ISSN 2237-8960 (Online)

RESEARCH PAPER

Prediction of external corrosion rate in Oil and Gas platforms using ensemble learning: a Maintenance 4.0 approach

Fernanda Ramos Elmas[1,2], Marina Polonia Rios[1], Eduardo Rocha de Almeida Lima[2], Rodrigo Goyannes Gusmão caiado[1], Renan Silva Santos[1]

[1]Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Gávea, Rio de Janeiro, RJ, Brazil.

[2]State University of Rio de Janeiro (UERJ), Rio de Janeiro, RJ, Brazil.

How to cite: Elmas, F. R. et al. (2023), "Prediction of external corrosion rate in Oil and Gas platforms using ensemble learning: a Maintenance 4.0 approach", *Brazilian Journal of Operations & Production Management,* Vol. 20, No. 3 special edition, e20231952. https://doi.org/10.14488/BJOPM.1952.2023

ABSTRACT

Goal: This study aims to use artificial intelligence, specifically a random forest model, to predict the annual corrosion rate on FPSO offshore platforms in the oil and gas industry. Corrosion is a significant cause of equipment failure, leading to costly replacements. The random forest model, a machine learning technique, was developed using climatic and other relevant data to forecast corrosion trends based on selected variables.

Design/methodology/approach: The methodology involved four steps: identifying influential factors affecting corrosion, selecting factors based on reliability and accessibility of measurements, applying the machine learning model to predict annual corrosion progression, and comparing the random forest model with other ML models.

Results - The results showed that the random forest regression model successfully predicted corrosion rates, indicating an average yearly increase of 2.43% on the analyzed platforms. The main factors influencing this increase were wind speed, percentage of measured corrosion, and platform operating time. Regions with higher incidence of these factors are likely to experience higher corrosion rates, necessitating more frequent maintenance.

Limitations of the investigation - The research sample consisted exclusively of 4 platforms located in the offshore region of Rio de Janeiro, Brazil. Thus, the results obtained must be interpreted as representative of these platforms and respective climate conditions.

Practical implications – The use of data science tools to improve corrosion management allows managers to have knowledge of which areas has a greater or lesser tendency to corrode, helping to prioritize maintenance activities over time.

Originality/value - This study aims to fill gaps regarding the use of random forest techniques for regression focused on predicting the rate of increase in corrosion. It offers a novel approach to assist decision-making in maintenance planning, providing insights into influential corrosion factors and facilitating more effective painting plans to preserve industrial unit integrity.

Keywords Corrosion; Maintenance Plan; Random Forest Regressor; Corrosion rate.

*This article is from Call for Papers, "When titans meet –Industry 4.0, Lean and Circular Economy Opportunities towards operational excellence and Sustainability". Due to their relevance and contribution to the field, the articles were published in the September/2023 Special Issue of the Brazilian Journal of Operations & ProductionManagement.

Financial support: None.
Conflict of interest: The authors have no conflict of interest to declare.
Corresponding author: fernandaelmas@tecgraf.puc-rio.br
Received: 04 August 2023.
Accepted: 17 August 2023.
Editor: Daniel Nascimento, Guilherme Tortorella, Rodrigo Goyannes Gusmão Caiado, Juan Manuel Maqueira.

1. INTRODUCTION

In the oil and gas (O&G) industry, maintenance costs represent 40% of total costs, and most of these costs result from inadequate or unscientifically based planned maintenance activities (Mobley, 2002). Maintenance shutdowns, for example, are an important strategic process for maintenance planning within an industrial plant, being essential in the management of equipment that requires more prolonged and in-depth inspections (Caiado et al., 2015). Among the various degradation mechanisms of a platform, corrosion is one of the main causes of equipment failure (Wu et al., 2013). This can also lead to safety risks, which in turn involve increased costs (Muniz et al., 2018)

From the perspective of offshore facilities, corrosion is the primary factor that affects the longevity and reliability of assets, consuming up to 80% of the total maintenance cost in the O&G exploration industry (Koch et al., 2002). The corrosion process is a complex physical-chemical phenomenon that is influenced by various factors (environmental and climatic conditions, material characteristics and compositions, among others) (Mishra et al., 2019). From the point of view of asset management (Caiado et al., 2022; Lima et al., 2023; Nascimento et al., 2019), many of the damages that affect the equipment's integrity over a project are related to corrosion (Dawson et al., 2010). The worldwide market for corrosion monitoring equipment (excluding inspection) is estimated to be around 25 million dollars, including auxiliary accessories and associated tools (Britton, 1990). According to Britton (1990), corrosion monitoring usually has one or more of the following objectives: i) diagnose corrosion problems in operational equipment; ii) monitor and control the effectiveness of corrosion risk mitigation processes; iii) facilitate shutdown scheduling.

In general, the information necessary for the evaluation of corrosion includes a list of items of interest in the asset register (structures, vessels, pipelines, storage tanks, etc.), historical data (inspection, monitoring, maintenance), theoretical analysis (new data-based systems/models), and informed opinions (Dawson, 2010). Risk classification tools aim to focus attention on critical areas, allowing an assessment team to concentrate on items in a plant or process that have varying levels of corrosion risks. Corrosion rate data, models, or field information can even be used to assess the risk of corrosion in an asset or component (Dawson et al., 2010).

In the oil and gas industry, in addition to the inspection of external corrosion, several other activities must be closely monitored to maintain the integrity and safety of units. In this context, Maintenance 4.0, which extends predictive maintenance to use data science techniques such as Machine Learning is of utmost importance(Orrù et al., 2020) . (Sircar et al., 2021) report that the use of artificial intelligence in the oil and gas industry is rapidly advancing precisely because it infiltrates various areas, from planning, exploration, and production, with the main objective of refining the development plan using historical data.

Since the development of data analysis techniques, machine learning, which is a type of artificial intelligence, has shown significant advantages in data modeling and mining(de Paula Vidal et al., 2022a) and can be combined with other techniques, such as nonparametric models or multicriteria decision aid (da Cruz et al., 2022; Kaiser et al., 2022). Machine learning-based methods have also been proposed in the field of corrosion research, under controlled conditions (Diao et al., 2021). For example, (Wei et al., 2021) established a relationship model between the corrosion potential of low-alloy steel and its influencing factors through an artificial neural network and visualized the influence of various alloying elements on the corrosion potential. (Lv et al., 2020) quantified the steel's forming parameters, such as input features, and applied the methods to accurately predict the sectional corrosion rate.

The emerging application of machine learning in the field of corrosion offers a variety of tools to categorize and prioritize the influence of certain parameters that affect its progression, allowing for more effective predictions of corrosive processes. In an initial study, (Cai et al., 2018) built a machine learning model to predict corrosion of steel and zinc. The results showed that the developed model could account for predicting over 70% of the variation in corrosion data, outperforming linear regression models developed using the same dataset.

Few models using machine learning for prediction conduct a directed study on how a certain set of factors can influence the rate of corrosion over time for the oil and gas industry, in field conditions. In this environment, corrosion is greatly accelerated and directly depends on the marine atmosphere, which enhances the corrosion process due to its environmental conditions. Lyon (2010) states that while corrosion begins to occur in the earth's atmosphere from a relative humidity of 75%, in marine atmospheres, it begins to occur from 33% relative humidity. Therefore, the use of machine learning prediction models for predicting corrosion and how the factors present in an oil and gas platform directly impact its progression is of fundamental importance.

The random forest model, a machine learning model that can be used to create estimates and predictions through a process that aggregates information over a series of decision trees, implies less overfitting than a single decision tree model. However, unlike tree models that are easy to visualize, Random Forest models are not easily visualized but can produce an importance ranking for each possible predictor and can be easily displayed graphically (Buskirk, 2018). Like tree methods, the random forest model can handle predictors that are continuous, categorical, skewed or sparse data; missing data must be handled before applying the model (Strobl et al., 2007). It can also be very effective at estimating results that are complex functions of data with many interactions (Mendez et al., 2008).

The application of the Random Forest model in the field of machine learning has a wide range of research conducted for various areas, such as the medical field, where, for example, Parameswari et al.(2022) conducted studies related to Alzheimer's disease with Random Forests for classification, or studies related to pollution, where (He et al., 2022) conducted studies using Random Forests for classification as well. Nevertheless, few studies use the Random Forest model for regression related to the study of corrosion progression, providing a numerical value of the corrosion rate, as shown by Diao et al. (2021), in the article about improving the corrosion rate prediction and Zhi et al. (2021) about improving corrosion prediction, taking into account environmental factors. Thus, it is evident that knowledge of the behavior of corrosion progression over time is quite relevant information because it allows for the development of a more assertive and realistic painting plan for each oil and gas platform. Understanding how the factors that influence corrosion progression operate is crucial for maintaining the integrity of the platforms, minimizing maintenance costs.

In this context, the purpose of this study is to develop a model based on the random forest to predict the annual progression of the corrosion rate and evaluate which of these identified factors had the most influence on this result. To this end, the specific objectives of this paper consist of: (i) identifying the main factors that can influence the behavior of corrosion in FPSO offshore platforms; (ii) selecting factors based on the availability/accessibility of measurements and the degree of reliability/precision of these measurements; (iii) proposing and applying a machine learning model that can predict the annual progression of corrosion in FPSO offshore platforms from the selected factors; and (iv) comparing the Random Forest model with the XG Boost model. In addition, it also aims to provide data to assist companies that plan maintenance focused on mitigating corrosion, promoting understanding of how influencing factors in corrosion act so that it is possible to maintain the integrity of industrial units and develop more accurate painting plans.

2. BACKGROUND

In general terms, Machine Learning (ML) is the evolution of computational algorithms that can mimic human intelligence through learning provided by the environment (El Naqa and Murphy, 2015). ML emerges from the intersection of computer science and statistics and is at the center of artificial intelligence and data science. It is notably one of the technically relevant fields nowadays, given its usefulness in various areas, such as marketing, investments, manufacturing, and telecommunications, for example (Fayyad et al., 1996).

Machine learning systems are generally divided into three categories: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the learning content usually includes labeled inputs and outputs and predicts new outputs from new inputs. In unsupervised learning, the algorithm does not learn from labeled data; it tries to find patterns in the dataset. Reinforcement learning is closer to supervised learning. The difference is that the reinforcement learning program does not learn from labeled output but gives feedback on the decision (Rajendra et al., 2022). Figure 1 shows a generic step-by-step description of a machine learning model construction described by (Vieira et al., 2020), as follows:

- Formulation, which consists of defining the problem and the guiding questions of the study;
- Preparation of the database, which consists of selecting and processing the available data of the problem (e.g., transforming numeric variables into categorical and vice versa);
- Creation of variables, at this stage, new variables (or features) are generated from the available data;
- Division of the data into subsets of training, testing, and validation;
- Training, this stage is dedicated to finding relationships between the variables being modeled from the selected available data for training;
- Model validation, this stage is dedicated to evaluating the model's performance from the selected available data for model validation, therefore seeking to identify the accuracy of the model;
- Analysis, which consists of analyzing the result obtained through pre-established metrics.

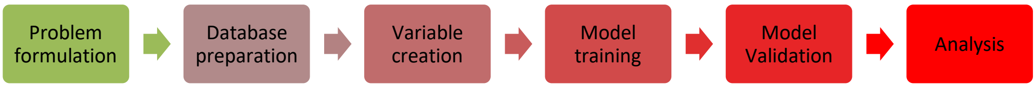


Figure 1 - Machine learning steps. Adapted from (Vieira et al., 2020)

In summary, supervised machine learning can be operationalized through several techniques, the most commonly used being classification and regression. For each of these techniques, there are specific calculation rules associated with them, which are called algorithms. Therefore, the choice of an algorithm is a crucial part of the machine learning process. There are many types of learning algorithms, some of the most common being Naive Bayes, Decision Trees, Random Forests, Artificial Neural Networks, Support Vector Machines (SVM), among others (Rajendra et al., 2022).

The algorithms known as decision trees and random forests are highlighted in the present study. According to (Gupta et al., 2017), the decision tree model is very intuitive, so that in the tree-building process, all features are traversed and the feature with the highest importance is selected as the splitting node until there are no remaining features in the dataset for further splitting. Thus, the impurity from the root node to the leaf node gradually decreases. Analysis using the decision tree model is based on the construction of trees that can be used to visually and explicitly represent decisions. In order to minimize the high variance generated in the data from the decision tree method, the Random Forest method is used, which considers a set of different decision trees for calculation.

Random Forest (RF) is a method that uses a combination of many decision trees and selects the average prediction of the results as the output variable (Breiman, 2001). The method involves training different decision trees on separate generated data samples, combining the learning of these trees to produce the final result. In the decision-making process, the trees do not interact with each other during training. Therefore, different samples are generated to train the decision trees, dividing a set of features into several subsets while keeping a percentage of features in each subset. This step ensures that the model does not depend on any individual feature and uses all features productively (Liaw and Wiener, 2002).

The use of machine learning and random forest models is growing in the determination of corrosion rates over time, as demonstrated by Diao et al. (2021) in their article on improving corrosion rate prediction and Zhi et al. (2021) on improving corrosion prediction while considering environmental factors.

In the article by Diao et al. (2021), marine corrosion data of low-alloy steels were collected and prediction models for corrosion rates were developed using machine learning algorithms. According to them, both the chemical composition of the low-alloy steel and environmental factors were used as input features, and the Random Forest algorithm was used for modeling and determining corrosion rates. Thus, two methods for creating features were proposed to convert the information of the steel's chemical composition into a set of atomic and physical property features.

As a result, the developed method created a model no longer limited to materials with specific chemical compositions. Therefore, machine learning-based corrosion rates showed good accuracy in predicting corrosion rates. The study improved the generalization ability of the model and proved the feasibility of machine learning in evaluating corrosion behavior.

In the article by Zhi et al. (2021), models for atmospheric corrosion prediction were constructed based on the corrosion rates of carbon steel and twelve environmental factors from long-term exposure tests. A hybrid method combining the Random Forest model and Spearman Correlation was used, compared with the maximum information coefficient (MIC) and principal component analysis (PCA). Then, the support vector machine (SVM) method was applied using the identified key environmental factors, presenting higher accuracy than those with dimensionality reduction by MIC and PCA. Dimensionality reduction also significantly improved the accuracy and generalization of the SVM model.

Table 1 - Literature review of factors influencing the advancement of corrosion

Factors		References
Influence of time	Age of the component	(American Petroleum Institute, 2009; Wu et al., 2013)
	Failure rate	(American Petroleum Institute, 2009)
	Maintenance history	(American Petroleum Institute, 2016)
Element information	Component geometry	(American Petroleum Institute, 2009)
	Dissimilar metals	(St. Clair and Sinha, 2014)
	Material	(American Petroleum Institute, 2016, 2009; Cai et al., 2018; Dawson, 2010; Wu et al., 2013)
	Piece orientation	(Lyon, 2010)
	Equipment function	(Dawson, 2010)
	Acting stresses	(American Petroleum Institute, 2016; St. Clair and Sinha, 2014)
	Operating temperature	(American Petroleum Institute, 2009; Wu et al., 2013)

	Surface temperature	(American Petroleum Institute, 2016; Macha et al., 2019)
	Production interruption	(Wu et al., 2013)
Information about the corrosive	Ambient temperature	(Cai et al., 2018; Lyon, 2010)
	Air contaminants	(Bento et al., 2009; Cai et al., 2018; Lyon, 2010; Macha et al., 2019)
	Humidity	(Bento et al., 2009; Cai et al., 2018; Lyon, 2010; Macha et al., 2019)
Information about the fluid	Pressure	(American Petroleum Institute, 2016, 2009; Dawson, 2010; Wu et al., 2013)
	Potential energy	(Dawson, 2010))
	Flow rate	(American Petroleum Institute, 2009)
	Fluid composition	(American Petroleum Institute, 2016, 2009)
	Fluid physical state	(American Petroleum Institute, 2009)
Painting	Painting area	(Cho, 2020)
	Surface roughness	(Bento et al., 2009; Nor Asma et al., 2011)
	Paint type	(American Petroleum Institute, 2009)
Location	Lighting condition	(Chen et al., 2010)
	Sun exposure	(American Petroleum Institute, 2009)
	Wind incidence	(Lyon, 2010)
	Equipment access	(Dawson, 2010; Wu et al., 2013)
	Influence area (other equipment)	(American Petroleum Institute, 2009; Dawson, 2010)
	Influence area (personal safety)	(American Petroleum Institute, 2009; Dawson, 2010)

	Influence area (environmental impact)	(American Petroleum Institute, 2009; Dawson et al., 2010)
	Corrosion position	(Lyon, 2010)
Cost	Equipment value	(American Petroleum Institute, 2009; Dawson, 2010)

Source: the authors themselves.

3. METHODOLOGY

Building a model to determine the progression of corrosion over time requires a thorough study of the main factors that influence this behavior, as recommended in ISO 9223 (British Standards Institution, 2012). This standard state that the nature and speed with which corrosion occurs in metals, alloys, and coatings, among other factors, also depend on the properties of the electrolytes formed on the surface, particularly with respect to the level and type of gaseous and particulate pollutants in the atmosphere and the duration of their action on the metal surface.

Based on the mentioned issues, a case study was carried out using data from four ship-type offshore platforms, whose design configurations were similar, meaning that the overall arrangement of all modules was the same. This study consisted of the development an experiments using the Random Forest model, and a comparison was between the model used and the XG Boost model.

3.1 Data acquisition and description

To conduct the study, it was necessary to survey several factors that influence the progression of corrosion, shown in Table 1. In addition, as illustrated in Table 2, interviews were also conducted with maintenance professionals with a focus on corrosion, who reported the main factors considered in the study, such as those that have the greatest impact on the progression of corrosion.

Table 2 - Experts who participated in the interview about factors that influence the rate of corrosion

Experts	Position	Experience
Expert 1	Paint Inspector	Has been working as a paint inspector for more than 5 years
Expert 2	Paint inspector	Has been working for more than 5 years as a paint inspector
Expert 3	Planning Technician	Has worked for more than 5 years planning painting works
Expert 4	Equipment Inspector	Has worked for more than 5 years as an equipment inspector
Expert 5	Planning Technician	Has worked for more than 5 years as service planning and management

Expert 6	Petroleum Technologist	Has been working for more than 5 years with technical engineering documentation
Expert 7	Chemical Engineer	Has been working for more than 5 years in corrosion management
Expert 8	Civil Engineer - Consultant	Has worked for more than 5 years as coordinator in the inspection of naval structures
Expert 9	Planning Engineer	Has been working for more than 5 years in corrosion management
Expert 10	Manager	Has been working for more than 5 years managing maintenance teams
Expert 11	Electrical Engineer - Coordinator	Has been working for 2 years as coordinator in the painting area
Expert 12	Oceanographer	Has been working for more than 5 years in the climatic factors monitoring area
Expert 13	Economist - Consultant	Has been working for 2 years in the area of costs related to corrosion

Source: the authors themselves.

The factors listed in Table 1 were validated via interviews with the experts listed in table 2. However, not all of them could be easily obtained in practical terms, so that only a few were actually considered for the study. One of the main reasons for this difficulty was because of the lack of historical data records for all the variables surveyed at the time of painting.

The group of factors considered were: (i) ambient temperature, relative humidity, wind speed, and preferential wind direction - obtained through a corporate portal that monitors these information hourly for each platform - , (ii) platform, id (location data on the platform, represented by module and sector), and system - specific design data for each platform that are obtained through their general arrangements -, (iii) and inspection year, percentage of corrosion, and annual corrosion rate - data collected from annual inspections conducted through ships and are collected by painting system.

The information on the percentage and advancement of corrosion was obtained at the system level, which is the smallest subdivision of the platform (platform > module > sector > system), while the data obtained from the corporate portal, such as ambient temperature, relative humidity, and wind speed, were generally obtained for the platform. Therefore, for all systems on a platform, it was considered that the data on ambient temperature, relative humidity, and wind speed remained constant throughout the year for the same platform.

Regarding the design data of the platform, it was only necessary to consult the general arrangement of each one of them, because in this drawing it is indicated the division of the existing modules and sectors within a platform. In addition to the arrangement data, the names of the systems were also considered, because according to the interviews conducted with the experts cited in Table 2, they corrode differently, either by their conformation, which causes difficulty in ensuring that the paint film can fill the part completely, as is the case of supports, valves and flanges, or because it is a region that has a greater tendency to suffer mechanical damage, as is the case of the loading floor

Regarding the inspection data, all four platforms considered in the study have data on the percentage of corrosion, which were obtained through annual inspections from 2014 to 2018, following the physical division of the platform. Finally, since the predicted variable is the corrosion advancement, the difference between the corrosion in the inspection year and the previous year was calculated to list how much the corrosion has advanced over the year. Below is a summary of the information that was used as input data for the model:

- Platform: classified as categorical variables such as Platform A, Platform B, Platform C, and Platform D, all of them FPSO type;
- Id: classified as categorical variables indicating the location of each module on the platform;
- System: classified as categorical variables, following the following classification: ceiling, floor, support, metal structures, handrail, stairs, equipment, bulkhead, and PVF (pipes, valves, and flanges);
- Inspection year: are the years that have information on corrosion inspection, namely 2014, 2015, 2016, and 2017;
- Percentage of corrosion: varies from 0 to 100%;
- Corrosion advancement: is calculated from the percentage of corrosion, so that in the first year considered in the study (2014), all advances were 0, and from then on, in the following years, the advancement was calculated as follows: $advance = advance_{year\ i+1} - advance_{year\ i}$. It is important to note that this information was only present in case 04 of the case study;
- Ambient temperature, Wind speed, and Relative humidity: to determine this data on each platform over the year, a corporate portal was consulted, which stores the historical data, whose information was collected on the first day of each month per year, and the values considered were the average of the annual data;
- Wind direction: for the use of this information in the model, the wind directions were divided according to the compass rose into four classes: North, West, East, and South, and this information was collected on the first day of each month per year, and the value considered for analysis was the direction that had the highest frequency in that year.

3.2 Machine learning approach and analysis

The research used the supervised ML approach with ensemble learning, which uses multiple learning algorithms to achieve better prediction performance compared to using any single learning algorithm. Supervised learning can be seen as a search in a hypothesis space to find a suitable hypothesis that will make good predictions for a specific problem. The ensemble learning model used in this study was the random forest (RF) method, which is a popular method used to generate classification and regression models (Zhao et al., 2022).

The RF method constructs a multitude of decision trees using the training set and produces the mode of the class (for a classification problem) and the average prediction of the individual trees (for a regression problem). The goal of reducing variance comes at the cost of a small increase in bias and some loss of interpretability. However, this greatly increases the final model performance and also corrects the problem of overfitting, which is a common occurrence in decision trees. The general bootstrap or bagging aggregation technique is implemented in the training algorithm.

Breiman (2001) introduced the concept of RF models, comprising multiple decision trees that are randomly selected and combined to produce a more accurate prediction output. The aim of the RF model is to reduce the variance of bagging by minimizing the correlation between the trees without increasing the variance too much, which is achieved mainly through the random selection of features, also known as input variables or features, mentioned when splitting the database. Another advantage of RF is its ability to measure the importance of these variables, which is calculated by how much each one contributes to reducing the variance.

In RF applications for regression (random forest regressor - RFR), bootstrap samples are created from the original dataset to build a network of decision trees, whose outputs are calculated to form the estimate of the response variable. Random subsets of predictor variables are used to split each decision node, which, in addition to the randomness of the bootstrap sampling, leads to better predictive performance. In addition, the large number of decision trees (n = 100 here) created in each model avoids overfitting, which means that the ML model can get a good fit effect by training data in the database but cannot fit data well by training data outside the database. To avoid the overfitting phenomenon, test sets and cross-validation tests are introduced. The proportion of training set, cross-validation test, and test set can be changed according to ML and R^2 models.

R^2 is employed to evaluate the predictive ability of various ML models ($0 \leq R^2 \leq 1$), and represents the quality that an ML model trains its own dataset, as the ML models need to be used to predict new data. The model error is quantified by the average of the prediction results of all trees using out-of-bag (OOB) data (i.e., data not included in the bootstrap sample). The importance of the predictor variable can be determined by permuting a particular variable in the OOB data (and keeping all others constant) and calculating the increase in root mean square error (RMSE) compared to the original data.

Initially, there was the generation of the database, consisting of five stages: platform selection, historical data collection, calculation of annual progression, identification and selection of factors that affect corrosion, and input of data related to the factors. After data collection was completed, the next step was to preprocess the dataset to transform raw data into a comprehensible format. The dataset contained both categorical and continuous variables that required additional processing to be converted into a meaningful and standard form. Subsequently, the dataset was split into training and testing sets using an 80% vs. 20% ratio (Mokarian et al., 2022).

Because different machine learning models exhibit various advantages and different predictive abilities on different datasets, even for a similar dataset, due to their respective characteristics, it is important to train the dataset with the appropriate ML model (de Paula Vidal et al., 2022b; Kaiser et al., 2022; Tang et al., 2022). In this research, RFR was selected because it provides not only predictive models, but also a deeper understanding and valuable information about the relative importance of different variables that affect overall accuracy. The RFR model provides a fair assessment of importance among parameters. The importance score of the input features (independent variables) calculated by the RFR is represented; it defines the importance of independent variables in estimating the dependent variable. The importance score indicates the predictive power of the parameters. Evaluating the feature importance score helps maximize the efficiency of a predictive model by providing sufficient understanding of the data and the model, as well as providing useful information necessary for dimensionality reduction and feature selection (Mokarian et al., 2022).

For the case studied, data was preprocessed, varying the encoding of categorical variables and performing normalization. Also, two methodologies for optimizing hyperparameters related to the Random Forest algorithm were also analyzed, namely, Grid Search and eXtreme Gradient Boosting (XGBoost),in addition to evaluating the differences in the model's behavior when evaluating the entire dataset together or separately for each platform. Finally, an evaluation of the importance of factors related to corrosion progression was conducted in order to understand the behavior of corrosion over time.

For the data preprocessing steps, algorithm training, and results evaluation, the Scikit Learn library (Pedregosa et al., 2012) developed in Python 3 (van Rossum and Drake, 2009) was used. A computer with an Intel(R) Core(TM) i5-8400 CPU and 16GB of RAM was used for testing purposes.

4. CASE STUDY

4.1 Case description

A major bottleneck in the planning process for painting the operational units is the prediction of corrosion behavior over time. Currently, an average advancement value is considered for the entire platform, depending solely on the type of corrosion, 0.5% in cases of mild or moderate corrosion, and 3% in cases of generalized corrosion. However, this model is not consistent with the reality of an operational unit, as it does not consider the specificities of each region of the platform, subject to different climatic conditions, with diverse function and location characteristics.

The oil company used in this study has been systematically visual inspecting the percentage of corrosion since 2016, presenting a vast history of data indicating the average percentage of corrosion by region of the platform, divided into module/sector/system. Therefore, it was decided to explore this data history to generate an artificial intelligence model capable of predicting the advancement of corrosion over time for each system of the platform, considering the inspection data history carried out over the years and factors that influence the behavior of corrosion, such as climate, location, and function factors, raised in a technical report from the company that operates the platform (Tecgraf Institute/PUC-Rio, 2021).

This study can be classified as a regression problem, in which a specific value is desired to be estimated based on various factors. In the literature, there are various algorithms for regression, such as Naive Bayes, Decision Trees, Random Forests, Artificial Neural Networks, among others (Rajendra et al., 2022). In this study, the random forest (RF) method (Breiman, 2001) was chosen, as mentioned in the methodology section.

A case study considering four platforms was conducted. It is worth noting that although the model was limited to evaluating FPSO platforms, the methodology developed in this study can be replicated for other platform models.

4.2 Experiment

The case study conducted contemplated an experiment, with the main objective to to compare the model that was applied from Random Forests with XG Boost.

Based on the factors that influence corrosion, as described in the previous section, a database was built to predict the advance of corrosion. The predicted variable was the advancement of corrosion. A variability in one of the factors that influence the advance of corrosion was used, improving the accuracy of the model versus initial tests conducted. Thus, in the following items the experiment that comprised the case study will be presented.

One factor was crucial to the increase in performance metrics: the inclusion of annual advance information and its use for corrosion prediction. The variable predicted is the advancement of corrosion, using 2014 as the base year (e.g. advance = 0%), and the percentage of corrosion column served only as a variable that impacts the advance of corrosion, as do all the others. With the results obtained, other tests were made considering this same database. The Grid Search function was run, which aims to find the best hyperparameters of the sklearn.ensemble.randomforestregressor function in order to obtain better performance.

It is important to note that in preliminary tests, the predicted variable was the percentage of corrosion, which did not bring such good results. In the experiment shown in this work, the predicted variable became the corrosion progression, which brought better results. Other tests were performed, however the results shown in this paper represent well the main milestones in the construction and application of the Random Forest model.

Finally, a comparative analysis between the Random Forests and the XG Boost models was performed. To obtain the best performance of the Random Forests model, the Grid Search function was run with the data from the experiment, aiming to obtain the best sckitlearn hyperparameters. Then, the experiment was run again, but with the new hyperparameters, and a comparison with the XG Boost model was performed

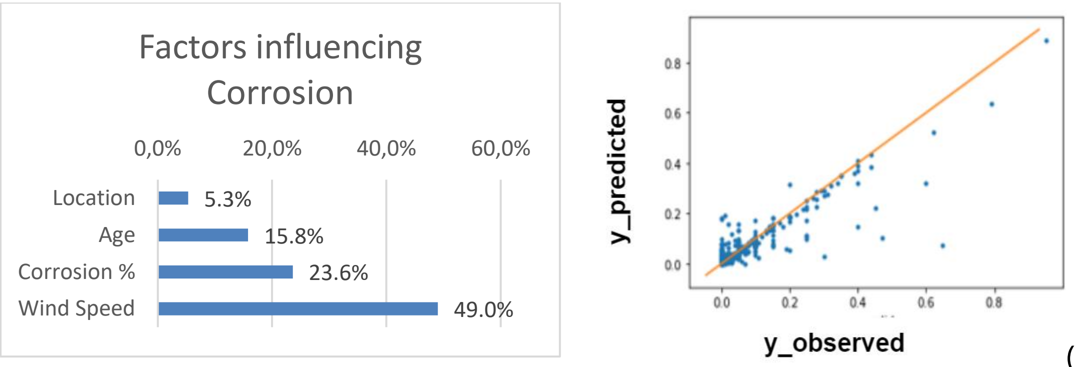
5. RESULTS AND DISCUSSION

In the experiment, data contained no missing data and a data proportion of 80% for testing and 20% for validation. The predicted value, as stated in previous sections, was the corrosion advancement. This was made precisely to determine how corrosion advances from the factors that impact it, rather than necessarily on the corrosion percentage information itself.

Figure 2 demonstrates the results obtained in the experiment. Analyzing Figure 2a, the model provides information regarding variable importance, indicating wind speed as the most influential variable in corrosion percentage and the corrosion percentage itself. Such results are consistent because, according to the described in tables 1, 2 and the methodology section, wind speed information directly impacts corrosion percentage.

In Figure 2b, the model makes a comparison between the observed y, which is the actual corrosion percentage, and the predicted y, which is the predicted corrosion percentage based on a 45° line, representing a perfect linear relation. It can be seen that the points are close to the line, which indicates a high precision in the prediction of corrosion advancement. In this case, Figure 2c indicates that both the R^2value (0.975) for training and validation (0.812) are closer to 1, and the RMSE values also approached zero (0.0112 for training and 0.0305 for validation).

The value of R^2 for the OOB set improved (0.810), and the RMSE was closer to zero (0.0309). In this case, the predicted value was already the corrosion advancement.



(a) (b)

Set	RMSE	R^2
Testing	0.0112	0.975
Validation	0.0305	0.812
OOB (Out of Bag)	0.0309	0.81
Computation time: 1.24 s		

(c)

y_validation x y_predicted	
MAE (Mean Absolute Error)	0.00856
MSE (Mean Square Error)	0.00932

(d)

Figure 2 - Results from experiment
Source: the authors themselves.

Caption: (a) Percentage of variable importance results;
(b) Comparison between the predicted value of y and the value used in the model;
(c) Model performance metrics;
(d) Performance metrics of the predicted value.

5.1 Comparative analysis

Considering the same input dataset used in the experiment, which produced the best results, another adjustment was made. Aiming to improve the accuracy of the model, the

Grid Search function was run to find the best hyperparameters for the sklearn.ensemble.randomforestregressor function. As indicated in Table 3, in the "Default" column we have the default values used; in the GRID column, we have the results suggested by the model.

Table 3 - Suggested hyperparameters by the GRID function with an 80%-20% proportion

Hyperparameters	Standard	GRID
n_estimators	100	500
min_samples_leaf	1	1
min_samples_split	2	2

Source: the authors themselves.

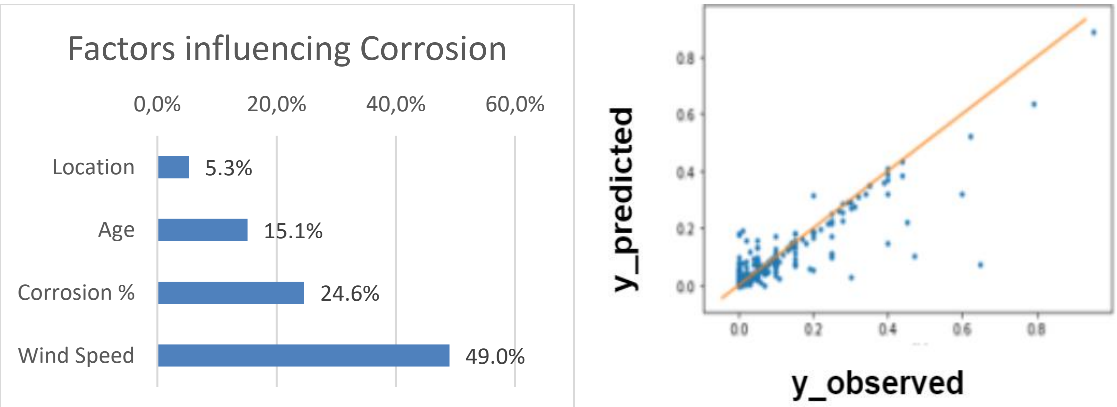
Using the hyperparameters suggested by the GRID function, the results obtained were slightly better than when the default hyperparameters were used, making the change of n_estimators to 500, as illustrated in Figure 3. The only downside was the increase in computational time.

Finally, a comparative analysis was performed between the Random Forest model with the hyperparameters from the Grid Search function and the XG Boost model, using the regression prediction. Table 4 describes the comparison of the results obtained, and it can be seen that the Random Forest model performed slightly better than XG Boost when analyzing R^2. Anyway, even if the result of the XG Boost model with the hyperparameters from the Grid Search function was not compared, the Random Forest model would still have a better performance, despite requiring more computational time (3.07 s).

Table 4 - Comparative analysis between Random Forest and XGBoost models

Model	RMSE	R^2
Random Forest	0.0305	0.812
XGBoost	0.0300	0.804

Source: the authors themselves.



(a)

(b)

Set	RMSE	R^2
Testing	0.0113	0.975
Validation	0.0305	0.812
OOB (Out of Bag)	0.0302	0.818
Computation time: 3.07 s		

(c)

y_validation x y_predicted	
MAE (Mean Absolute Error)	0.00855
MSE (Mean Square Error)	0.00932

(d)

Figure 3 - Results from the GRIDSearch function
Source: the authors themselves.

Caption: (a) Percentage of variable importance results;
(b) Comparison between the predicted value of y and the value used in the model;
(c) Model performance metrics;
(d) Performance metrics of the predicted value.

6. CONCLUSIONS

Based on the results obtained in this dissertation, it is concluded that the use of the Random Forest (Regression) machine learning model for the problem in question provided better accuracy, based on the analysis of the R^2, RMSE, MAE, MSE, and OOB metrics.

After a literature review, several studies were found in the literature using the Random Forest model for classification, however, few cases used the regression model for studies related to corrosion, and even those who did so considered controlled laboratory conditions, while the data used in this study were measured on oil platforms. Thus, the proposed work brings two contributions in the academic scope: (i) the use of Random Forest models (regression) for corrosion prediction, and (ii) the use of field data to conduct the study.

When compared with other models such as XGBoost, the Random Forest model performed better when analyzing the R^2 and RMSE metrics. Thus, for the Random Forest model, $R^2 = 0.812$ and RMSE = 0.0305 were obtained when the GRID Search function hyperparameters were used. Therefore, the model requires little data pre-processing and, once constructed, it is only necessary to feed it with updated data so that the saved model can generate a new updated prediction.

The objectives listed in the introduction were achieved, which made it possible to identify several factors that impact the advancement of corrosion, in addition to understanding how each factor acts more incidentally. These factors were, in descending order of importance: wind speed, corrosion percentage, and inspection year, where the higher the year, the longer that region has gone without any maintenance.

Thus, it becomes possible to develop a more assertive maintenance plan for corrosion, since by predicting the painting of a particular region, it is only necessary to have knowledge of the incidence of the factors listed in this dissertation to better understand how corrosion will behave over the next year. If that region cannot be painted for any reason, it is possible to have knowledge of the progression of corrosion in that area for the next year. According to the result in this dissertation, it can be affirmed that for the next year, there will be an average annual progression of 2.43%, if there is no painting.

Finally, it is important to emphasize that the constructed progression model has already been incorporated into a module of a prototype program of a large oil company that is in the finalization phase, however, it includes additional variables not mentioned in the dissertation, as it takes into account customized variables according to other types of platforms. Due to the structure with which the model was built, by using the same logic, several other factors can be selected and placed in the database. The model will return the same type of result, always indicating how corrosion will behave over time based on the factors selected by the user as influencing in corrosion advancement.

6.1 Suggestions for further research

Although an extensive survey was conducted on the factors that influence corrosion progression, many of them could not be used primarily due to the difficulty in acquiring information. This situation brought some limitations to the research, as other factors may also influence corrosion as much or even more than those indicated by the model as priorities. Therefore, to obtain better results, it is important to have an extensive database with information on the various factors that influence corrosion progression, such as system-specific information, not just general information on wind speed, direction, and relative humidity, as well as information indicating when the last painting was done in the inspected area.

Another issue is the subjectivity of the corrosion inspection process, as many corrosion percentages decreased from one year to the next, and without historical information on the regions that were painted, it was impossible to guarantee if that inspection evaluation had some kind of error due to the process itself or not. Therefore, in these cases, this information was excluded from the database, reducing the input base and providing less information to the model. As an alternative to this subjectivity issue, the model can also be improved by using Fuzzy techniques.

Furthermore, another factor that negatively impacted the results generation was the low variability of wind incidence, relative humidity, and temperature information, as the corporate portal from where the information was consulted only provided the measured value for the entire platform, while inspection information is recorded by the platform's system. In order to use the information, an approximation was made with this data, using the annual average of these data. For example, it is noticed that the application of variability in wind speed data brought an improvement in the evaluated metrics.

ACKNOWLEDGMENTS

This article was developed within the research and extension projects 'MI3D - 3D Environment for Maintenance, Integrity and Monitoring of Surface Installations' and 'Technological evolution of the Environ/Petrobras System for monitoring and managing the integrity of UEPs'. The authors would like to thank the Tecgraf Institute/PUC-Rio, the Padre Leonel Franca Foundation, the National Petroleum Agency (ANP), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brazil—CAPES (Finance Code 001), the Brazilian National Council for Scientific and Technological Development—CNPq (grant number307173/2022-4), and the Carlos Chagas Filho Foundation for Research Support of the State of Rio de Janeiro—FAPERJ (grant numbers E-26/201.363/2021; E-26/211.298/2021).

REFERENCES

American Petroleum Institute (2016), “API 579-1”, Fitness for Service.

American Petroleum Institute (2009)”, API RP 580: Risk-based inspection”.

Bento, M.P.; Ramalho, G.L.B.; Medeiros, F.N.S.; Medeiros, L.C.L. and Ribeiro, E.S. (2009), “Automatic identification of corrosion damage using image processing techniques“, in: Rio Pipeline Conference. Brazilian Petroleum, Gas and Biofuels Institute [IBP], Rio de Janeiro.

Breiman, L. (2001), “Random Forests”, Mach Learn 45, pp. 5–32. Doi: https://doi.org/10.1023/A:1010933404324

British Standards Institution (2012), “BS EN ISO 9223:2012: Corrosion of metals and alloys — Corrosivity of atmospheres — Classification, determination and estimation”.

Britton, C.F. (1990), “Corrosion monitoring and inspection, in: Microbiology in Civil Engineering”, CRC Press, pp. 382–390.

Buskirk, T.D. (2018), “Surveying the Forests and Sampling the Trees: An overview of Classification and Regression Trees and Random Forests with applications in Survey Research”, Surv Pract 11, pp. 1–13. Doi: https://doi.org/10.29115/SP-2018-0003

Cai, Y.K.; Zhao, Y.; Zhang, Z.K., Ma; X.B. and Cheng, B. (2018), “Atmospheric and Marine Corrosion: Influential Environmental Factors and Models”, in: Proceedings of the International Workshop on Environmental Management, Science and Engineering. SCITEPRESS - Science and Technology Publications, pp. 178–186. Doi: https://doi.org/10.5220/0007558601780186

Caiado, R.G.G.; Lima, G.B.A. and Quelhas, O.L.G. (2015), “Aspectos da aplicação da manutenção centrada em confiabilidade”, in: xi congresso nacional de excelência em gestão, Niterói.

Caiado, R.G.G.; Scavarda, L.F.; Azevedo, B.D.; Nascimento, D.L.M. and Quelhas, O.L.G. (2022), “ Challenges and Benefits of Sustainable Industry 4.0 for Operations and Supply Chain Management—A Framework Headed toward the 2030 Agenda”, Sustainability Vol. 14, pp. 830. Doi: https://doi.org/10.3390/su14020830

Chen, P.-H.; Yang, Y.-C.; Chang, L.-M. (2010), “Illumination adjustment for bridge coating images using BEMD-Morphology Approach (BMA)”, Autom Constr, Vol. 19, pp. 475–484. Doi: https://doi.org/10.1016/j.autcon.2009.12.017

Cho, D.Y. (2020), “Development of the Paint Amount Estimating Software for Pipe Supports of Ship and Offshore Structures Using 3D Cad Models”, International Journal of Advanced Research in Engineering and Technology (IJARET), Vol. 11, pp. 334–345.

Cruz, M.M.; Caiado, R.G.G.; Santos, R.S. (2022), "Industrial Packaging Performance Indicator Using a Group Multicriteria Approach: An Automaker Reverse Operations Case", Logistics, Vol. 6, pp. 58. Doi: https://doi.org/10.3390/logistics6030058

Dawson, J.L. (2010), "Corrosion management overview", in: Richardson, T.J.A. (Ed.), Shreir's Corrosion. Elsevier Science, Durham, United Kingdom, pp. 3001–3039.

Dawson, J.L. and John, G., Oliver, K. (2010), "Management of corrosion in the oil and gas industry", in: Richardson, T.J.A. (Ed.), Shreir's Corrosion. Elsevier Science, Durham, United Kingdom, pp. 3230–3269.

Paula Vidal, G.H.; Caiado, R.G.G.; Scavarda, L.F.; Ivson, P. and Garza-Reyes, J.A. (2022a), "Decision support framework for inventory management combining fuzzy multicriteria methods, genetic algorithm, and artificial neural network", Comput Ind Eng, Vol. 174, 108777. Doi: https://doi.org/10.1016/j.cie.2022.108777

de Paula Vidal, G.H.; Caiado, R.G.G.; Scavarda, L.F.; Santos, R.S. (2022b), "MRO Inventory Demand Forecast Using Support Vector Machine – A Case Study", pp. 221–233. Doi: https://doi.org/10.1007/978-3-031-14763-0_18

Diao, Y.; Yan, L.; Gao, K. (2021), "Improvement of the machine learning-based corrosion rate prediction model through the optimization of input features", Mater Des, Vol. 198, 109326. Doi: https://doi.org/10.1016/j.matdes.2020.109326

El Naqa, I. and Murphy, M.J. (2015), "What Is Machine Learning?", in: Machine Learning in Radiation Oncology. Springer International Publishing, Cham, pp. 3–11. Doi: https://doi.org/10.1007/978-3-319-18305-3_1

Fayyad, U.; Piatetsky-Shapiro, G. and Smyth, P. (1996), "From Data Mining to Knowledge Discovery in Databases", AI Mag, Vol. 17, pp. 37. Doi: https://doi.org/10.1609/aimag.v17i3.1230

Gupta, B.; Rawat, A.; Jain, A.; Arora, A. and Dhami, N. (2017), "Analysis of Various Decision Tree Algorithms for Classification in Data Mining", Int J Comput Appl, Vol. 163, pp. 15–19. Doi: https://doi.org/10.5120/ijca2017913660

He, S.; Wu, J.; Wang, D. and He, X., (2022), "Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest", Chemosphere, Vol. 290, 133388. Doi: https://doi.org/10.1016/j.chemosphere.2021.133388

Kaiser, B.C.S.; Santos, R.S.; Caiado, R.G.G.; Scavarda, L.F. and Netto, P.I. (2022), "Efficiency Assessment of Public Transport Vehicles Using Machine Learning and Non-parametric Models", pp. 207–220. Doi: https://doi.org/10.1007/978-3-031-14763-0_17

Koch, G.H.; Brongers, M.P.H.; Thompson, N.G.; Virmani, Y.P. and Payer, J.H. (2002), "Corrosion costs and preventive strategies in the US". Houston.

Liaw, A. and Wiener, M. (2002), "Classification and Regression by randomForest", R News 2, pp. 18–22.

Lima, B.F. and Neto, J.V.; Santos, R.S. and Caiado, R.G.G. (2023), "A Socio-Technical Framework for Lean Project Management Implementation towards Sustainable Value in the Digital Transformation Context. Sustainability", Vol. 15, 1756. Doi: https://doi.org/10.3390/su15031756

Lv, Y.; Wang, Jun-wei; Wang, Julian; Xiong, C.; Zou, L.; Li, L. and Li, D. (2020), "Steel corrosion prediction based on support vector machines", Chaos Solitons Fractals, Vol. 136, 109807. Doi: https://doi.org/10.1016/j.chaos.2020.109807

Lyon, S.B. (20100, "Corrosion of carbon and low alloy steels", in: Richardson, T.J.A. (Ed.), Shreir's Corrosion. Elsevier Science, Durham, United Kingdom, pp. 1693–1736.

Macha, E.N.; Dante, J.F. and DeCarlo, E. (2019), "Development of a Methodology to Predict Atmospheric Corrosion Severity Using Corrosion Sensor Technologies", in: CORROSION 2019. NACE International, Nashville.

Mendez, G.; Buskirk, T.D.; Lohr, S. and Haag, S. (2008), "Factors Associated With Persistence in Science and Engineering Majors: An Exploratory Study Using Classification Trees and Random Forests", Journal of Engineering Education, Vol. 97, pp. 57–70. Doi: https://doi.org/10.1002/j.2168-9830.2008.tb00954.x

Mishra, M.; Keshavarzzadeh, V. and Noshadravan, A. (2019), "Reliability-based lifecycle management for corroding pipelines", Structural Safety, Vol. 76, pp. 1–14. Doi: https://doi.org/10.1016/j.strusafe.2018.06.007

Mobley, R.K. (2002), "An introduction to predictive maintenance". BUTTERWORTH-HEINEMANN LTD.

Mokarian, P.; Bakhshayeshi, I.; Taghikhah, F.; Boroumand, Y.; Erfani, E. and Razmjou, A. (2022), "The advanced design of bioleaching process for metal recovery: A machine learning approach", Sep Purif Technol, Vol. 291, 120919. Doi: https://doi.org/10.1016/j.seppur.2022.120919

Muniz, M.V.P.; Lima, G.B.A.; Caiado, R.G.G. and Quelhas, O.L.G. (2018), "Bow tie to improve risk management of natural gas pipelines", Process Safety Progress, Vol. 37, pp. 169–175. Doi: https://doi.org/10.1002/prs.11901

Nascimento, D.L. de M.; Goncalvez Quelhas, O.L.; Gusmão Caiado, R.G.; Tortorella, G.L.; Garza-Reyes, J.A. and Rocha-Lona, L. (2019), "A lean six sigma framework for continuous and incremental improvement in the oil and gas sector", International Journal of Lean Six Sigma, No. 11, pp. 577–595. Doi: https://doi.org/10.1108/IJLSS-02-2019-0011

Nor Asma, R.B.A.; Yuli, P.A. and Mokhtar, C.I. (2011), "Study on the Effect of Surface Finish on Corrosion of Carbon Steel in CO2 Environment", Journal of Applied Sciences, No. 11, pp. 2053–2057. Doi: https://doi.org/10.3923/jas.2011.2053.2057

Orrù, P.F.; Zoccheddu, A.; Sassu, L., Mattia, C.; Cozza, R. and Arena, S. (2020), "Machine Learning Approach Using MLP and SVM Algorithms for the Fault Prediction of a Centrifugal Pump in the Oil and Gas Industry", Sustainability, Vol. 12, 4776. Doi: https://doi.org/10.3390/su12114776

Parameswari, A.; Kumar, K.V. and Gopinath, S. (2022), "Thermal analysis of Alzheimer's disease prediction using random forest classification model", Mater Today Proc, Vol. 66, pp. 815–821. Doi: https://doi.org/10.1016/j.matpr.2022.04.357

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. and Duchesnay, É. (2012), "Scikit-learn: Machine Learning in Python", The Journal of Machine Learning Research, Vol. 12, pp. 2825–2830.

Rajendra, P.; Girisha, A. and Gunavardhana Naidu, T. (2022), "Advancement of machine learning in materials science", Mater Today Proc 62, pp. 5503–5507. Doi: https://doi.org/10.1016/j.matpr.2022.04.238

Sircar, A.; Yadav, K.; Rayavarapu, K.; Bist, N. and Oza, H. (2021), "Application of machine learning and artificial intelligence in oil and gas industry", Petroleum Research, Vol. 6, pp. 379–391. Doi: https://doi.org/10.1016/j.ptlrs.2021.05.009

St. Clair, A.M.; Sinha, S. (2014), "Development of a Standard Data Structure for Predicting the Remaining Physical Life and Consequence of Failure of Water Pipes", Journal of Performance of Constructed Facilities, Vol. 28, pp. 191–203. Doi: https://doi.org/10.1061/(ASCE)CF.1943-5509.0000384

Strobl, C.; Boulesteix, A.-L.; Zeileis, A. and Hothorn, T. (2007), "Bias in random forest variable importance measures: Illustrations, sources and a solution|", BMC Bioinformatics, Vol. 8, No. 25. Doi: https://doi.org/10.1186/1471-2105-8-25

Tang, Y.; Wan, Y.; Wang, Z.; Zhang, C.; Han, J.; Hu, C. and Tang, C. (2022), "Machine learning and Python assisted design and verification of Fe–based amorphous/nanocrystalline alloy", Mater Des, Vol. 219, 110726. Doi: https://doi.org/10.1016/j.matdes.2022.110726

Tecgraf Institute/PUC-Rio (2021). "RELATÓRIO TÉCNICO 02 – Projeto Ambiente 3D para Manutenção, Integridade e Monitoramento de Instalações de Superfície". Rio de Janeiro.

Van Rossum, G. and Drake, F.L. (2009), "Python 3 Reference Manual", CreateSpace, Scotts Valley, CA.

Vieira, S.; Lopez Pinaya, W.H. and Mechelli, A. (2020), "Introduction to machine learning", in: Machine Learning. Elsevier, pp. 1–20. Doi: https://doi.org/10.1016/B978-0-12-815739-8.00001-8

Wei, X.; Fu, D.; Chen, M.; Wu, W.; Wu, D. and Liu, C. (2021), "Data mining to effect of key alloying elements on corrosion resistance of low alloy steels in Sanya seawater environment Alloying Elements", J Mater Sci Technol, Vol. 64, pp. 222–232. Doi: https://doi.org/10.1016/j.jmst.2020.01.040

Wu, W.; Cheng, G.; Hu, H.; Zhou, Q. (2013), "Risk analysis of corrosion failures of equipment in refining and petrochemical plants based on fuzzy set theory", Eng Fail Anal, Vol. 32, pp. 23–34. Doi: https://doi.org/10.1016/j.engfailanal.2013.03.003

Zhao, H.; Gunardi, W.; Liu, Y.; Kiew, C.; Teng, T.-H. and Yang, X.B. (2022), "Prediction of Traffic Incident Duration Using Clustering-Based Ensemble Learning Method", J Transp Eng A Syst, Vol. 148. Doi: https://doi.org/10.1061/JTEPBS.0000688

Zhi, Y.; Jin, Z.; Lu, L.; Yang, T.; Zhou, D.; Pei, Z.; Wu, D.; Fu, D.; Zhang, D. and Li, X. (2021), "Improving atmospheric corrosion prediction through key environmental factor identification by random forest-based model", Corros Sci, Vol. 178, 109084. Doi: https://doi.org/10.1016/j.corsci.2020.109084

Author contributions: FRE: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Funding Acquisition; MPR: Validation, Data Curation, Writing – Review &Editing, Visualization, Software; ERAL: Validation, Data Curation, Writing – Review &Editing, Visualization, Software; RGGC: Methodology, Validation, Resources, Writing – Review &Editing, Supervision, Project administration, Funding Acquisition; RSS: Software, Validation, Data Curation, Writing – Review &Editing, Visualization.