XAIRE: An ensemble-based methodology for determining the relative importance of variables in regression tasks. Application to a hospital emergency department

A.J. Rivera[b,*], J. Cobo Muñoz[a], M.D. Pérez-Goody[b], B. Sáez de San Pedro[a], F. Charte[b], D. Elizondo[c], C. Rodríguez[a], M.L. Abolafia[a], A. Perea[a], M.J. del Jesus[b]

[a]*Emergency Department. University Hospital of Jaén (Spain)*
[b]*Computer Science Department. University of Jaén (Spain)*
[c]*Department of Computer Science and Informatics, De Montfort University (UK)*

Abstract

Nowadays it is increasingly important in many applications to understand how different factors influence a variable of interest in a predictive modeling process. This task becomes particularly important in the context of Explainable Artificial Intelligence. Knowing the relative impact of each variable on the output allows us to acquire more information about the problem and about the output provided by a model.

This paper proposes a new methodology, XAIRE, that determines the relative importance of input variables in a prediction environment, considering multiple prediction models in order to increase generality and avoid bias inherent in a particular learning algorithm. Concretely, we present an ensemble-based methodology that promotes the aggregation of results from several prediction methods to obtain a relative importance ranking. Also, statistical tests are considered in the methodology in order to reveal significant differences between the relative importance of the predictor variables. As a case study, XAIRE is applied to the arrival of patients in a Hospital Emergency Department, which has resulted in one of the largest sets of different predictor variables in the literature. Results show the extracted knowledge related to the relative importance of the predictors involved in the case study.

Keywords: Relative importance of variables, Hospital emergency department, Time series forecasting, Regression analysis, Explainable artificial intelligence

1. Introduction

Regression analysis (Chatterjee & Hadi, 2012) is an area dedicated to studying relationships among variables in a predictive context. Regression methods, developed in this area, have achieved successful results in multiple fields (Chatterjee & Hadi, 2012) such as financial, industrial, medicine, and energy fields, etc. In fact, multiple regression is perhaps the most widely used tool for data analysis according to (Tonidandel & LeBreton, 2011).

A model for predicting a dependent variable from a set of predictor variables (or predictors) is obtained when a regression method is applied. For explanatory purposes, an important question in the field (Tonidandel & LeBreton, 2011), is to what extent a predictor or exogenous variable influences the output or predicted variable. This question is answered by calculating the so-called relative importance of the predictor variable. Currently, diverse techniques are used to obtain this parameter.

Statistical methods (James et al., 2017) are classically used in regression analysis. Over the last few years, machine learning methods (Maimon & Rokach, 2010) have been successfully used for regression tasks. Determining the relative importance of predictor variables with machine learning regression algorithms is an open research field. Some studies, such as (Bi, 2012), focus on the importance of taking these kinds of algorithms into account in order to solve this problem. Furthermore, interest in this field is growing due to an increasing demand in the area of Explainable Machine Learning or Explainable Artificial Intelligence (XAI) to understand the inner workings of a specific model (Roscher et al., 2020). A key point is the relative importance of the predictor variables in the regression process. This knowledge endorses the scientific value of the research, deducing causal relationships from the data of the input-output problem, or even reaching new scientific conclusions.

Different methods (Kuhn, 2008) have been proposed to obtain the relative importance of predictor variables for different (classical or machine learning) regression algorithms. However, each regression method usually returns a different influence order for the predictors, which depends on the learning characteristics and bias of the learning method. This fact poses a challenge that needs to be addressed: how to process or analyze these different variable sets to obtain a relative order of importance of the global predictors (that are not specific to a learning method).

*Corresponding author.
Email addresses: arivera@ujaen.es (A.J. Rivera), jocomu001@gmail.com (J. Cobo Muñoz), lperez@ujaen.es (M.D. Pérez-Goody), blanca.saenzdesanpedro@gmail.com (B. Sáez de San Pedro), fcharte@ujaen.es (F. Charte), elizondo@dmu.ac.uk (D. Elizondo), drdoncesareo@gmail.com (C. Rodríguez), abolafiaml@gmail.com (M.L. Abolafia), alvaroperea.65@gmail.com (A. Perea), mjjesus@ujaen.es (M.J. del Jesus)

The XAIRE (eXplainable Artificial Intelligence Regression Ensemble) methodology is proposed to address this issue. XAIRE, based on ensemble techniques (Zhou, 2012), aggregates the relative importance of each predictor variable obtained by each regression method, obtaining a general ranking. Additionally, significant differences, among these influences, are calculated by applying statistical tests to the predictor variables.

As a case study, the importance of predictors of the time series of Emergency Department (ED) arrivals at the University Hospital of Jaén has been analyzed. Hospital Emergency Departments (HED) are considered key to the functioning of the National Health System (Schuur et al., 2013) of any country. ED saturation represents a very important problem within the entire hospital system (Fernandes et al., 2020). The lack of foresight in the fluctuations of ED demand (Kadri et al., 2014) has been identified as one of the causes of ED overcrowding. The saturation of this service can lead to delays in patient care and hospitalization, which can prove to be fatal.

The rest of the paper is organized as follows: Section 2 discusses the usefulness of establishing the relative importance of predictor variables and details different approaches within this area. Section 3 describes XAIRE. Section 4 explains the problem of arrivals at the ED of the University Hospital of Jaén. Section 5 shows and analyses the results obtained after applying XAIRE. In Section 6 the performance of the methodology is analyzed by comparing it with others methods in the same field. Section 7 outlines the main conclusions obtained.

2. Using relative importance of predictor variables to improve the explainability of regression models

As has been said, it is increasingly important to understand the influence that different factors have on an interesting variable and explain how to determine the relative importance of the predictor variables handled in a modeling process.

XAI encompasses a set of techniques 1) that produce more explainable models while maintaining their efficiency, as well as 2) allowing humans to understand, and therefore trust AI techniques (Barredo et al., 2020). In a broader sense, the ultimate aim of explainability (Roscher et al., 2020) is to elucidate the decision rationale of a model. A system is interpretable when a user can understand how the input variables influence the output.

Interest in this discipline is growing rapidly due to the significant presence that AI has and will have in society as whole, and in areas such as industry, economics or automotive among others (Tjoa & Guan, 2021). Obtaining the operation roots of a model is important in many fields because of the importance of the decisions made in these areas, specifically in medicine, as it can increase confidence in the decisions made (Ploug & Holm, 2020; Pennisi, 2021).

One of the most important lines of research in XAI is determining how predictors influence a problem's output variable (Barredo et al., 2020). A literature review of methods for this task is provided in the following subsections.

2.1. *Classical methods for determining the relative importance of variables*

Trying to understand how a model works is not a new task. Multiple regression (Tonidandel & LeBreton, 2011) is one of the most widely used statistical tools for data analysis. It has two different objectives: prediction and inference. Prediction uses a regression equation to obtain the output values from a set of predictor variables. Inference determines the influence of each predictor variable on the output. In (Johnson & LeBreton, 2004), the relative importance for one predictor variable is defined as the contribution that this variable makes to the prediction of the output variable both by itself and together with the other predictor variables.

There are multiple ways to calculate the relative importance of predictor variables in the context of predictive models. Some of the most noteworthy are (Bi, 2012; Tonidandel & LeBreton, 2011): to use correlation coefficients or regression weights as measures, to obtain the variance decomposition based on the average over orderings, or to use machine learning methods that work recursively by partitioning the input space of a problem.

One of the most recommended proposals (Bi, 2012; Johnson & LeBreton, 2004) is the one described in Johnson (Johnson, 2000). Johnson's method involves performing the regression of the output variable on the orthogonalized predictor variables by obtaining β coefficients. In addition, λ coefficients are produced from the regression equations of each predictor variable on the orthogonalized predictor variables. The important of the predictor variable is then drawn with the coefficients β and λ.

In any case, these individual proposals may yield a list of variables of different relative importance, which depends on the heuristics and bias of the learning method used, and the values of their parameters. To overcome these drawbacks, a possible solution is to aggregate the relative importance of the variables generated using different methods. This aggregation is the basis of the ensemble-based techniques.

2.2. *Variable selection ensembles for regression*

Variable Selection Ensembles (VSEs) (Xin & Zhu, 2012) are a type of methodology that determines the relative importance of predictors (Beyene et al., 2009) using variable selection lists returned by base methods. VSEs comprise the following phases: generating the variable groups, determining the importance of each variable in the groups, calculating the final importance of each variable by means of some aggregation strategy and, finally, choosing the selected variables. These phases are explained as follows:

1. In the first phase of a VSE algorithm, the sets of variables must be generated. A very common alternative is to use classical selection methods (Zhang et al., 2019b), for example Lasso with or without subsampling, to obtain sets of important variables. Another option is to use stochastic, forward-backward (Xin & Zhu, 2012) or genetic algorithms (Zhu & Chipman, 2006)(PGA), and to use some type of information criterion, Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), to determine the quality of the selected variables.

2. Next, the importance of each variable in the groups selected in the previous step is usually determined. Most methods will assign a value of 1 if a variable is selected, or 0 otherwise (Xin & Zhu, 2012; Zhang et al., 2019a). Occasionally, the weighting calculated by the model for the corresponding variable in the regression equation is also used (Ye et al., 2018).
3. In order to calculate the final influence of each variable, the average of the importance that each variable has obtained in the previous step is usually calculated. In the last step, variables with an importance value above a certain threshold are selected.

There are different proposals based on the inclusion of variations of the phases described above. Well known representative alternatives in the literature are:

- (Xin & Zhu, 2012) presents the ST2E method, an ensemble of their proposed ST2 algorithm. ST2 uses the forward/backward paradigm to add or remove groups of variables, whose size is defined stochastically. Each candidate group is evaluated, and it is added to or removed from the model according to the behavior of the objective function (typically AIC). The importance of the variables is determined by assigning 1 if the variable belongs to the selected group, and 0 otherwise. Finally, the importance of each variable is determined by averaging these values. The variables selected will be those whose importance value exceeds a given threshold. The experimentation included in the paper only considers datasets with a low number of predictors and linear regression methods.

- (Ye et al., 2018) defines the SOIL methodology that builds an ensemble of models generated using base methods such as Lasso, Adaptive Lasso, SCAD (Smoothly Clipped Absolute Deviation) or MCP (Minimax Concave Penalty). The importance of a variable is calculated as the accumulated sum of weights that each model assigns to this variable. These weights are calculated based on the coefficients obtained in the equations of the linear regression equation determined by the models. This operation mode restricts the use of this methodology to methods that obtain this linear equation.

- In (Zhang et al., 2019a) the SSLasso algorithm is presented. It is a modification of the StabSel method (Meinshausen & Bühlmann, 2010), a variable selection algorithm based on subsampling. SSLasso operates by estimating the frequency with which each variable is selected (1 = selected, 0 = not selected) after repeatedly applying Lasso. This frequency will determine the importance of the variable. Additionally, a threshold is set that will define the variables to be selected. One drawback of this algorithm is the internal use of a single regression method.

- (Zhang et al., 2021) presents the DPP-VSE algorithm that uses the DPP (Determinantal Point Processes) technique to determine the number of variables to be selected in each group. Then, the L2Boosting algorithm is used to choose the variables themselves. This process is repeated multiple times, adding 1 to the importance of the variables if they are selected, or 0 if they are not. The final importance of the variables will be obtained as the average of these values (0 or 1). As the authors acknowledge, in high-dimensional environments, it is inefficient to apply DPP-VSE directly because large sets of variables will be assigned to a very low or zero probability. In addition, a high computational cost would be required to build the ensemble.

Therefore, it can be stated that most classical techniques tend to obtain the relative importance of the predictor variables in a process that is often biased by the heuristic and the tendency of the specific learning algorithm being considered. VSE techniques have been proposed to address this, however, these techniques still present some drawbacks, such as:

- They use either a classical feature selection method or a limited and not very diverse set of regression methods (or perhaps even only one). This strategy does not eliminate the aforementioned bias and does not consider sufficient variability to determine the overall impact of the predictors regardless of the method used to finally predict the output variable.

- In addition, VSE methodologies are based on variable selections returned by classical feature selection or regression methods. These variable selections are of type zero or one, which polarizes the results to groups of variables that are always or never selected. These groups of variables do not differentiate between the levels of influence of variables in the output prediction.

- In any case, VSE methodologies offer a list of variable importance in which no statistical differences can be established between the influence of these variables.

Taking into account these drawbacks it is of interest to examine heuristics in order to tackle the bias on influence of predictors resulted from different regression algorithms. It is expected the combination of more diverse information will provide the degree of importance or influence of a predictor variable in a more global and generalized way.

3. XAIRE: eXplainable Artificial Intelligence Regression Ensemble

The objective of XAIRE is to determine the global importance of the predictor variables on the output in a regression process.

Normally, the relative importance depends on the method used to obtain it. Thus, the influence variable list extracted from a regression method represents how this method uses the predictor variables to obtain output values. To mitigate the bias yielded by the application of a single method, XAIRE proposes how to integrate the results on influence of predictors provided by different regression methods. These results are aggregated using an ensemble-based technique in order to obtain a single and global relative importance variable list. Furthermore, the possible significant differences present in the relative importance of the variables are detected.

The main steps of XAIRE methodology are as follows:

A. Preparation and exploratory analysis of the data
B. Selection of the prediction methods
C. Execution of the prediction methods
D. Obtaining the global importance of predictors
E. Determining significant differences in the influence of the predictors on output

Figure 1 depicts their interconnection.

3.1. Detailed description of XAIRE

The main steps of the proposed methodology are detailed below:

A. Preparation and exploratory analysis of the data
In this step, the data must be prepared. To avoid any bias, it is recommended having sufficient and also significant amount of data in order to make the required prediction. Then, a classical exploratory analysis is performed to extract the main data characteristics and determine whether any pre-processing of the data is necessary. Next, the data has to be transformed into an input prediction method format. Finally, for prediction purposes, data are divided into training and test datasets.

B. Selection of prediction methods
The prediction methods used to calculate the relative importance of the predictor variables ares now selected. Any prediction method proposed in the literature that provides information about the variables' importance can be chosen. Better results will be achieved by considering a broad set of different types of base algorithms. In this manner the diversity of the information considered in the combination would be increased. In addition, the quality of these methods must be taken into account, avoiding, for instance, simple correlation methods or basic linear regression methods (Bi, 2012).

C. Execution of prediction methods
The results obtained from the prediction methods are analyzed in order to find and eliminate methods with a poor quality performance measure (e.g. those whose MAPE is 50% higher than the median MAPE). It should be noted that the influence of the predictors on the output obtained by these models is probably not reliable.

Then, the relative importance returned by each prediction method is collected for each variable.

D. Obtaining the global importance of predictors
In this step, a ranking of variables is obtained for each method based on the input variables' influence on the output variable. Then, the average position (Equation 1) of predictors in the individual rankings are calculated. The global influence is computed by ordering the previously calculated positions from lowest to highest. This ranking represents the relative importance of each input variable. Other basic statistics, such as standard deviation (SD) (Equation 2) or coefficient of variation (CV) (Equation 3), are computed in order to draw conclusions. For example, a low SD implies that regression methods agree on the relative importance of a predictor, however, a high SD implies a disagreement in it.

$$Mean(p_k) = \frac{\sum_{i=1}^{m} p_{ki}}{m} \tag{1}$$

$$SD(p_k) = \frac{\sum_{i=1}^{m}(Mean(p_k) - p_{ki})^2}{m-1} \tag{2}$$

$$CV(p_k) = \frac{Mean(p_k)}{SD(p_k)} \tag{3}$$

E. Determining significant differences in the influence of the predictors on the output
By applying statistical tests, able to detect significant differences between the variables, additional knowledge is obtained in this step. For this task, the Friedman test with pair-wise comparisons is used. The Friedman test (Demšar, 2006) is best suited when data is dependent, non-parametric and conclusions about the order need to be drawn. To obtain pair-wise comparisons post-hoc raw p-values are computed. For the sake of reproducibility, the *friedmanPost* function of the *scmamp* R package (Calvo & Santafé, 2016) is recommended, as this step provides additional comparative information on the global influence of the variables.

For the sake of reproducibility, the XAIRE code is available in the footnote link[1].

3.2. Differences between XAIRE and VSE methods

XAIRE is related to the proposals in the VSEs area. However, some key differences have been identified:

- To begin with, the objective of XAIRE is not to select a subset of variables, as other methods do, but to determine their individual influence. It should be kept in mind that the importance of the variables may depend on the regression model chosen. Therefore, there is no single ordering of the variables. In this sense, an interesting challenge is to find out the importance given to the variables when using a certain set of regression methods. This approach is in line with new areas of research in the field of XAI (Barredo et al., 2020).

- VSEs (Ye et al., 2018) and (Zhang et al., 2019a) use a fixed set of regression methods that internally select variables, such as Lasso, SCAD, MCP, etc. In fact, the proposals reviewed usually only consider classical regression methods. It should be noted that each learning method has a certain bias. In XAIRE any kind of regression method is allowed. In the case study presented here, machine learning methods such as support vector machines or neural networks are introduced in the ensemble, and results obtained by other ensembles are also included.

[1] https://ajriverar.github.io/XAIRE/XAIREcode.R

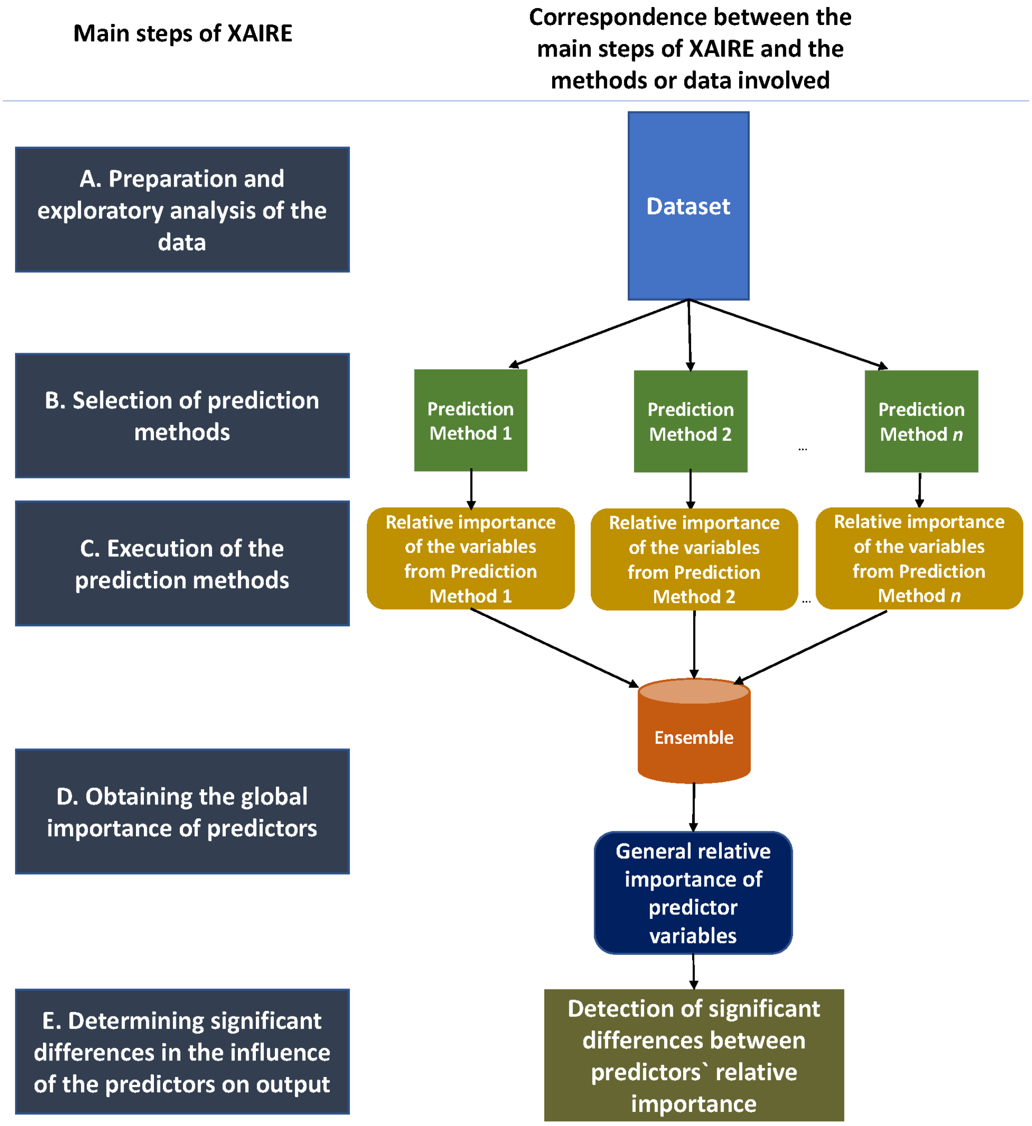

Figure 1: XAIRE methodology for extracting the relative importance of predictor variables

XAIRE determines the importance of variables in a more global way as it allows the combination of different methods.

- With respect to the variable ensemble generation phase, according to (Zhang et al., 2019a), the use of techniques such as forward-backward (Xin & Zhu, 2012) could have robustness problems and be unstable when there are a considerable number of variables. XAIRE does not present these problems when dealing with a high number of predictors, such as the one considered in the case study.

- The base methods belonging to the VSE methodologies assign a discrete importance value, 0 or 1, depending on whether the variable has been selected or not (Xin & Zhu, 2012; Zhu & Chipman, 2006; Zhang et al., 2021). This mode of operation may lack the necessary precision in variables that are selected either regularly or rarely, since they both have the same final importance value and no difference can be established between them. According to (Zhang et al., 2021), this fact could be aggravated in high-dimensionality environments. In XAIRE, this problem is avoided by assigning a real ranking position in the interval [0, 1] to every variable involved in the regression process.

- In addition, XAIRE provides statistically significant differences in the influence of the predictors on the output variable. However, the VSE methods reviewed as well as, in general, the traditional methods that determine the importance of variables, only return a list of variables with associated importance.

3.3. Limitations

In order to obtain a global and unbiased ranking of the importance of the variables involved in a regression process, the use of a broad and diverse range of regression methods is recommended. The less this recommendation is followed, the less global and more biased the list returned by the proposed set will be.

Another limitation in achieving good results has to do with the characteristics of the problem data used. For instance, if in the first step (exploratory analysis) of the XAIRE methodology a high random component is detected, the prediction task of the methods used will be more difficult. This fact is reflected in the moderate R^2 values returned by the methods. Logically, this fact can also influence the study of the importance of predictor variables in a prediction process.

4. Case study

XAIRE is applied to the regression analysis of arrivals at an ED. The service offered by an ED must meet a series of functional, structural and organizational requirements to ensure quality emergency care. The characteristics of the service offered by an ED include: the significant and variable demand to which it is subjected, its influence on the operation of the hospital in which it is located, or that is the main admission point for patients in the corresponding hospital.

We performed a descriptive observational study of the daily time series of patient arrivals at the ED of the University Hospital in Jaén (comprising two centers) from June 1, 2015 to May 31, 2019. The area of influence of this hospital includes two districts, which correspond to about 300,000 people. The number of annual visits to the ED is about 60,000 and the number of admissions is around 18,000 patients per year. Bearing in mind that the total number of hospital admissions is about 26,000 patients per year, this implies that emergency admissions account for 70% of all hospital admissions.

In the experimentation phase of the study, a period of training has been defined between June 1, 2015 and May 31, 2018, and a period of model testing from June 1, 2018 and May 31, 2019. The prediction horizon set out in this paper is 1 day, i.e., emergency room arrival is predicted one day in advance.

A large and varied set of predictor variables has been considered in this study, the largest to date to the authors' knowledge. The predictor variables (taken in time series mode) considered to analyze the influence on the prognosis of the ED series were selected according to the literature on this subject (Gul & Celik, 2018) and the expert knowledge of medical doctors involved in the authorship of this paper. Table 1 depicts the exogenous variables considered, including the variable of ED arrivals (variable to be predicted), as it is also considered to be an input variable. Four groups of predictor variables were identified, specifically: `Calendar` variables, `Meteorological` variables, `Air Quality` variables and `Pollen Concentration` variables. For each one of these, the temporal displacement (lags) used in the prediction of the arrival variable is given. Thus, when predicting tomorrow (lag 1), both today's lag (lag 0) and the remaining values of the variables from the last two weeks (lags from -1 to -13) will be used as input for the models. As we already have the value of the `Calendar` variables for the day to be predicted, this value will be used directly (lag 1). In this manner, a total of 217 variables were used as the input set for the prediction methods.

5. Experimentation and analysis of results

The XAIRE methodology is applied in this section, and the following subsections correspond to the steps described in Section 3.

5.1. Preparation and exploratory analysis of the data

The time series of patient arrivals at the ED of the University Hospital of Jaén, from June 1, 2015 to May 31, 2019, is shown in Figure 2.

The results of a basic exploratory analysis are described in Table 2. As can be seen, the series has a high standard deviation and coefficient of variation. This implies important variations in the arrival series.

Figure 3 displays the additive decomposition of the time series into its trend, seasonal and random components. As can be

Variable name	Variable lags
Arrivals in ED (output)	1
Arrivals in ED (input)	-13, -12, . . . , -1, 0
Calendar	
WeekDay	1
Month	1
Day (of the month)	1
Working day	1
Saturday	1
Holiday	1
PostHoliday	1
Meteorological	
Maximum Temperature	-13, -12, . . . , -1, 0
Average Temperature	-13, -12, . . . , -1, 0
Minimum Temperature	-13, -12, . . . , -1, 0
Air quality (maximum)	
Particles	-13, -12, . . . , -1, 0
Nitrogen Dioxide	-13, -12, . . . , -1, 0
Carbon Monoxide	-13, -12, . . . , -1, 0
Ozone	-13, -12, . . . , -1, 0
Pollen (concentration)	
Cupress	-13, -12, . . . , -1, 0
Chenopodium	-13, -12, . . . , -1, 0
Olea europaea	-13, -12, . . . , -1, 0
Plantago	-13, -12, . . . , -1, 0
Platanus	-13, -12, . . . , -1, 0
Poaceae	-13, -12, . . . , -1, 0
Urticaceae	-13, -12, . . . , -1, 0

Table 1: Predictor variables considered in the study of arrivals at the ED of the University Hospital of Jaén

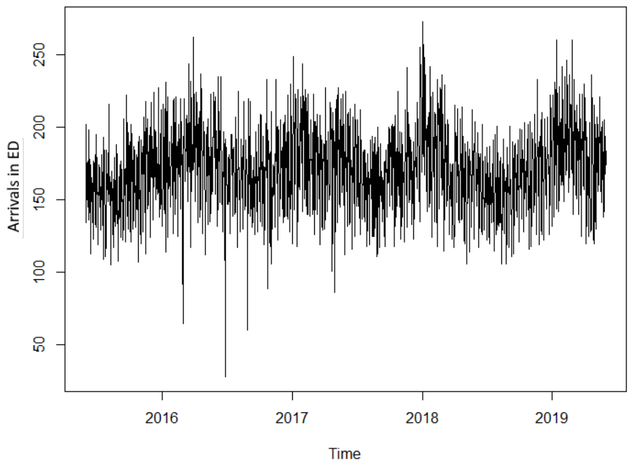

Figure 2: Time series of patient arrivals at the ED

observed, the values of the trend component oscillate within a narrow range between 167 and 174. This is corroborated by the Dickey-Fuller stationarity test (Box et al., 2008), which returns a p-value = 0.01, confirming the absence of trend and that the series maintains a constant variance.

In terms of seasonality, the graph displays an increase in the number of ED arrivals around January and February, and a decrease in the months of July and August. Specifically, the mean value of the seasonal component of the series of ED arrivals in January and February is 13,645 patients, while in July and August this seasonal component has an average value of -12,012 patients. To further study the seasonality of the series, a t-test is applied (see Table 3), confirming that there are significant differences between the averages for the periods of January-February (winter) and July-August (summer) in the years 2016, 2017 and 2018.

Measure	Value
Min.	28
1st Quantile	149
Median	172
Mean	170.8
3rd Quantile	191
Maximum	273
Skewness	-0.035
Standard Deviation	29.9
Coefficient of Variation	0.175

Table 2: Exploratory time-series data for ED arrivals at the University Hospital of Jaén

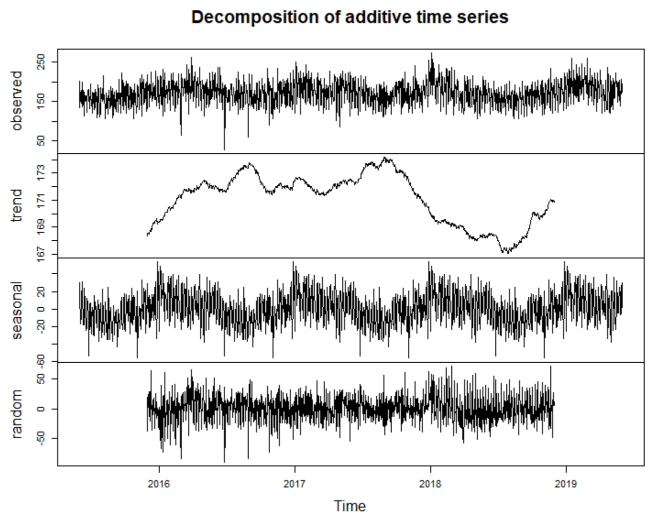

Figure 3: Additive decomposition of series components

Figure 3 also displays that the series has a high random component. A study of the series of this component shows that 50% of the values are outside the interquartile range defined between the first and third quartiles. The average of these values

Year	Mean (JanuaryFebruary)	Mean (JulyAugust)	p-value
2016	175.702	164.177	0.02842
2017	186.474	159.226	5.299e07
2018	192.175	154.484	4.8e-10

Table 3: t-test for the mean of the ED arrivals between the periods of January-February and July-August

is 29.23 and represents 17% with respect to the average of the series of arrivals, with a difference of up to 90 patients, which is 53% higher than the previously mentioned average.

Figure 4 depicts the autocorrelation function of the ED arrival series, showing the correlation between a specific ED arrival value and the previous days' values. In this case, a correlation of 0.653 is observed between the arrival value and the arrival value for the same day one week earlier. This indicates the importance of the day of the week that patients arrived at the ED, as displayed in Figure 5. In this graph the boundaries of the box represent the number of arrivals that are between the first and third quartile. The central bar of the box represents the median. In this sense, it is clear that the number of ED arrivals is lower on weekends, while Monday is the day with the highest number of arrivals. The chi-square analysis test on the ED arrival and day of the week variables yields a p-value of less than 2.2e-16. Thus, the alternative hypothesis that the days of the week are not independent of the ED arrival variable was fulfilled.

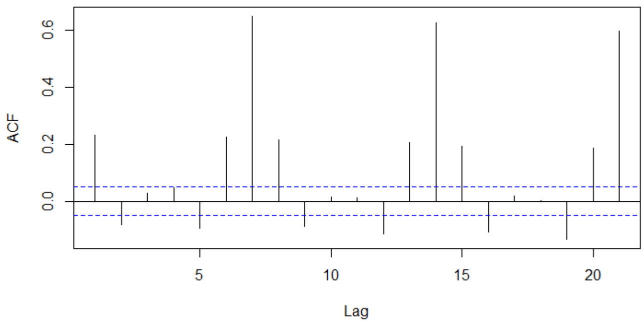

Figure 4: Emergency series autocorrelation function

5.2. *Selection of the prediction methods*

The methods used to predict the arrivals at an ED are listed in Table 4. For the sake of diversity, the selected methods belong to different and representative areas of classical mathematical regression (Chatterjee & Hadi, 2012) and machine learning (Maimon & Rokach, 2010) available in the *caret* R package (Kuhn, 2008) .

With the aim of using a complete set of time series forecasting methods, statistic time series prediction methods (Box et al., 2008) such as ARIMA and ARIMAX, from the *forecast* R package (Hyndman & Khandakar, 2008), are also included in the experimentation. However, since ARIMA does not support

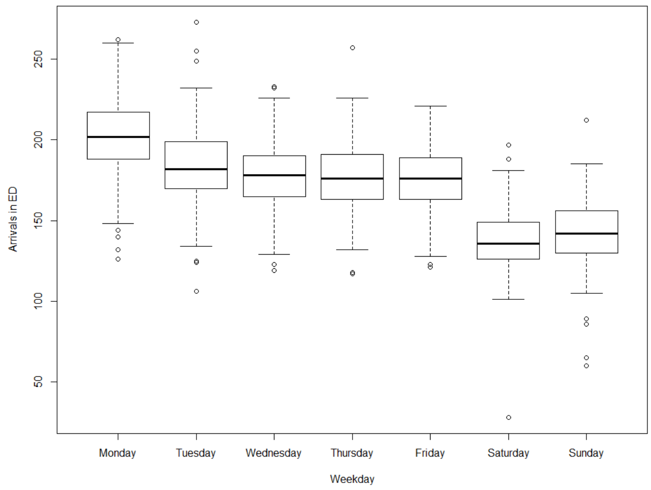

Figure 5: Analysis of emergency room arrivals by day of the week

Mathematical methods
Bayesian Ridge Regression (BRIDGE) (Murphy, 2012)
Least Absolute Shrinkage and Selection Operator (LASSO) (Murphy, 2012)
Machine learning methods
Conditional Random Forest(Hothorn et al., 2006)
Monotone Multi-Layer Perceptron Neural Network
Quantile Random Forest (Meinshausen, 2006)
Random Forest (Breiman, 2001)
Support Vector Machines with Polynomial Kernel (A. et al., 2011)
Hibrydation methods
Bayesian Additive Regression Trees (BARTMachine) (Kapelner & Bleich, 2016)
Bayesian Ridge Regression Model Averaged (BLASSOAveraged) (Murphy, 2012)
Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016)

Table 4: Prediction methods considered in the study to be included in the XAIRE methodology

exogenous predictors and ARIMAX does not offer reliable information about the relative importance of the exogenous predictors, they do not include the "extraction of relative importance" step. The default parameters defined in the corresponding R packages have been used for all models.

The quality measures used to evaluate the models obtained are: the Root of the Mean Square Error (RMSE, Equation 4), the Mean Absolute Percentage Error (MAPE, Equation 5) and the Mean Absolute Error (MAE, Equation 6).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{4}$$

$$MAPE = \frac{\sum_{i=1}^{n} |\frac{\hat{y}_i - y_i}{\hat{y}_i}|}{n} \tag{5}$$

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{6}$$

where y_i is the real value and $\hat{y}_i$ is the predicted one. The coefficient of determination (R^2) is also used, indicating what proportion of the total variation in the variable to be predicted is explained by the estimated model (Equation 7)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2} \quad (7)$$

where $\bar{y}$ is the mean of the y_i values.

5.3. Execution of the prediction methods

The prediction of the time series is addressed with the methods in Table 4, which will be considered later in the ensemble.

Once the models have been applied to predict the arrival series, the test data set is used to obtain the errors in the predictions shown in Table 5.

	MAPE	MAE	RMSE	R^2
LASSO	7.834	12.926	16.265	0.5410
BLASSOAveraged	7.994	13.182	16.713	0.6412
BRIDGE	8.060	13.292	16.823	0.6421
SVMPoly	8.334	13.722	17.382	0.6230
QRF	8.588	14.157	17.957	0.8880
ARIMA(0,0,5)(0,1,1)	8.621	14.468	19.229	0.6015
RF	8.666	14.196	17.975	0.6600
CForest	8.669	14.215	18.015	0.5690
XGBoost	8.788	14.711	18.783	0.9496
BARTMachine	8.841	14.651	18.458	0.7036
MONMLP	9.459	15.863	20.095	0.7298
ARIMAX(1,1,1)(0,0,1)	9.956	16.644	21.924	0.7675

Table 5: Quality measures for prediction methods. The results are ordered from lowest to highest MAPE error

The prediction methods used have an average daily error range of 12.9 to 16.6 patients, which represents a value between 7.8% and 9.95% of the average number of ED arrivals. The Lasso mathematical regression method obtained the best results, and the mathematical regression-based methods ranked among the top three. The first machine learning regression method is a Support Vector Machine, SVMPoly, which comes fourth in the overall ranking of the methods. The ARIMA time series prediction method is placed in an intermediate position in the table, and depends on the error measure considered, while ARIMAX comes last in the ranking. The poor performance of ARIMAX as compared to other studies (Díaz-Hierro et al., 2012) where it obtained the best results in the comparison, may be due to the higher number of variables considered. Random Forest based methods rank somewhere in the middle. It should be noted from the residual analysis carried out, in general, the methods have a moderate R^2 due to the high random component of the series.

In summary, all the methods used have obtained an error percentage of less than 10% of the mean, which is consistent with other studies (Sudarshan et al., 2021; Wargon et al., 2009), and are considered noteworthy according to (Liu et al., 2016), despite the significant random component of the series.

5.4. Obtaining the global importance of predictors

This subsection displays and analyzes the relative importance of the predictors obtained by the proposed methodology, XAIRE.

Table 6 shows a summary of the ranking of the importance of the 217 external variables that may influence the series of ED arrivals. Specifically, the first 10 positions of this ranking order are shown, followed by the first appearance of a variable belonging to a group of lags. For each variable its ranking, mean, standard deviation and coefficient of variation of its position in the lists of importance are depicted.

Pos	Variable	Mean (Ranking)	SD (Ranking)	CV (Ranking)
1	WeekDay	1.000	0.000	0.000
2	Arrivals -6	2.500	0.850	0.340
3	PostHoliday	3.400	0.966	0.284
4	Arrivals -13	5.300	4.596	0.867
5	Holiday	5.600	0.516	0.092
6	TempMin -9	9.800	2.700	0.275
7	TempMin -13	12.400	4.904	0.395
8	TempMin -11	14.100	4.771	0.338
9	TempMean -13	15.300	7.119	0.465
10	TempMin -6	20.100	13.908	0.692
11	TempMax -10	21.700	2.669	0.123
15	WorkingDay	27.500	43.482	1.581
34	Saturday	47.700	89.231	1.871
35	Month	48.800	25.209	0.517
38	COMax -13	50.900	18.472	0.363
41	O3Max -7	53.800	24.179	0.449
51	NO2Max -6	61.300	14.974	0.244
75	Poaceae 0	78.700	32.891	0.418
80	Urticaceae 0	85.200	19.124	0.224
91	Platanus -2	99.100	32.306	0.326
95	Olea -3	106.400	33.672	0.316
99	MaxPart 0	111.100	24.574	0.221
100	Cupress -5	111.600	33.374	0.299
105	MonthDay	116.000	60.178	0.519
108	Plantago -1	118.700	43.818	0.369
142	Chenopo -9	140.800	24.666	0.175

Table 6: List of important variables returned by XAIRE. The mean, SD and CV of the ranking is offered for each variable

According to the analysis of the ranking the five most influential variables in the prediction of the series are: `WeekDay`, `Arrival -6`, `PostHoliday`, `Arrival 13` and `Holiday`. Basically, it can be said that the arrival of patients at an ED depends on the `Calendar` variables and the values of the same-day arrival variable from previous weeks.

Continuing with the ranking, `Meteorological` variables from the two weeks prior are shown below, which may indicate that extreme temperatures influenced patient ED arrivals two weeks later. Then come the variables related to `Air Quality` that also affected patient ED arrivals two weeks later. The `Pollen Concentration` variables rank last, having immediately affected patient ED arrivals, between 0 and 3 days after there was a high concentration of these particles.

In this ranking it should be noted that predictor variables such as working days or even Saturdays are not among the top positions of ranking and have a high standard deviation. This implies that some prediction methods consider these variables to be more important than other. The minor relative importance attributed to these variables may be due to the fact that some prediction methods extract the corresponding information from other `Calendar` variables. In any case, this fact supports the robustness of the (ensemble-based) methodology used and the quality of the information extracted from it, since if this relative importance is observed in only one or two methods, erroneous or confusing conclusions may be deduced.

5.5. Determining significant differences in the influence of the predictors on output

Table 7 depicts the p-values of the Friedman test in pairs, calculated for the variables with the highest importance in the ranking, and considering each group of lags. Since a symmetrical matrix is being represented, for reasons of clarity and space, the part above the main diagonal is not shown. The total number of variables is 22, and comparisons are made between each pair of variables. For example, between the variable `Arrival -6` (column number 2) and `Max_NO2 -6` (row number 13) a p-value of 0.036 was obtained, indicating that there is a significant difference between these variables. Given the size of the table, no comparisons appear from column 15 onwards. These columns correspond to p-values higher than 0.05, so there are no significant differences between the variables compared. For a better understanding, Figure 6 displays a heat map for all of Table 7.

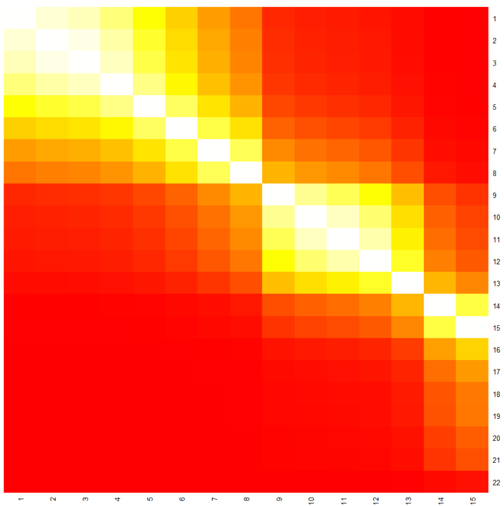

Figure 6: Heat map for the p-values of the Friedman test. The numbers represent the variables described in the caption in Table 7

Regarding Table 7, an analysis of the significant differences between the relative importance of the predictor variables is conducted. Thus, between the most important predictor variables (`Calendar` and `Arrivals`) and the less important predictor variables (`Air Quality` and `Pollen Concentration`) significant differences appear. Significant differences are also found between `Air Quality` and `Pollen Concentration` categories. The low importance of the `Pollen Concentration` variables may be due to the fact that these variables have a temporal or seasonal effect on arrivals at an ED, although they are considered all year round.

The results obtained after applying XAIRE have been analyzed by the medical doctors involved in the authorship of this paper. The knowledge extracted is of significant value and coincides with the information available to them/their expertise, confirming hypotheses already put forward and suggesting new trends to be studied.

6. Discussion

The results obtained by XAIRE were compared with those achieved by other representative methodologies to determining the influence of the variables in regression processes. Specifically, XAIRE was compared with the VSE type methodology SOIL (Ye et al., 2018) and with the Johnson technique (Johnson, 2000). A brief description of these techniques can be found in Section 2.

Table 8 depicts the 30 most influential variables obtained using the methodologies discussed above. The first column represents the global rank number of the variables returned by these methods. Two columns are shown for each variable in XAIRE: the name and the relative importance value. For each variable in SOIL or Johnson, the columns displayed are: the variable name, the relative importance value and the difference between the rank that variable holds in the corresponding methodology (SOIL or Johnson) and in XAIRE. The complete comparison table can be found in the footnote link[2].

It should be noted that these results refer to the influence assigned by a proposal to the predictors involved in a regression process. Therefore, they will depend on the methodology used and more specifically on the base regression methods employed.

The analysis of the results is structured as follows: first the study focuses on the top important variables and then the complete list of variables returned by the methodologies are analyzed.

The first variables are considered to be the most influential. SOIL is not able to distinguish, considering the level of influence, between the variables that hold the top 6 positions in the ranking. In these top positions can be observed that the methodologies coincide in variables such as `WeekDay`, `Arrivals -6`, `PostHoliday`, `Holiday`. `Arrivals -13` also occupies these first 6 positions in XAIRE and Johnson, and ranks 7th for SOIL. For XAIRE and for SOIL the variable `TempMin -9` also ranks in the top 6, but for the SOIL method, it ranks 24th. The Saturday variable ranks in the top 6 positions with the SOIL and Johnson methods; however, it ranks 34th with XAIRE. This is because XAIRE has been able to extract this information from the variable `WeekDay`. A similar explanation holds for the variable `WorkDay`, which is placed 7th by Johnson, 10th by SOIL and 15th by XAIRE.

For the sake of facilitating the analysis of the entire list, Table 9 shows an analysis considering the position in the ranking

[2]https://ajriverar.github.io/XAIRE/MethodComparison.xlsx

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	.957	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	.932	.974	-	-	-	-	-	-	-	-	-	-	-	-	-
4	.870	.912	.938	-	-	-	-	-	-	-	-	-	-	-	-
5	.754	.795	.820	.881	-	-	-	-	-	-	-	-	-	-	-
6	.611	.648	.672	.730	.845	-	-	-	-	-	-	-	-	-	-
7	.461	.494	.515	.566	.672	.820	-	-	-	-	-	-	-	-	-
8	.345	.373	.391	.435	.528	.664	.836	-	-	-	-	-	-	-	-
9	.117	.130	.138	.161	.210	.290	.407	.533	-	-	-	-	-	-	-
10	.089	.099	.106	.124	.165	.233	.334	.448	.892	-	-	-	-	-	-
11	.076	.085	.091	.107	.143	.205	.298	.405	.834	.940	-	-	-	-	-
12	.060	.068	.073	.086	.117	.170	.253	.349	.754	.859	.918	-	-	-	-
13	.032	.036	.039	.047	.067	.101	.158	.229	.562	.656	.711	.789	-	-	-
14	.006	.007	.007	.009	.014	.024	.042	.068	.230	.287	.322	.375	.535	-	-
15	.003	.003	.004	.005	.007	.013	.024	.040	.152	.195	.222	.263	.395	.817	-
16	.000	.001	.001	.001	.001	.003	.006	.011	.054	.073	.086	.107	.178	.468	.621
17	.000	.000	.000	.000	.001	.001	.003	.005	.029	.040	.048	.061	.108	.324	.450
18	.000	.000	.000	.000	.000	.001	.001	.003	.019	.027	.032	.041	.076	.249	.356
19	.000	.000	.000	.000	.000	.001	.001	.003	.018	.025	.031	.040	.073	.241	.347
20	.000	.000	.000	.000	.000	.000	.001	.002	.011	.017	.020	.027	.051	.184	.273
21	.000	.000	.000	.000	.000	.000	.001	.001	.009	.013	.016	.021	.041	.154	.233
22	.000	.000	.000	.000	.000	.000	.000	.000	.001	.001	.001	.002	.005	.027	.048

Table 7: p-values obtained the Friedman test to detect significant differences. (1: WeekDay, 2: Arrivals -6, 3: PostHoliday, 4: Holiday, 5: TempMin -9, 6: TempMean -13, 7: TempMax -10, 8: WorkingDay, 9: Saturday, 10: Month, 11: COMax -13, 12: O3Max -7, 13: NO2Max -6, 14: Poaceae 0, 15: Urticaceae 0, 16: Platanus -2, 17: Olea -3, 18: MaxPart 0, 19: Cupress -5, 20: MonthDay, 21: Plantago -1, 22: Chenopo -9)

	XAIRE		SOIL			JOHNSON		
Ranking	Variable	Importance	Variable	Importance	Differences	Variable	Importance	Differences
1	WeekDay	1	PostHoliday	1	-2	PostHoliday	0.118	-2
2	Arrivals -6	2.5	Holiday	1	-3	WeekDay	0.118	1
3	PostHoliday	3.4	Saturday	1	-31	Arrivals -6	0.101	1
4	Arrivals -13	5.3	WeekDay	1	3	Arrivals -13	0.092	0
5	Holiday	5.6	TempMin -9	1	-1	Holiday	0.084	0
6	TempMin -9	9.8	Arrivals -6	1	4	Saturday	0.071	-28
7	TempMin -13	12.4	Arrivals -13	1	3	WorkDay	0.052	-8
8	TempMin -11	14.1	TempMin -11	0.993	0	Arrivals 0	0.021	-13
9	TempMean -13	15.3	TempMin -6	0.993	-1	Arrivals -5	0.019	-15
10	TempMin -6	20.1	WorkDay	0.993	-5	Arrivals -12	0.018	-18
11	TempMax -10	21.7	Arrivals -3	0.993	-85	Arrivals -7	0.015	-16
12	TempMin -7	23.1	Arrivals -5	0.992	-12	Arrivals -11	0.012	-33
13	TempMax -13	23.2	Arrivals 0	0.564	-8	Arrivals -8	0.010	-36
14	TempMean -7	26.5	TempMin -13	2.79E-06	7	Arrivals -1	0.010	-32
15	WorkDay	27.5	Olea -1	2.67E-06	-96	Arrivals -4	0.010	-28
16	TempMean -6	27.6	Olea -2	2.67E-06	-85	Arrivals -3	0.006	-80
17	TempMin -10	27.7	Poaceae 0	2.97E-08	-58	Max_NO2 -13	0.005	-35
18	TempMean -11	28.7	Arrivals -12	4.68E-09	-10	Month	0.005	-17
19	TempMin -12	30.1	Olea 0	1.61E-11	-97	Arrivals -9	0.004	-166
20	TempMean -10	30.5	TempMin -5	1.61E-11	-11	Arrivals -2	0.004	-114
21	Arrivals 0	30.7	Arrivals -7	7.68E-14	-6	MonthDay	0.004	-84
22	TempMin -8	32.5	MonthDay	3.04E-20	-83	TempMin -9	0.004	16
23	TempMean -9	32.9	Olea -4	3.04E-20	-74	TempMin -11	0.003	15
24	Arrivals -5	33	Olea -6	3.04E-20	-119	Poaceae 0	0.003	-51
25	TempMax -11	36.2	Arrivals -2	1.78E-24	-109	Arrivals -10	0.003	-129
26	TempMean -12	36.8	Arrivals -10	1.78E-24	-128	Max_O3 -2	0.003	-35
27	Arrivals -7	39.9	Urticaceae -5	1.78E-24	-146	TempMin -13	0.003	20
28	Arrivals -12	42.1	Platanus -13	4.89E-32	-171	TempMin -6	0.003	18
29	TempMean -8	43.4	Cupress 0	4.89E-32	-86	Platanus -2	0.003	-62
30	TempMax -12	43.9	Max_O3 -2	4.89E-32	-31	TempMean -9	0.003	7

Table 8: Influence of Predictors as determined by the proposed methodology, XAIRE, and the SOIL (Ye et al., 2018) and Johnson (Johnson, 2000) methods

of predictors by groups (`Arrivals`, `Calendar`, `Meteorological`, `Air Quality` and `Pollen Concentration`). In this Table, the statistics represent the mean and standard deviation of the positions of the variables of the different categories in the list of variables obtained with the three methodologies. From these results the following conclusions can be drawn:

	XAIRE		SOIL		JOHNSON	
Variable Category	Mean	SD	Mean	SD	Mean	SD
Calendar	28.286	36.664	25.714	50.049	8.571	7.807
Meteorological	34.786	20.799	149.119	65.542	49.143	14.919
Arrivals	61.286	57.827	51.857	64.492	12.786	6.066
Air Quality	101.554	42.031	106.946	57.272	114.571	47.455
Pollen Concetration	157.643	38.165	107.092	52.112	152.388	43.170

Table 9: Exploratory analysis of positions by variable categories

- XAIRE ranks `Calendar` variables first, followed by the `Meteorological` category. From the statistical tests carried out in Section 5 it can be deduced that the variable to be predicted had a greater dependence on the day of the week and, therefore, on `Calendar` variables. As a result, this category was ranked first with XAIRE, and the `Meteorological` variables category was ranked second. The medical team agrees on the importance of these variables, which continue to have an influence throughout the year. `Arrivals`-related variables come in third place, which is in line with the knowledge of experts in the field, as there are variables in this category (those for the same day of the week in the previous two weeks) that are very important when predicting ED arrivals. However, some values of emergency arrivals from previous days are not very important, since the variable to be predicted is very dependent on `Calendar` variables. Next comes the `Air Quality` category, although it ranks quite a lot lower. This type of variable may be important for the arrival of patients at the emergency department, since it is influential throughout the year, except during the summer season. Finally, there are variables related to `Pollen Concentration` that, although they influence patient arrivals at the emergency department, are more seasonal.

- For SOIL, the `Calendar` variable category is the most important, as expected, followed by the `Arrival` variables with an average value similar to that of XAIRE. From here the list differs considerably, the `Air quality` and `Pollen Concentration` variables occupy the 3rd and 4th place, with a very similar mean, which indicates uncertainty in the positioning of these variables by the methodology. The `Meteorological` variables rank last, whilst they come second for XAIRE. This result for SOIL is very uncommon, since this type of variable is usually considered important. In fact, this category is among the most used to predict emergency arrivals (Gul & Celik, 2018).

- The most important variable category for Johnson's method is the `Calendar` variables with a value very similar to that of the `Arrival` category. Overall, this method ranked the `Arrival` variables in relatively top positions. However, it cannot be said that any `Arrival` variables are so important in predicting what is going to happen in the next day's arrival prediction. This fact can be deduced from different results obtained by other methods (Gul & Celik, 2018), where the category of `Calendar` variables is the one that most influences this prediction, since depending on the day of the week, there is a significant variability in the number of ED arrivals. Continuing with Johnson's results, the next most important category is `Meteorological` variables, which XAIRE ranks second in terms of importance. Finally, and in agreement with XAIRE, come the `Air Quality` and `Pollen Concentration` categories.

As an additional comparison of the methods, the mean of the absolute value of the difference between the positions of the variables obtained by XAIRE and that returned by the SOIL and Johnson methodologies (Differences columns in Table 8), have been determined. SOIL's mean is 69.24 while Johnson's mean is 32.55, indicating that, in general, Johnson's positions are more similar to those obtained by XAIRE than those obtained by SOIL. It corroborates some differences expressed in Section 3.2, where VSE methodologies (such as SOIL) were compared with XAIRE. Thus, certain VSE methodologies draw information from methods such as Lasso that return groups of variables that work well together to achieve good prediction results. This objective influences the calculation of the true influence of a predictor on the output variable in a regression process.

The general conclusions that can be drawn from the results obtained by the methods are as follows:

- There is a group of 5 variables (`WeekDay`, `Arrivals -6`, `PostHoliday`, `Holiday` and `Arrivals -13`) that the three methodologies rank in their top positions.

- The `Calendar` variable category ranks first on average in all methodologies.

- The importance of the category of `Meteorological` variables, which have year-round effects, has been recognized by XAIRE and by the medical team. However, the SOIL method has placed it last in terms of importance.

- The category of `Arrival` variables contains variables that vary considerably in importance. Variables such as `Arrivals -6` or `Arrivals -13` are ranked higher in a list of variables, however other lags of these variables should rank much lower. This is because of the demonstrated dependence of the data in this problem on `Calendar` variables. The Johnson method encountered the most problems when ranking these variables.

- Finally, the `Air Quality` and `Pollen concentration` categories tend to be ranked, in the same order, at the bottom of the list of variables. This is because they do not have the same influence throughout the year. SOIL had more problems in correctly placing these categories.

In summary, the results of this comparison, according to the medical team and the bibliography in the area, support that XAIRE has achieved a more consistent influence ranking for the predictors than the other methods. The use of disparate regression methods along global aggregation and statistical analysis, lets XAIRE produce a ranking of variable importance less biased to specific regression methods, thus increasing the knowledge on the regression problem.

7. Conclusions

This paper presents a new methodology, XAIRE, to determine the global importance of the predictor variables in a regression analysis context. Concretely, an ensemble-based methodology is proposed in which the relative importance of the different methods is aggregated in order to obtain an overall relative importance ranking. This fact mitigates the subjectivity or bias of applying a single method, since the results of different methods are analyzed and aggregated. Furthermore, the statistical tests integrated into XAIRE reveal significant differences in the overall ranking of the predictor variables.

To complete this methodology, in the first stage different basic exploratory analysis techniques are used to improve the knowledge extraction process. For the sake of the applicability and reproducibility of XAIRE, and without loss of generality, specific methods and environments are used throughout the process.

XAIRE is applied to the time series of ED arrivals at a university hospital that serves a population of 300,000 people. One of the main contributions is that the relative importance of the predictor variables was analyzed, obtaining a ranking headed by the `Calendar` variables and the input variable's own weekly lags. Next come the `Meteorological` variables, mainly those that have a lag of about two weeks. Then come the variables related to `Air Quality`, also with a lag of approximately two weeks. The `Pollen Concentration` variables with lags of zero or very few days are the least important among the significant results found. The results have been validated by a team of medical doctors who participated in the proposed study.

In summary, using XAIRE, explainable knowledge can be obtained at the arrivals stage of an ED prediction process that facilitates the understanding of the way the models operate, and also helps validate the models.

Moreover its comparison with well-known methods from the literature has resulted that XAIRE methodology has returned more congruent results for the case study of patient arrivals at an ED.

Acknowledgements

The research carried out in this study is part of the project "ToSmartEADs: Towards intelligent, explainable and precise extraction of knowledge in complex problems of Data Science" financed by the Ministry of Science, Innovation and Universities with code PID2019-107793GB-I00 / AEI / 10.13039 / 501100011033.

References

A., S., Aliferis, C., Hardin, D., & Guyon, I. (2011). *A gentle introduction to support vector machines in biomedicine: Volume 1: Theory and methods*. World Scientific Publishing Co.

Barredo, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82–115.

Beyene, J., Atenafu, E., Hamid, J., To, T., & Sung, L. (2009). Determining relative importance of variables in developing and validating predictive models. *BMC Med Res Methodol*, *9*.

Bi, J. (2012). A review of statistical methods for determination of relative importance of correlated predictors and identification of drivers of consumer liking. *Journal of Sensory Studies*, *27*, 87–101.

Box, G., Jenkins, G., & Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control*. Wiley-Blackwell.

Breiman, L. (2001). Random forests. *Mach. Learn.*, *45*, 5–32.

Calvo, B., & Santafé, G. (2016). scmamp: Statistical Comparison of Multiple Algorithms in Multiple Problems. *The R Journal*, *8*, 248–256. URL: `https://doi.org/10.32614/RJ-2016-017`. doi:`10.32614/RJ-2016-017`.

Chatterjee, S., & Hadi, A. S. (2012). *Regression Analysis by Example*. Wiley.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (p. 785–794).

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn.*, *7*, 1–30.

Díaz-Hierro, J., Martín, J., Vilches, A., López del Amo, M., Patón, J., & Varo, C. (2012). Evaluación de modelos de series temporales para la previsión de la demanda de emergencias sanitarias. *Emergencias*, *24*, 181–188.

Fernandes, M., Vieira, S. M., Leite, F., Palos, C., Finkelstein, S., & Sousa, J. (2020). Clinical decision support systems for triage in the emergency department using intelligent systems: a review. *Artificial Intelligence in Medicine*, *102*, 101762.

Gul, M., & Celik, E. (2018). An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. *Health Systems*, *9*, 263–284.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*, 651–674.

Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for r. *J Stat Softw*, *26*, 1–22.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics.

Johnson, J. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, *35*, 1–19.

Johnson, J., & LeBreton, J. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods*, *7*, 238–257.

Kadri, F., Harrou, F., Chaabane, S., & Tahon, C. (2014). Time series modelling and forecasting of emergency department overcrowding. *J Med Syst*, *38*, 107–127.

Kapelner, A., & Bleich, J. (2016). bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software*, *70*, 1–40.

Kuhn, M. (2008). Building predictive models in r using the caret package. *J Stat Softw*, *28*, 1–26.

Liu, L., Wang, Q., Wang, J., & Liu, M. (2016). A rolling grey model optimized by particle swarm optimization in economic prediction. *Comput Intell*, *32*, 391–419.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook*. Springer.

Meinshausen, N. (2006). Quantile regression forests. *Mach. Learn. Res.*, *7*, 983–999.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*, 417–473.

Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Pennisi, M. e. a. (2021). An explainable ai system for automated covid-19 assessment and lesion categorization from ct-scans. *Artificial Intelligence in Medicine*, *118*, 102114.

Ploug, T., & Holm, S. (2020). The four dimensions of contestable ai diagnostics - a patient-centric approach to explainable ai. *Artificial Intelligence in Medicine*, *107*, 101901.

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, 42200–42216.

Schuur, J., Hsia, R., H., B., Schull, M., & Pines, J. (2013). Quality measurement in the emergency department: Past and future. *Health Affairs*, *32*, 2129–2138.

Sudarshan, V., Brabrand, M., Range, T., & Wiil, U. (2021). Performance evaluation of emergency department patient arrivals forecasting models by including meteorological and calendar information: A comparative study. *Computers in Biology and Medicine*, *135*.

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, *32*, 4793—4813. doi:`10.1109/tnnls.2020.3027314`.

Tonidandel, S., & LeBreton, J. (2011). Relative importance analysis: A useful supplement to regression analysis. *J Bus Psychol*, *26*, 1–9.

Wargon, M., Guidet, B., Hoang, T., & Hejblum, G. (2009). A systematic review of models for forecasting the number of emergency department visits. *Emerg Med J*, *26*, 395–399.

Xin, L., & Zhu, M. (2012). Stochastic stepwise ensembles for variable selection. *Journal of Computational and Graphical Statistics*, *21*, 275 – 294.

Ye, C., Yang, Y., & Yang, Y. (2018). Sparsity oriented importance learning for high-dimensional linear regression. *Journal of the American Statistical Association*, *113*, 1797 – 1812.

Zhang, C., Liu, J., Wang, G., & Li, G. (2021). Dpp-vse: Constructing a variable selection ensemble by determinantal point processes. *Expert Systems with Applications*, *178*. doi:`10.1016/j.eswa.2021.115025`.

Zhang, C., Wu, Y., & Zhu, M. (2019a). Pruning variable selection ensembles. *Statistical Analysis and Data Mining*, *12*, 168 – 184.

Zhang, Y., Luo, L., Yang, J., Liu, D., Kong, R., & Feng, Y. (2019b). A hybrid arima-svr approach for forecasting emergency patient flow. *J Amb Intel Hum Comp*, *10*, 3315–3323.

Zhou, Z. (2012). *Ensemble Methods.Foundations and Algorithms*. CRC Press.

Zhu, M., & Chipman, H. A. (2006). Darwinian evolution in parallel universes: A parallel genetic algorithm for variable selection. *Technometrics*, *48*, 491–502.