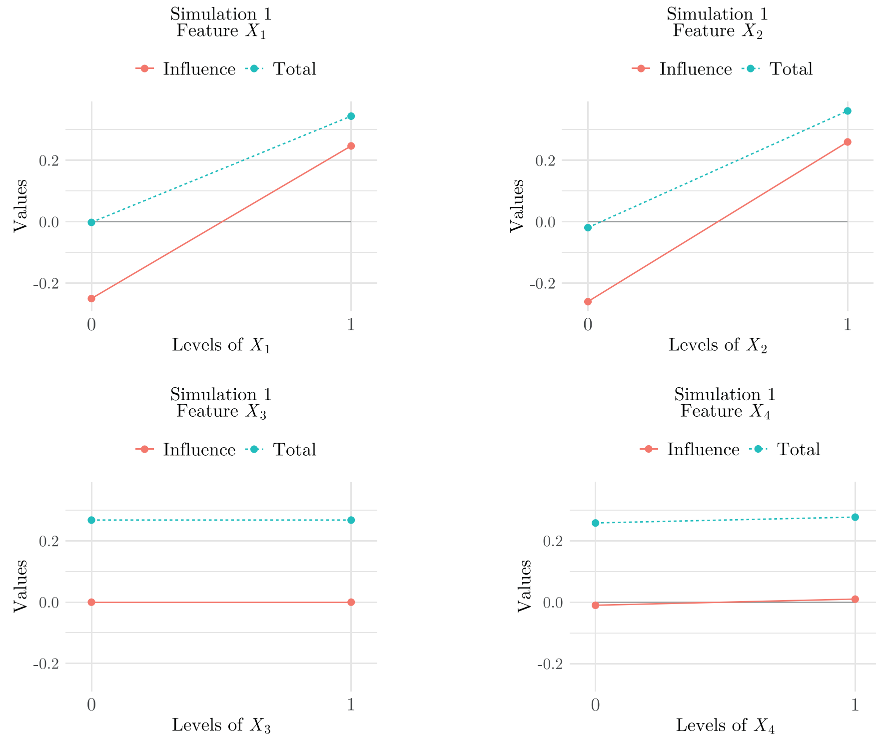

Fig. 1. Influence and total influence for the features (Simulation 1).

Table 1
Results for simulation 1.

$X_j, j \in K$	a_j	$\sum_{l \in K} I_l^{\Phi}(a_j, b, K)$	$I_j^{\Phi}(a_j, b, K)$
X_1	0	-0.002	-0.250
	1	0.344	0.247
X_2	0	-0.019	-0.260
	1	0.361	0.260
X_3	0	0.268	0.000
	1	0.268	0.000
X_4	0	0.258	-0.010
	1	0.277	0.010

Table 2
Results for simulation 2.

$X_j, j \in K$	a_j	$\sum_{l \in K} I_l^{\Phi}(a_j, b, K)$	$I_j^{\Phi}(a_j, b, K)$
X_1	0	-0.001	-0.009
	1	0.017	0.011
X_2	0	-0.012	-0.019
	1	0.026	0.023
X_3	0	0.007	-0.002
	1	0.010	0.006
X_4	0	0.002	-0.003
	1	0.014	0.005

feature X_j takes the value a_j. In view of the results obtained, for features X_1 and X_2 the quantities $I_j^{\Phi}(a_j, b, K)$ and $\sum_{l \in K} I_l^{\Phi}(a_j, b, K)$ are closer when $a_j = 1$ than when $a_j = 0$. Thus, the total influence on the response being 1 when either X_1 or X_2 are 1, is in fact due to these specific attributes taking the value 1. In the case of features X_3 and X_4, their influence is near 0 whatever value they take.

Applying the procedure in Datta et al. (2015) to the previous experiment, we obtain the measure $(0.50, 0.50, 0.25, 0.25)$. As expected, features X_1 and X_2 present a higher influence than X_3 and X_4. Just as we have already mentioned, Datta et al.'s procedure measures the number of times that a change in a specific attribute produces a different response. Thus, it only takes positive values, which prevents us from knowing the direction of the influence. In our case, setting features X_1 and X_2 to 0 works against the response being 1, and this is made clear by the negative sign of their influences.

The second experiment differs from the previous one in the procedure to assign the class to the instances. The response is now generated as a binary vector which takes the values 0 and 1 with probability 0.5, independently of the attributes. The goal of this simulation is to show that the influence of the features in the classification of the instances with response $b = 1$ does not depend on the features' values. Table 2 and Fig. 2 present the results obtained for this simulation. The computational time was 12.4 minutes.

Again, the outcomes are as expected: for each feature, there are barely differences in the values $I_j^{\Phi}(a_j, b, K)$ and $\sum_{l \in K} I_l^{\Phi}(a_j, b, K)$ when a_j changes. In this case, Datta et al.'s measure resulted in (0.375,0.375,0.375,0.375). The response is not influenced by any one attribute more than the others. However, because the class was generated independently of the features, one would expect their influence to be zero.

Finally, we have considered the non-binary case. Now, the four attributes can take the values 0, 1 and 2 with equal probability, and the class of the response is computed as follows: in $1/3$ of the instances, it is the value of attribute X_1 that determines the response; while in the remaining $2/3$, it is attribute X_2 that determines it. Table 3 and Fig. 3 illustrate the results. This took a runtime of 13.3 minutes.

The outcomes obtained show that changes in features X_3 and X_4 do not affect to the response being $b = 1$, and their influence is almost zero whatever their values. Nevertheless, the value 1 of attributes X_1 and X_2 has a positive influence, which is larger in the case of the latter. On the contrary, when these attributes take the values 0 and 2, their influence is negative. This speaks against the class resulting in 1. In this case, the influence measure of Datta

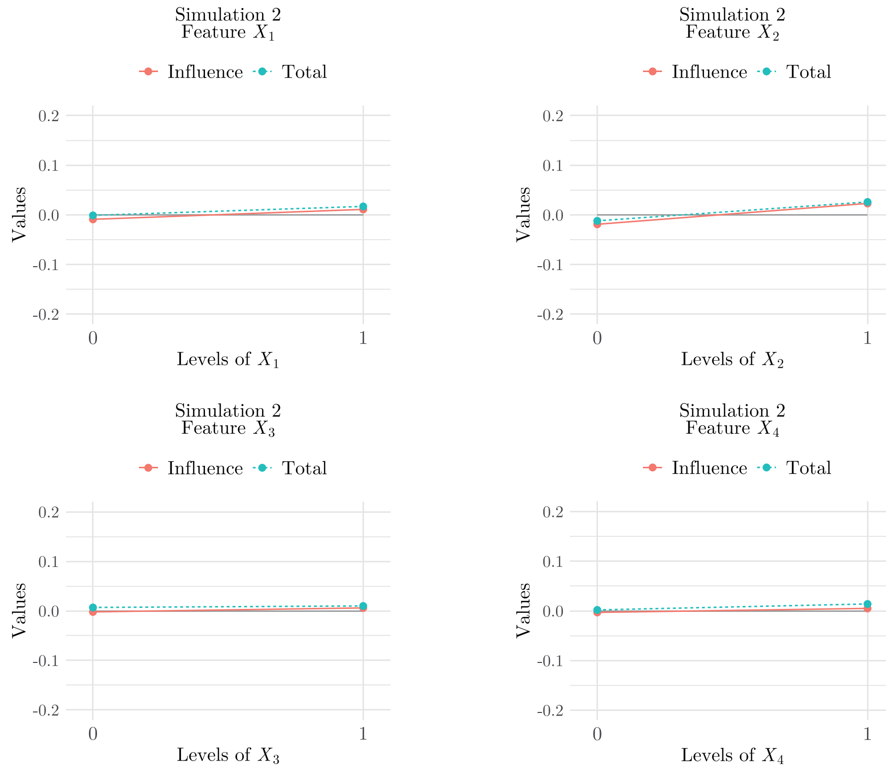

Fig. 2. Influence and total influence for the features (Simulation 2).

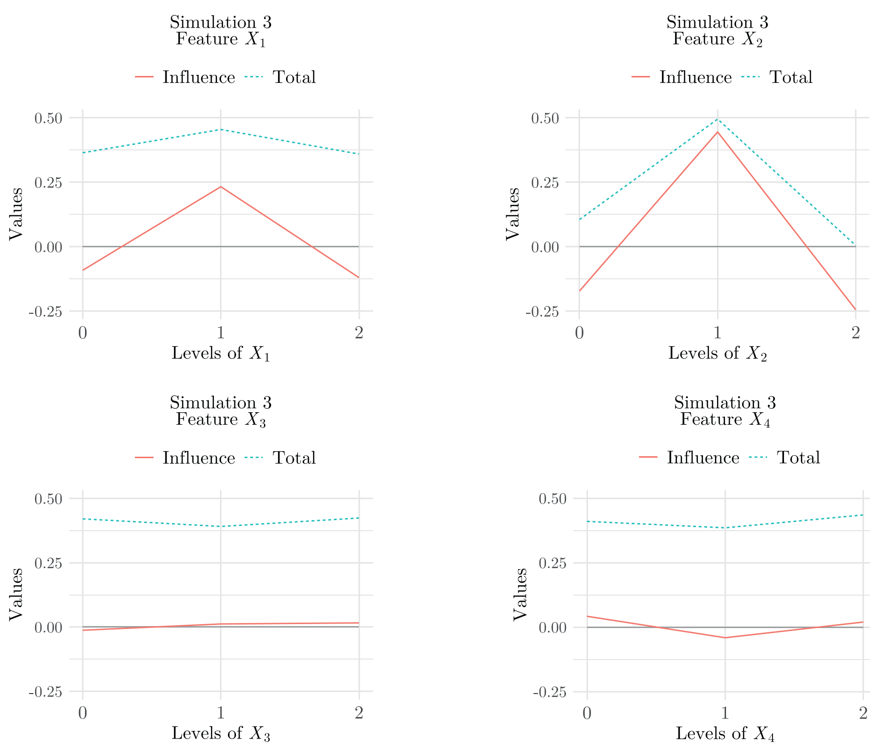

Fig. 3. Influence and total influence for the features (Simulation 3).

et al. is $(0.321, 1.827, 0.296, 0.296)$. This result shows that X_2 is the most influential feature, and that X_1 is more relevant than X_3 and X_4. Nevertheless, this measure does not properly capture the magnitude of how much more influential attribute X_1 is in comparison to X_3 and X_4.

In view of the previous results, our methodology seems to be appropriate to study the influence that the different feature values have on the classification of individuals. Since the experiments are satisfactory, this analytic tool can be applied to real-life problems. Consequently, this procedure has been employed on a real dataset

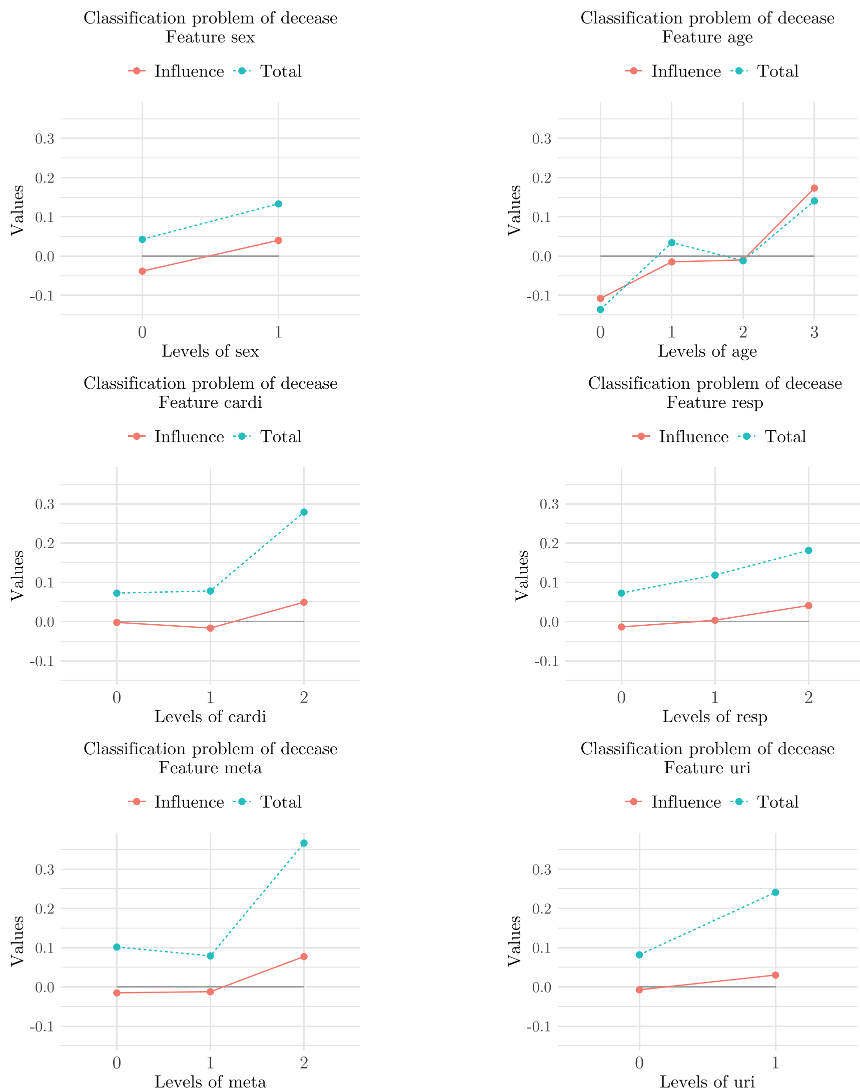

Fig. 4. Influence and total influence for the features on the decease.

Table 3
Results for simulation 3.

$X_j, j \in K$	a_j	$\sum_{l \in K} I_l^{\Phi}(a_j, b, K)$	$I_j^{\Phi}(a_j, b, K)$
X_1	0	0.364	-0.091
	1	0.455	0.233
	2	0.360	-0.120
X_2	0	0.105	-0.172
	1	0.495	0.445
	2	0.005	-0.245
X_3	0	0.421	-0.012
	1	0.391	0.012
	2	0.424	0.016
X_4	0	0.410	0.042
	1	0.385	-0.041
	2	0.435	0.020

concerning COVID-19 patients, whose results are presented in the next section.

4. Application of our influence measure to COVID-19 data

This section analyses a database of 10,454 patients from Galicia (a region in the northwest of Spain) infected with COVID-19 from March 6, 2020 to May 7, 2020. The objective is to study the influence of various patients' characteristics in three binary response variables of special interest: the need for hospitalisation, the need for ICU admission, and the eventual decease. The emphasis is not on the predictive classification of new patients, but on the analysis of the characteristics that influenced the patients whose complete history is known to have a positive response in the binary variables indicated. On the other hand, what follows is not intended to be an exhaustive study of these data to draw definitive conclusions about the evolution of COVID-19, but simply an illustration of some of the uses of the measure of influence we introduced in Section 2.

The features that have been considered in this study are the following:

- **Sex**: 0 (woman), 1 (man).
- **Age**: 0 (0–49 y/o), 1 (50–64 y/o), 2 (65–79 y/o), 3 (80 y/o and over).
- **Cardiovascular diseases**: 0 (without diseases), 1 (mild diseases), 2 (severe diseases: ischaemia with angina, infarction, stroke).
- **Respiratory diseases**: 0 (no diseases), 1 (mild diseases), 2 (severe diseases: malignancy, COPD, pneumonia).

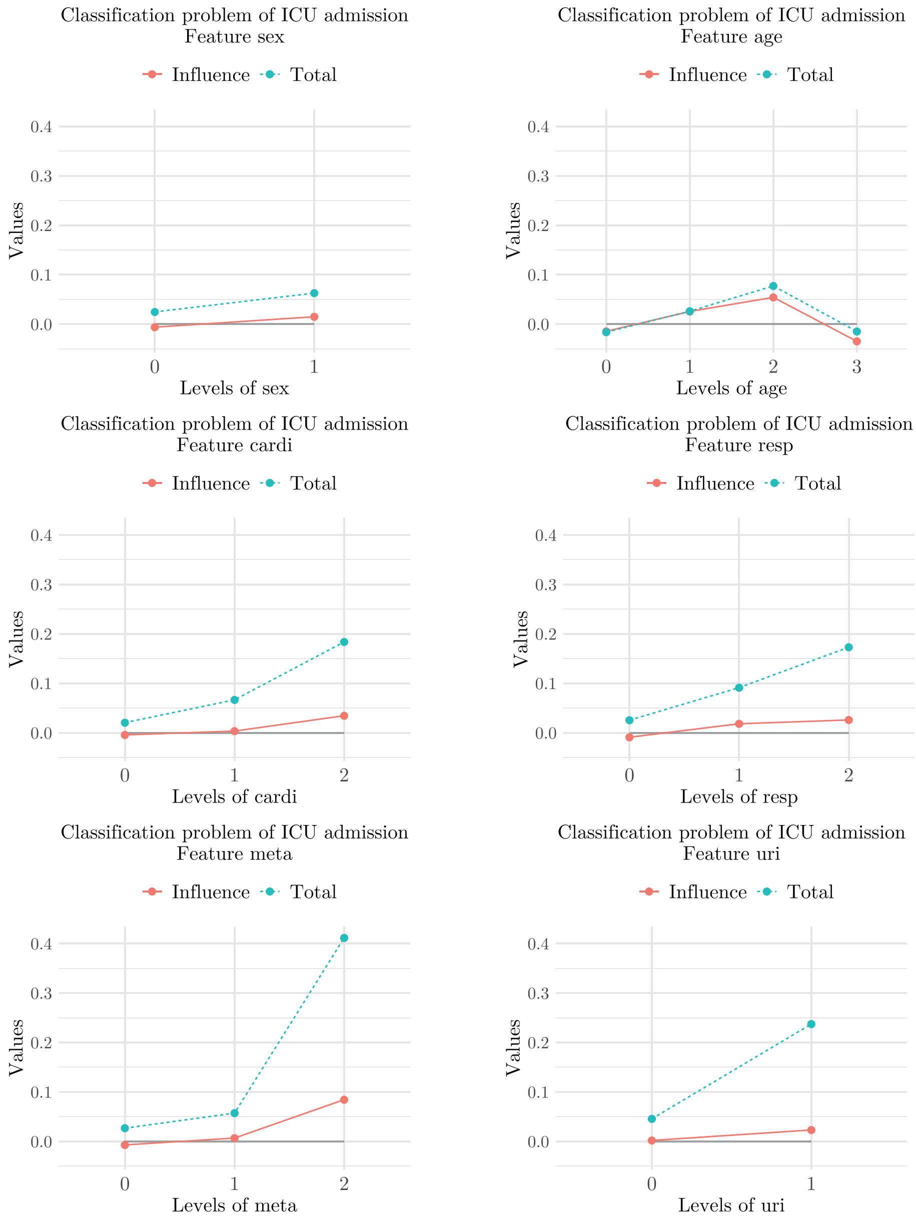

Fig. 5. Influence and total influence for the features on the ICU admission.

- **Metabolic diseases**: 0 (no diseases), 1 (mild diseases), 2 (severe diseases: malignancy, insulin-dependent diabetes).
- **Urinary diseases**: 0 (none or mild diseases), 1 (severe diseases: malignancy, kidney failure).

The binary response variables considered in this application are:

- **Decease (exitus)**: 0 (no), 1 (yes).
- **ICU admission**: 0 (no), 1 (yes).
- **Need for hospitalisation**: 0 (no), 1 (yes).

Next, we applied the methodology outlined in Section 2 to measure the influence of the features in the classification with respect to the binary response variables. For instance, the interest would reside in selecting those individuals who resulted in decease (that is, $\texttt{decease} = 1$) when our purpose is to know the most influential attributes for the exitus. Note that to estimate the influence of feature X_j on Y, we use the influence that X_j has in the classification of the elements of the sample $\mathcal{M}$ using an excellent classifier, since it is precisely trained with the sample $\mathcal{M}$. As in the previous section, we use the random forest classifier introduced by Breiman (2001) and implemented in R through the RWeka library.

Let $\{X_1 = \texttt{sex}, X_2 = \texttt{age}, X_3 = \texttt{cardi}, X_4 = \texttt{resp}, X_5 = \texttt{meta}, X_6 = \texttt{uri}\}$ be the set of features. We start the analysis by presenting Figs. 4–6, which display the influence and total influence of the different features' values on the three classification problems. Let us explain in more detail what the graphics in the figures show. In each of the graphics a response variable is chosen and set its value to 1, and also a feature is chosen. The graphic shows in red (solid) the measure of influence of the chosen feature when we set its value to each of the possible values it can take (feature influence), and in blue (dashed) the sum of the measures of influence of all the features (total influence). The objective of these figures is to identify what we call *influence scenarios*. An influence scenario is detected when the total influence shown in the corresponding graphic deviates noticeably from zero.

For example, in Fig. 4 several influence scenarios can be identified. The first is the case of age, both when it is worth 0 and when it is worth 3. There are two influence scenarios here that al-

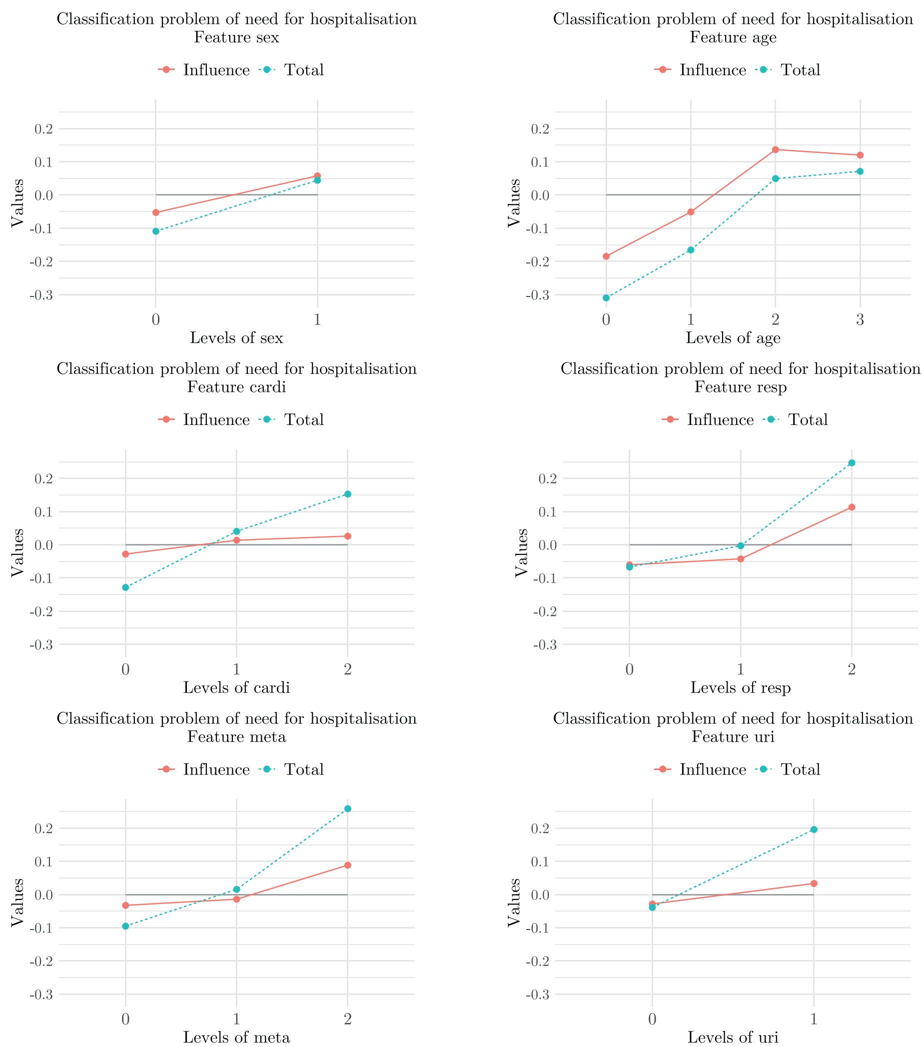

Fig. 6. Influence and total influence for the features on the need for hospitalisation.

Table 4
Influence measure. Decease = 1.

	`sex`	`age`	`cardi`	`resp`	`meta`	`uri`	Total
`cardi` = 2	0.025	**0.151**	**0.049**	0.016	0.014	0.023	0.279
`meta` = 2	0.035	**0.142**	0.034	0.062	**0.078**	0.017	0367

Table 5
Influence measure. ICU admission = 1.

	`sex`	`age`	`cardi`	`resp`	`meta`	`uri`	Total
`meta` = 2	0.064	**0.098**	0.072	0.071	**0.084**	0.022	0.412

low us to state that in the case of young individuals (`age` = 0) and in the case of old individuals (`age` = 3) we detect an important influence of the features on mortality, negative in the first case and positive in the second. We can observe that in this graphic the red and blue lines (`age` influence and total influence, respectively) are very close, which means that this total influence is mainly due to age.

Other influence scenarios that can be inferred from the figure are those corresponding to the feature `cardi` being 2 and the feature `meta` being 2. Note, however, that in such scenarios the red and blue lines are noticeably separated, which means that the significant total influence detected is not primarily due to the features chosen in each case. Therefore, for each of these two scenarios, Table 4 presents the value of the influence measure for all the features, in order to identify which ones are influencing the most.

From Table 4 it can be seen that `age` is the most influential feature in these two scenarios, although the features chosen in each case (`cardi` and `meta`, respectively) are the second most influential.

Fig. 5 shows, surprisingly, a minor influence of age on ICU admissions. This is probably because in the first wave of COVID-19 in Spain, a considerable number of elderly died in nursing homes before they could even be hospitalised or admitted to ICU. In any case, `age` generates an influence scenario when it is worth 2. As in Fig. 4, in the case of age the blue and red lines are very close, showing that the total influence in this particular situation is mainly due to age.

Another influence scenario presented in Fig. 5 is the one corresponding to the `meta` feature being equal to 2. In that case, the blue and red lines are far apart, so we show in Table 5 the value of the influence measure for all features. It can be observed that all features are influential, although the most influential are, in this order, age and metabolic diseases.

Fig. 6 allows us to identify other influence scenarios, among which we highlight those corresponding to `age` equal to 0, `meta`

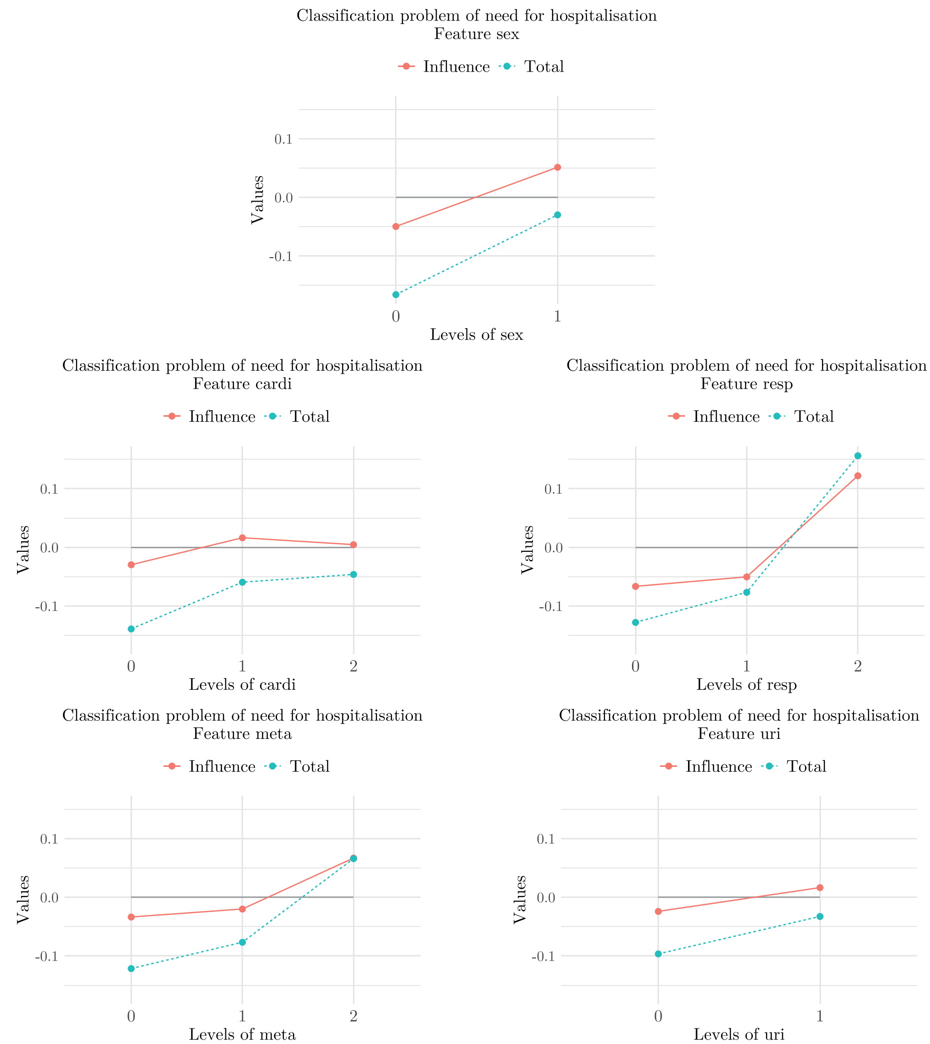

Fig. 7. Influence and total influence for features $K \setminus \{2\}$ on the need for hospitalisation.

Table 6
Influence measure. Need for hospitalisation = 1.

	sex	age	cardi	resp	meta	uri	Total
age = 0	-0.001	**-0.184**	-0.013	**-0.047**	-0.030	-0.034	-0.310
resp = 2	0.039	**0.085**	-0.009	**0.113**	0.014	0.005	0.247
meta = 2	0.026	**0.095**	0.054	0.012	**0.089**	-0.017	0.247

equal to 2 and resp equal to 2. In this case, although the blue and red lines tend to coincide more in the age feature, they are considerably separated in all the influence scenarios. Therefore, we show in Table 6 the value of the influence measure for all features in the three scenarios.

Again, age remains a highly influential feature in the occurrence of hospitalisation in all the influence scenarios we have detected. In the first scenario, when age is 0, what we observe is that the marked tendency towards less hospitalisation when patients are young is mainly due to their youth, although we also detect an important influence of good health in terms of respiratory ailments. In the influence scenario when resp = 2, the measure indicates that respiratory diseases are the most influential in the need for hospitalisation, even more so than age. Somehow we detect that respiratory pathologies, in addition to age, are considerably influential in the need for hospitalisation of COVID-19 patients.

In light of the above, it is evident that the most influential feature in all the response variables considered is age: young people are less likely to need hospitalisation and admission to the ICU, as well as to die from COVID-19; the only exception we detected is that elderly people who die have a tendency to die quickly, even before being admitted to the ICU.

With this in mind, we could look further for other influential features by eliminating the age effect. That is, we can remove age from the list of features (i.e., following the notation in Section 2, $T = K \setminus \{2\}$, where $X_2 =$ age) and calculate the corresponding measure of influence. Through this approach, the expectation is that fewer influential scenarios will be detected; but in detected cases, the most influential features after age may come to light. We perform this analysis for the sub-sample in which we have the largest number of observations: the one corresponding to need for hospitalisation equal to 1.

Fig. 7 seems to confirm the considerable influence of respiratory diseases on the need for hospitalisation of COVID-19 patients. Indeed, the only positive influence scenario detected occurs when resp = 2. Note also that, in this case, the blue and red lines are close, so that the total influence detected is mostly due to respiratory pathologies.

There is another scenario of influence when cardi = 0. In this case, it is striking that the red line is close to the point (0,0).

Table 7
Influence measure without considering age. Need for hospitalisation = 1.

	sex	cardi	resp	meta	uri	Total
cardi = 0, resp = 0	0.006	-0.028	**-0.063**	-0.025	-0.046	-0.156
cardi = 0, meta = 0	0.009	-0.038	**-0.052**	-0.033	-0.037	-0.151
cardi = 0, uri = 0	0.005	-0.030	**-0.054**	-0.022	-0.042	-0.143

This seems to indicate that in healthy individuals regarding cardiac functions an important influence on the decrease in hospitalisations is detected, but however such a decrease is not due to the feature cardi. To detect which is the most influential feature in this case, we show in Table 7 the value of the measure of influence when cardi = 0 and any other of the pathologies considered is also 0. Notice that in these three cases, feature resp is the most influential by far. Once again, the data we handle seem to confirm the important influence of the presence of respiratory pathologies on the need for hospitalisation of COVID-19 patients.

5. Conclusions

This paper introduces a new general measure of the influence that various features of a set of individuals have on their classification. For the construction of such measure, we consider several ideas taken from game theory. In particular, we define a cooperative game (whose players are the features considered) and apply a game theoretical solution, known as the Shapley value. An axiomatic characterisation theorem for the proposed influence measure is stated and mathematically proved. The properties used in this result are adaptations of Shapley value's properties in the general context of cooperative games, and are highly desirable from the exploratory data analysis point of view. To test the scope and adequacy of the proposed influence measure, a control experiment that provides a very satisfactory result is designed. Our proposal is also compared with the influence measure defined in Datta et al. (2015), which also uses ideas from game theory. Section 4 provides an application of our measure to the study of a Spanish database of patients infected with COVID-19 from the first wave of the pandemic, between March and May 2020. The aim of this application is to determine which demographic features, as well as previous pathologies, are the most influential in the classification of a patient regarding their potential need for hospitalisation, admission to the intensive care unit, or death. Initial results obtained present a promising future for the technique proposed here as a decision support tool. It serves, in particular, to alert medical professionals of the importance of certain patient characteristics, such as age or prior pathologies, as opposed to the lesser importance or influence of others. Such characteristics potentially pose an added difficulty in patients with a given disease, which should be taken into account both in the care and treatment that these patients should receive and in the planning of resources destined for them.

As for future lines of research, we believe that additional work on the recently introduced measure of influence is worthwhile. We cite, for example, the desirability of further analysing the sensitivity of the results provided by the measure of influence according to the classifiers used, exploring its relation with other statistical techniques of multivariate analysis, as well as extending it to continuous scenarios. It would also be possible to complete the application presented using data from successive waves of the COVID-19 pandemic and taking into account the existence of different virus strains.

Acknowledgments

The authors are grateful to Ricardo Cao Abad and to the Dirección Xeral de Saúde Pública of the Xunta de Galicia in Spain. This work has been supported by the ERDF, the Government of Spain/AEI [grants MTM2017-87197-C3-1-P and MTM2017-87197-C3-3-P]; the Xunta de Galicia [Grupos de Referencia Competitiva ED431C2016-015, ED431C2017/38, and ED431C 2021/24, and Centro Singular de Investigación de Galicia ED431G/01]; and by the collaborative research project of the IMAT "Mathematical, statistical and dynamic study of the epidemic COVID-19", subsidized by the Vice-Rector's Office for Research and Innovation at the University of Santiago de Compostela, Spain. The research of Laura Davila-Pena has been funded by the Government of Spain [grant FPU17/02126]. We would also like to thank the three anonymous referees and the editor for their constructive comments and suggestions, which helped us to improve the final version of this paper.

References

Algaba, E., Fragnelli, V., & Sánchez-Soriano, J. (2019). *Handbook of the Shapley value* (1st). CRC Press, Taylor & Francis.
Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.
Burkart, N., & 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., & Morales, D. R. (2021). Mathematical optimization in classification and regression trees. *TOP, 29*, 5–33.
Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the feature importance for black box models. In M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, & G. Ifrim (Eds.), *Machine learning and knowledge discovery in databases* (pp. 655–670). Springer.
Datta, A., Datta, A., Procaccia, A. D., & Zick, Y. (2015). Influence in classification via cooperative game theory. In *Proceedings of the twenty-fourth international joint conference on artificial intelligence* (pp. 511–517).
Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*, 3133–3181.
Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research, 265*, 993–1004.
González-Díaz, J., García-Jurado, I., & Fiestras-Janeiro, M. G. (2010). *An introductory course on mathematical game theory* (1st). American Mathematical Society.
Jothi, N., Husain, W., & Rashid, N. A. (2021). Predicting generalized anxiety disorder among women using Shapley value. *Journal of Infection and Public Health, 14*, 103–108.
Li, T., & Chen, J. (2020). Alliance formation in assembly systems with quality-improvement incentives. *European Journal of Operational Research, 285*(3), 931–940.
Liu, D., Ji, X., Tang, J., & Li, H. (2020). A fuzzy cooperative game theoretic approach for multinational water resource spatiotemporal allocation. *European Journal of Operational Research, 282*(3), 1025–1037.
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems, volume 30* (pp. 4765–4774). Curran Associates Inc.
Myerson, R. B. (1980). Conference structures and fair allocation rules. *International Journal of Game Theory, 9*, 169–182.
Saavedra-Nieves, A., & Saavedra-Nieves, P. (2020). On systems of quotas from bankruptcy perspective: The sampling estimation of the random arrival rule. *European Journal of Operational Research, 285*(2), 655–669.
Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn, & A. W. Tucker (Eds.), *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307–318). Princeton University Press.
Smith, M., & Alvarez, F. (2021). Identifying mortality factors from Machine Learning using Shapley values–a case of COVID19. *Expert Systems with Applications, 176*, 114832.
Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research, 11*, 1–18.
Štrumbelj, E., & Kononenko, I. (2011). A general method for visualizing and explaining black-box regression models. In A. Dobnikar, U. Lotrič, & B. Šter (Eds.), *Adaptive and natural computing algorithms* (pp. 21–30). Springer.