



Application of artificial intelligence models to predict the compressive strength of concrete

Lucas Elias de Andrade Cruvinel¹ · Wanderlei Malaquias Pereira Junior¹ · Amanda Isabela de Campos² · Rogério Pinto Espíndola² · Antover Panazzolo Sarmento¹ · Daniel de Lima Araújo³ · Gustavo de Assis Costa⁴ · Roberto Viegas Dutra¹

Received: 28 June 2023 / Revised: 30 January 2024 / Accepted: 10 March 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract

The concrete mixture design and mix proportioning procedure, along with its influence on the compressive strength of concrete, is a well-known problem in civil engineering that requires the execution of numerous tests. With the emergence of modern machine learning techniques, the possibility of automating this process has become a reality. However, a significant volume of data is necessary to take advantage of existing models and algorithms. Recent literature presents different datasets, each with its own unique details, for training their models. In this paper, we integrated some of these existing datasets to improve training and, consequently, the models' results. Therefore, using this new dataset, we tested various models for the prediction task. The resulting dataset comprises 2358 records with seven input variables related to the mixture design, while the output represents the compressive strength of concrete. The dataset was subjected to several pre-processing techniques, and afterward, machine learning models, such as regressions, trees, and ensembles, were used to estimate the compressive strength. Some of these methods proved satisfactory for the prediction problem, with the best models achieving a coefficient of determination (R^2) above 80%. Furthermore, a website with the trained model was created, allowing professionals in the field to utilize the AI technique in their everyday problem-solving.

Keywords Artificial intelligence · Trees · Ensembles · Regressions · Concrete · Prediction · Mix proportioning · Compressive strength

1 Introduction

The characteristic value of compressive strength of concrete (f_c) is one of the most critical variables for designing concrete structures, and knowledge of this attribute is essential for ensuring the quality and safety control of structural systems (Rauecker et al. 2019). As a rule, the f_{ck} is defined as

✉ Wanderlei Malaquias Pereira Junior
wanderlei_junior@ufcat.edu.br

Lucas Elias de Andrade Cruvinel
lucascruvinel@discente.ufcat.edu.br

Amanda Isabela de Campos
amanda.campos@coc.ufrrj.br

Rogério Pinto Espíndola
rogerio.espindola@coc.ufrrj.br

Antover Panazzolo Sarmento
antoverps@ufcat.edu.br

Daniel de Lima Araújo
dlaraujo@ufg.br

Gustavo de Assis Costa
gustavo.costa@ifg.edu.br

Roberto Viegas Dutra
roberto.dutra@discente.ufcat.edu.br

¹ Federal University of Catalão, Dr. Lamartine Pinto de Avelar Avenue, Bloco O, St. Universitário, Catalão, GO CEP: 75705-22, Brazil

² Federal University of Rio de Janeiro, Horácio Macedo Avenue, s/n-Cidade Universitária, Rio de Janeiro, RJ CEP: 21941-450, Brazil

³ Federal University of Goiás, Universitária, Avenue, Block 86, Lot Area 1488, Setor Leste Universitário, Goiânia, GO CEP: 74605-220, Brazil

⁴ Federal Institute of Goiás–Campus Jataí, Alameda Santos Dumont, 775-Residencial Flamboyant, Jataí, GO CEP: 75804-714, Brazil

the compressive strength measure with 95% confidence at an age of 28 days. This measurement plays a crucial role in various stages of building design, including (a) determining when to remove supports during the construction planning of a structure and (b) assessing the strength of structural designs.

This measurement is determined experimentally through the compression testing of cylindrical specimens as described in NBR 5739 (2018). It is influenced by several factors (Erdal 2013; Abbass et al. 2019), such as water–cement ratio, type of cement, specimen model, and testing speed. Due to the large number of parameters involved, ensuring concrete quality control is a complex task, especially in variations of conventional concrete where additional parameters come into play. For example, in fiber-reinforced concrete (FRC), the volume of fibers, and in high-performance concrete (HPC), the types of additives become significant factors. Therefore, using Machine Learning techniques for predictions has shown great potential in various applications. The fundamental idea behind machine learning-based concrete mix design is to optimize the mix proportions and reduce the time spent in the semi-empirical design process based on the relationships between design factors and concrete compressive strength. Several studies have explored Machine Learning techniques to predict the engineering properties of various construction materials, including recycled aggregate concrete (RCA) (Zhang et al. 2020a; Behnood and Golafshani 2020), normal strength concrete (Feng et al. 2020; Chou et al. 2014), and ultra-high-performance concrete (UHPC) (Al-Shamiri et al. 2020; Fan et al. 2020; Alabduljabbar et al. 2023).

Computational learning capability (Mirjalili et al. 2020) has been highlighted in various application areas (Madabhushi and Lee 2016; Komura and Ishikawa 2018; Milhomem and Dantas 2020; Zhang et al. 2020b; Isinkaye et al. 2015). This article aims to contribute to popularizing these techniques for predicting the mechanical properties of concrete and mortars, as such techniques are still in full development. Additionally, the availability of a dataset related to the concrete mixtures and compressive strength of concrete, resulting from a data curatorial work, can contribute to improving other authors' works, since these data were enhanced.

We can summarize the main contributions of this paper as follows:

- (a) The creation and consolidation of an extensive dataset comprising data from various works in the literature that utilize concrete mixture design procedures and their respective compressive strength (f_c);
- (b) The demonstration and analysis of results from experiments involving the application of different regression models in the context of different concrete mixtures, ranging from low- to high-strength concrete and their respective compressive strength.

2 Theoretical reference

In addition to the traditional data sources, the advent of different technologies for data extraction, like sensors or even Web scraping, has been increasing the amount of available data for processing. This scenario is expected for machine learning scientists and professionals who intend to create and improve their models. From another perspective, this unprecedented amount of data can also bring some problems known to the database community, like noise and data inconsistency.

In this sense, it is essential that these data would be pre-processed and analyzed before being submitted to subsequent phases of an ML pipeline. However, the most critical aspect before constructing a pipeline is identifying the problem and the most adequate task. There are different tasks that can be applied to a class of problems, like clustering, classification, and regression. Our approach is based on the regression task, where the main objective is to make predictions regarding the future state of data, considering that unknown outcomes adhere to patterns identified in previous observations.

In Fig. 1, we can observe an overview of the machine learning pipeline where we first define the task, validate the scenario, and, finally, decide which model to employ. Still, between steps 1 and 2, we had to integrate and process all collected data from the literature. All the data were submitted to pre-processing and, thus, to feature engineering and exploratory data analysis.

The regression task is a classical problem of supervised learning, involving the modeling of a predictive function that generates a response, also known as target variables or dependent variables, that can be obtained from a combination of one or more independent variables. In addition to predictions, it also allows analyzing the behavior of data, i.e., the relation between the response and the variables (Al-Shamiri et al. 2020; Igual and Seguí 2017).

These models are used to fit the data points along the line generated from the function as a best fit and minimize the distance between the data points and the line by least-squares methods. Therefore, in this way, the prediction Machine Learning problems can be described generically according to Eq. (1), where (y) is the vector containing the observed measures or responses and (\hat{y}) is the vector containing the numerical measures obtained by the prediction model

$$\text{minimize} |\hat{y} - y|. \quad (1)$$

The following sections present the formulations of the Machine Learning methods used in this article's context.

2.1 Machine learning

The concept of Machine Learning has stood out a lot in recent years due to the significant advances obtained in different

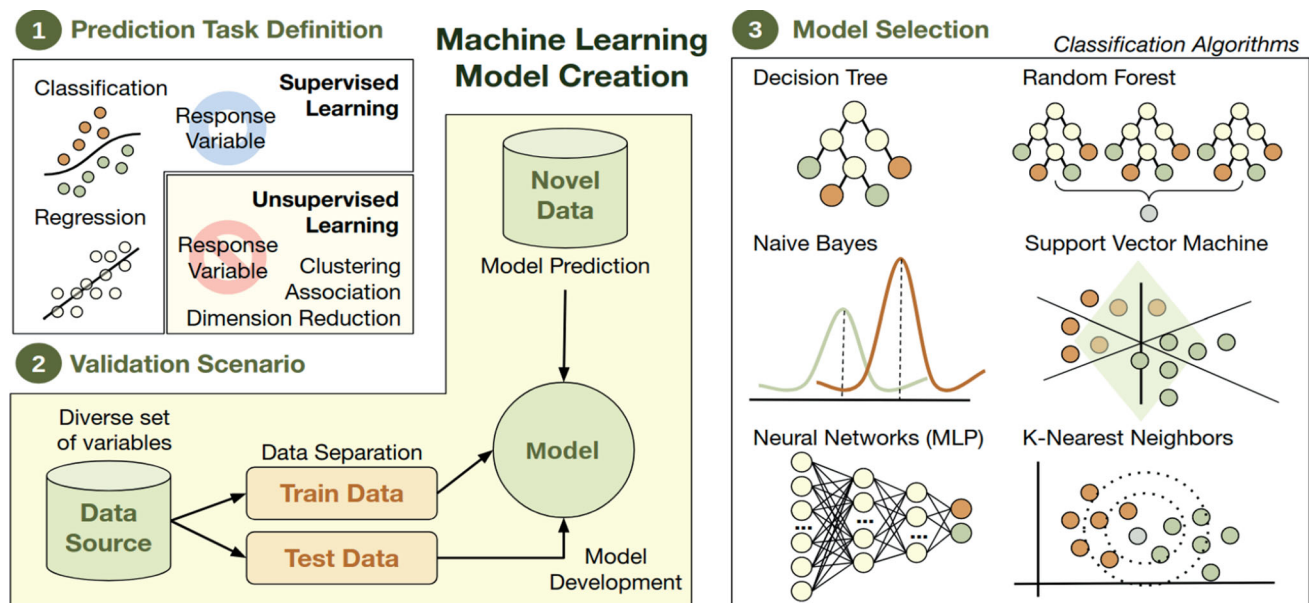


Fig. 1 Machine learning (ML) workflow and main ML models for predictive tasks (Hild-Aono et al. 2022)

areas of computing. For instance, databases, artificial intelligence, distributed computing, and advances obtained in other areas that, directly or indirectly, were also primordial for this evolution, contributing to the opportunity to generate, make available, and access ever-increasing volumes of data.

Learning is a vast concept, but when discussed in this work, it concerns the ability to automatically learn by machines capable of performing or simulating such a skill. Since computers perform all their procedures through calculations, nothing is more natural than modeling the learning process mathematically. In this case, the main idea behind this modeling is based on the pattern recognition process, since incorporating prior knowledge is the main influencing factor in the learning process.

In this case, some tools from areas, such as statistics, optimization, and information theory, for example, are essential to “train” the algorithms based on the patterns observed in data extracted in previous moments. A central theme of machine learning theory is developing solutions to express knowledge of a particular domain based on a learning process (Shalev-Shwartz and Ben-David 2014).

In the scientific literature, one can find some proposals for the categorization of types of machine learning. In the vast majority, the division into three main categories prevails, which are (Russell and Norvig 2016):

- (i) **Supervised Learning:** In supervised learning, the agent is presented with labeled input–output pairs, enabling it to learn a mapping function from input to output. This

method is crucial in tasks like classification and regression, where the algorithm is trained to make accurate predictions based on the provided examples.

- (ii) **Unsupervised Learning:** Unlike supervised learning, the agent does not receive explicit input–output pairs in unsupervised learning. Instead, it aims to identify patterns, structures, or intrinsic relationships within the input data. Clustering techniques, for instance, enable the algorithm to automatically detect clusters or segments within the data without explicit guidance.
- (iii) **Reinforcement Learning:** The agent learns through rewards and punishments while interacting with an environment. It makes sequential decisions and knows to maximize cumulative rewards over time. This method is often employed in gaming, robotics, and process optimization, allowing the agent to learn optimal behaviors through direct interaction and experimentation within its environment.

The following sections present the formulations of the supervised Machine Learning methods used in this article’s context. The choice for supervised methods was due to the characteristics of the context in question, where measurement techniques that originated the datasets were employed. The data are of a continuous numerical nature, and the task performed was predicting values.

2.2 Regressions

Regression is a technique to investigate the relationship between the input space (independent variables) and the output space (dependent variable). According to Kang et al. (2021), regressions are one of the most common techniques for performing prediction tasks. In these models, numerical values, known as regression coefficients (w), are used as parameters in the predictive functions to describe the relationship between a predictor variable and its corresponding response. Thus, this method seeks to produce a linear or non-linear predictive function (\hat{y}) that minimizes the loss function given by Eq. (2) (Tai 2021). In this process, y represents the vector of observations and w the vector of weights that minimize the loss function. The loss function is alternatively known as Mean Squared Error (MSE) for this scenario (Igual and Seguí 2017; Russell and Norvig 2016)

$$\text{minimize Loss}(x, w) = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i(x, w) - y_i)^2 \quad (2)$$

$$\hat{y}_i(x, w) = w_0 + \sum_i w_i \cdot x_i. \quad (3)$$

Despite their predictive capacity, the generated models are likely to be affected by overfitting, despite their predictive capacity, the generated models are likely to be affected by the effect of overfitting, a situation characterized by the difficulty of generalizing the model, causing the model to overfit the data set. To reduce the possibility of overfitting the data set, it is possible by creating regularization rules in the loss function. The most commonly employed techniques are Ridge (L2) and Lasso (L1) (Géron 2019). Such techniques were employed in this article.

2.3 Regressive decision tree

Regressive decision tree models, or regression trees, were first introduced by Breiman (1998). In general terms, the decision tree procedure divides the data system into hierarchies. According to Géron (2019), Decision Trees comprise nodes representing the attributes and branches originating from these nodes and receiving the possible values for these attributes (each descending branch corresponds to a possible value of this attribute). In trees, there are leaf nodes (leaf of the tree) representing the different values of a training set. That is, each leaf is associated with a class. Each path in the tree (from root to leaf) corresponds to a regression rule and can be represented as sets of if-then rules. The rules are written considering the path from the root node to a leaf in the tree.

In the same way that the regression process minimizes the residuals (Scikit-Learn 2024), the Regression tree will seek to

reduce the impurities in each subset formed. For a Regression tree problem, the cost function is given by Eq. (4)

$$\text{Loss}(Q_m, \theta) = \frac{n_m^{\text{left}}}{n_m} \cdot H(Q_m^{\text{left}}(\theta)) + \frac{n_m^{\text{right}}}{n_m} \cdot H(Q_m^{\text{right}}(\theta)) \quad \theta = (j, t_m) \quad (4)$$

$$H(Q_m(\theta)) = \frac{1}{n_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2 \quad (5)$$

$$\bar{y}_m = \frac{1}{n_m} \sum_{y \in Q_m} y, \quad (6)$$

where m represents a node, j is a feature, t_m represents the threshold, $H()$ measures the impurity of the subsets, n_m is the number of instances in the subsets (n_m^{left} is the number of instances in the left subset, and n_m^{right} is the number of instances in the right subset), Q_m is a subset (left subset— Q_m^{left} and right subset— Q_m^{right}), y represents the observed value in i node, and \bar{y}_m represents the average value in each region. The partitions are given by Eqs. (7) and (8)

$$Q_m^{\text{left}}(\theta) = \{(x, y) | x_j \leq t_m\} \quad (7)$$

$$Q_m^{\text{right}}(\theta) = \{(x, y) | x_j > t_m\}. \quad (8)$$

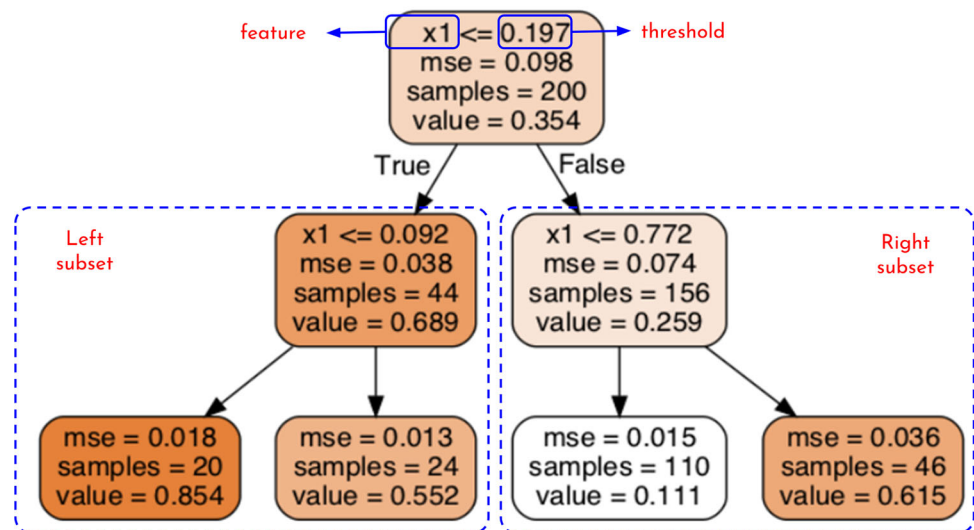
In regression trees, the main difference is that instead of predicting a class in each node, it predicts a value. The example results of regression trees are presented in Fig. 2. For example, to predict a new instance with $x_1 = 0.6$. You traverse the tree starting at the root, and you eventually reach the leaf node that predicts value = 0.1106. This prediction is simply the average target value of the 110 training instances associated with this leaf node. This prediction results in a Mean Squared Error (MSE) equal to 0.0151 over these 110 instances (Géron 2019).

Therefore, these trees can perform analyses among the proposed data and find patterns that can be organized into different series of prediction rules (Kang et al. 2021). Such a model is usually used as an alternative when linear models fail to return an accuracy within the acceptable level (Güçlüer et al. 2021).

2.4 Ensembles

Ensemble-type learning methods train combinations of models, which can be decision trees, neural networks, or others traditionally used in supervised learning. Ensemble methods have gained popularity, because many researchers have demonstrated superior prediction performance over single models in various problems (Oza 2000).

Fig. 2 Example of a regression tree (Géron 2019)



In the case of this article, the Gradient Boosting technique will be used, which has the general idea of sequentially training the predictive mode (F), and, at each iteration, correcting its predecessor model (F_{m-1}) (Géron 2019; Natekin and Knoll 2013). The model correction is given as a function of Eq. (9) where h_m represents the result of training a tree. In this case, the h_m portion is given by minimizing the loss function

$$F_m(x) = F_{m-1}(x) + h_m(x). \quad (9)$$

3 Machine learning workflow

This section demonstrates the necessary procedures for constructing predictive models of the material's mechanical strength. The Machine Learning workflow in this paper is similar to several other AI papers, such as Yassen et al. (2018), Pakzad et al. (2023), and Alabduljabbar et al. (2023).

3.1 Task definition and features analysis

Following the procedure shown in Fig. 1, the first part of this work begins with constructing a database related to concrete mixtures and defining the task related to regression. Therefore, the dataset studied results from the collection and integration of data from different sources (Beck 2009—24 samples, Bilim et al. 2009—225 samples, Bouzoubaâ and Fournier 2003—68 samples, Chopra et al. 2016—228 samples, Demirboğa et al. 2004—29 samples, Duran Atiş 2005—69 samples, Durán-Herrera et al. 2011—114 samples, Jiang and Malhotra 2000—54 samples, Lee et al. 2006—53 samples, Oner and Akyuz 2007—224 samples, Pala et al. 2007—90 samples, Pitroda 2014—10 samples,

Sonebi 2004—62 samples, Yeh 1998—1030 samples, and Yen et al. 2007—80 samples), forming a new base with 2358 records related to very-low- to high-strength concrete (1.76–113.2 MPa). Before generating the statistics of the data set, the Database (BD) was cleaned to form a single reference of samples with seven input attributes referring to the concrete mix proportioning and an output attribute referring to the cylinder compressive strength (f_c).

The cleaning carried out consists of two changes to the database. The first refers to creating the water/cement column, since this rate is critical in concrete technology, directly influencing the strength variable (Singh et al. 2015). The second change relates to merging the filler additions (Fly Ash and Blast Furnace) into a single variable. Addition is an important factor, especially in high-strength concrete (Abbass et al. 2019), but the bases used do not give details about these variables, so this combination was chosen. Also, possible duplicates have been eliminated. DB was missing values before cleanup.

Table 1 presents a brief statistical description of the dataset, and Fig. 3 shows the histogram of the attributes that form the database

Mixture design is essential for all types of concrete and involves defining the proportions of materials that constitute the concrete composite. The objective of mixture design is to determine the quantities of materials needed to achieve specific mechanical properties, such as f_c .

Mixture design plays a critical role in projects ranging from small buildings to large-scale structures. Tutikian and Helene (Tutikian and Helene 2011) explain that mixture design aims to find the ideal and most cost-effective mix proportions using available materials while meeting design requirements. Various mixture design procedures exist, including semi-empirical methods like those provided

Table 1 Statistical characteristics of the analyzed dataset attributes

UNITY	<i>c</i> kg/m ³	<i>sp</i> kg/m ³	<i>cg</i> kg/m ³	<i>fg</i> kg/m ³	<i>t</i> days	<i>f_c</i> MPa	<i>w/c</i> –	add kg/m ³
MEAN	285.71	3.75	1000.16	724.24	62.59	40.06	0.79	128.68
SD	115.04	5.06	121.30	122.82	86.44	20.14	0.44	102.75
MIN	38.00	0.00	723.00	175.95	1.00	1.76	0.27	0.00
25%	194.68	0.00	932.00	645.00	7.00	24.90	0.50	0.00
50%	275.00	0.60	1005.60	742.00	28.00	38.95	0.70	122.58
75%	375.00	6.86	1088.55	798.90	90.00	52.74	0.96	200.00
MAX	643.00	32.20	1419.30	1066.00	365.00	113.10	4.87	594.00

Cement (*c*); superplasticizer (*sp*); coarse aggregate (*cg*); fine aggregate (*fg*); age (*t*); compressive strength (*f_c*); water–cement ratio (*w/c*); additions (add)

Table 2 Result of the sensitivity analysis of the database attributes concerning the response variable *f_c* (value ± standard deviation)

Variable	SRC	SRRC
<i>c</i>	0.76 ± 0.04	– 0.03 ± 0.06
<i>sp</i>	0.14 ± 0.02	0.03 ± 0.01
<i>cg</i>	0.32 ± 0.02	0.11 ± 0.02
<i>fg</i>	0.14 ± 0.02	– 0.01 ± 0.02
<i>t</i>	0.46 ± 0.01	0.61 ± 0.01
<i>w/c</i>	– 0.13 ± 0.03	– 0.81 ± 0.05
add	0.55 ± 0.03	0.44 ± 0.03

by the American Concrete Institute (ACI) and Brazilian Portland Cement Association (BPCA), as well as more complex procedures like the packaging method used for Ultra-High-Performance Concrete (UHPC).

The proportions of cement, mineral admixtures, and water content directly influence chemical reactions within the paste, such as the formation of Calcium Silicates (C–S–H). Therefore, successful mixture design is crucial for controlling the durability of concrete in its hardened state (Ribeiro et al. 2021).

To verify the correlation between the database attributes, the problem's correlation matrix was verified (Fig. 4). It is observed that the highest correlation value is between the variables' compressive strength (*f_c*) and age (*t*), with a correlation coefficient of 0.46. There is also a negative correlation between the attribute *w/c* and the *f_c* of -0.40, which aligns with experimental findings showing that mechanical strength decreases as the water content in the mixture increases. Another experimentally proven factor is the strong positive correlation (0.46) between the amount of the cement and the compressive strength.

It was also verified (Table 2) that the sensitivity of the database attributes as a function of the response variable, *f_c*, using the standardized regression coefficient (SRC) and

standardized rank regression coefficient (SRRC) methods. The SRC method is widely used to check the sensitivity of linear models, and its ranked version (SRRC) is often used for non-linear models (Homma and Saltelli 1996).

Using the Stepwise Regression Coefficient (SRC) method, it was determined that the three most important variables concerning *f_c*, are *c*. (cement), add (additions), and *t* (age). However, when applying the Stepwise Robust Regression Coefficient (SRRC) method, the order shifted to *w/c*, *t*, and add. The prominence of attribute *c* in the SRC method aligns with its linear trend behavior (Figure), which is consistent with experimental observations.

Figure 5f illustrates why the attribute *w/c* (water–cement ratio) ranked first in importance by the SRRC method, since it did not present a linear behavior, as observed in experimental observations that used the Abrams law. Additionally, Fig. 5 showcases the regression lines that demonstrate the relationship between the inputs and the compressive strength *f_c*.

The add attribute appeared among the top three in importance in both the SRC and SRRC sensitivity analyses despite its very-low correlation with *f_c* (– 0.09, as shown in Fig. 4). This unexpected finding may be explained by the correlation between *add* and the attributes *c* and *w/c*, indicating an existing interaction among these variables.

Therefore, sensitivity analysis is justified in multiple ways, including assessing response variable behavior concerning input attributes and their interrelationships through plotting and visual inspection.

To apply the numerical modeling of Data Mining, the Python 3 language was used with the use of the following libraries: (a) Scikit-learn, (b) Pandas, (c) Numpy (d), Matplotlib, and (e) Seaborn. The studied problem consists of a prediction case using.

In this step, corrupted records and duplicates were eliminated from the database. This process resulted in a reduction of 1.21% in the original database. The pre-processing phase is one of the most important for machine learning, through

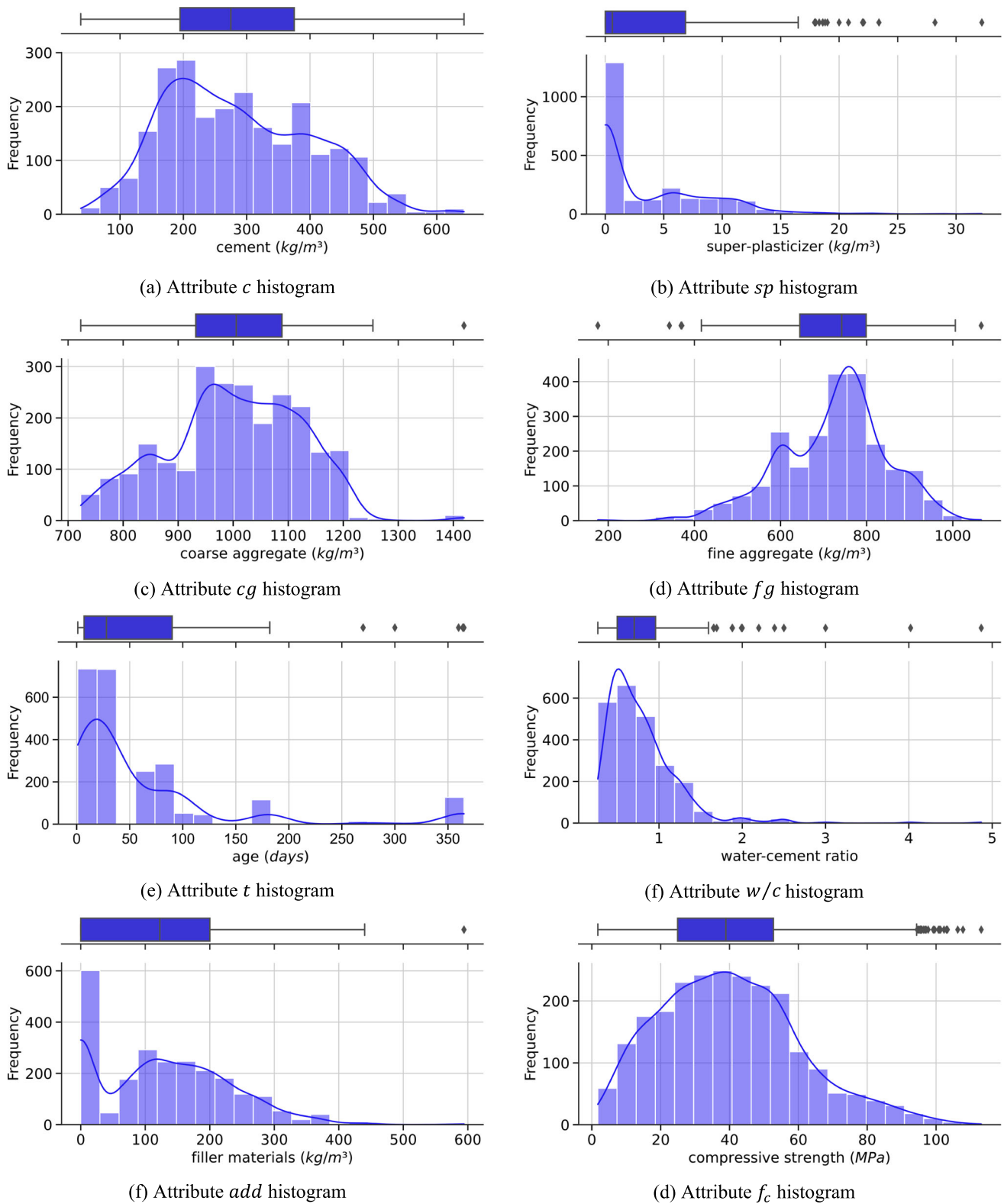
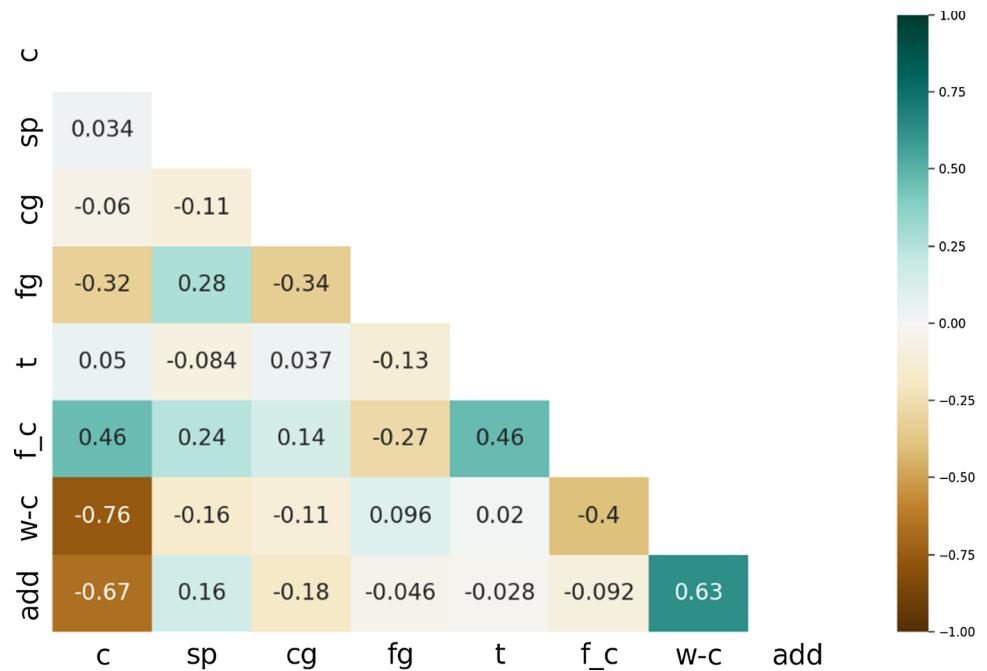


Fig. 3 Distribution of some of the database variables. **a** Attribute c histogram, **b** attribute sp histogram, and **c** attribute cg histogram, **d** attribute fg histogram, **e** attribute t histogram, **f** attribute w/c histogram, **f** attribute add histogram, and **d** attribute f_c histogram

Fig. 4 Database correlation matrix

which it is possible to amplify the performance or even guarantee it (McCabe et al. 2012; Oliveri et al. 2019).

As magnitudes of different units formed the database, it was necessary to normalize it. This process aims to place the attributes in a standard range of values, thus reducing model accuracy errors due to the weight of the units in the analysis (Sola and Sevilla 1997). In this line of reasoning, the z-score normalization was used in the database. As described in Eq. (10), where μ is the mean of the analyzed attribute, σ is the standard deviation of the analyzed attribute, x is the vector containing the original values of the analyzed attribute, and z is the vector containing the normalized attribute values

$$z = \frac{x - \mu}{\sigma}. \quad (10)$$

This is valid in this work, since there are scales regarding material consumption, time, and a dimensionless scale regarding the water–cement ratio.

3.2 Validation scenario

Twelve computational models of machine learning were selected, and these techniques were chosen due to their effectiveness against learning problems of this nature. Table 3 presents the characteristics of each of the adopted models. The other characteristics of the model follow the standardized settings of the Scikit-learn library.

The complete assembly pipeline of the machine learning model followed the guideline of Fig. 6. In the conception used in this article, the data were separated in the proportions of

Table 3 Computational models applied

Model	Type	Description	Algorithm characteristics
Linear regression	Regression	RL1	1st-degree approximation
Polynomial regression		RL2	2nd-degree approximation
		RL3	3rd-degree approximation
Ridge linear regression		RR1	1st-degree approximation
Polynomial ridge regression		RR1	2nd-degree approximation
		RR2	3rd-degree approximation
Decision tree	Tree	AR2	Tree depth: 2
		AR5	Tree depth: 5
		AR10	Tree depth: 10
		AR20	Tree depth: 20
Gradient boosting	Ensemble	GB10	Max depth: 10, number of estimators: 100
		GB20	Max depth: 20, number of estimators: 100

80%/20% for training and testing, respectively. The separation percentage was defined as a function of the learning curve given in Fig. 7, where it was possible to detect that the Root Mean Square Error (RMSE) measure already presented

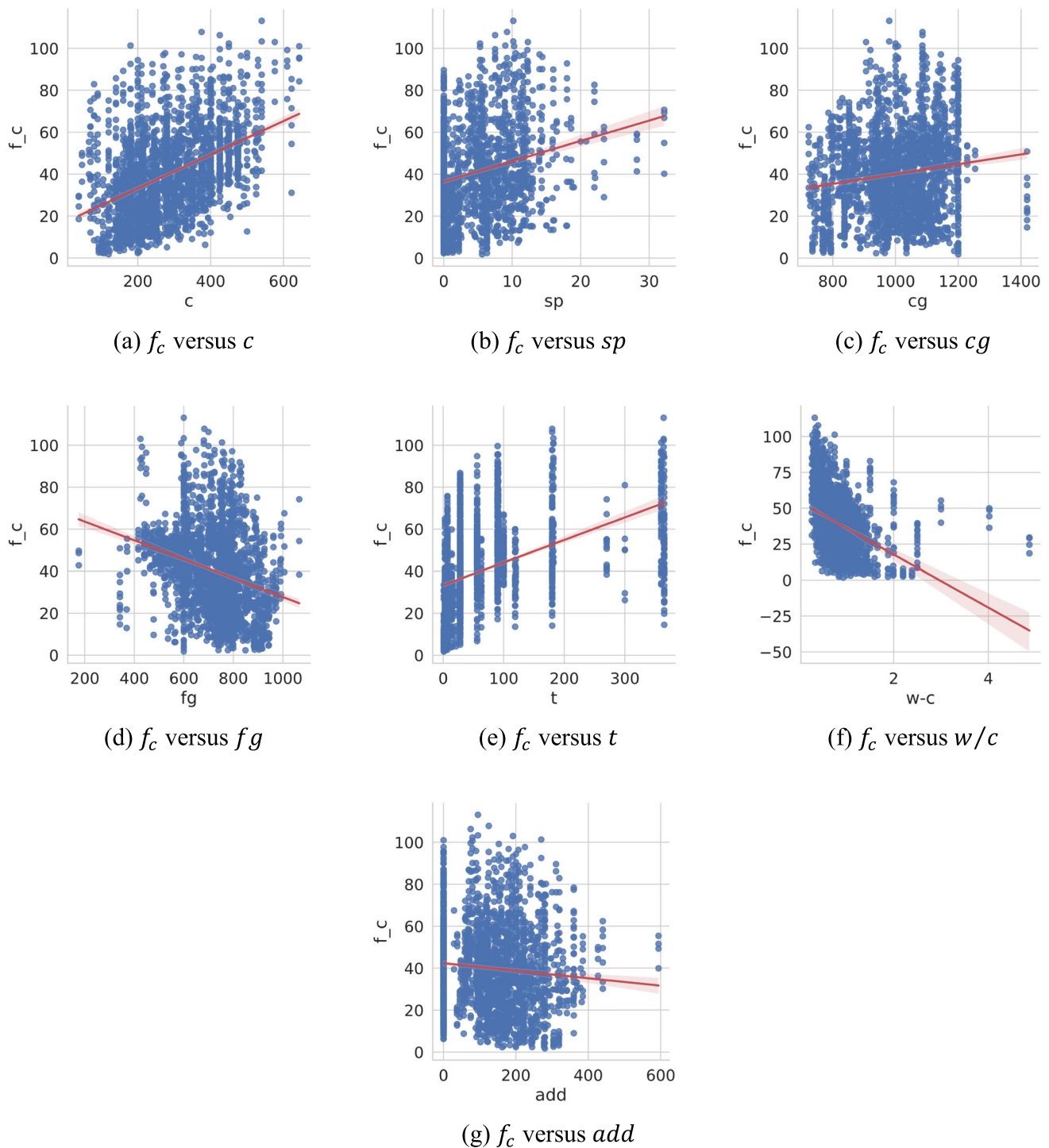


Fig. 5 Scatter plot with a regression line for each database attribute as a function of f_c . **a** f_c versus c , **(b)** f_c versus sp , **(c)** f_c versus cg , **(d)** f_c versus fg , **(e)** f_c versus t , **(f)** f_c versus w/c , and **(g)** f_c versus add

a satisfactory convergence value for an 80% separation of the training data.

Once the learning models were defined, cross-validation tests were applied before training, thus verifying whether the models performed satisfactorily to start the learning

task. Figure 8 presents the cross-validation model used in this work: the Kfold model. In the case of this article, a division with ten sections ($cv = 10$) was applied. This cross-validation process was performed 30 times to verify the consistency of the models concerning the dataset.

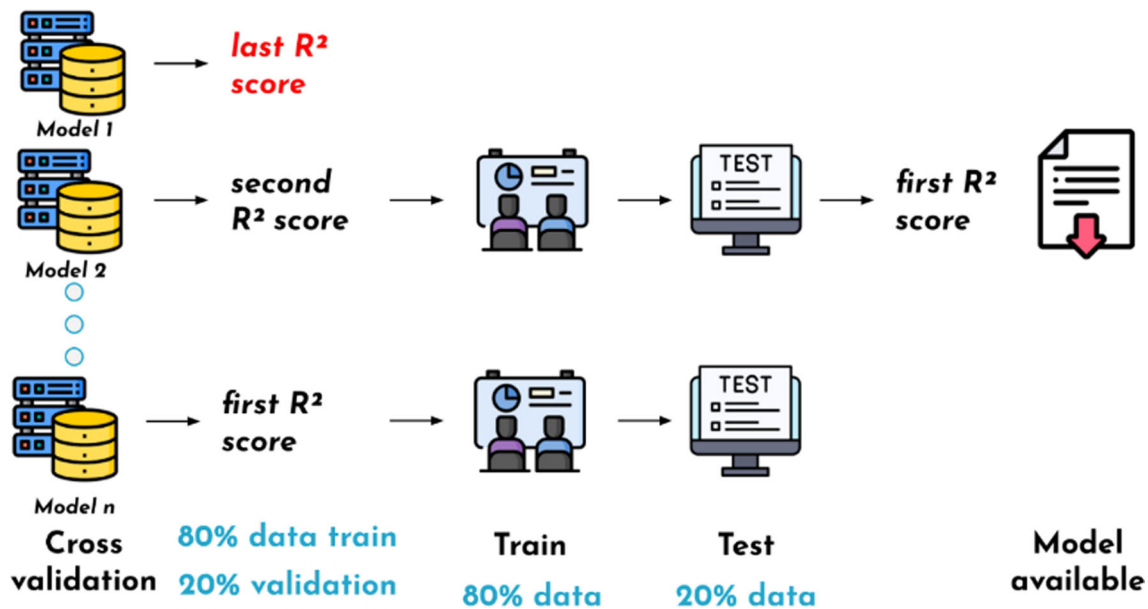


Fig. 6 AI training and testing pipeline

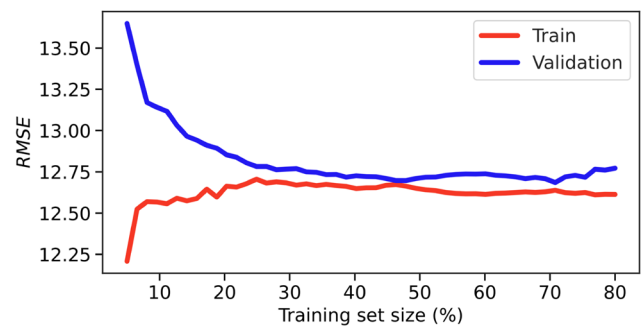
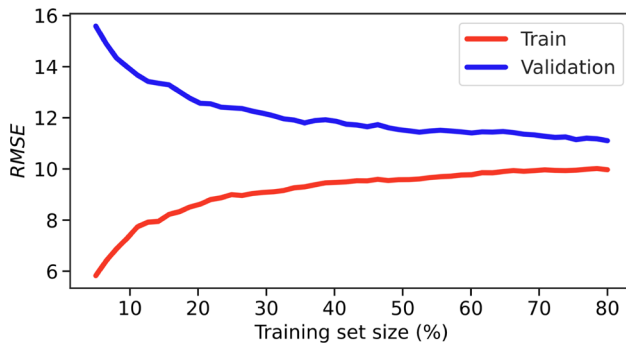
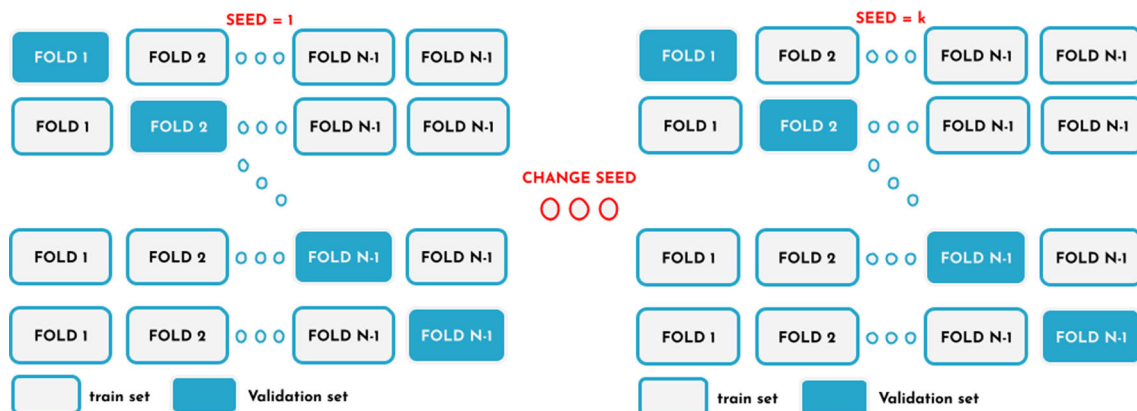
Fig. 7 Learning curve percentage of the decision tree model in separating training and testing with error. **a** Learning curve for the tree models and **b** learning curve for the linear regression model

Fig. 8 Kfold cross-validation strategy

The accuracy measure for both procedures performed in this work was the Coefficient of Determination (R^2) as shown

in Eq. (11). In this equation, \hat{y} is the vector that includes the numerical measurements estimated by the model, \bar{y} is

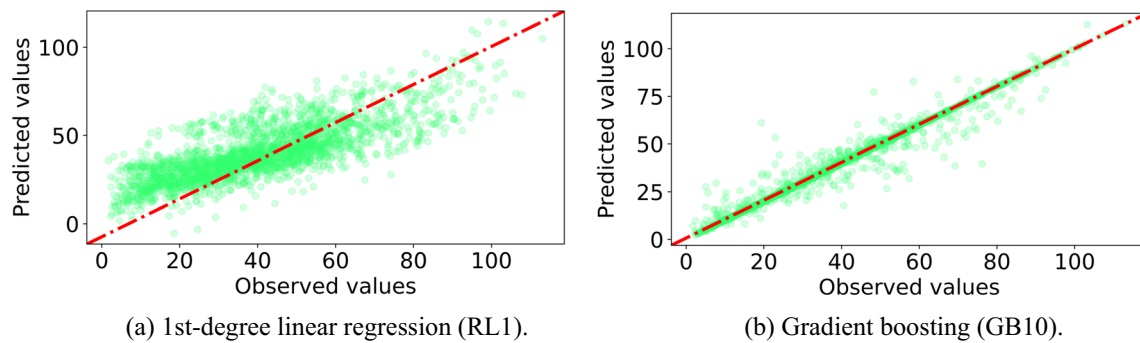


Fig. 9 Predicted values *versus* observed values: **a** 1st-degree linear regression (RL1) and **b** Gradient boosting (GB10)

Table 4 R^2 of the models in the iterative cross-validation phase

Model	Typo	Description	Cross-validation (%) ^a	Train (%)	Test (%)
Linear regression	Regression	RL1	60.98 ± 0.06	61.91	55.85
Polynomial regression		RL2	77.74 ± 0.07	79.03	74.41
		RL3	84.92 ± 0.17	87.66	85.40
Ridge linear regression	Regression	RR1	60.98 ± 0.06	61.92	55.84
Ridge polynomial regression		RR2	77.74 ± 0.07	79.02	74.38
		RR3	85.05 ± 0.12	87.56	85.15
Decision tree	Tree	AR2	45.09 ± 0.29	47.30	44.82
		AR5	68.97 ± 0.23	75.01	67.95
		AR10	79.06 ± 0.34	96.48	76.65
		AR20	79.06 ± 0.35	99.83	77.14
Gradient boosting	Ensemble	GB10	87.35 ± 0.19	99.83	86.33
		GB20	80.43 ± 0.31	99.83	76.95

^aConfidence level = 95%

the average of the observations in the database, and y is the vector of the observations in the database. In addition, other evaluation metrics, such as Model Error (ε_{mod}) and Mean Absolute Error (MAE) given by Eqs. (12) and (13) are used

$$R^2 = 1 - \frac{(\hat{y} - y)^2}{(y - \bar{y})^2} \quad (11)$$

$$\varepsilon_{\text{mod}} = y / \hat{y} \quad (12)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (13)$$

At the end of the simulation and analysis of the results, a Python algorithm was built that loads the trained model into a Jupyter notebook, allowing users of this platform to download and use the Artificial Intelligence model.

3.3 Model selection

The selected model for deployment on the World Wide Web must have an accuracy greater than 80%. Just one machine learning model will be chosen. In Civil Engineering problems, for a model to be considered accurate, it must present a Coefficient of Determination more incredible than 70%, and the closer to 100%, the more precise it is (Arroyo et al. 2020, 2023; Montgomery 2013).

4 Results

Based on Fig. 8, the cross-validation results are presented (Table 4). It is possible to notice that the data have a non-linearity, since the linear regressions could not consistently represent the data. For these models, in particular, the value of the coefficient R^2 was lower than 65%. The curvilinear versions of the regressions presented an R^2 greater than

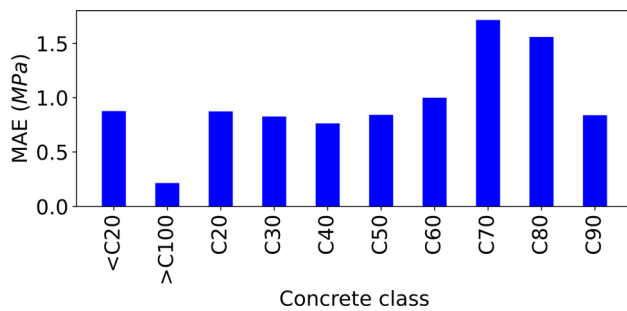


Fig. 10 MAE as a function of concrete strength class

75%, showing that non-linear parameters were necessary to improve the accuracy of the predictive model.

In the test stage, the three models with the highest accuracy were the Ensemble GB10 type model and the RL3 and RR3 regressions. In this work, the model with the best accuracy for the tested examples was Gradient Boosting GB10, with an accuracy of 86.33%. Although other models present an R^2 not much lower than the previously mentioned models, it was possible to detect overfitting in data, since some models, such as the AR20 tree model, presented an accuracy greater than 99% in training, while in the test, this value reduced to the 75% range. In this work, the GB10 model was selected for uploading and making available online (<https://wmpjrufg.github.io/Concreta/>).

Figure 9 shows the comparison between the forecasted values and the actual values based on trained models. It is possible to observe that the 1st-degree linear regression (RL1) has the most dispersed values when compared to the Gradient Boosting model (GB10), which has a concentration of values around the diagonal line $\hat{y} = y$ ($\varepsilon_{\text{mod}} = 1.00$), showing the efficiency of the latter model in predicting the data.

To predict the bias of the predictive model, the model error variable (Eq. (12)) for the predictions was calculated. In the case of this Artificial Intelligence model, the model error variable (ε_{mod}) has a mean value of 0.9989 for the bias factor. Therefore, it can be concluded that the predictive model tends to overestimate the compressive strength (f_c) slightly compared to the observed result.

Still evaluating the predictive model, it is possible to calculate the Mean Absolute Error (MAE). This analysis categorized the concretes into strength classes ranging from 5 to 115 MPa. The MAE was higher in the higher strength classes of concrete, particularly in the 70 MPa to 90 MPa range. However, the MAE values did not exceed 2 MPa for these concrete classes, which is significantly lower than the compressive strength. Figure 10 illustrates these MAE values across all concrete classes.

The confidence interval of the predictive model response was calculated to present the result of the strength prediction

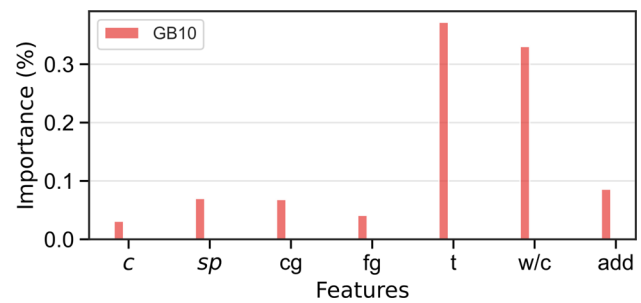


Fig. 11 Feature importance in GB10 model

on the platform. For this, a confidence level of 95% was used, reaching an error of only ± 0.48 MPa.

Additionally, an analysis of the importance of the variables in the Gradient Boosting model was carried out. Gradient Boosting models capture complex non-linear connections among variables, and their variable importance scores are based on how much each variable contributes to reducing the model's loss function. The importance analysis of the GB10 model is presented in Fig. 11 and shows results similar to Table 2, which classified the variables t , w/c , and add as the most important.

In addition to the numerical results presented above, a web platform was built using the Python Streamlit framework. On this website, it is possible to have access to the AI created in this article. User can enter his mixture design and get the f_c value based on the AI predictive model. The framework's interface presents the process results and the model's error rate. Figure 12 shows the program's interface. It should be noted here that the program is online and has a desktop and mobile version.

5 Concluding remarks

This work aimed to evaluate data mining methods for studying compressive concrete strength. It was possible to observe that non-linear models were more effective in extracting information from the concrete database, which comprised 2358 records.

The experimental data used in this research have good coverage; however, these values are unbalanced regarding compressive strength, as seen in Fig. 3d. Furthermore, the age values are concentrated on a date less than 50 days away (Fig. 3e). This situation may imply a more significant error for these input conditions, since the model does not have good accuracy outside this region.

The initial data treatment necessitated prior cleaning of the database, enabling the creation of crucial variables in the mixture design, such as the water–cement ratio. Furthermore, visualizing the data before training allowed the validation of the authenticity of the database by confirming the negative



Hello, I am Concreta, the first Brazilian AI built to determine the concrete strength (fck) based on dosage data. I am one of the creations of the GPEE (Research and Studies Group in Civil Engineering) of UFCAT, and I am here to help you, Civil Engineer and engineering student who wants to determine the properties of concrete.

Inputs

Please, fill in the table with the mixture parameters. For your consideration, here is an example:

- Cement consumption (kg/m³): 222.36
- Superplasticizer consumption (kg/m³): 4.46
- Coarse aggregate consumption (kg/m³): 967.08
- Fine aggregate consumption (kg/m³): 870.32
- Curing time (days): 3
- Water-cement ratio: 0.85
- Filler materials consumption (kg/m³): 96.67

What is your platform?

- ☒ PC
☐ mobile

Fig. 12 CONCRETA prediction framework interface

correlation between the water–cement ratio and compressive concrete strength, as observed in the experiments.

During the application of the methods, the iterative cross-validation technique was used to ensure comprehensive testing of the dataset, ensuring that the selected model possesses the ability to generalize. This factor is of paramount importance in Artificial Intelligence (AI) tasks. In this case, 30 repetitions were used in the validation phase.

Simple models were utilized in this work, and the results proved satisfactory. The model with the highest generalization ability was an Ensemble-type model called Gradient Boosting. With this model, an accuracy greater than 85% was achieved, and subsequently, a predictive model was developed, which is available for download on the portal: <https://wmpjrufg.github.io/Concreta/>.

This research contribution provides individuals and organizations access to state-of-the-art technology based on Machine Learning, enabling them to analyze mixture designs even before conducting experimental tests with cylinder specimen ruptures. Consequently, this work streamlines and reduces the time and resources expended in semi-empirical mixture design. It is important to note that AI should not replace the traditional compression test regulated by NBR 5739 (2018) but rather be used as an additional tool to increase productivity in concrete production.

As a suggestion for future work, we recommend adding new databases to broaden the coverage of the predictive model, especially in strength ranges with the highest mean absolute error, as shown in Fig. 9. Additionally, creating specific models for cement mortars and permeable concrete could be valuable. Including data on new cementitious

materials could expand the potential applications of these techniques in civil engineering.

Authorship contributions L. E. A. Cruvinel: writing, data curation, and formal analysis; W. M. Pereira Junior: conceptualization, funding acquisition, writing, methodology, and formal analysis; A. I. Campos: conceptualization, funding acquisition, writing, methodology, and formal analysis; R. P. Espíndola: machine learning analysis and formal analysis. A. P. Sarmiento: methodology and statistical analysis; D. L. Araújo: methodology and revision; G. A. Costa: machine learning analysis; R. V. Dutra online framework, writing, and revision.

Funding This research received no external funding.

Data availability All data used in the analysis are available on website: <https://wmpjrufg.github.io/Concreta/>.

Declarations

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this paper.

Ethical approval This article contains no studies with human participants or animals performed by any authors.

References

- Abbass W, Khan M, Mourad S (2019) Experimentation and predictive models for properties of concrete added with active and inactive SiO₂ fillers. *Materials* 12(2):299
- Alabduljabbar H, Khan M, Awan HH et al (2023) Predicting ultra-high-performance concrete compressive strength using gene expression programming method. *Case Stud Constr Mater* 18:e02074
- Al-Shamiri AK, Yuan T-F, Kim JH (2020) Non-tuned machine learning approach for predicting the compressive strength of high-performance concrete. *Materials* 13(5):1023

- Arroyo FN, Christoforo AL, Salvini VR et al (2020) Development of plaster foam for thermal and acoustic applications. *Constr Build Mater* 262:120800
- Arroyo FN, Borges JF, Junior WMP et al (2023) Estimation of flexural tensile strength as a function of shear of timber structures. *Forests* 14(8):1552
- ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (2018) ABNT NBR 5739 - Concreto - Ensaios de compressão de corpos-de-prova cilíndricos. ABNT, Rio de Janeiro
- Beck SM (2009) Efeitos nas propriedades mecânicas, elásticas e de deformação em concretos com altos teores de escória e cinza volante. Mestrado em Engenharia Civil, Universidade Federal de Santa Maria, Santa Maria
- Behnood A, Golafshani EM (2020) Machine learning study of the mechanical properties of concretes containing waste foundry sand. *Constr Build Mater* 243:118152
- Bilim C, Atiş CD, Tanyildizi H et al (2009) Predicting the compressive strength of ground granulated blast furnace slag concrete using artificial neural network. *Adv Eng Softw* 40(5):334–340
- Bouzoubaâ N, Fournier B (2003) Optimization of fly ash content in concrete. *Cem Concr Res* 33(7):1029–1037
- Breiman L (1998) Classification and regression trees, 1st edn. Chapman & Hall/CRC, Boca Raton
- Chopra P, Sharma RK, Kumar M (2016) Prediction of compressive strength of concrete using artificial neural network and genetic programming. *Adv Mater Sci Eng* 2016:1–10
- Chou J-S, Tsai C-F, Pham A-D et al (2014) Machine learning in concrete strength simulations: multi-nation data analytics. *Constr Build Mater* 73:771–780
- Demirboğa R, Türkmen İ, Karakoç MB (2004) Relationship between ultrasonic velocity and compressive strength for high-volume mineral-admixed concrete. *Cem Concr Res* 34(12):2329–2336
- Duran-Atiş C (2005) Strength properties of high-volume fly ash roller compacted and workable concrete, and influence of curing condition. *Cem Concr Res* 35(6):1112–1121
- Durán-Herrera A, Juárez CA, Valdez P et al (2011) Evaluation of sustainable high-volume fly ash concretes. *Cem Concr Compos* 33(1):39–45
- Erdal HI (2013) Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction. *Eng Appl Artif Intell* 26(7):1689–1697
- Fan DQ, Yu R, Shui ZH et al (2020) A new design approach of steel fibre reinforced ultra-high performance concrete composites: experiments and modeling. *Cem Concr Compos* 110:103597
- Feng D-C, Liu Z-T, Wang X-D et al (2020) Machine learning-based compressive strength prediction for concrete: an adaptive boosting approach. *Constr Build Mater* 230:117000
- Géron A (2019) Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, 2nd edn. Sebastopol, CA, O'Reilly Media, Inc, Beijing [China]
- Güçlüer K, Özbeyaz A, Göymen S et al (2021) A comparative investigation using machine learning methods for concrete compressive strength estimation. *Mater Today Commun* 27:102278
- Hild-Aono A, Gonzaga-Pimenta RJ, Francisco FR et al (2022) Machine learning for crop science: applications and perspectives in maize breeding. *Rev Bras Milho Sorgo*. <https://doi.org/10.18512/rbms2022vol21e1257>
- Homma T, Saltelli A (1996) Importance measures in global sensitivity analysis of nonlinear models. *Reliab Eng Syst Saf* 52(1):1–17
- Igual L, Seguí S (2017) Introduction to data science. Springer International Publishing, Cham
- Isinkaye FO, Folajimi YO, Ojokoh BA (2015) Recommendation systems: principles, methods and evaluation. *Egypt Inform J* 16(3):261–273
- Jiang LH, Malhotra VM (2000) Reduction in water demand of non-air-entrained concrete incorporating large volumes of fly ash. *Cem Concr Res* 30(11):1785–1789
- Kang M-C, Yoo D-Y, Gupta R (2021) Machine learning-based prediction for compressive and flexural strengths of steel fiber-reinforced concrete. *Constr Build Mater* 266:121117
- Komura D, Ishikawa S (2018) Machine learning methods for histopathological image analysis. *Comput Struct Biotechnol J* 16:34–42
- Lee KM, Lee HK, Lee SH et al (2006) Autogenous shrinkage of concrete containing granulated blast-furnace slag. *Cem Concr Res* 36(7):1279–1285
- Madabhushi A, Lee G (2016) Image analysis and machine learning in digital pathology: challenges and opportunities. *Med Image Anal* 33:170–175
- McCabe KO, Mack L, Fleeson W (2012) A guide for data cleaning in experience sampling studies. In: Mehl MR, Conner TS (eds) *Handbook of research methods for studying daily life*, pp 321–338, The Guilford Press
- Milhomem DA, Dantas MJP (2020) Analysis of new approaches used in portfolio optimization: a systematic literature review. *Production* 30:e20190144
- Mirjalili S, Faris H, Aljarah I (2020) Evolutionary machine learning techniques: algorithms and applications. Springer Singapore, Singapore
- Montgomery DC (2013) Design and analysis of experiments, 8th edn. John Wiley & Sons Inc, Hoboken, p 103
- Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. *Front Neurobot*. <https://doi.org/10.3389/fnbot.2013.00021>
- Oliveri P, Malegori C, Simonetti R et al (2019) The impact of signal pre-processing on the final interpretation of analytical outcomes—a tutorial. *Anal Chim Acta* 1058:9–17
- Oner A, Akyuz S (2007) An experimental study on optimum usage of GGBS for the compressive strength of concrete. *Cem Concr Compos* 29(6):505–514
- Oza NC (2000) “Online Ensemble Learning”, em Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, Austin, Texas, USA, p 1109
- Pakzad SS, Roshan N, Ghalehnovi M (2023) Comparison of various machine learning algorithms used for compressive strength prediction of steel fiber-reinforced concrete. *Sci Rep* 13(1):3646
- Pala M, Özbay E, Öztaş A et al (2007) Appraisal of long-term effects of fly ash and silica fume on compressive strength of concrete by neural networks. *Constr Build Mater* 21(2):384–394
- Pitroda J (2014) Prediction of strength for fly ash cement concrete through soft computing approaches. *Int J Adv Res Eng, Sci Manag* 1:1–11
- Rauecker JCN, Pereira Junior WM, Pituba JJDC et al (2019) Uma abordagem experimental e numérica para determinação de curvas de compressão para concreto simples e reforçados com fibras de aço. *Matéria (Rio De Janeiro)* 24(3):e12476
- Ribeiro DV, Pinto SA, Amorim Júnior NS et al (2021) Effects of binders characteristics and concrete dosing parameters on the chloride diffusion coefficient. *Cem Concr Compos* 122:104114
- Russell SJ, Norvig P (2016) Artificial intelligence: a modern approach, 3rd edition, Global edition, Boston Columbus Indianapolis New York San Francisco Upper Saddle River Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo, Pearson
- Scikit-Learn (2024) Decision trees. <https://scikit-learn.org/stable/modules/tree.html#id2>
- Shalev-Shwartz S, Ben-David S (2014) Understanding machine learning: from theory to algorithms. Cambridge University Press, New York

- Singh SB, Munjal P, Thammishetti N (2015) Role of water/cement ratio on strength development of cement mortar. *J Build Eng* 4:94–100
- Sola J, Sevilla J (1997) Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Trans Nucl Sci* 44(3):1464–1468
- Sonebi M (2004) Medium strength self-compacting concrete containing fly ash: modelling using factorial experimental plans. *Cem Concr Res* 34(7):1199–1208
- Tai Y (2021) A survey of regression algorithms and connections with deep learning. *arXiv*
- Tutikian BF, Helene P (2011) *Dosagem dos Concretos de Cimento Portland*. Concreto: Ciência e Tecnologia, Ibracon, 2011.
- Yaseen ZM, Tran MT, Kim S et al (2018) Shear strength prediction of steel fiber reinforced concrete beam using hybrid intelligence models: a new approach. *Eng Struct* 177:244–255
- Yeh I-C (1998) Modeling of strength of high-performance concrete using artificial neural networks. *Cem Concr Res* 28(12):1797–1808
- Yen T, Hsu T-H, Liu Y-W et al (2007) Influence of class F fly ash on the abrasion–erosion resistance of high-strength concrete. *Constr Build Mater* 21(2):458–463
- Zhang J, Huang Y, Aslani F et al (2020a) A hybrid intelligent system for designing optimal proportions of recycled aggregate concrete. *J Clean Prod* 273:122922
- Zhang Z, Zohren S, Roberts S (2020b) Deep learning for portfolio optimization. *J Financ Data Sci* 2(4):8–20

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.