Stochastic Environmental Research and Risk Assessment (2023) 37:4359–4383
https://doi.org/10.1007/s00477-023-02512-2

ORIGINAL PAPER

# Forecasting air pollutants using classification models: a case study in the Bay of Algeciras (Spain)

M. I. Rodríguez-García[1] · M. C. Ribeiro Rodrigues[2,3] · J. González-Enrique[1] · J. J. Ruiz-Aguilar[4] · I. J. Turias[1]

Accepted: 29 June 2023 / Published online: 22 July 2023

## Abstract

The main goal of this work is to obtain reliable predictions of pollutant concentrations related to maritime traffic ($SO_2$, $PM_{10}$, $NO_2$, $NO_X$, and NO) in the Bay of Algeciras, located in Andalusia, the south of Spain. Furthermore, the objective is to predict future air quality levels of the principal maritime traffic-related pollutants in the Bay of Algeciras as a function of the rest of the pollutants, the meteorological variables, and vessel data. In this sense, three scenarios were analysed for comparison, namely Alcornocales Park and the cities of La Línea and Algeciras. A database of hourly records of air pollution immissions, meteorological measurements in the Bay of Algeciras region and a database of maritime traffic in the port of Algeciras during the years 2017 to 2019 were used. A resampling procedure using a five-fold cross-validation procedure to assure the generalisation capabilities of the tested models was designed to compute the pollutant predictions with different classification models and also with artificial neural networks using different numbers of hidden layers and units. This procedure enabled appropriate and reliable multiple comparisons among the tested models and facilitated the selection of a set of top-performing prediction models. The models have been compared using several quality classification indexes such as sensitivity, specificity, accuracy, and precision. The distance ($d_1$) to the perfect classifier (1, 1, 1, 1) was also used as a discriminant feature, which allowed for the selection of the best models. Concerning the number of variables, an analysis was conducted to identify the most relevant ones for each pollutant. This approach aimed to obtain models with fewer inputs, facilitating the design of an optimised monitoring network. These more compact models have proven to be the optimal choice in many cases. The obtained sensitivities in the best models were 0.98 for $SO_2$, 0.97 for $PM_{10}$, 0.82 for $NO_2$ and $NO_X$, and 0.83 for NO. These results demonstrate the potential of the models to forecast air pollution in a port city or a complex scenario and to be used by citizens and authorities to prevent exposure to pollutants and to make decisions concerning air quality.

**Keywords** Air pollution forecasting · Classification models · Minimum redundancy maximun relevance · Maritime traffic · Artificial neural networks

## 1 Introduction

Air pollution is a real threat in today's world according to the World Health Organization (WHO). The European Directive 2008/50/EC regulates several key atmospheric pollutants, including particulate matter (PM), nitrogen dioxide ($NO_2$), sulfur dioxide ($SO_2$), ozone ($O_3$), and carbon monoxide (CO). Vessels-related atmospheric pollutants encompass sulfur dioxide ($SO_2$), nitrogen oxides ($NO_x$), and particulate matter (PM). Exposure to hazardous air pollutants emissions can lead to a range of human health problems, including respiratory disorders, cardiovascular disease, and increased risk of stroke. Manisalidis et al. (2020) showed an overview of the effects of air pollution on human health. A large number of scientific work has demonstrated that particulate matter directly affects human health by reducing air quality (Adeyemi et al. 2022). Air pollution in urban areas is a complex mixture of toxic components that have unhealthy effects on residents, especially sensitive populations such as children and people with cardiac and respiratory diseases (Kolehmainen et al. 2001). From an environmental point of view, the

Extended author information available on the last page of the article conduct of a study on the prediction of air pollutant levels or concentrations (inmisions) is crucial for the protection of human health and the environment. This research pretends to provide valuable insights into the factors influencing the distribution, temporal variations, and potential exposure risks associated with ambient pollutants. Accurate prediction models can be developed to forecast pollution levels, identify pollution hotspots and assess compliance with regulatory standards. These predictive models play a vital role in urban planning, industrial siting and the formulation of effective emission control strategies. By proactively predicting and mitigating high pollution episodes, air pollution forecasting research contributes to protecting public health, reducing environmental impact and promoting sustainable communities (Pope and Dockery 2006; Stieb et al. 2009; Kloog et al. 2013). A review of models to forecast air pollution health outcomes is presented by Oliveri et al. (2017), where different huge cities were compared regarding different pollutants. Besides, in Savouré et al. (2021), Subramaniam et al. (2022), Traina et al. (2022) artificial intelligence is applied to forecast air pollution related to human health.

Different studies show the air pollutants related to vessel traffic (Miola and Ciuffo 2011; Moreno-Gutiérrez et al. 2015; Ekmekçioğlu et al. 2020), and estimate the amount of pollution associated with ships in port areas (Lu et al. 2006; Liu et al. 2014; Fameli et al. 2020). These pollutants are sulphur dioxide ($SO_2$), nitrogen oxides ($NO_x$) and Particulate Matter (PM). Marine pollution is regulated by the International Maritime Organisation (IMO) through the Marine Pollution Protocol (MARPOL). Decarbonisation is the main purpose of the IMO and the reduction of emissions of Greenhouse gas emissions. An energy efficiency index is applied to vessels to indicate their classification (A, B, C, D, E) (MARPOL, Annex VI). The aim of the IMO is to achieve zero emissions by 2050 (IMO 2021). The air pollutants responsible for acid rain are sulphur dioxide ($SO_2$) and nitrogen oxides ($NO_x$) in the atmosphere, which react with water, oxygen, and other chemicals to form sulphuric acid and nitric acid. $NO_2$ is primarily responsible for the formation of smog and acid rain in urban areas, causing both acute and chronic effects (Menezes and Popowicz 2022). These pollutants are emitted from the combustion of fossil fuels in industrial processes, power generation, and transport. The main pollutants associated with port activity are presented in (Yang et al. 2022; Yeh et al. 2022; Mueller et al. 2023).

In recent decades, artificial neural networks (ANNs) have been applied in the field of air quality forecasting in a wide range of literature (Kukkonen et al. 2003; Fernando et al. 2012; Hu et al. 2021; Muruganandam and Arumugam 2023). Numerous studies have been developed using artificial intelligence (AI) and machine learning techniques in monitoring the air quality (Bai et al. 2018; Mclean et al. 2019; Baklanov and Zhang 2020; Liu et al. 2021; Masood and Ahmad 2021). Bai et al. (2018) analysed the three classical methods for forecasting air pollution (statistical, artificial intelligence, and numerical prediction methods). There is literature on air quality in urban areas using different statistical methods to forecast air quality (Mavroidis et al. 2007; Ilacqua et al. 2007; Lu et al. 2014). Considering meteorological aspects, in Mavroidis et al. (2007) a successful methodology was suggested for assessing the impact of different emission reduction scenarios on the attainment of air quality standards for CO and $NO_2$ in the Athens area. Furthermore, in Ribeiro and Gonçalves, (2022), in Portugal, $NO_2$ is classified as a binary objective using a benchmark model. In Durão et al. (2016), classification and regression tree techniques were successfully used to predict ozone in Sines (Portugal). For $NO_2$, Prati et al. (2015) provided an insight into the relevance of a spatial analysis of data that provides knowledge on how ship emissions affect the air in a port city. To forecast air quality in urban areas, Lu et al. (2014) proposed different semi-parametric regression models. Particulate matter (PM) sources in three European cities (Athens, Basle, and Helsinki) are described and analysed using structural equation modelling in parallel with traditional principal components (Ilacqua et al. 2007). Similar machine learning techniques are used by Lakra and Avishek, (2022) to forecast fog, which is also related to meteorological factors. Other techniques are used to construct air quality models. In García-Nieto et al. (2015) air quality in Oviedo (Spain) was modelled using multivariate adaptive regression splines (MARS) and subsequently, support vector regression (SVR), multilayer perceptron (MLP), were specifically used to forecast $PM_{10}$ concentrations in the same city by García-Nieto et al. (2018). In addition, meteorological variables are considered by Luna et al. (2019), where low-cost electrochemical sensors are used to quantify air pollution exposure, prediction, and control of $CO_2$ and $SO_2$ concentrations using ANNs. The most relevant information extracted from this study was that pollution prediction is sensitive to humidity, wind speed, and temperature. Therefore, the use of ANNs could predict and impute missing values or re-evaluate doubtful values. A method for predicting $SO_2$ emissions in several cities is shown by Ju et al. (2023), which is of great help for accurate control of this pollutant. Applied to megacities, He et al. (2014) provided an ANN-based method, in particular a multilayer perceptron (MLP), that predicts fine particles suggesting that particulate matter concentrations are generated by traffic and controlled by weather conditions.

Air quality assessment, from an operational point of view, requires the characterisation of atmospheric quality

(Corani and Scanagatta 2016; Méndez et al. 2023). The aim of this work is to predict future values of the levels of each pollutant. Machine learning methods based on classification models have been used for this purpose. A comprehensive comparison of classification models was developed. The classifiers tested were trees, support vector machines (SVMs), artificial neural networks (ANNs), ensembles, K-nearest neighbours (KNNs), discriminant, and naïve Bayes. Most of them have already been successfully used by authors in different papers (Turias et al. 2008; Ruiz-Aguilar et al. 2020; Song and Fu 2020; González-Enrique et al. 2021; Moscoso-López et al. 2022). Regarding local studies, the impact of ship propulsion systems on air pollution in the Strait of Gibraltar in 2017 is presented in Durán-Grados et al. (2022). This study is based on an inventory of ships crossing the Strait and calling at the ports of Algeciras, Tarifa, and Ceuta. In Martín et al. (2008) air pollution was modelled with classification techniques in the Bay of Algeciras (Spain). Additionally, Rodríguez-García et al. (2022) conducted an extensive analysis of statistical, risk, and trends developed in the area of the Bay of Algeciras from 2010 to 2015. Furthermore, due to the large number of inputs used to build the models, the problem of the curse of dimensionality (Bishop 2006) could arise. Therefore, a feature selection stage was applied using the Minimum Redundancy Maximum Relevance (mRMR) method, which has been successfully tested by the authors previously in air pollution forecasting problems (González-Enrique et al. 2021).

The main motivation of this manuscript is to provide citizens with reliable information on air pollution forecasts. This challenge is achieved through a data-driven approach using historical data and machine learning techniques, which will be explained in more detail in the next sections. Improving the air quality in populated cities is another of the main motivations for this study, which is carried out in the Bay of Algeciras (southern Spain), where the most important port in Spain and the fourth in Europe in terms of cargo traffic is located. The importance of maritime traffic in Algeciras, which has experienced a massive increase in the last ten years, in terms of air pollution, lies in the fact that this increase in the number of vessels in the port of Algeciras may affect the air quality in the area and in the nearest city (Algeciras). Since there have been few studies on air pollution in this strategic area of port activities in terms of pollution, this research can make a specific contribution.

Another main contribution of this work is the use of a classification-based machine learning scheme to predict the next level of a pollutant, including an analysis of the most relevant variables (using mRMR) for each of the pollutants and sites studied. In addition, many different classification methods were used and compared. This research has allowed us to develop a procedure for predicting future pollution levels, both on an hourly basis for nitrogen oxides ($NO_2$, $NO_x$, and NO) and, on a daily basis for $SO_2$ and $PM_{10}$. The results obtained are suitable for the design of air pollution forecasting system that can be used by citizens or institutions to support decision making.

The rest of this article is organised as follows: Sect. 2 describes the database, the site, the case study and the regulations to be applied, Sect. 3 presents the methodology including the classification models tested in the study together with the feature selection process and the experimental procedure used to achieve the objectives, Sect. 4 presents and discusses the results and, finally, Sect. 5 draws the main conclusions.

## 2 Materials

The importance of environmental studies in this area is due to the fact that the Port of Algeciras is located in this area, handling more than 100 million tonnes of goods per year since 2017, and is located in an area with special meteorological and orographic conditions, the Strait of Gibraltar, as well as in a highly industrialised region where the Port of Algeciras coexists with numerous industries (a refinery, several chemical and thermal power plants, a stainless steel factory, etc.), together with several highways and the Gibraltar airport.), together with several motorways and Gibraltar airport, contribute to a very complex air pollution scenario. Maritime traffic in Algeciras has increased dramatically over the last decade. It is logical to think that the increase in the number of vessels in the Port of Algeciras could affect the air quality in the area.

In order to develop this study, the main pollutants related to port activities were selected as shown in (Yang et al. 2022; Yeh et al. 2022; Mueller et al. 2023). Immission data of $SO_2$, $NO_2$, $NO_X$, NO and, $PM_{10}$ concentrations, meteorological data (relative humidity, solar radiation, temperature, atmospheric pressure, wind speed, wind direction, and rainfall) were provided through the Andalusian Government's monitoring network, and the vessel gross tonnage (GT) database was provided by the Algeciras Bay Port Authority, all for the years 2017 to 2019. Similar studies, such as López-Aparicio et al. (2017), analysed all these pollutants in a Nordic port and concluded that the main emission contributions come from berthed vessels and manoeuvres.

The Andalusian Government's system of sensors in the Bay of Algeciras includes a total of sixteen air pollutant monitoring stations and five specialised meteorological sensors ($W_{1\text{-}5}$) distributed throughout the bay (see Fig. 1), which record hourly data of each pollutant and

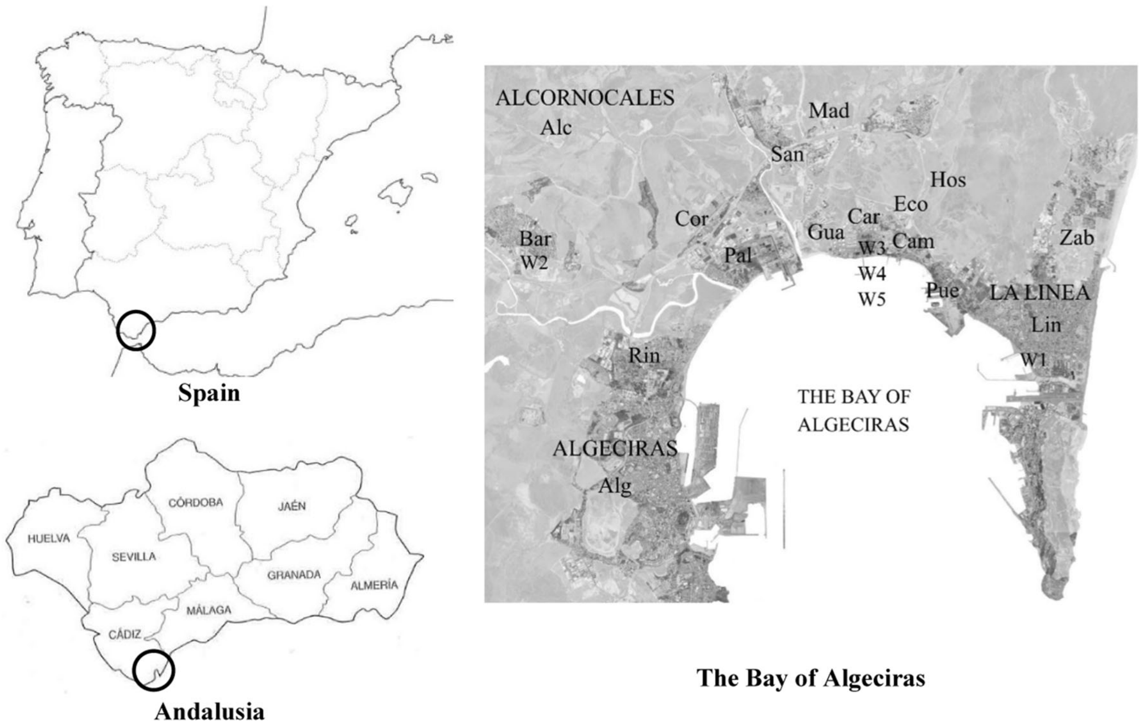

**Fig. 1** Location of the area of study. Spain, Andalusia and The Bay of Algeciras in the Strait of Gibraltar. The three studied monitoring stations in the cities of Algeciras and La Línea and Alcornocales Park and the rest of sensors over the Bay

meteorological values over a three-year period, from 1st January 2017 to 31st December 2019 (see Table 1). The meteorological sensors $W_3$, $W_4$, and $W_5$ are located in the chimney of a refinery at different heights, 10 m, 15 m, and 60 m. The data analysed are recorded at stations in the towns of Algeciras and La Línea and in the Alcornocales Park, in order to compare three distant locations. The importance of Algeciras and La Línea spots is due to their coastal areas and the huge port of Algeciras, with massive truck traffic, and Alcornocales Park is an unspoilt area far from anthropogenic activity. In addition, La Línea and Algeciras are two cities located opposite each other, thus studying both can shed more light on air pollution immissions. Algeciras is the most populated city in the bay with 122,982 inhabitants in 2021 and La Línea is the second most populated city with 63,365 inhabitants.[1] The entire database consists of 131 variables. In each experiment, the output variable is the concentration of each pollutant in each of the monitoring stations according to the rest of the study variables described in Table 1 (pollutant concentrations in the rest of the monitoring stations, meteorological information and vessel data).

**Table 1** Monitoring stations codes. Meteorological variables codes. Pollutant variables

| Code | Monitoring stations | Code | Variables of the study |
|---|---|---|---|
| Alg | Algeciras | Ws | Wind speed (km/h) |
| Cam | Campamento | Wd | Wind direction (degree) |
| Cor | Los Cortijillos | RH | Relative humidity (%) |
| Hos | Hostelería | RF | Rainfall (l/m$^2$) |
| Alc | Alcornocales | T | Temperature (°C) |
| Car | Carteya | AP | Atmospheric pressure (hPa) |
| Rin | Rinconcillo | SR | Solar radiation (W/m$^2$) |
| Pal | Palmones | $SO_2$ | Sulphur dioxide (μg/m$^3$) |
| San | San Roque | $NO_2$ | Nitrogen dioxide (μg/m$^3$) |
| Zab | El Zabal | $NO_X$ | Nitrogen oxides (μg/m$^3$) |
| Eco | Economato | NO | Nitrogen monoxide (μg/m$^3$) |
| Gua | Guadarranque | $PM_{2.5}$ | Particulate matter ($\leq$ 2.5 μm) |
| Lin | La Línea | $PM_{10}$ | Particulate matter ($\leq$ 10 μm) |
| Mad | Madrevieja | CO | Carbon monoxide (μg/m$^3$) |
| Bar | Los Barrios | $O_3$ | Ozone (μg/m$^3$) |
| Pue | Puente Mayorga | Tol | Toluene (μg/m$^3$) |
| $W_{1\text{-}5}$ | Meteo stations | Ben | Benzene (μg/m$^3$) |
| | | Ves | Vessels (GT/h) |

This study has been developed in three stages: preprocessing of the data, classification stage and the stage of feature selection to reduce the number of variables. Among the wide range of feature selection methods, the mRMR method was used in this work to rank the variables considered as inputs. Feature selection, one of the fundamental problems in pattern recognition and machine learning,

[1] http://www.ine.es.

**Table 2** Simulation scenarios and Directive 2008/50/EC limit values for pollutants of the study

| Pollutant | Limit value | Upper assessment threshold (μg/m$^3$) |
|---|---|---|
| $SO_2$ | Daily mean | 75 |
| $PM_{10}$ | | 35 |
| $NO_2$ | Hourly mean | 140 |
| $NO_X$ | | |
| NO | | |

involves identifying subsets of data that are relevant to the parameters used, usually referred to as maximum relevance. These subsets often contain material that is relevant but redundant, and mRMR attempts to address this problem by eliminating these redundant subsets. In this paper, the ten most relevant features were selected as inputs to the different models to test whether there are significant differences when all variables are used in the models.

## 3 Methodology

The main objective of this work is to predict the future air quality levels of the main maritime pollutants in the Bay of Algeciras as a function of other pollutants, meteorological variables, and vessel data. In order to achieve this objective, the time series were considered according to the limits marked in the European Directive 2008/50/EC (Table 2), and the outputs were transformed into disjoint quartiles (Q1–Q4).

The predictions are calculated using pollutant concentrations in each station (Algeciras, Alcornocales, and La Línea) as outputs and the rest of the variables as inputs (pollutants in other stations, meteorological parameters, and the vessel data). Different classification techniques are compared together with ANN models in order to find improvements and the best model. The performance of the tested models is calculated using hourly and daily mean data time series.

$$\widehat{y}_q(t+1) = f_{classification}(\widetilde{x}(t), y(t)) \tag{1}$$

Equation 1 shows mathematically the prediction approach, where $t$ is the time and $t + 1$ is one step ahead to be predicted. In the case of hourly data, the next 1 h-mean period concentration value is predicted and in the case of daily data, the next day mean concentration value is predicted. Inputs $\widetilde{x}(t)$ consist of all other pollutants measured at the monitoring stations together with meteorological variables and vessel time series. The scheme of the process is shown in Fig. 2.

Three stages were developed. The first step is the pre-processing of the data. On the one hand, the imputation of missing values was done using a previous algorithm successfully proposed by the authors (González-Enrique et al. 2019a, 2019b; Rodríguez-García et al. 2022). On the other hand, the standarisation of the database. A transformation of the vessel data, given as incoming and outgoing vessels in the bay into hourly data was also performed. Once the databases are transformed and unified, the data consist of 26,280 hourly records × 131 variables (130 inputs and 1 output) of a unique database. Each row is a record of hourly data for the three years from 2017 to 2019. The database has been normalised and the output has been divided into disjoint quartiles. The second stage of classification is described in Sect. 3.1 and the third stage is a feature selection procedure using the mRMR approach proposed by Peng et al. (2005), which is a feature selection algorithm that ranks a set of features according to their relevance to the target variable. It also penalises redundant features. The best features are those with the highest trade-off between maximum relevance with the target variable and minimum redundancy with the remaining features.

### 3.1 Classification

In this stage, 29 classification models (Table 3) were tested to select the best classifier. Classification is a type of supervised machine learning where an algorithm learns to classify new observations from labelled data samples. In this work, the database is labelled in quartiles, as shown in Table 3. The different classification schemes are briefly explained below.

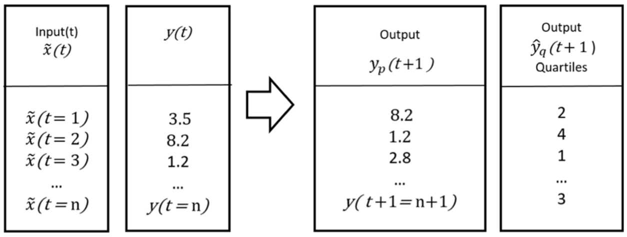

**Fig. 2** Methodology scheme. The output data was transformed into quartiles ($Q_1$–$Q_4$). The inputs and output at the timestamp $t$ are the predictors of the quartile at timestamp $t + 1$

**Table 3** Classification models

| Models | Enumeration |
|---|---|
| Trees | 1–3 |
| Discriminant | 4–5 |
| Naïve Bayes | 6–7 |
| SVM | 8–13 |
| KNN | 14–19 |
| Ensembles | 20–24 |
| ANNs | 25–29 |

### 3.1.1 Trees

Trees are a hierarchical non-parametric supervised learning algorithm consisting of a root node, branches, internal nodes, and leaf nodes. It is based on classification principles that predict the outcome of a decision for both classification and regression tasks (Breiman et al. 1984). Three types of trees were used depending on the maximum number of splits (100, 20, 4). The maximum number of splits equal to 100 is when many leaves are used to make many fine distinctions between classes. When the number of leaves is equal to 4, the distinctions that can be made are stronger.

### 3.1.2 Discriminant analysis

Discriminant analysis is a statistical transformation technique that produces a function capable of classifying phenomena (Fisher 1936). The objective is to maximise the between-group variance and minimise the within-group variance through these linear (or quadratic) combinations. The procedure is to discover the autovalues and autovectors of a quotient matrix of the interclass distance matrix and the intraclass distance matrix. For linear discriminant analysis, the model has the same covariance matrix for each class; only the means vary. For quadratic discriminant analysis, both the means and the covariances of each class vary.

### 3.1.3 Naïve Bayes

Naive Bayes models assume that observations have a multivariate distribution with regard to class membership, although the predictors or features that make up the observation are independent. This framework can accommodate a full set of features, so that an observation is a set of multinomial counts (Mitchell 1997). Normal (Gaussian) distribution is appropriate for predictors that have normal distributions in each class. The Naïve Bayes classifier estimates a separate normal distribution for each class by calculating the mean and standard deviation of the training data in that class. The kernel distribution is suitable for predictors that have a continuous distribution. It does not require a strong assumption such as a normal distribution, and you can use it in cases where the distribution of a predictor may be skewed or have multiple peaks or modes.

### 3.1.4 Support Vector Machines (SVMs)

The goal of SVM is to find out a hyperplane that best separates two different classes of data points with the widest margin between the two classes. The algorithm can only find this hyperplane in problems that allow linear separation; in most practical problems, the algorithm maximises the flexible margin by allowing a small number of misclassifications. The support vectors refer to a subset of the training observations that identify the location of the separation hyperplane. SVMs can use a kernel function to transform the features. Kernel functions map the data into a different, usually higher dimensional space, with the expectation that it will be easier to separate the classes after this transformation (Vapnik and Chervonenkis 1971; Cortes and Vapnik 1995). The types tested are Linear SVM (makes a simple linear separation between classes), Quadratic SVM, Cubic SVM, and three categories of Gaussian SVM (fine, with kernel scale set to $\sqrt{P}/4$; medium, with kernel scale set to $\sqrt{P}$; and coarse, with kernel scale set to $\sqrt{P} \cdot 4$, where $P$ is the number of predictors).

### 3.1.5 KNN

The k-nearest neighbour algorithm, also known as KNN or k-NN, is a non-parametric supervised learning classifier, that uses proximity to make classifications or predictions about the clustering of a single data point. While it can be used for regression or classification problems, it is generally used as a classification algorithm, based on the assumption that similar points will be found close together. Usually, the number $k$ is an odd number (1,3,5…) (Silverman and Jones 1989). The types of trees tested were Fine KNN (the number of neighbours is set to 1), Medium KNN (the number of neighbours is set to 10), Coarse KNN (the number of neighbours is set to 100), Cosine KNN, using a cosine distance metric (the number of neighbours is set to 10), Cubic KNN, using a cubic distance metric (the number of neighbours is set to 10), Weighted KNN, using a distance weight (the number of neighbours is set to 10).

### 3.1.6 Ensemble learning

Classification ensemble learning uses multiple learning algorithms to obtain a better predictive model, which is aa weighted combination of several classification models. In general, the combination of several classification models increases the predictive power. The types of ensembles tested were: Subspace with discriminant learners, Subspace with nearest neighbour learners, and RUSBoost, Random Forest Bag, and AdaBoost, with decision tree learners (Breiman 1996, 2001; Hastie et al. 2008; Freund 2009).

### 3.1.7 Artificial neural networks (ANNs)

ANNs were also included in the second stage. A feedforward fully connected ANN can be arbitrarily well suited to multidimensional mapping problems, given consistent data and enough neurons in its hidden layer (Hornik et al. 1989). The authors have successfully used ANNs in similar prediction problems (Gonzalez-Enrique et al., 2019b; Ruiz-Aguilar et al. 2020; Moscoso-López et al. 2022). ANNs were trained with the backpropagation algorithm (Rumelhart et al. 1986) using the Levenberg–Marquardt optimisation procedure. Finally, the obtained results were statistically analysed and compared using a resampling procedure in order to select the model with the best generalisation capabilities. ANN models with different hidden units were compared to determine the effect of adding non-linear processing capabilities on model performance. Each model is a feedforward fully connected neural network with a different number of fully connected layers and hidden units. A ReLU activation function was used in each model. The rectified linear activation function, or ReLU, is a non-linear or piecewise linear function that directly outputs the input if it is positive, otherwise, it outputs zero (Glorot et al. 2011). It is the most commonly used activation function in neural networks since 2017 (Ramachandran et al. 2017). The types of tested ANNs were: One hidden layer with 10, 25, and 100 neurons; two hidden layers with 10 *x* 10 neurons and three hidden layers with 10 *x* 10 *x* 10 neurons.

## 3.2 Feature selection

The third stage is a feature selection procedure. The Minimum Redundancy Maximum Relevance (mRMR) approach (Peng et al. 2005) is a feature selection algorithm that ranks a set of features according to their relevance to the target variable. It also penalises redundant features. The best features are those with the highest trade-off between maximum relevance with the target variable and minimum redundancy with the remaining features.

Among the wide range of feature selection methods, the mRMR method has been used in this work to rank the variables considered as inputs. This method has been successfully used by the authors in other studies related to air pollution (González-Enrique et al. 2021). Feature selection, one of the fundamental problems in pattern recognition and machine learning, involves identifying subsets of data that are relevant to the parameters used, usually referred to as maximum relevance. These subsets often contain material that is relevant but redundant, and mRMR attempts to address this problem by eliminating these redundant subsets. In this paper, the top ten relevant features were selected as inputs to the different models to test whether there are significant differences when all variables are used in the models.

## 3.3 Experimental procedure

A resampling procedure was used to reduce the prediction error of a test set and to reduce the effects of overfitting. The strategy randomly divided the database into three parts (training 70%, validation 10%, and test sets 20%) and the performance results were collected only for the test set in order to estimate the generalisation error of each model using unseen data, as the authors have successfully implemented in other papers (Turias et al. 2008; González-Enrique et al. 2019a; Ruiz-Aguilar et al. 2020; Moscoso-López et al. 2022). In this research, all of the simulations were developed and tested in Matlab © software.

The whole system can be seen as a mapping from a set of input features to an output variable. The mathematical form of the mapping is determined by the data (training set). Of course, we need to build a system that is capable of making good predictions on unseen data. In order to measure this generalisation ability, cross-validation is used with another set of samples (test set) is used. We adopted five-fold cross-validation to select the best model based on the generalisation performance of each model. The available data were divided into three different groups (training, validation, and test sets). The parameters of each model were estimated using one of the groups (the training set). A validation set is used for early stopping and to avoid overfitting. Finally, the test set is used to test the classification quality indexes (sensitivity, specificity, accuracy, and precision), simulating the real performance of the model. This process is repeated 20 times and the results are averaged over these runs. To visualise the obtained results with a classification model, the confusion matrix is used (Ting 2010). Each row (i) of the matrix (C) represents the number of predicted values for each class and each column (j) represents the number of real values for each class (C(i,j)). In this case, four classes are considered, one for each of the quartiles of the output. Once an air pollutant has been considered, its values are divided into four quartiles, each containing 25% of the total distribution. The confusion matrix is calculated and then the quality indexes of sensitivity, specificity, accuracy, and precision are also calculated. The Euclidean distance ($d_1$) to a perfect classifier in terms of the quality indexes (sensitivity = 1, specificity = 1, accuracy = 1, precision = 1) is also calculated (expressed by Eq. 2).

$$d_1 = \sqrt{(1 - \mathit{sensitivity})^2 + (1 - \mathit{specificity})^2 + (1 - \mathit{accuracy})^2 + (1 - \mathit{precision})^2)} \tag{2}$$

In this case, the confusion matrix has a 4 × 4 dimension due to data are divided into four disjoint quartiles (classes). In order to obtain individual classification results for each quartile, the matrix was sequentially transformed, quartile by quartile, into an equivalent 2 × 2 confusion matrix (Table 4), which was used to calculate the well-known and above-mentioned classification measures (sensitivity, specificity, accuracy, and precision, see Eqs. 3–6). The lower $d_1$ distance is chosen to indicate the best classification model for each quartile. Quartiles are the statistical values that divide the dataset into four equal parts or quarters, each containing 25% of the data, resulting in lower, lower-middle, middle-high, and upper divisions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

True-positive (TP) and true-negative (TN) results are correctly classified, while false-negative (FN) and false-positive (FP) results are two types of errors calculated according to the literature (Ting 2010).

All the calculations are performed separately. The air pollutants ($SO_2$, $PM_{10}$, $NO_2$, $NO_X$, and NO) as outputs in the three locations (Algeciras, Alcornocales, and La Línea), using all variables or only the ten most relevant variables, in a total of 30 scenarios, repeated 20 times each, following the resampling procedure explained above. The time series of $SO_2$ and $PM_{10}$ concentrations are calculated as daily averages and $NO_2$, $NO_X$, and NO on an hourly basis. Once the experiments have been developed, the results are presented in the next section.

## 4 Results

Simulations and prediction experiments were computed for five pollutants directly related to maritime traffic: $SO_2$, $PM_{10}$, $NO_2$, $NO_X$, and NO. The models were tested at three different locations, in the cities of Algeciras and La Línea, and at a third location at a certain distance in the remote area of the Alcornocales Park. As explained above in Table 2, the averages were calculated hourly or daily to comply with the European Directive 2008/50/EC. Figures 3, 4 show the time series graphs with their upper assessment thresholds of the pollutants analysed on an hourly or daily basis according to the Directive measured in μg/m$^3$. These graphs show average concentrations and it is worth noting that in 2017 the average $SO_2$ concentrations in La Línea, where a refinery is located, are very high compared to the rest of the years, which seems to be due to the installation of a desulphurisation unit in 2018 in this refinery. Considering particulate matter, the lowest concentrations are found at the Alcornocales station, and the highest at Algeciras, although overall concentrations are very similar in both Algeciras and La Línea. On the other hand, the pollutant $NO_2$ (and nitrogen oxides in general) clearly shows very high average concentrations in Algeciras compared to La Línea and Alcornocales, which are quite similar. This increase could be an indication of the high presence of diesel engines in Algeciras, which is consistent with the heavy truck traffic in and out of the port, the ships berthed in the Port of Algeciras and the higher traffic density, since it is the most densely populated city in the Bay.

Since the pollutant thresholds are defined in the regulations in terms of hourly and daily values, and in order to better understand the behaviour of each pollutant, weekly average graphs of each air pollutant at the different stations have been calculated (Fig. 5). The pollutant $SO_2$ shows a higher concentration in La Línea, probably because the prevailing winds carry the pollution from ships and the surrounding industries more towards La Línea (westerly situations), and in easterly situations $SO_2$ seems to move towards Los Alcornocales, the remote area 30 km from the bay, which paradoxically has a higher concentration than Algeciras. In the case of the $PM_{10}$ averages, it can be seen that concentrations decrease during the night, and from the early hours of the morning, when anthropogenic activity begins, the values increase until late in the day. At weekends there is not much difference compared to the rest of the week. In the case of nitrogen oxides, there is a daily decrease in the early hours of the morning, then an increase to a maximum around midday, and then a downward trend with a slowdown around mid-afternoon, which coincides with the pace of human activity and therefore traffic, especially vehicle traffic. The trend is higher in the cities of La Línea and Algeciras. At the remote Los Alcornocales

**Table 4** Equivalent multi-class confusion matrix

| | Real class | |
|---|---|---|
| Predicted class | Positive | Negative |
| Positive | TP = C(i,i); | FP = sum(C(i,:))-C(i,i); |
| Negative | FN = sum(C(:,i))-C(i,i); | TN = sum(sum(C(:,i))-(TP + FP + FN); |

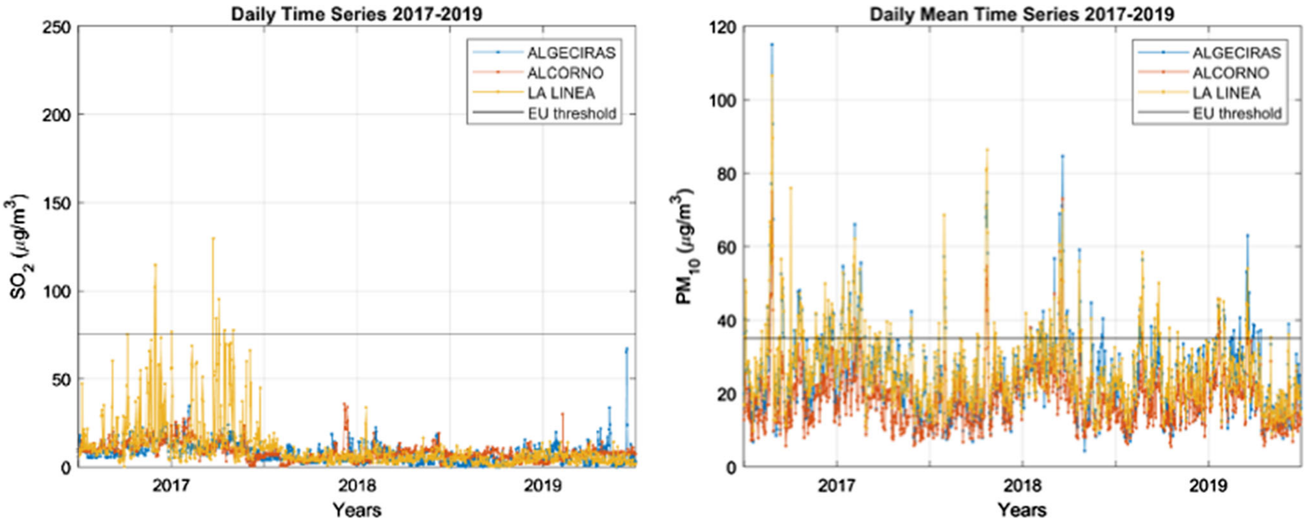

**Fig. 3** Daily mean time series of $SO_2$ and $PM_{10}$ from 2017 to 2019 with the Directive 2008/50/EC limit thresholds

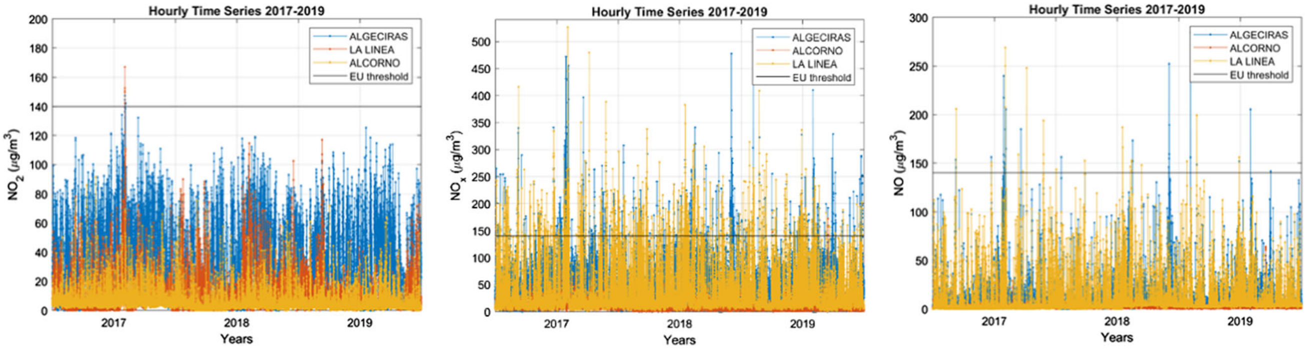

**Fig. 4** Hourly mean time series of $NO_2$, $NO_X$ and NO from 2017 to 2019 with the Directive 2008/50/EC limit thresholds

station, there is only a slight increase at midday. In terms of daily averages, maximum values are observed on Tuesdays and Fridays, with a significant decrease at weekends. It should be noted that $NO_2$ is higher in Algeciras than in La Línea, probably due to road traffic. Nitrogen oxides have two peaks per day, which suggests that they are related to human activity, and especially to diesel engines whereas particulate matter and $SO_2$ have only one peak per day.

As explained above, one-step-ahead prediction models have been developed with the aim of predicting the next value of a time series of quartile concentrations in order to contrast with exceedances of the thresholds set in the Directive. Several classification models, including ANNs, were tested and compared for their performance using the resampling procedure explained in Sect. 3. In each case, two experiments were calculated, one using all variables as inputs and another one with only the ten most relevant variables. It should be noted that the results shown in Tables 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14 are always calculated for test sets (unseen data). In general, the obtained results are quite adequate, with higher values for the classification quality indexes. Results of around 90% indicate that the prediction for the next timestamp-ahead or the next daily/hourly mean is quite accurate and represents a very reliable prediction. The results are collected for the different separated quartiles to achieve a more detailed picture of the prediction.

In Tables 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14, the best prediction model for each air pollutant, location, and quartile is shaded underline (the combination with the smallest distance $d_1$). The best counterpart model (same model, location, quality index, and quartile) is shown in bold between Tables 5, 7, 9, 11, and 13 which present the results of the models using all variables, and Tables 6, 8,

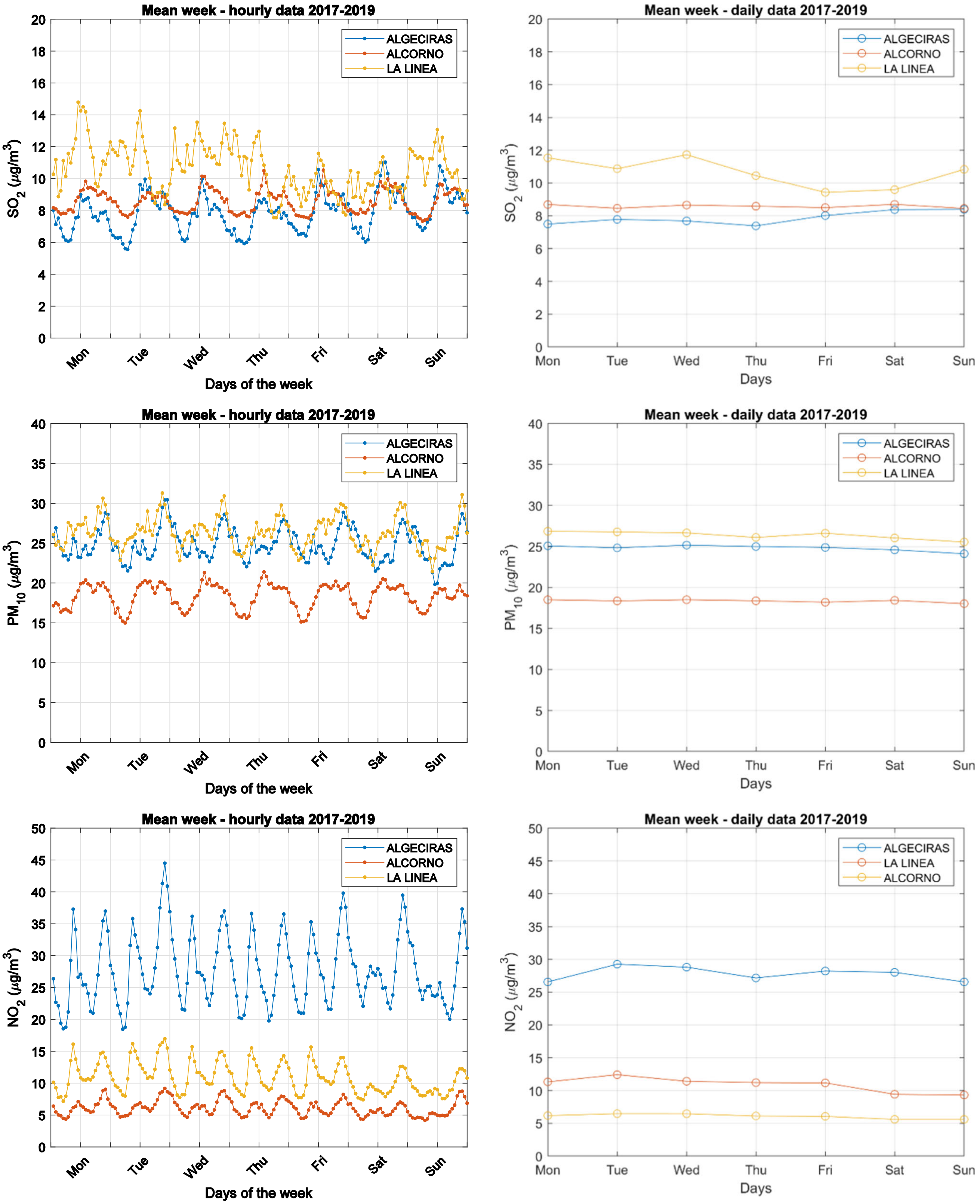

**Fig. 5** Hourly and daily mean week diagrams for pollutants from 2017 to 2019

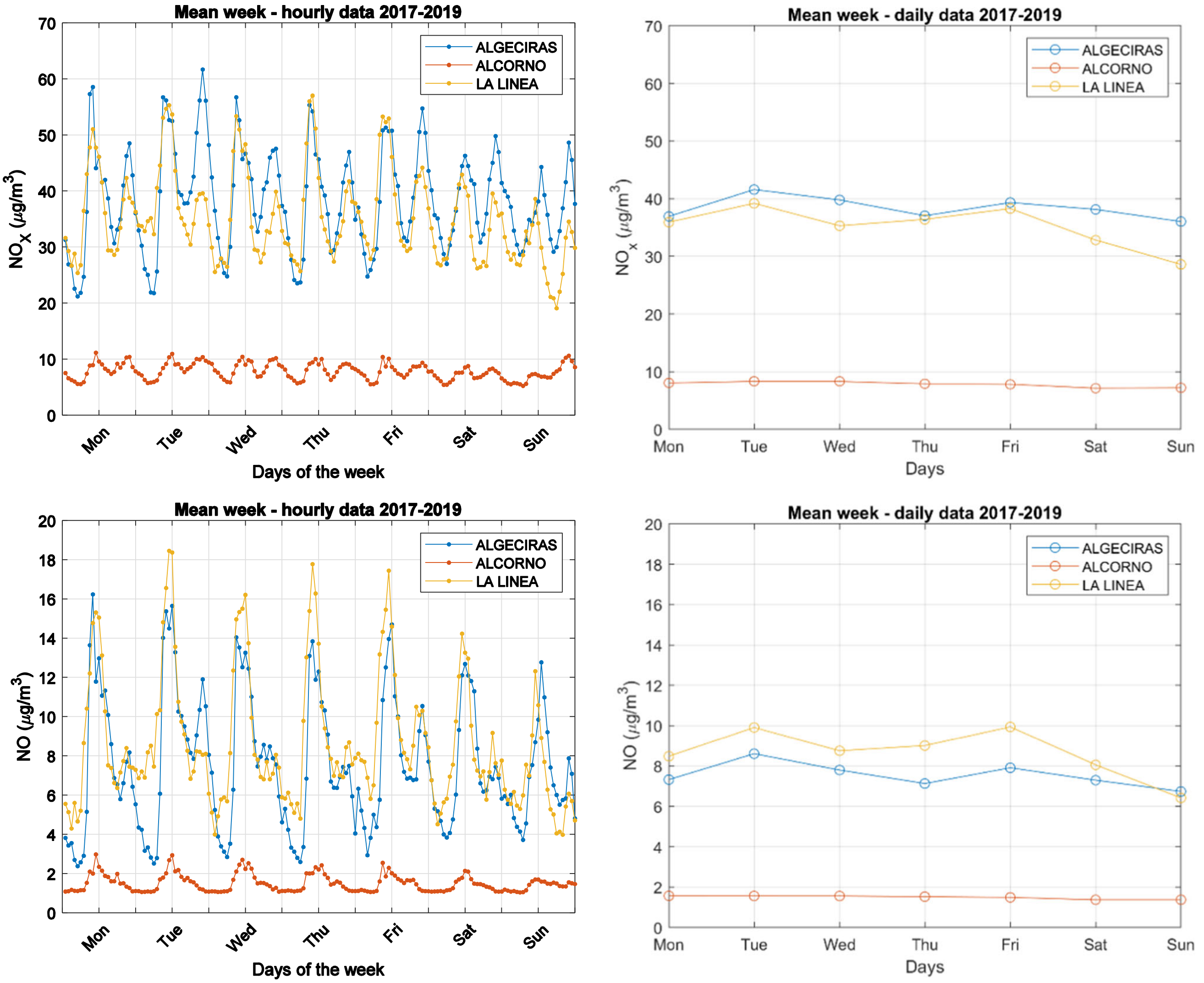

**Fig. 5** continued

10, 12, and 14, which use only the ten most relevant variables. The distance $d_1$ was used to compare models of the same location and to select the best model in each case.

Comparing Table 5 with Table 6 for the pollutant $SO_2$, prediction models using all variables with the models using only the relevant variables, it can be seen that in all cases the ANN models significantly improve their prediction performance when using only relevant variables, although the tree classifiers predict better than the ANNs as their distance $d_1$ is the smallest. For this pollutant, tree classifiers are the best predictors in all cases. For $SO_2$ in Algeciras, quartiles, $Q_1$ and $Q_2$ are best predicted by the tree classifiers using only the ten most relevant variables. However, quartile $Q_3$ is also best predicted by tree classifiers using all 130 variables and $Q_4$ is best predicted by ANNs models using the ten most relevant variables. The best performing quartile (with the lowest $d_1$) in Algeciras is the $Q_1$ with sensitivity, specificity, accuracy, and precision above 0.90. For $SO_2$ in Alcornocales, better predictions are obtained in quartiles $Q_1$ and $Q_2$ with tree-type classifiers and using only the ten most relevant variables. In the case of quartiles $Q_3$ and $Q_4$, better predictions are obtained with ensemble classifiers using all variables. The best results for Alcornocales are obtained for $Q_1$, $Q_3$, and $Q_4$, all with values up to 0.97. For $SO_2$ in La Línea, better predictions were obtained for quartiles $Q_1$ and $Q_4$ using relevant variable

**Table 5** Best prediction model results for daily $SO_2$ ($t + 1$) concentrations using all variables at $t$

| Pollutant | Location | Best Model/hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_I$ |
|---|---|---|---|---|---|---|---|---|---|
| $SO_2$ | Algeciras | Best Classifier | Tree (Max. 4 splits) | Q1 | 0.8954 | 0.9664 | 0.9465 | 0.9153 | 0.1418 |
| | | | | Q2 | 0.7579 | **0.9255** | 0.8884 | **0.7430** | 0.3777 |
| | | | | Q3 | 0.7049 | 0.9087 | **0.8555** | **0.7316** | **0.4340** |
| | | | | Q4 | 0.8207 | 0.9336 | 0.9065 | 0.8040 | 0.2893 |
| | | Best ANN | Neural Network (10 × 10x10 neurons) | Q1 | 0.8954 | 0.9664 | 0.9465 | 0.9121 | 0.1505 |
| | | | | Q2 | 0.7328 | 0.9215 | 0.8790 | 0.7307 | 0.4058 |
| | | | | Q3 | 0.6969 | 0.9005 | 0.8487 | 0.7051 | 0.4600 |
| | | | | Q4 | 0.8207 | 0.9336 | 0.9065 | 0.7960 | 0.2948 |
| | Alcornocales | Best Classifier | Ensemble (Bagged Tree) | Q1 | 0.8383 | 0.9389 | 0.9145 | 0.8149 | 0.2673 |
| | | | | Q2 | 0.6761 | 0.8975 | 0.8405 | 0.6959 | 0.4830 |
| | | | | Q3 | **0.9817** | **0.9927** | **0.9900** | **0.9781** | **0.0311** |
| | | | | Q4 | ***0.9783*** | *0.9951* | ***0.9909*** | *0.9854* | ***0.0282*** |
| | | Best ANN | Neural Network(10 × 10 neurons) | Q1 | 0.8873 | 0.9634 | 0.9443 | 0.8905 | 0.1707 |
| | | | | Q2 | 0.7500 | 0.9215 | 0.8776 | 0.7664 | 0.3718 |
| | | | | Q3 | 0.8039 | 0.9190 | 0.8922 | 0.7509 | 0.3445 |
| | | | | Q4 | 0.9018 | 0.9790 | 0.9589 | 0.9380 | 0.1250 |
| | La Línea | Best Classifier | Ensemble (Bagged Tree) | Q1 | 0.9673 | 0.9902 | 0.9845 | 0.9708 | 0.0475 |
| | | | | Q2 | **0.9455** | **0.9829** | **0.9735** | **0.9489** | **0.0811** |
| | | | | Q3 | **0.9631** | **0.9854** | **0.9799** | 0.9560 | **0.0625** |
| | | | | Q4 | 0.9818 | **0.9939** | 0.9909 | **0.9818** | 0.0280 |
| | | Best ANN | Neural Network (25 neurons) | Q1 | 0.8607 | 0.9595 | 0.9342 | 0.8796 | 0.1997 |
| | | | | Q2 | 0.7239 | 0.9033 | 0.8594 | 0.7080 | 0.4366 |
| | | | | Q3 | 0.7138 | 0.9072 | 0.8584 | 0.7216 | 0.4337 |
| | | | | Q4 | 0.8930 | 0.9612 | 0.9443 | 0.8832 | 0.1723 |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

models with tree classifiers, and $Q_2$ and $Q_3$ were better predicted with ensemble classifiers using all variables. In La Línea the $Q_4$ is the one with the best results, with quality indexes above 0.98.

For the $PM_{10}$ pollutants, Tables 7 and 8 show that the use of relevant variables improves the results of the classification models in almost all cases. The results of the best models for each quartile are those shaded in underline, regardless of whether they include all variables or only the relevant ones, and turn out to be the trees with slightly better results than the ANNs models. The obtained results for the best models for $PM_{10}$ pollutants have the highest quality indexes. For instance, in Algeciras, the $Q_1$ is the best predicted with tree classifiers using relevant variables, obtaining quality indexes above 0.95. In Alcornocales, quartile $Q_4$ is also the best predicted with tree classifiers using relevant top ten variables with quality indexes above 0.97. In La Línea, the best predicted quartile is $Q_4$ with quality indexes up to 0.96.

The results for $NO_2$ are shown in Tables 9 and 10. In this case, the relevant variables give better results only for Alcornocales and the quartile $Q_1$ of La Línea. Tree-type models are also the best predictors for Alcornocales and La Línea, especially when all the variables are used, and only for quartiles $Q_3$ and $Q_4$ of Alcornocales do neural network models perform better when the relevant variables are used. In the case of Algeciras, all quartiles are predicted equally well by SVM classifiers using all variables. For $NO_2$, the best results are obtained in the case of quartile $Q_1$ in Algeciras with all quality indexes above 0.82, $Q_1$ in Alcornocales with ensemble models using all variables and quality indexes above 0.82, and $Q_1$ in La Línea with quality indexes above 0.82 with ensemble tree classifiers using the top ten variables.

**Table 6** Best prediction model results for daily $SO_2$ ($t + 1$) concentrations using top ten relevant features at $t$

| Pollutant | Location | Best Model/hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_I$ |
|---|---|---|---|---|---|---|---|---|---|
| $SO_2$ | Algeciras | Best Classifier | Tree (Max. 4 splits) | Q1 | **0.9026** | **0.9697** | **0.9508** | **0.9208** | **0.1382** |
| | | | | Q2 | **0.7604** | 0.9253 | **0.8889** | 0.7420 | **0.3767** |
| | | | | Q3 | **0.7059** | **0.9069** | 0.8549 | 0.7254 | 0.4377 |
| | | | | Q4 | **0.8296** | **0.9368** | **0.9111** | **0.8060** | **0.2803** |
| | | Best ANN | Neural Network (10 × 10 neurons) | Q1 | **0.9004** | **0.9699** | **0.9503** | **0.9215** | **0.1395** |
| | | | | Q2 | **0.7594** | **0.9242** | **0.8881** | **0.7382** | **0.3804** |
| | | | | Q3 | **0.7070** | **0.9070** | **0.8554** | **0.7255** | **0.4367** |
| | | | | Q4 | ***0.8307*** | ***0.9375*** | ***0.9118*** | ***0.8081*** | ***0.2777*** |
| | Alcornocales | Best Classifier | Tree (Max. 4 splits) | Q1 | **0.9852** | **0.9927** | **0.9909** | **0.9779** | **0.0290** |
| | | | | Q2 | **0.9636** | **0.9890** | **0.9826** | **0.9672** | **0.0531** |
| | | | | Q3 | 0.9704 | 0.9866 | 0.9826 | 0.9597 | 0.0546 |
| | | | | Q4 | 0.9748 | **0.9963** | **0.9909** | **0.9891** | 0.0292 |
| | | Best ANN | Neural Network (25 neurons) | Q1 | **0.9673** | **0.9903** | **0.9845** | **0.9708** | 0.0475 |
| | | | | Q2 | **0.9517** | **0.9783** | **0.9717** | **0.9343** | 0.0890 |
| | | | | Q3 | **0.9355** | **0.9853** | **0.9727** | **0.9560** | 0.0840 |
| | | | | Q4 | **0.9708** | **0.9878** | **0.9836** | **0.9638** | 0.0508 |
| | La Línea | Best Classifier | Tree (Max. 4 splits) | Q1 | **0.9674** | **0.9915** | **0.9854** | **0.9745** | **0.0447** |
| | | | | Q2 | 0.9446 | 0.9782 | 0.9699 | 0.9343 | 0.0936 |
| | | | | Q3 | 0.9493 | **0.9866** | 0.9772 | **0.9597** | 0.0700 |
| | | | | Q4 | ***0.9890*** | ***0.9939*** | ***0.9927*** | ***0.9818*** | ***0.0233*** |
| | | Best ANN | Neural Network (10 × 10 neurons) | Q1 | **0.9590** | **0.9794** | **0.9744** | **0.9380** | **0.0813** |
| | | | | Q2 | **0.8811** | **0.9728** | **0.9489** | **0.9197** | **0.1547** |
| | | | | Q3 | **0.8993** | **0.9613** | **0.9461** | **0.8828** | **0.1682** |
| | | | | Q4 | **0.9451** | **0.9805** | **0.9717** | **0.9416** | **0.0872** |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

In the case of $NO_X$, reasonably equivalent behaviour is observed between models using all variables and models using only the relevant variables. By using fewer but more relevant variables, a large number of models improve their overall performance. In La Línea, the best models are ANNs using the relevant variables for quartiles $Q_2$-$Q_4$. In Algeciras, the performance of the ANNs is similar for the quartiles $Q_2$ and $Q_4$, and in Alcornocales for $Q_1$. The rest of the best models use all variables and correspond to SVM and ensembles. The values are somewhat lower than for other pollutants, reaching sensitivities above 80% and higher specificities above 93%. In the case of NO in Alcornocales (Tables 13, 14), no values have been obtained for $Q_2$ because most of the data available in the database for this pollutant are at such low values that they correspond for the most part to $Q_1$, except for some peaks of exceedances found in the $Q_3$ and $Q_4$ quartiles. For NO, ANNs seem to be the models that best predict the quartiles using the relevant variables. In fact, the best result is obtained for the $Q_1$ quartile with more than 94% precision for Alcornocales.

Tables 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14 show that ensemble boosted trees and tree classifiers produce better results than ANN models in most cases, but by reducing the number of variables to the best 10, ANNs improve quite a lot. Tables 15, 16, 17, 18, and 19 have also been included, highlighting the most leveraged variables used for each prediction model using the mRMR method. In these tables, only the best ten most relevant variables are shown. Using these variables, similar prediction results were obtained to those shown in Tables 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14 for the models using all the variables. Therefore, using only these top ten variables, a more efficient monitoring system could be designed, saving economic and time resources in

**Table 7** Best prediction model results for daily $PM_{10}$ ($t + 1$) concentrations using all variables at $t$

| Pollutant | Location | Best Model | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_1$ |
|---|---|---|---|---|---|---|---|---|---|
| $PM_{10}$ | Algeciras | Best Classifier | Tree (Max. 4 splits) | Q1 | **0.9705** | 0.9867 | 0.9826 | 0.9599 | 0.0544 |
| | | | | Q2 | 0.9124 | **0.9708** | 0.9562 | **0.9124** | 0.1346 |
| | | | | Q3 | 0.9118 | 0.9696 | 0.9553 | 0.9084 | 0.1382 |
| | | | | Q4 | 0.9568 | 0.9902 | 0.9817 | 0.9708 | 0.0561 |
| | | Best ANN | Neural Network (100 neurons) | Q1 | 0.8603 | 0.9514 | 0.9288 | 0.8540 | 0.2197 |
| | | | | Q2 | 0.7183 | 0.9137 | 0.8630 | 0.7445 | 0.4133 |
| | | | | Q3 | 0.7333 | 0.8976 | 0.8594 | 0.6850 | 0.4479 |
| | | | | Q4 | 0.8345 | 0.9544 | 0.9233 | 0.8650 | 0.2315 |
| | Alcornocales | Best Classifier | Tree (Max.100 splits) | Q1 | **0.9560** | 0.9842 | 0.9772 | 0.9526 | 0.0704 |
| | | | | Q2 | 0.9262 | 0.9721 | 0.9607 | 0.9161 | 0.1217 |
| | | | | Q3 | 0.9283 | 0.9828 | 0.9689 | 0.9487 | 0.0950 |
| | | | | Q4 | 0.9706 | 0.9878 | 0.9836 | 0.9635 | 0.0511 |
| | | Best ANN | Neural Network (10 neurons) | Q1 | 0.8078 | 0.9423 | 0.9078 | 0.8285 | 0.2796 |
| | | | | Q2 | 0.6742 | 0.8845 | 0.8338 | 0.6496 | 0.5195 |
| | | | | Q3 | 0.7201 | 0.9033 | 0.8584 | 0.7070 | 0.4400 |
| | | | | Q4 | 0.8404 | 0.9545 | 0.9251 | 0.8650 | 0.2267 |
| | La Línea | Best Classifier | Ensemble (Boosted Trees) | Q1 | **0.9667** | 0.9842 | **0.9799** | 0.9526 | **0.0634** |
| | | | | Q2 | 0.9206 | **0.9768** | 0.9626 | **0.9307** | **0.1143** |
| | | | | Q3 | 0.9307 | 0.9781 | 0.9662 | 0.9341 | 0.1038 |
| | | | | Q4 | 0.9672 | 0.9890 | 0.9836 | 0.9672 | 0.0505 |
| | | Best ANN | Neural Network (10 × 10x10 neurons) | Q1 | 0.8421 | 0.9397 | 0.9160 | 0.8175 | 0.2625 |
| | | | | Q2 | 0.6804 | 0.9055 | 0.8457 | 0.7226 | 0.4602 |
| | | | | Q3 | 0.7722 | 0.9127 | 0.8795 | 0.7326 | 0.3815 |
| | | | | Q4 | 0.8925 | 0.9694 | 0.9498 | 0.9088 | 0.1528 |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

the sensor network by measuring fewer variables to store and transmit, thus designing a more energy sustainable system with a lower carbon footprint. Tables 15, 16, 17, 18, and 19 show the ten most relevant variables for each pollutant ($SO_2$, $PM_{10}$, $NO_2$, $NO_X$, and NO) and monitoring station (Algeciras, La Línea, and Alcornocales). In these tables, the meteorological variables for each pollutant are marked in yellow, and the rest of the relevant pollutants, different from those analysed and repeated in at least two stations, are marked in other colours. In the Tables 15, 16, 17, 18, and 19, it is expected that each pollutant's own time series ($SO_2$(t), $PM_{10}$(t), $NO_2$(t), $NO_X$(t), and NO(t)) will always appear, and this is indeed the case. For instance, Table 15 shows the most relevant meteorological variables for $SO_2$, namely wind direction (WD) and rainfall (RF). For $SO_2$, $O_3$ and nitrogen oxides are the most relevant air pollutants, as expected. Table 16 shows the relevant variables for the $PM_{10}$ pollutant, indicating that the most relevant meteorological variables are wind speed (WS), rainfall (RF), and relative pressure (RP), and the most relevant pollutants are nitrogen oxides. Similarly, Table 17 for the pollutant $NO_2$ indicates that the most relevant meteorological variables are related to wind (wind direction (WD) and wind speed (WS)) and rainfall (RF), and the most relevant pollutants are particulate matter ($PM_{10}$ and $PM_{2.5}$), $O_3$ and $SO_2$. Table 18 for each $NO_X$ case, shows the same relevant meteorological variables as for $NO_2$ and includes relative humidity (RH) and the same relevant pollutants except $SO_2$. In the case of the NO pollutant, Table 19 shows that the relevant meteorological variables are related to the wind (wind direction (WD) and wind speed (WS)), solar radiation (SR), and rainfall (RF).

The best models for each pollutant and location for the fourth quartile are shown in italics. Results are given for all

**Table 8** Best prediction model results for daily $PM_{10}$ (*t* + *1*) concentrations using top ten relevant features at *t*

| Pollutant | Location | Best Model /hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_1$ |
|---|---|---|---|---|---|---|---|---|---|
| $PM_{10}$ | Algeciras | Best Classifier | Tree (Max. 4 splits) | Q1 | 0.9539 | **0.9938** | **0.9836** | **0.9818** | **0.0526** |
| | | | | Q2 | **0.9321** | 0.9675 | **0.9589** | 0.9015 | **0.1307** |
| | | | | Q3 | **0.9185** | **0.9697** | **0.9571** | **0.9084** | **0.1334** |
| | | | | Q4 | ***0.9568*** | ***0.9902*** | ***0.9817*** | ***0.9708*** | ***0.0561*** |
| | | Best ANN | Neural Network (10 × 10x10 neurons) | Q1 | **0.9348** | **0.9805** | **0.9689** | **0.9416** | **0.0949** |
| | | | | Q2 | **0.8759** | **0.9586** | **0.9379** | **0.8759** | **0.1907** |
| | | | | Q3 | **0.8806** | **0.9553** | **0.9370** | **0.8645** | **0.1965** |
| | | | | Q4 | **0.9314** | **0.9804** | **0.9680** | **0.9416** | **0.0976** |
| | Alcornocales | Best Classifier | Tree (Max. 4 splits) | Q1 | 0.9536 | **0.9914** | **0.9817** | **0.9745** | **0.0567** |
| | | | | Q2 | **0.9580** | **0.9724** | **0.9689** | **0.9161** | **0.1026** |
| | | | | Q3 | **0.9395** | **0.9889** | **0.9763** | **0.9670** | **0.0737** |
| | | | | Q4 | ***0.9816*** | ***0.9915*** | ***0.9890*** | ***0.9745*** | ***0.0344*** |
| | | Best ANN | Neural Network (100 neurons) | Q1 | **0.9585** | **0.9759** | **0.9717** | **0.9270** | **0.0918** |
| | | | | Q2 | **0.8741** | **0.9703** | **0.9452** | **0.9124** | **0.1655** |
| | | | | Q3 | **0.9222** | **0.9709** | **0.9589** | **0.9121** | **0.1277** |
| | | | | Q4 | **0.9708** | **0.9903** | **0.9854** | **0.9708** | **0.0449** |
| | La Línea | Best Classifier | Tree (Max. 4 splits) | Q1 | 0.9496 | **0.9878** | 0.9781 | **0.9635** | 0.0671 |
| | | | | Q2 | **0.9398** | 0.9710 | **0.9635** | 0.9124 | 0.1160 |
| | | | | Q3 | **0.9317** | **0.9829** | **0.9699** | **0.9487** | **0.0922** |
| | | | | Q4 | ***0.9707*** | ***0.9891*** | ***0.9845*** | ***0.9672*** | ***0.0479*** |
| | | Best ANN | Neural Network (10 × 10x10 neurons) | Q1 | **0.9234** | **0.9744** | **0.9616** | **0.9234** | **0.1178** |
| | | | | Q2 | **0.8638** | **0.9596** | **0.9352** | **0.8796** | **0.1972** |
| | | | | Q3 | **0.9167** | **0.9627** | **0.9516** | **0.8864** | **0.1535** |
| | | | | Q4 | **0.9460** | **0.9865** | **0.9763** | **0.9599** | **0.0726** |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

quartiles, but we assume that the fourth quartile is the most important for prediction as it represents the most dangerous concentration levels.

## 5 Conclusions

In this work, an experimental procedure using a resampling strategy with five-fold cross-validation allowed the statistical comparison of the different classification models tested. The proposed approach is based on classification modelling, since the desired output is the next level (quartile at *t* + *1*) of an air pollutant as a function of the other variables at a given time *t*. Two approaches have been used, one with hourly mean data for nitrogen oxides ($NO_2$, $NO_X$, and NO) and another one with daily mean data (for $SO_2$ and $PM_{10}$), due to the thresholds established in the European Directive 2008/50/EC, in order to obtain more reliable information in the study area. The approaches were developed in three different and separate locations: the main city of Algeciras, the city of La Línea, and the unspoilt remote area of Alcornocales, in order to contrast them and obtain more details on the behaviour of the air pollutants.

The main conclusions of this study are as follows:

- The classification models can be adequately used to provide very good air quality prediction results with quality indexes up to 90% in most cases.
- In general, the use of the ten relevant variables improves the results in most cases.
- Ensemble boosted trees, SVM, trees, and ANNs classifiers tend to be the best prediction models in most cases.

**Table 9** Best prediction model results for hourly $NO_2$ ($t + 1$) concentrations using all variables at $t$

| Pollutant | Location | Best Model/hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_1$ |
|---|---|---|---|---|---|---|---|---|---|
| $NO_2$ | Algeciras | Best Classifier | SVM (Medium Gaussian) | $Q_1$ | **0.8248** | **0.9405** | **0.9113** | **0.8243** | **0.2701** |
| | | | | $Q_2$ | **0.6513** | 0.8846 | **0.8270** | 0.6494 | **0.5365** |
| | | | | $Q_3$ | **0.6047** | **0.8700** | **0.8026** | 0.6132 | **0.6014** |
| | | | | $Q_4$ | *0.7349* | ***0.9100*** | ***0.8669*** | ***0.7271*** | ***0.4130*** |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | 0.8160 | 0.9383 | 0.9073 | 0.8180 | 0.2817 |
| | | | | $Q_2$ | 0.6368 | **0.8835** | 0.8213 | **0.6484** | 0.5487 |
| | | | | $Q_3$ | **0.5914** | 0.8596 | 0.7941 | 0.5769 | 0.6388 |
| | | | | $Q_4$ | 0.7270 | **0.9105** | 0.8647 | **0.7300** | 0.4168 |
| | Alcornocales | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | 0.7861 | **0.9373** | **0.8954** | **0.8276** | **0.3006** |
| | | | | $Q_2$ | **0.5950** | 0.8731 | **0.8031** | 0.6119 | 0.6079 |
| | | | | $Q_3$ | **0.6127** | **0.8654** | **0.8043** | **0.5922** | **0.6105** |
| | | | | $Q_4$ | **0.8123** | **0.9279** | **0.9014** | 0.7705 | **0.3206** |
| | | Best ANN | Neural Network (10 × 10 neurons) | $Q_1$ | 0.7794 | 0.9291 | 0.8885 | 0.8037 | 0.3235 |
| | | | | $Q_2$ | 0.5757 | 0.8634 | 0.7925 | 0.5791 | 0.6473 |
| | | | | $Q_3$ | 0.5880 | 0.8605 | 0.7932 | 0.5804 | 0.6388 |
| | | | | $Q_4$ | 0.7943 | 0.9282 | 0.8967 | 0.7732 | 0.3310 |
| | La Línea | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | **0.8265** | 0.9381 | **0.9099** | 0.8190 | 0.2736 |
| | | | | $Q_2$ | 0.6482 | **0.8871** | **0.8257** | **0.6651** | **0.5283** |
| | | | | $Q_3$ | **0.6160** | 0.8780 | **0.8125** | 0.6277 | **0.5797** |
| | | | | $Q_4$ | *0.8030* | ***0.9262*** | ***0.8967*** | ***0.7744*** | ***0.3253*** |
| | | Best ANN | Neural Network (10 × 10 neurons) | $Q_1$ | 0.8105 | 0.9332 | 0.9021 | 0.8048 | 0.2968 |
| | | | | $Q_2$ | 0.6312 | 0.8774 | 0.8155 | 0.6335 | 0.5652 |
| | | | | $Q\ _3$ | 0.6016 | 0.8691 | 0.8039 | 0.5968 | 0.6139 |
| | | | | $Q_4$ | 0.7765 | **0.9287** | 0.8904 | **0.7853** | 0.3364 |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

- The results obtained with ANNs are always improved by reducing the number of variables to the ten relevant ones.
- Variable selection models can be used to rank the importance of leverage variables.
- By selecting fewer variables, it is possible to design a more energy sustainable system with a lower carbon footprint.
- All forecasts can be useful to the citizens, institutions, businesses in the port area, and the cities surrounding the port.
- There is background radiation (averages that are constantly repeated) that does not provide useful or accurate information from the ships. The conclusion that can be drawn from the data is that we need more sensors close to the dock area where the ships are

**Table 10** Best prediction model results for hourly $NO_2$ $(t + 1)$ concentrations using top ten relevant features at $t$

| Pollutant | Location | Best Model/hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_1$ |
|---|---|---|---|---|---|---|---|---|---|
| $NO_2$ | Algeciras | Best Classifier | SVM (Medium Gaussian) | $Q_1$ | 0.8226 | 0.9388 | 0.9096 | 0.8190 | 0.2759 |
| | | | | $Q_2$ | 0.6437 | 0.8850 | 0.8244 | **0.6523** | 0.5403 |
| | | | | $Q_3$ | 0.5885 | 0.8698 | 0.7959 | **0.6170** | 0.6121 |
| | | | | $Q_4$ | **0.7450** | 0.9032 | 0.8662 | 0.7020 | 0.4255 |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | **0.8218** | **0.9411** | **0.9107** | **0.8263** | **0.2709** |
| | | | | $Q_2$ | **0.6511** | 0.8822 | **0.8259** | 0.6405 | **0.5434** |
| | | | | $Q_3$ | 0.5903 | **0.8677** | **0.7960** | **0.6085** | **0.6167** |
| | | | | $Q_4$ | **0.7377** | 0.9088 | **0.8672** | 0.7227 | **0.4143** |
| | Alcornocales | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | **0.7928** | 0.9305 | 0.8936 | 0.8066 | 0.3106 |
| | | | | $Q_2$ | 0.5853 | **0.8760** | 0.7999 | **0.6259** | **0.6061** |
| | | | | $Q_3$ | 0.6110 | 0.8650 | 0.8035 | 0.5913 | 0.6125 |
| | | | | $Q_4$ | 0.8118 | 0.9268 | 0.9005 | **0.7664** | 0.3244 |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | **0.7912** | **0.9327** | **0.8944** | **0.8134** | **0.3067** |
| | | | | $Q_2$ | **0.5990** | **0.8726** | **0.8045** | **0.6088** | **0.6069** |
| | | | | $Q_3$ | **0.6136** | **0.8687** | **0.8058** | **0.6048** | **0.6003** |
| | | | | $Q_4$ | ***0.8084*** | ***0.9312*** | ***0.9025*** | ***0.7823*** | ***0.3136*** |
| | La Línea | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | 0.8242 | **0.9389** | 0.9097 | **0.8216** | **0.2731** |
| | | | | $Q_2$ | **0.6487** | 0.8864 | 0.8256 | 0.6625 | 0.5297 |
| | | | | $Q_3$ | 0.6148 | **0.8780** | 0.8120 | **0.6277** | 0.5807 |
| | | | | $Q_4$ | **0.8033** | 0.9255 | 0.8963 | 0.7721 | 0.3269 |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | **0.8225** | **0.9393** | **0.9095** | **0.8231** | **0.2732** |
| | | | | $Q_2$ | **0.6500** | **0.8846** | **0.8254** | **0.6557** | **0.5337** |
| | | | | $Q_3$ | **0.6152** | **0.8755** | **0.8113** | **0.6182** | **0.5873** |
| | | | | $Q_4$ | **0.7925** | 0.9279 | **0.8947** | 0.7811 | **0.3274** |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

located in order to be able to deduce the direct effect of pollutants coming directly from the ships.

The logistical activity of a port has an impact on air quality. Therefore, it is necessary to implement predictive models to provide reliable forecasts that help citizens, companies and institutions, to make decisions and drive policy changes to ensure a healthier and cleaner environment for present and future generations.

**Table 11** Best prediction model results for hourly $NO_X$ ($t + 1$) concentrations using all variables at $t$

| Pollutant | Location | Best Model/hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_1$ |
|---|---|---|---|---|---|---|---|---|---|
| $NO_X$ | Algeciras | Best Classifier | SVM (Medium Gaussian) | $Q_1$ | **0.8178** | **0.9410** | **0.9100** | **0.8234** | **0.2756** |
| | | | | $Q_2$ | **0.6492** | **0.8841** | **0.8248** | **0.6543** | **0.5354** |
| | | | | $Q_3$ | **0.6103** | **0.8690** | **0.8045** | **0.6074** | **0.6011** |
| | | | | $Q_4$ | *0.7190* | ***0.9048*** | ***0.8590*** | ***0.7118*** | ***0.4370*** |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | 0.8153 | 0.9384 | 0.9076 | **0.8152** | 0.2838 |
| | | | | $Q_2$ | **0.6377** | 0.8831 | **0.8201** | 0.6531 | 0.5455 |
| | | | | $Q_3$ | 0.5954 | **0.8621** | 0.7965 | **0.5849** | **0.6296** |
| | | | | $Q_4$ | **0.7073** | 0.9017 | 0.8536 | 0.7028 | 0.4529 |
| | Alcornocales | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | **0.7885** | 0.9275 | **0.8907** | 0.7967 | 0.3213 |
| | | | | $Q_2$ | **0.5960** | 0.8768 | **0.8040** | 0.6287 | 0.5955 |
| | | | | $Q_3$ | 0.6115 | **0.8717** | **0.8091** | **0.6017** | **0.6021** |
| | | | | $Q_4$ | **0.7982** | 0.9218 | **0.8927** | 0.7586 | **0.3415** |
| | | Best ANN | Neural Network (10 × 10 neurons) | $Q_1$ | 0.7657 | 0.9304 | 0.8848 | **0.8078** | 0.3316 |
| | | | | $Q_2$ | 0.5851 | 0.8644 | 0.7960 | 0.5835 | 0.6369 |
| | | | | $Q_3$ | 0.5921 | 0.8643 | 0.7994 | 0.5776 | 0.6352 |
| | | | | $Q_4$ | 0.7871 | **0.9227** | 0.8901 | **0.7626** | 0.3460 |
| | La Línea | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | **0.7985** | 0.9321 | **0.8979** | 0.8021 | 0.3079 |
| | | | | $Q_2$ | **0.5740** | **0.8606** | **0.7897** | **0.5753** | **0.6523** |
| | | | | $Q_3$ | **0.5856** | **0.8584** | **0.7928** | **0.5673** | **0.6496** |
| | | | | $Q_4$ | 0.7628 | 0.9267 | 0.8849 | 0.7815 | 0.3502 |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | 0.7930 | 0.9307 | 0.8953 | 0.7981 | 0.3153 |
| | | | | $Q_2$ | **0.5828** | 0.8549 | 0.7916 | 0.5492 | 0.6646 |
| | | | | $Q_3$ | 0.5688 | 0.8625 | 0.7869 | 0.5891 | 0.6474 |
| | | | | $Q_4$ | 0.7617 | 0.9242 | 0.8830 | 0.7731 | 0.3574 |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

**Table 12** Best prediction model for hourly $NO_X$ ($t + 1$) concentrations using top ten relevant features at $t$

| Pollutant | Location | Best Model/hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_1$ |
|---|---|---|---|---|---|---|---|---|---|
| $NO_X$ | Algeciras | Best Classifier | SVM (Medium Gaussian) | $Q_1$ | 0.8171 | 0.9390 | 0.9085 | 0.8169 | 0.2812 |
| | | | | $Q_2$ | 0.6304 | 0.8818 | 0.8169 | 0.6502 | 0.5536 |
| | | | | $Q_3$ | 0.5929 | 0.8645 | 0.7962 | 0.5950 | 0.6242 |
| | | | | $Q_4$ | **0.7196** | 0.8999 | 0.8565 | 0.6945 | 0.4500 |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | **0.8238** | **0.9369** | **0.9091** | 0.8096 | **0.2821** |
| | | | | $Q_2$ | 0.6325 | **0.8853** | 0.8190 | **0.6622** | **0.5432** |
| | | | | $Q_3$ | **0.6058** | 0.8586 | **0.7992** | 0.5679 | 0.6343 |
| | | | | $Q_4$ | 0.7012 | **0.9076** | **0.8545** | **0.7243** | **0.4416** |
| | Alcornocales | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | 0.7842 | **0.9292** | 0.8903 | **0.8024** | **0.3204** |
| | | | | $Q_2$ | 0.5938 | **0.8785** | 0.8036 | **0.6355** | **0.5926** |
| | | | | $Q_3$ | **0.6150** | 0.8672 | 0.8088 | 0.5829 | 0.6135 |
| | | | | $Q_4$ | 0.7937 | **0.9221** | 0.8917 | **0.7600** | 0.3435 |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | **0.7855** | **0.9310** | **0.8918** | 0.8076 | **0.3154** |
| | | | | $Q_2$ | **0.6066** | **0.8735** | **0.8072** | **0.6132** | **0.5979** |
| | | | | $Q_3$ | **0.6075** | **0.8740** | **0.8084** | **0.6116** | **0.5979** |
| | | | | $Q_4$ | ***0.7982*** | *0.9224* | ***0.8931*** | *0.7604* | ***0.3400*** |

**Table 12** (continued)

| Pollutant | Location | Best Model/hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_1$ |
|---|---|---|---|---|---|---|---|---|---|
| | La Línea | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | 0.7948 | **0.9334** | 0.8975 | **0.8066** | **0.3073** |
| | | | | $Q_2$ | 0.5681 | 0.8598 | 0.7870 | 0.5737 | 0.6583 |
| | | | | $Q_3$ | 0.5854 | 0.8562 | 0.7920 | 0.5583 | 0.6565 |
| | | | | $Q_4$ | **0.7659** | **0.9270** | **0.8860** | **0.7821** | **0.3473** |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | **0.7961** | **0.9345** | **0.8986** | **0.8100** | **0.3037** |
| | | | | $Q_2$ | 0.5807 | **0.8577** | **0.7916** | **0.5614** | **0.6572** |
| | | | | $Q_3$ | **0.5822** | **0.8640** | **0.7930** | **0.5904** | **0.6353** |
| | | | | $Q_4$ | ***0.7753*** | ***0.9255*** | ***0.8880*** | ***0.7760*** | ***0.3447*** |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

**Table 13** Best prediction model results for hourly NO (*t* + *1*) concentrations using all variables at *t*

| Pollutant | Location | Best Model/hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_1$ |
|---|---|---|---|---|---|---|---|---|---|
| NO | Algeciras | Best Classifier | SVM (Medium Gaussian) | $Q_1$ | 0.8197 | **0.9425** | 0.9115 | **0.8286** | **0.2702** |
| | | | | $Q_2$ | **0.7101** | 0.8859 | **0.8402** | 0.6864 | **0.4700** |
| | | | | $Q_3$ | 0.5649 | **0.8693** | 0.7977 | **0.5708** | **0.6569** |
| | | | | $Q_4$ | **0.7298** | **0.9143** | **0.8679** | **0.7411** | **0.4060** |
| | | Best ANN | Neural Network (10 × 10 neurons) | $Q_1$ | 0.8103 | **0.9410** | 0.9077 | 0.8245 | 0.2807 |
| | | | | $Q_2$ | 0.6894 | 0.8836 | 0.8318 | 0.6829 | 0.4887 |
| | | | | $Q_3$ | **0.5745** | 0.8611 | **0.7993** | 0.5320 | 0.6780 |
| | | | | $Q_4$ | 0.7140 | 0.9188 | 0.8649 | 0.7583 | 0.4062 |
| | Alcornocales | Best Classifier | Ensemble (Bagged Tree) | $Q_1$ | **0.8302** | 0.8656 | **0.8410** | 0.9336 | **0.2768** |
| | | | | $Q_2$ | – | – | – | – | – |
| | | | | $Q_3$ | 0.3925 | 0.8733 | **0.8418** | 0.1788 | **1.0414** |
| | | | | $Q_4$ | ***0.7572*** | *0.9255* | ***0.8852*** | *0.7619* | ***0.3666*** |
| | | Best ANN | Neural Network (10 × 10 neurons) | $Q_1$ | **0.8291** | 0.8525 | 0.8364 | 0.9260 | 0.2884 |
| | | | | $Q_2$ | – | – | – | – | – |
| | | | | $Q_3$ | 0.3333 | **0.8696** | 0.8334 | **0.1563** | **1.0959** |
| | | | | $Q_4$ | **0.7451** | 0.9242 | 0.8808 | 0.7586 | 0.3785 |
| | La Línea | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | 0.8272 | **0.9502** | **0.9161** | **0.8644** | **0.2404** |
| | | | | $Q_2$ | 0.6791 | **0.8913** | **0.8430** | **0.6481** | **0.5131** |
| | | | | $Q_3$ | **0.6890** | 0.8834 | 0.8385 | 0.6392 | **0.5163** |
| | | | | $Q_4$ | 0.7422 | **0.9305** | **0.8807** | **0.7931** | **0.3582** |
| | | Best ANN | Neural Network (10 neurons) | $Q_1$ | 0.8152 | 0.9479 | 0.9108 | 0.8587 | 0.2545 |
| | | | | $Q_2$ | 0.6724 | 0.8839 | 0.8373 | 0.6208 | 0.5395 |
| | | | | $Q_3$ | 0.6708 | 0.8793 | 0.8308 | 0.6274 | 0.5389 |
| | | | | $Q_4$ | 0.7304 | 0.9292 | 0.8760 | 0.7902 | 0.3702 |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

**Table 14** Best prediction model results for hourly NO (*t* + *1*) concentrations using top ten relevant features at *t*

| Pollutant | Location | Best Model/hidden neurons | | Quartile/class | Sensitivity | Specificity | Accuracy | Precision | $d_1$ |
|---|---|---|---|---|---|---|---|---|---|
| NO | Algeciras | Best Classifier | Ensemble (RUSBoosted Tree) | $Q_1$ | **0.8238** | 0.9413 | **0.9119** | 0.8243 | 0.2704 |
| | | | | $Q_2$ | 0.6853 | **0.8917** | 0.8342 | **0.7099** | 0.4716 |
| | | | | $Q_3$ | **0.5740** | 0.8625 | **0.7995** | 0.5384 | 0.6731 |
| | | | | $Q_4$ | 0.7246 | 0.9132 | 0.8656 | 0.7381 | 0.4124 |
| | | Best ANN | Neural Network (25 neurons) | $Q_1$ | **0.8205** | 0.9393 | **0.9096** | 0.8181 | **0.2778** |
| | | | | $Q_2$ | **0.7103** | **0.8864** | **0.8405** | 0.6880 | **0.4687** |
| | | | | $Q_3$ | 0.5703 | **0.8653** | 0.7989 | 0.5518 | **0.6664** |
| | | | | $Q_4$ | ***0.7149*** | ***0.9202*** | ***0.8659*** | *0.7631* | ***0.4022*** |
| | Alcornocales | Best Classifier | Tree (Max. 4 splits) | $Q_1$ | 0.8081 | **0.8866** | 0.8296 | **0.9498** | **0.2850** |
| | | | | $Q_2$ | – | – | – | – | – |
| | | | | $Q_3$ | **0.8081** | **0.8866** | 0.8296 | **0.9498** | 1.1416 |
| | | | | $Q_4$ | 0.7282 | **0.9312** | 0.8791 | **0.7849** | **0.3734** |
| | | Best ANN | Neural Network (10 × 10x10 neurons) | $Q_1$ | 0.8202 | **0.8807** | **0.8376** | **0.9443** | **0.2758** |
| | | | | $Q_2$ | – | – | – | – | – |
| | | | | $Q_3$ | **0.3553** | 0.8635 | **0.8459** | 0.0853 | 1.1378 |
| | | | | $Q_4$ | 0.7338 | **0.9308** | **0.8808** | **0.7832** | **0.3700** |
| | La Línea | Best Classifier | Ensemble (Boosted Tree) | $Q_1$ | **0.8264** | 0.9499 | 0.9156 | 0.8636 | 0.2417 |
| | | | | $Q_2$ | **0.6782** | 0.8910 | 0.8425 | 0.6473 | 0.5144 |
| | | | | $Q_3$ | 0.6848 | **0.8836** | **0.8373** | **0.6409** | 0.5180 |
| | | | | $Q_4$ | **0.7436** | 0.9291 | 0.8805 | 0.7884 | 0.3603 |
| | | Best ANN | Neural Network (10 neurons) | $Q_{\ 1}$ | **0.8171** | **0.9520** | **0.9139** | **0.8703** | **0.2440** |
| | | | | $Q_2$ | **0.6942** | **0.8885** | **0.8461** | **0.6348** | **0.5128** |
| | | | | $Q_3$ | **0.6902** | **0.8864** | **0.8404** | **0.6500** | **0.5068** |
| | | | | $Q_4$ | ***0.7457*** | ***0.9323*** | ***0.8829*** | ***0.7987*** | ***0.3515*** |

In bold: the best model; in underline: with all/relevant variables; in italics: the best model for $Q_4$

**Table 15** The ten most relevant variables for each $SO_2$ (*t* + *1*) level prediction

| $SO_2$ Algeciras daily concentrations | | $SO_2$ Alcornocales daily concentrations | | $SO_2$ La Línea dail y concentrations | |
|---|---|---|---|---|---|
| Variable | Monitoring Station | Variable | Monitoring Station | Variable | Monitoring Station |
| $SO_2(t)$ | Algeciras | $SO_2(t)$ | Alcornocales | $SO_2(t)$ | La Línea |
| WD | $W_1$ (La Línea) | RF | $W_4$ (CEPSA 15 m high) | WD | $W_3$ (CEPSA 60 m high) |
| NO | Alcornocales | $NO_2$ | Alcornocales | NO | Carteya |
| $O_3$ | Cortijillos | CO | Escuela Hostelería | $NO_X$ | Alcornocales |
| $NO_2$ | Algeciras | WD | $W_1$ (La Línea) | $PM_{2.5}$ | Rinconcillo |
| $PM_{2.5}$ | Economato | NO | Los Barrios | $O_3$ | Alcornocales |
| $SO_2$ | Los Barrios | $SO_2$ | Algeciras | $NO_X$ | Carteya |
| $SO_2$ | Alcornocales | $O_3$ | Cortijillos | $NO_X$ | Rinconcillo |
| $SO_2$ | Palmones | NO | Alcornocales | $SO_2$ | Algeciras |
| $PM_{10}$ | Palmones | $SO_2$ | Madrevieja | RF | $W_1$ (La Línea) |

**Table 16** The ten most relevant variables for each $PM_{10}$ (*t* + *1*) level prediction

| $PM_{10}$ Algeciras daily concentrations | | $PM_{10}$ Alcornocales daily concentrations | | $PM_{10}$ La Línea dail y concentrations | |
|---|---|---|---|---|---|
| Variable | Monitoring Station | Variable | Monitoring Station | Variable | Monitoring Station |
| $PM_{10}$(t) | Algeciras | $PM_{10}$(t) | Alcornocales | $PM_{10}$(t) | La Línea |
| Tolueno | Puente Mayorga | WS | $W_4$ (CEPSA 15 m high) | RF | $W_4$ (CEPSA 15 m high) |
| WS | $W_4$ (CEPSA 15 m high) | $SO_2$ | Puente Mayorga | NOx | Rinconcillo |
| $PM_{10}$ | Alcornocales | $PM_{10}$ | Palmones | $PM_{10}$ | Alcornocales |
| $NO_2$ | Algeciras | $PM_{2.5}$ | Alcornocales | $PM_{10}$ | El Zabal |
| $PM_{10}$ | La Línea | $PM_{10}$ | Carteya | $PM_{2.5}$ | Economato |
| $PM_{10}$ | Carteya | RF | $W_4$ (CEPSA 15 m high) | $PM_{10}$ | Algeciras |
| $PM_{10}$ | Palmones | RP | $W_1$ (La Línea) | $PM_{10}$ | Carteya |
| $PM_{2.5}$ | Alcornocales | $PM_{10}$ | Los Barrios | $PM_{2.5}$ | El Zabal |
| $PM_{10}$ | El Zabal | $PM_{2.5}$ | Economato | $PM_{10}$ | Palmones |

**Table 17** The ten most relevant variables for each $NO_2$ (*t* + *1*) level prediction

| $NO_2$ Algeciras hourly concentrations | | $NO_2$ Alcornocales hourly concentrations | | $NO_2$ La Línea hourly concentrations | |
|---|---|---|---|---|---|
| Variable | Monitoring Station | Variable | Monitoring Station | Variable | Monitoring Station |
| $NO_2$(t) | Algeciras | $NO_2$(t) | Alcornocales | $NO_2$(t) | La Línea |
| WD | $W_1$ (La Línea) | WD | $W_3$ (CEPSA 60 m high) | RF | $W_2$ (CEPSA 10 m high) |
| $NO_X$ | Los Barrios | NO | Alcornocales | $NO_2$ | Hostelería |
| RF | $W_2$ (CEPSA 10 m high) | $NO_X$ | Los Barrios | Benceno | Campamento |
| $O_3$ | Algeciras | $PM_{2.5}$ | Alcornocales | $NO_x$ | Campamento |
| $PM_{10}$ | Rinconcillo | $O_3$ | Cortijillos | $PM_{2.5}$ | San Roque |
| $NO_2$ | Cortijillos | $NO_X$ | Alcornocales | $NO_X$ | El Zabal |
| $NO_X$ | Algeciras | $NO_X$ | Carteya | $NO_2$ | San Roque |
| WS | $W_4$ (CEPSA 15 m high) | $SO_2$ | Algeciras | $NO_X$ | Economato |
| WD | $W_3$ (CEPSA 60 m high) | RF | $W_1$ (La Línea) | $SO_2$ | Puente Mayorga |

**Table 18** The ten most relevant variables for each $NO_X$ (*t* + *1*) level prediction

| $NO_X$ Algeciras hourly concentrations | | $NO_X$ Alcornocales hourly concentrations | | $NO_X$ La Línea hourly concentrations | |
|---|---|---|---|---|---|
| Variable | Monitoring Station | Variable | Monitoring Station | Variable | Monitoring Station |
| $NO_X$ (t) | Algeciras | $NO_X$ (t) | Alcornocales | $NO_X$ (t) | La Línea |
| WD | $W_3$ (CEPSA 60 m high) | RF | $W_2$ (CEPSA 10 m high) | WD | $W_3$ (CEPSA 60 m high) |
| WS | $W_4$ (CEPSA 15 m high) | $O_3$ | Cortijillos | $NO_2$ | Hostelería |
| $NO_X$ | Los Barrios | $NO_2$ | Los Barrios | $NO_X$ | Economato |
| $PM_{10}$ | Palmones | NO | Alcornocales | HR | $W_2$ (CEPSA 10 m high) |
| $O_3$ | Algeciras | $PM_{10}$ | Los Barrios | NO | La Línea |
| $NO_2$ | Algeciras | $NO_2$ | Alcornocales | WD | $W_1$ (La Línea) |
| $NO_X$ | Cortijillos | $SO_2$ | Algeciras | $NO_2$ | La Línea |
| WD | $W_1$ (La Línea) | $NO_X$ | Carteya | $NO_X$ | Guadarranque |
| NO | Algeciras | $SO_2$ | Alcornocales | WD | $W_4$ (CEPSA 15 m high) |

**Table 19** The ten most relevant variables for each NO (*t* + *1*) level prediction

| NO Algeciras hourly concentrations | | NO Alcornocales hourly concentrations | | NO La Línea hourly concentrations | |
|---|---|---|---|---|---|
| Variable | Monitoring Station | Variable | Monitoring Station | Variable | Monitoring Station |
| NO(t) | Algeciras | NO(t) | Alcornocales | NO(t) | La Línea |
| RF | $W_2$ (CEPSA 10 m high) | $SO_2$ | Los Barrios | WS | $W_3$ (CEPSA 60 m high) |
| NO | Cortijillos | $PM_{10}$ | Carteya | CO | Hostelería |
| Benceno | Campamento | $NO_X$ | Madrevieja | WD | $W_1$ (La Línea) |
| $NO_2$ | Algeciras | NO | La Línea | $NO_X$ | El Zabal |
| SR | $W_4$ (CEPSA 15 m high) | $SO_2$ | Carteya | $PM_{2.5}$ | Economato |
| Tolueno | Cortijillos | $PM_{2.5}$ | Hostelería | NO | Madrevieja |
| $PM_{10}$ | Palmones | $SO_2$ | San Roque | $NO_X$ | La Línea |
| NO | Rinconcillo | $NO_2$ | Alcornocales | $SO_2$ | La Línea |
| $O_3$ | Algeciras | NO | Hostelería | SR | $W_2$ (CEPSA 10 m high) |

**Author contributions** Conceptualization, M.I.R.-G, C.R., and I.J.T.; data curation, M.I.R.-G, J.G.-E., and J.J.R.-A.; formal analysis, M.I.R.-G, J.G.-E. and I.J.T.; funding acquisition, I.J.T.; investigation, M.I.R.-G and J.G.-E..; methodology, M.I.R.-G, C.R., and I.J.T.; project administration, C.R., and I.J.T.; software, M. I.R.-G, J.G.-E., and I.J.T.; supervision, C.R., J.J.R.-A., and I.J.T.; validation, M.I.R.-G, C.R., and I.J.T.; visualization, M.I.R.-G. and C.R.; writing—original draft, M.I.R.-G, C.R., and I.J.T.; writing—review and editing, M.I.R.-G., C.R. and I.J.T. All authors have read and agreed to the published version of the manuscript.

**Funding** Funding for open access publishing: Universidad de Cádiz/CBUA. This work is part of the research project RTI2018-098160-B-I00 supported by 'MICINN. 'Programa Estatal de I+D+i Orientada a los Retos de la Sociedad' and besides, it is partially financed by national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/00006/2020. This research is supported by 'Plan Propio de la Universidad de Cádiz'. Data used in this work have been kindly provided by the Andalusian Regional Government.

## Declarations

**Conflict of interest** The authors have not got relevant conflicts of interest to declare to the content of this article.

## References

Adeyemi A, Molnar P, Boman J, Wichmann J (2022) Particulate matter ($PM_{2.5}$) characterization, air quality level and origin of air masses in an urban background in pretoria. Arch Environ Contam Toxicol 83(1):77–94. https://doi.org/10.1007/s00244-022-00937-4

Bai L, Wang J, Ma X, Lu H (2018) Air pollution forecasts: an overview. Int J Environ Res Public Health 15(4):780. https://doi.org/10.3390/ijerph15040780

Baklanov A, Zhang Y (2020) Advances in air quality modeling and forecasting. Global Transitions 2:261–270. https://doi.org/10.1016/j.glt.2020.11.001

Bishop CM (2006) Pattern Recognition and Machine Learning. Springer, Berlin

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Routledge, p 368. ISBN 978-0-412-04841-8. https://doi.org/10.1201/9781315139470

Breiman L (1996) Bagging predictors. Mach Learn 26:123–140

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

Corani G, Scanagatta M (2016) Air pollution prediction via multi-label classification. Environ Model Softw 80:259–264

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018

Durán-Grados V, Rodríguez-Moreno R, Calderay-Cayetano F, Amado-Sánchez Y, Pájaro-Velázquez E, Nunes RAO, Alvim-Ferraz M, Sousa S, Moreno-Gutiérrez J (2022) The influence of emissions from maritime transport on air quality in the strait of gibraltar (Spain). Sustainability 14(19):12507. https://doi.org/10.3390/su141912507

Durão RM, Mendes MT, Pereira JM (2016) Forecasting O3 levels in industrial area surroundings up to 24 h in advance, combining classification trees and MLP models. Atmos Pollut Res 7(6):961–970

Ekmekçioğlu AS, Levent K, Ünlügençoğlu K, Çelebi UB (2020) Assessment of shipping emission factors through monitoring and modelling studies. Sci Total Environ 743:140742. https://doi.org/10.1016/j.scitotenv.2020.140742

EU (2008) Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe

Fameli KM, Kotrikla AM, Psanis C, Biskos G, Polydoropoulou A (2020) Estimation of the emissions by transport in two port cities of the northeastern Mediterranean, Greece. Environ Pollut 257:113598. https://doi.org/10.1016/j.envpol.2019.113598

Fernando HJSF, Mammarella MC, Grandoni G, Fedele P, Marco RD, Dimitrova R, Hyde P (2012) Forecasting $PM_{10}$ in metropolitan areas: efficacy of neural networks. Environ Pollut 163:62–67. https://doi.org/10.1016/J.ENVPOL.2011.12.018

Fisher RA (1936) The use of multiple measurements in taxanomic problems. Ann Eugen 7(2):179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Freund Y (2009) A more robust boosting algorithm. Vol. 1. https://doi.org/10.48550/arXiv.0905.2138

García-Nieto PJ, Álvarez Antón JC, Vilán Vilán JA, García-Gonzalo E (2015) Air quality modeling in the Oviedo urban area (NW Spain) by using multivariate adaptive regression splines. Environ Sci Pollut Res 22:6642–6659. https://doi.org/10.1007/s11356-014-3800-0

García-Nieto PJ, Sánchez Lasheras F, García-Gonzalo E, de Cos Juez FJ (2018) $PM_{10}$ concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: a case study. Sci Total Environ 621:753–761. https://doi.org/10.1016/j.scitotenv.2017.11.291

Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS). Rectifier and softplus activation functions. The second one is a smooth version of the first. Journal of Machine Learning Research

González-Enrique J, Turias IJ, Ruiz-Aguilar JJ, Moscoso-López JA, Franco L (2019a) Spatial and meteorological relevance in $NO_2$ estimations: a case study in the Bay of Algeciras (Spain). Stoch Environ Res Risk Assess 33(3):801–815. https://doi.org/10.1007/s00477-018-01644-0

González-Enrique J, Turias IJ, Ruiz-Aguilar JJ, Moscoso-López JA, Jerez- Aragonés J, Franco L (2019b) Estimation of $NO_2$ concentration values in a monitoring sensor network using a fusion approach. Fresenius Environ Bull 28:681–686

González-Enrique J, Ruiz-Aguilar JJ, Moscoso-López JA, Urda D, Turias IJ (2021) A comparison of ranking filter methods applied to the estimation of $NO_2$ concentrations in the Bay of Algeciras (Spain). Stochastic Environ Res Risk Assessment 35(10):1999–2019. https://doi.org/10.1007/s00477-021-01992-4

Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning. Data mining, inference, and prediction, 2nd edn. Springer, New York

He H-d, Lu W-Z, Xue Yu (2014) Prediction of particulate matter at street level using artificial neural networks coupling with chaotic particle swarm optimization algorithm. Build Environ 78:111–117. https://doi.org/10.1016/j.buildenv.2014.04.011

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2(5):359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Hu L, Yan G, Duan Z, Chen C (2021) Intelligent modeling strategies for forecasting air quality time series: a review. Appl Soft Comput 102:106957. https://doi.org/10.1016/j.asoc.2020.106957

Ilacqua V, Hänninen O, Saarela K, Katsouyanni K, Künzli N, Jantunen M (2007) Source apportionment of population representative samples of $PM_{2.5}$ in three European cities using structural equation modelling. Sci Total Environ 384(1–3):77–92. https://doi.org/10.1016/j.scitotenv.2007.06.020

IMO (International Maritime Organization) (2021) The International Convention for the Prevention of Pollution from Ships (MARPOL), annex VI. London

Ju T, Lei M, Guo G, Xi J, Zhang Y, Xu Y, Lou Q (2023) A new prediction method of industrial atmospheric pollutant emission intensity based on pollutant emission standard quantification. Front Environ Sci Eng. https://doi.org/10.1007/s11783-023-1608-1

Kloog I, Ridgway B, Koutrakis P, Coull BA, Schwartz JD (2013) Long-and short-term exposure to $PM_{2.5}$ and mortality: using novel exposure models. Epidemiology 24(4):555–561

Kolehmainen M, Martikainen H, Ruuskanen J (2001) Neural networks and periodic components used in air quality forecasting. Atmos Environ 35(5):815–825. https://doi.org/10.1016/S1352-2310(00)00385-X

Kukkonen J, Partanen L, Karppinen A, Ruuskanen J, Junninen H, Kolehmainen M, Niska H, Dorling S, Chatterton T, Foxall R, Gavin C (2003) Extensive evaluation of neural networks models for the prediction of $NO_2$ and $PM_{10}$ concentrations, compared with a deterministic modelling system and measurement in central Helsinki. Atmos Environ 37:4539–4550. https://doi.org/10.1016/S1352-2310(03)00583-1

Lakra K, Avishek K (2022) A review on factors influencing fog formation, classification, forecasting, detection and impacts. Rendiconti Lincei-Scienze Fisiche e Naturali 33(2, SI):319–353

Liu H, Yan G, Duan Z, Chen C (2021) Intelligent modeling strategies for forecasting air quality time series: a review. Appl Soft Comput J 102:106957. https://doi.org/10.1016/j.asoc.2020.106957

Liu TK, Sheu HY, Tsai JY (2014) Sulfur dioxide emission estimates from merchant vessels in a Port area and related control strategies. Aerosol Air Quality Res 14(1):413–421. https://doi.org/10.4209/aaqr.2013.02.0061

López-Aparicio S, Tønnesen D, Thanh TH, Neilson H (2017) Shipping emissions in a Nordic port: assessment of mitigation strategies. Transp Res Part D Transp Environ 53:205–216. https://doi.org/10.1016/j.trd.2017.04.021

Lu G, Brook JR, Rami Alfarra M, Anlauf K, Richard Leaitch W, Sharma S, Wang D, Worsnop DR, Phinney L (2006) Identification and characterization of inland ship plumes over Vancouver, BC. Atmos Environ 40(15):2767–2782. https://doi.org/10.1016/j.atmosenv.2005.12.054

Lu H, Zhang Y, Wang X, He L (2014) A semiparametric statistical approach for forecasting $SO_2$ and $NO_X$ concentrations. Environ Sci Pollut Res 21(13):7985–7995. https://doi.org/10.1007/s11356-014-2748-4

Luna A, Talavera A, Navarro H, Cano L (2019) Monitoring of air quality with low-cost electrochemical sensors and the use of artificial neural networks for the atmospheric pollutants concentration levels prediction. Commun Computer Inf Sci 898:137–150. https://doi.org/10.1007/978-3-030-11680-4_15

Masood A, Ahmad K (2021) A review on emerging artificial intelligence (AI) techniques for air pollution forecasting: fundamentals, application and performance. J Cleaner Prod 322:129072. https://doi.org/10.1016/j.jclepro.2021.129072

Manisalidis I, Stavropoulou E, Stavropoulos A, Bezirtzoglou E (2020) Environmental and health impacts of air pollution: a review. Front Public Health. https://doi.org/10.3389/fpubh.2020.00014

MARPOL (Marine Pollution). Annex VI the International Convention for the Prevention of Pollution from Ships.

Martín ML, Turias IJ, González FJ, Galindo PL, Trujillo FJ, Puntonet CG, Gorriz JM (2008) Prediction of CO maximum ground level concentrations in the Bay of Algeciras, Spain using artificial neural networks. Chemosphere 70(7):1190–1195. https://doi.org/10.1016/j.chemosphere.2007.08.039

Mavroidis I, Gavriil I, Chaloulakou A (2007) Statistical modelling of CO and NO2 concentrations in the Athens area. Evaluation of emission abatement policies. Environ Sci Pollut Res 14(2):130–136. https://doi.org/10.1065/espr2006.04.299

Mclean S, Kaiser J, Ben Richard B (2019) A review of artificial neural network models for ambient air pollution prediction. Environ Model Softw 119:285–304. https://doi.org/10.1016/j.envsoft.2019.06.014

Méndez M, Merayo MG, Núñez M (2023) Machine learning algorithms to forecast air quality: a survey. Artif Intell Rev. https://doi.org/10.1007/s10462-023-10424-4

Menezes F, Popowicz GM (2022) Acid Rain and Flue Gas: Quantum Chemical Hydrolysis of $NO_2$. ChemPhysChem. https://doi.org/10.1002/cphc.202200395

Miola A, Ciuffo B (2011) Estimating air emissions from ships: Meta-analysis of modelling approaches and available data sources. Atmos Environ 45(13):2242–2251. https://doi.org/10.1016/j.atmosenv.2011.01.046

Mitchell T (1997) Machine learning. International Student Edition. McGraw-Hill, Maidenhead. ISBN: 0-07-115467-1, 414

Moreno-Gutiérrez J, Calderay F, Saborido N, Boile M, Rodríguez R, Durán-Grados V (2015) Methodologies for estimating shipping emissions and energy consumption: a comparative analysis of current methods. Energy 86:603–616. https://doi.org/10.1016/j.energy.2015.04.083

Moscoso-López JA, González-Enrique J, Urda D, Ruiz-Aguilar JJ, Turias IJ (2022) Hourly pollutants forecasting using a deep learning approach to obtain the AQI. Logic J IGPL. https://doi.org/10.1093/jigpal/jzac035

Mueller M, Westerby M, Nieuwenhuijsen M (2023) Health impact assessments of shipping and port-sourced air pollution on a global scale: A scoping literature review. Environ Res 216:114460. https://doi.org/10.1016/j.envres.2022.114460

Muruganandam NS, Arumugam U (2023) Dynamic ensemble multivariate time series forecasting model for $PM_{2.5}$. Comput Syst Sci Eng 44(2):979–989. https://doi.org/10.32604/csse.2023.024943

Oliveri G, Heibati B, Kloog I, Fiore M, Ferrante M (2017) A review of AirQ Models and their applications for forecasting the air pollution health outcomes. Environ Sci Pollut Res 24:6426–6445. https://doi.org/10.1007/s11356-016-8180-1

Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. https://doi.org/10.1109/TPAMI.2005.159

Pope CA, Dockery DW (2006) Health effects of fine particulate air pollution: lines that connect. J Air Waste Manag Assoc 56:709–742. https://doi.org/10.1080/10473289.2006.10464485

Prati MV, Costagliola MA, Quaranta F, Murena F (2015) Assessment of ambient air quality in the port of Naples. J Air Waste Manag Assoc 65(8):970–979. https://doi.org/10.1080/10962247.2015.1050129

Ramachandran P, Zoph B, Le QV (2017) Searching for activation functions. https://doi.org/10.48550/arXiv.1710.05941

Rodríguez-García MI, González-Enrique J, Moscoso-López JA, Ruiz-Aguilar JJ, Turias IJ (2022) Air pollution relevance analysis in the bay of Algeciras (Spain). Int J Environ Sci Technol. https://doi.org/10.1007/s13762-022-04466-4

Ribeiro VM, Gonçalves R (2022) Classification and prediction of nitrogen dioxide in a portuguese air quality critical zone. Atmosphere 13(10). In: 2nd international conference on cybernetics and intelligent system (ICORIS).

Ruiz-Aguilar JJ, Turias I, González-Enrique J, Urda D, Elizondo D (2020) A permutation entropy-based EMD–ANN forecasting ensemble approach for wind speed prediction. Neural Comput Appl. https://doi.org/10.1007/s00521-020-05141-w

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representation by error propagation. Parallel distributed processing: explorations in the microstructures of cognition, vol 1. MIT Press, Cambridge

Savouré M, Lequy E, Bousquet J, Chen J, de Hoogh K, Goldberg M, Vienneau D, Zins M, Nadif R, Jacquemin B (2021) Long-term exposures to $PM_{2.5}$, black carbon and NO2 and prevalence of current rhinitis in French adults: the Constances Cohort. Environ Int 157:106839. https://doi.org/10.1016/j.envint.2021.106839

Silverman BW, Jones MC (1989) E. Fix and J.L. Hodges (1951): an important contribution to nonparametric discriminant analysis and density estimation: commentary on Fix and Hodges (1951). Int Stat Rev 57(3):233–238. https://doi.org/10.2307/1403796

Song C, Fu X (2020) Research on different weight combination in air quality forecasting models. J Cleaner Prod 261:121169

Stieb DM, Burnett RT, Smith-Doiron M, Brion O, Shin HH, Economou V, Dales RE (2009) A new multipollutant, no-threshold air quality health index based on short-term associations observed in daily time-series analyses. J Air Waste Manag Assoc 59(3):299–307

Subramaniam S, Raju N, Ganesan A, Rajavel N, Maheswari Chenniappan M, Prakash C, Pramanik A, Basak AK, Dixit S (2022) Artificial intelligence technologies for forecasting air pollution and human health: a narrative review. Sustainability 14(16):9951. https://doi.org/10.3390/su14169951

Ting KM (2010) Confusion matrix. Encycl Mach Learn Data Min. https://doi.org/10.1007/978-1-4899-7687-1_50

Traina G, Bolzacchini E, Bonini M, Contini D, Mantecca P, Caimmi SME, Licari A (2022) Role of air pollutants mediated oxidative stress in respiratory diseases. Pediatr Allergy Immunol 22:38–40

Turias IJ, González FJ, Martin ML, Galindo PL (2008) Prediction models of CO, SPM and $SO_2$ concentrations in the Campo de Gibraltar Region, Spain: a multiple comparison strategy. Environ Monit Assess 143(1–3):131–146. https://doi.org/10.1007/s10661-007-9963-0

Vapnik VN, Chervonenkis A (1971) Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data. Avtomat i Telemekh 2:42–53

Yang L, Zhang Q, Lv Z, Zhang Y, Yang Z, Fu F, Lv J, Wu L, Mao H (2022) Efficiency of DECA on ship emission and urban air quality: a case study of China port. J Cleaner Prod 362:132. https://doi.org/10.1016/j.jclepro.2022.132556

Yeh CK, Lin C, Shen HC, Cheruiyot NK, Nguyen DH, Chang CC (2022) Real-time energy consumption and air pollution emission during the transpacific crossing of a container ship. Sci Rep 12:1. https://doi.org/10.1038/s41598-022-19605-7

## Authors and Affiliations

**M. I. Rodríguez-García**[1] · **M. C. Ribeiro Rodrigues**[2,3] · **J. González-Enrique**[1] · **J. J. Ruiz-Aguilar**[4] · **I. J. Turias**[1]

✉ M. I. Rodríguez-García
inma.rodriguezgarcia@gm.uca.es

M. C. Ribeiro Rodrigues
cribeiro@ualg.pt

J. González-Enrique
javier.gonzalezenrique@uca.es

J. J. Ruiz-Aguilar
juanjesus.ruiz@uca.es

I. J. Turias
ignacio.turias@uca.es

[1] Department of Computer Science Engineering, Algeciras School of Engineering and Technology (ASET), University of Cádiz, Algeciras, Spain

[2] Engineering Institute, University of Algarve, Campus da Penha, 8005-139 Faro, Portugal

[3] CEAUL - Centre of Statistics and Its Applications, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal

[4] Department of Industrial and Civil Engineering, Algeciras School of Engineering and Technology (ASET), University of Cádiz, Algeciras, Spain