Research, Society and Development, v. 11, n. 7, e27011729902, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i7.29902

A statistical analysis of the relationship of civil construction GDP to cement production in Brazil

Uma análise estatística da relação do PIB da construção civil com a produção de cimento no Brasil

Un análisis estadístico de la relación del PIB de la construcción civil com la producción de cemento em Brasil

Received: 05/03/2022 | Reviewed: 05/11/2022 | Accept: 05/18/2022 | Published: 05/23/2022

Ana Carolina Rodrigues da Rocha Souza
ORCID: https://orcid.org/0000-0003-2497-5091
Universidade Federal de Ouro Preto, Brazil
E-mail:profanacsouza@gmail.com
Helton Cristiano Gomes
ORCID: https://orcid.org/0000-0002-9432-5222
Universidade Federal de Ouro Preto, Brazil
E-mail: helton.gomes@ufop.edu.br
Irce Fernandes Gomes Guimarães
ORCID: https://orcid.org/0000-0002-6530-9434
Universidade Federal de Ouro Preto, Brazil
E-mail: irce@ufop.edu.br

Abstract
The ICC plays an important role in the Brazilian economy. This participation in the country's GDP remains, on average, above 5% per year. Cement, one of the main resources in this context, is used in almost all types of constructions in the country. The Brazil are among the 10 largest producers in the world and cement be the main component of concrete, makes widely used. The generation of different data is the starting point for decisions, optimization and forecasting of the activities of this communication network. To transform this data into information, many institutions use with tools such as Data Science. In this sense, this article presents the result of analysis of the behavior of cement production in Brazil, based on results generated through Machine Learning. Trends and seasonality periods were identified, as well as prediction models for future periods were proposed. Verified the existence of a strong positive correlation between cement production and the ICC GDP in Brazil. Machine Learning models were proposed and compared to predict the ICC GDP based on the annual cement production in Brazil, which showed high accuracy. It was concluded that the Ensemble Learning methods adapted better to the data, especially Random Forest.
Keywords: Construction; Cement production; GDP; Data science; Machine Learning.

Resumo
A ICC desempenha um papel importante na economia brasileira. Essa participação no PIB do país permanece, em média, acima de 5% ao ano. O cimento, um dos principais recursos neste contexto, é utilizado em quase todos os tipos de construções do país. O Brasil está entre os 10 maiores produtores do mundo e o cimento é o principal componente do concreto, faz com que seja amplamente utilizado. A geração de diferentes dados é o ponto de partida para decisões, otimização e previsão das atividades desta rede de comunicação. Para transformar esses dados em informações, muitas instituições utilizam ferramentas como Data Science. Nesse sentido, este artigo apresenta o resultado da análise do comportamento da produção de cimento no Brasil, com base em resultados gerados por meio de Machine Learning. Foram identificadas tendências e períodos de sazonalidade, bem como propostos modelos de previsão para períodos futuros. Verificou-se a existência de uma forte correlação positiva entre a produção de cimento e o PIB ICC no Brasil. Modelos de aprendizado de máquina foram propostos e comparados para prever o PIB do ICC com base na produção anual de cimento no Brasil, que apresentou alta precisão. Concluiu-se que os métodos de Ensemble Learning se adaptaram melhor aos dados, principalmente o Random Forest.
Palavras-chave: Construção civil; Produção de cimento; PIB; Ciência de dados; *Machine Learning*.

Resumen
ICC juega un papel importante en la economía brasileña. Esta participación en el PIB del país se mantiene, en promedio, por encima del 5% anual. El cemento, uno de los principales recursos en este contexto, se utiliza en casi todo tipo de construcciones en el país. Brasil se encuentra entre los 10 mayores productores del mundo y el cemento es el principal componente del hormigón, por lo que es ampliamente utilizado. La generación de diferentes datos es el

punto de partida para la toma de decisiones, optimización y previsión de las actividades de esta red de comunicación. Para transformar estos datos en información, muchas instituciones utilizan herramientas como Data Science. En ese sentido, este artículo presenta el resultado del análisis del comportamiento de la producción de cemento en Brasil, a partir de resultados generados a través de Machine Learning. Se identificaron periodos de tendencias y estacionalidad, así como se propusieron modelos de pronóstico para periodos futuros. Hubo una fuerte correlación positiva entre la producción de cemento y el PIB de ICC en Brasil. Se propusieron y compararon modelos de aprendizaje automático para predecir el PIB ICC basado en la producción anual de cemento en Brasil, que mostró una alta precisión. Se concluyó que los métodos de Ensemble Learning se adaptaron mejor a los datos, especialmente Random Forest.
Palabras clave: Construcción civil; Producción de cemento; PIB; Ciencia de datos; Aprendizaje Automático.

1. Introduction

An important sector in Brazil, the construction industry (ICC) has a strategic role in the country's economic growth. According to the Brazilian Construction Industry Chamber. (CBIC), the ICC is responsible for 6.2% of the GDP and corresponds to 34% of the total industry in the country. According to Teixeira and Carvalho (2005), the importance of the ICC for the economy occurs through the generation of income, arising from job offers, the need for inputs, heating up other sectors, and the collection of taxes. According to CBIC, approximately 13 million Brazilians work, directly or indirectly, in the sector. For the government, the return is great, 25% of what is invested in civil construction goes back to the public safes in the form of taxes. ICC also has a direct impact on the development of society, in which it is responsible for essential infrastructure. In the first quarter of 2021, the ICC GDP in Brazil grew 2.1%, compared to the fourth quarter of 2020, surpassing the rise in the national GDP and proving the strength of the sector in the country's economy.

One of the main inputs of the ICC, cement has a vast market, being used in buildings, roads and infrastructure works. According to De Melo et al. (2018), cement is the main component of concrete, the second most consumed product in the world, only behind water. According to Araújo (2020), the cement market moves around U$ 250 billion a year worldwide.

The annual consumption of cement signals different types of demands and capturing this data is essential for the planning of an industry, since the interpretation of different types of data generates reliable information, which helps in decision-making (CAO, 2017). Organizations are establishing more and more oriented strategies based on information and knowledge acquired through data processing, adopting the Data Driven culture. That is, the insertion of new technologies increases the generation, capture, transmission, storage and availability of data in a short period of time. According to Santovena (2013), the diversity of data, which can be complementary and fill existing gaps, can provide more accurate information and, with that, companies are able to foresee and improve their operations.

Vicario and Cleman (2020), present the Data Science (CD) as tools capable of processing and interpreting data, collected from different sources, and turning them into relevant information for organizations. The analyses performed by the CD can be divided into four classes: descriptive, diagnostic, predictive and prescriptive. In the descriptive analysis, the main trends existing in the data set and situations that lead to new facts are pointed out, allowing one to understand the events in real time. Diagnostic analysis aims to identify and understand the causes of variation in data behavior. In predictive analysis, future scenarios based on patterns found in the data set are estimated, enabling a more accurate decision making. The prescriptive analysis aims to verify the consequences of the actions taken, which makes it possible to know what should happen when opting for a certain alternative. With these analyses, it is possible to strategize to improve the organizations results.

According to Rezende and Abreu (2013), the main CD tools used to carry out the analyses are Data Mining (DM), a set of models capable of extracting information from a database and transforming it into useful and strategic knowledge for organizations, and Machine Learning (ML), a set of models capable of performing estimates and helping to predict

future scenarios, strategically offering support to corporations. Data analyses through these resources can bring great gains to different sectors of the economy, being able to help, for instance, in the planning and control of activities in the construction industry.

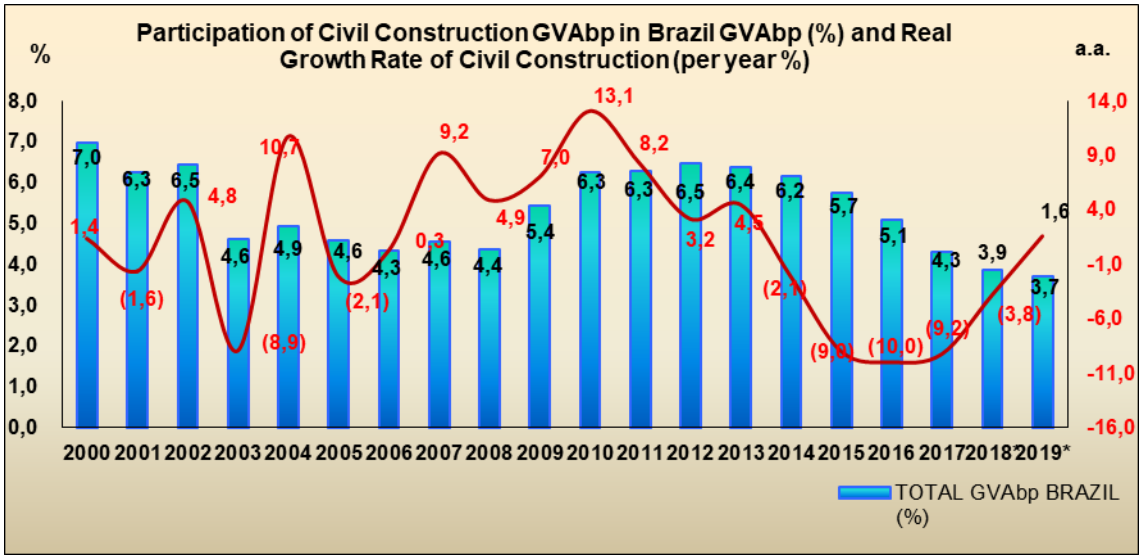
The ways that the generation of information obtained by these data analyzes are diverse, as prediction models can help to establish production, sales and inventory strategies, and can also direct investments more efficiently, taking advantage of the trends identified in the cement production throughout the year.

In this sense, this article aims to present an analysis of the behavior of cement production in Brazil, based on results generated through Machine Learning. For this purpose, the next sections were organized as follows: in section 2 the economic importance of the civil construction industry is presented. Section 3 presents the methodology used in this article. In section 4 the configuration of the experiment is specified, which includes data sets, algorithms and accuracy measures, it also presents the results and discussions and in section 5 the conclusions and guidelines for future work are pointed out.

2. The Economic Importance of the Brazilian Civil Construction Industry

The civil construction sector has a worldwide significance, represents an average of 13% participation in the global economy and boosts other industries through the provision of infrastructure and facilities, which entails a significant contribution to a country's GDP (AJAYI and OYEDELE, 2018). In Brazil, this is the second sector that generates the most jobs in the country, losing only to the wholesale and retail sector (CNI, 2019). In this sense, it is possible to observe the positive or negative impacts that the ICC activities can generate on the country's GDP. As an example, the year 2000 ended with a positive rate of 7.0% in the Brazilian GDP and growth of 1.4% in the ICC. In 2010, the Brazilian GDP grew 6.3% and had an expressive growth of 13.1% of the ICC GDP. The year 2019 had a growth of 1.6% compared to the previous year in the ICC GDP (CBIC, 2020). The graph in Figure 1 shows these behaviors.

Figure 1. Correlation of the Brazilian GDP with the Civil Construction Industry GD.



Source: CBIC (2020).

Normally, when there is an increase in cement consumption, there is a growth in the ICC GDP. According to Stafford et al. (2015) cement has a direct relationship with the growth of the IC, with the main producers in Latin America being Brazil, Argentina and Mexico. Regarding the Brazilian production in the year 2020, SNIC (2020) indicated that Brazil was responsible for the production of 60.8 million metric tons of cement, recording an increase of 10.9% compared

to the previous year. The quick reaction is justified by the sector's ability to capture relevant data for analysis and generation of important information. However, the construction industry sector, similar to other industries, generates a large amount of data throughout the project lifecycle, due to the advent of techniques such as Machine Learning and Data Mining, it is possible to understand hidden trends and patterns, predict future events, find correlations and insights. These perceptions will help to optimize construction costs, make faster and more assertive decisions and offer new products and services that meet the needs of the sector (Hussain et al. 2018).

3. Methodology

3.1 Data Science

Data Science (CD) is a multidisciplinary area dedicated to the study and analysis of data with the objective of formalizing knowledge, identifying patterns and obtaining subsidies for decision making. The application of the CD ranges from data preparation, including cleaning, aggregation, standardization, identification of outliers and treatment of missing values, to more advanced analyses. Among the CD tools, exploratory analysis, data mining and machine learning were used in this article.

3.1.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) analyzes data sets seeking to identify their main characteristics, using tables and graphs, as well as other visualization methods. According to Tukey (1977), EDA analyzes data freely, aiming to extract information and not limiting to testing only a single pre-defined hypothesis. EDA makes it possible to identify, visually or through statistical tests, important components such as, for instance, trends, cycles and seasonality in the data of a Time Series (ST), as well as explaining the reasons for these variations.

According to Bezerra et al. (2019), EDA has a wide variety of approaches, methodologies and techniques for its implementation in a simple form, such as statistical tools and open source computational libraries for data analysis.

3.1.2 Data Mining

DM is the process of finding anomalies, trends, patterns, and correlations in data sets to predict future results. Analyses made using DM techniques can be divided into two types, descriptive and predictive. According to Mccue (2007), descriptive analysis plays the role of analyzing past and/or current data, seeking to describe trends or patterns for information generation. Predictive analysis aims to determine the probable future result of an event or the probability of a current state that is unknown.

DM is based on three areas, statistics, which numerically studies the relationships between data, artificial intelligence, displayed by software and/or machines that simulate human reasoning ability in decisions, and machine learning, models capable of learning from data and make predictions. Machine learning is directly linked to DM, since it deals with models that seek to identify patterns in data sets and, in DM, these models are applied in search of information and knowledge (Ferrari & Silva, 2017).

3.2 Machine Learning

ML is a branch of artificial intelligence capable of automating the construction of analytical models, based on the idea that machines can learn from data. According to Evchenko et al. (2021), ML enables computers to learn without being explicitly programmed. Pacheco and Pereira (2018) and Kantardzic (2011) state that the ML models can be divided based on labeled data, that is, where the response variable is known. Being the models classified into: **a) supervised:** learning

takes place from a set of labeled data; **b) unsupervised:** the input data do not present labels and the model needs to automatically find patterns and group them properly according to their similarities; **c) semi-supervised:** similar to supervised learning, but a small amount of data is labeled and a larger amount does not have labels; **d) by reinforcement:** learning based on experience, in which the machine must deal with what went wrong before and search for the correct way.

ML models perform prediction, classification and/or grouping tasks. Prediction models are used to estimate future values based on past data. They also help to identify the existence of a correlation between two variables and define the cause-effect relationship. (Zhu, 2005)

Classification models define the class of new data based on input data. Regression and classification models are part of supervised learning models, that is, they require labeled data. Grouping models belong to unsupervised learning models, in which similar characteristics of the data are used to make a classification. (Kumar, 2014; Zhu, 2005; Pacheco e Pereira, 2018)

In this article, the following prediction models were used: Linear Regression, Decision Tree, Random Forest and Gradient Boosting.

3.2.1 Linear Regression

A Linear Regression analysis consists of determining a function to describe the statistical relationship between one or more predictor variables and a response variable. In this case, it is considered that the relationship between the variables is given by a linear function. Depending on the number of predictor variables considered, linear regression can be classified as simple or multiple. (Castro & Ferrari, 2016).

To Larose (2005), in Simple Linear Regression, the linear relationship between two variables, a predictor and a response, is evaluated. The relationship is represented by a line, creating a direct cause-effect relationship. In this way, it is possible to estimate values for the response variable based on the values of the predictor variable. This relationship is described by Equation (1), in which y is the response variable, and x the predictor.

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (1)$$

The parameter β_0 is the intercept on the y axis, that is, the expected value for y when x is equal to zero, β_1 is the slope of the line, that is, the expected variation for y when x varies by one unit and ε is a random error component used to include in the model the influences on the behavior of y that cannot be explained linearly by the behavior of the variable x (Montgomery et al., 2012).

In many cases, a single predictor variable is not able to explain the behavior of the response variable, making it necessary to include new predictor variables in the model. Thus, there is the Multiple Linear Regression, whose basic formulation is defined by Equation (2), in which y is the response variable and x_1 to x_n the predictors.

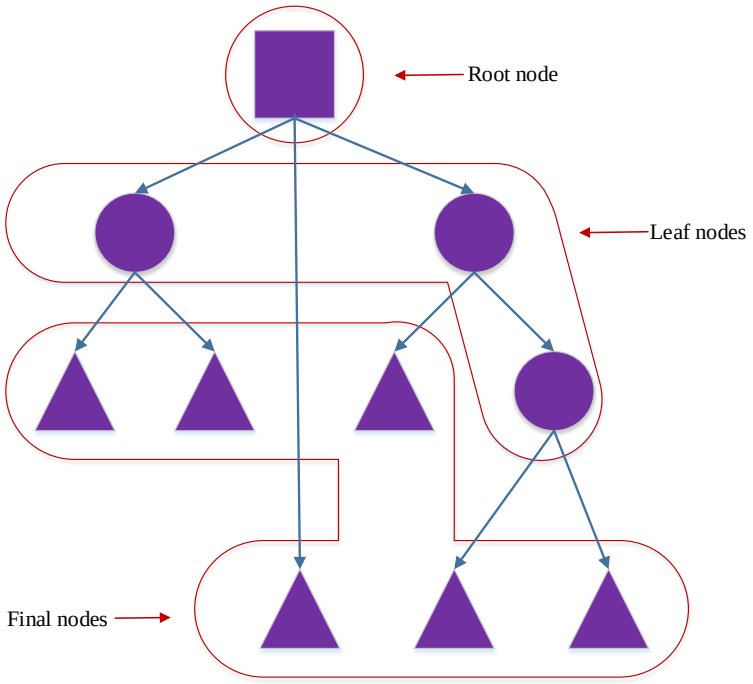
$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \quad (2)$$

The parameter β_0 is the expected value of y when all predictor variables are equal to zero, β_i, for all i different from zero, represents the variation of y when x_i varies by one unit, keeping all other predictor variables constant, and ε is a random error used to include influences on the behavior of variable y that cannot be explained linearly by the behavior of the predictor variables (Montgomery et al., 2012).

3.2.2 Decision Tree

Decision Trees (DT) are statistical models of supervised learning, used in prediction and classification tasks. Jijo and Abdulazeez (2021) define DT as "a successive model that efficiently and cohesively unites a series of basic tests, where a numerical resource is compared to a threshold value in each test". For these authors, conceptual rules are much easier to construct compared to numerical weights in the neural network of connections between nodes. This method uses tree-shaped graphs and visually presents the conditions and probabilities to arrive at results. The structure of a tree resembles a flowchart, where each node represents a test, the branches the possible test results, and the leaves represent classes. In Figure 2, it's possible to visualize an example of a decision tree.

Figure 2. Example of a decision tree.



Source: Adapted from Han et al., (2011)

It's possible to get better performance than using any ML model individually, using them together, creating new models called Ensemble Learning models (EL). The EL concept is based on the idea of combining simpler models, training them for a certain task and generating a more complex model, aiming to increase the accuracy of the prediction/classification. Examples of EL models are Random Forest and Gradient Boosting (Han et al., 2011).

3.2.3 Random Forest

Random Forest is an ML model used for prediction and classification tasks, combining several decision trees. According to Santos et al. (2019) this model aims to combine predictions from a set of complex classifiers (decision trees with many divisions) applied to bootstrap samples of the training set. The main difference is the random selection of predictors to be used, in order to reduce the correlation between the trees that will be aggregated to produce the final prediction. The result of a prediction is given by the average of the predictions of different trees. For the classification task, each tree determines a class and the most popular one is returned by the model. Increasing the number of trees, the precision of the result increases consecutively.

3.2.4 Gradient Boosting

Similarly to Random Forest, Gradient Boosting combines simpler supervised learning models, such as decision trees, and produces a more robust prediction model. According to Zhang and Haghani (2015), Gradient Boosting iteratively inserts additional decision trees into the model, correcting errors made by previous iteration models and improving prediction accuracy. Gradient Boosting is also used for prediction and classification tasks.

3.2.5 Time Series Analysis

According to Antunes and Cardoso (2015), time series (ST) can be defined as a sequence of quantitative data related to specific moments and studied according to their distribution over time. An ST is considered stationary when it develops in time randomly around a constant mean, reflecting some form of stable equilibrium. However, most of the ST have some form of non-stationarity, presenting components such as trend, seasonality and residue. An ST can be represented according to Equation (3).

$$Z(t) = T_t + S_t + N_t \quad (3)$$

Being $Z(t)$ the observed series, the trend (T_t) is defined as a pattern of increase/decrease that occurs gradually over time at the mean level of the ST. Seasonality (S_t) consists of behavior patterns in the ST values that are repeated at specific times of the year. The residual (N_t) is a component that contains all movements that do not belong to the trend or seasonality, that is, they are random movements that are not regular and are not repeated in a regular pattern.

The main objective when studying an ST is to identify its main characteristics/components, using the EDA, and, through specific prediction models, to estimate future values. The prediction models for ST used in this article are described below.

3.2.5.1 Moving average

The simple moving average model, or just moving average, estimates a value for the period t of an ST using the arithmetic mean of the last $t - r + 1$ observed values, according to Equation (4).

$$M_t = \frac{Z_{t-1} + Z_{t-2} + \ \dots \ + Z_{t-r+1}}{r} \quad (4)$$

In Equation (4), M_t is an estimate for Z_t and r is the number of observations included in the moving average. The values used to calculate M_t are the real values observed for the past r periods. The most common types are simple moving average, exponential moving average, and weighted moving average. In the exponential moving average, a greater weight is given to the most recent observations, unlike the simple moving average that considers all of them with equal weight. In the weighted moving average, factors are used to provide different weights for different observations.

3.2.5.2 ARIMA

The Autoregressive Integrated Moving Average (ARIMA) statistical model is one of the most used methodologies for ST analysis. According to Gujarati and Porter (2011), ARIMA is an autoregressive integrated moving average series and is

composed of three parameters (*p*, *d* and *q*): ***p*** denotes the number of autoregressive terms; ***d*** the number of times the series must be differentiated before becoming stationary; and ***q*** the number of moving average terms.

According to Box *et al.* (2015), the ARIMA model involves three statistical processes: autoregression (AR), integration (I) and moving averages (MA). The AR consists of expressing the autocorrelation of the observations, that is, how much the previous observation is capable of influencing the value of the next one. The value of I indicates the number of differences that will be necessary to ensure the stationarity of the ST.

3.2.5.3 PROPHET

PROPHET is a model developed to estimate future values for ST, having the capacity to deal with trends and seasonality. It is robust against missing data and trend changes and typically handles outliers well. According to Taylor and Letham (2017), PROPHET decomposes ST into its main components according to Equation (5).

$$y(t) = g(t) + s(t) + h(t) + \epsilon(t) \quad (5)$$

On which:

- $y(t)$ is the observed series;
- $g(t)$ is the trend that models the non-periodic changes in the ST value;
- $s(t)$ represents periodic changes (seasonality);
- $h(t)$ represents the effect of holidays, which occur on irregular dates for one or more days;
- $\epsilon(t)$ represents any peculiar changes that are not accommodated by the model.

3.3 Method

The construction of the models used in this article and the generation of the results for analyses were carried out using Python programming language. The methodology proposed in this work was divided into 7 phases. In the first phase, the cement production and ICC GDP in Brazil data were extracted from two databases, respectively the databases of the Institute for Applied Economic Research (IPEA) and the Brazilian Construction Industry Chamber (CBIC). The data refer to ST with monthly values, corresponding to the years 2003 to 2018. After extraction, in the second phase, a pre-processing was carried out, in which the data was cleaned and transformed. Missing values were not identified in the ST.

With the data prepared, in the third phase, a descriptive exploratory analysis was performed in order to identify patterns and trends in the behavior of ST of cement production in Brazil. To identify the causes of variations in the behavior of ST, a diagnostic analysis was performed, researching political, economic and ICC-related events.

In the fourth stage, to verify whether the cement production ST presents a stationary behavior, the Dickey-Fuller (DF) test was used, a unit root statistical test in which the following hypotheses are considered:

H_0: the ST has unit root

H_1: the ST does not have a unit root

According to the DF test, when the ST has a unit root, it is considered non-stationary. When performing the test, if the p-value found is less than or equal to 0.05 (95% of significance), H_0 is rejected and the ST can be considered stationary. Otherwise, H_0 is accepted and the ST is considered non-stationary.

After verifying the non-stationarity of the ST, predictive models were tested and compared to estimate future values for annual cement production in Brazil. The Moving Average, ARIMA and PROPHET models were used. The comparison between the models was made using the Root Mean Squared Error (RMSE), calculated by Equation (6):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta} - \theta)^2} \quad (6)$$

Where n is the number of test instances, $\hat{\theta}$ is the value found by the prediction model and θ is the real value obtained from the database (SAMMUT and WEBB, 2011).

After analyzing the ST of cement production in Brazil, in which the reasons for its behavior and the best prediction model for future values were identified, its relationship with the ST of the ICC GDP was analyzed.

In the fifth stage, to verify the influence of one variable on the behavior of another, it is first necessary to determine the degree of relationship between them. This statement is also valid for ST. Aiming to prove the existence of a relationship between the ST of the annual cement production and the GDP of the ICC in Brazil, the Pearson (ρ_p) and Spearman (ρ_s) correlation coefficients between them were calculated.

The ρ_p measures the degree of linear correlation between two quantitative variables $\boldsymbol{x}$ and $\boldsymbol{y}$, being calculated by Equation (7).

$$\rho_p = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (7)$$

In which x_i and y_i are pairs of n observations of the variables x and y and ρ_p varies between -1 and 1. The value zero indicates that there is no linear relationship between the two variables and the closer the absolute value is to 1, the stronger the linear relationship between the two variables. The sign indicates the direction of the relationship so that positive indicates that the two variables vary in the same direction, and negative indicates that they vary in the opposite direction.

ρ_s is a non-parametric correlation measure. Unlike ρ_p, it can indicate a not necessarily linear association between the variables, but it will indicate whether the relationship is increasing or decreasing. The value of ρ_s is calculated by Equation (8).

$$\rho_S = \frac{\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right)}{\frac{n(n^2-1)}{12}} \quad (8)$$

In which R_i and S_i are the variable ranks and n is the number of observations. The ρ_s also varies between -1 and 1, so it will indicate a decreasing or increasing relationship, respectively.

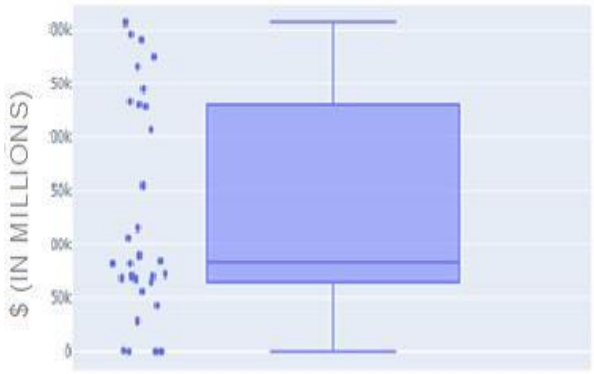
After confirming the strong correlation between the ST, four ML models were proposed to estimate the ICC GDP based on the annual cement production. To validate the models, a minimum accuracy was adopted, that is, a model will only be validated if its coefficient of determination (R^2) is greater than 80%. R^2 indicates the percentage of variability of the response variable that is explained by the predictor variable used in the model. The higher the R^2, the more explanatory the model is, that is, the better it fits the data. The models that reached the minimum R^2 were compared using the RMSE value.

In the sixth stage, to avoid overfitting in the prediction models, the data were divided into two sets, training and testing. Overfitting happens when a model performs well for the data set used in its creation and does not present the same accuracy for other sets. Therefore, the 192 ST inputs were divided as follows: 180 (93.75%) for training and 12 (6.25%) for testing. In the seventh stage, Linear Regression, Decision Tree, Random Forest and Gradient Boosting models were proposed to estimate future values for ICC GDP, based on annual cement production in Brazil.

4. Results and Discussion

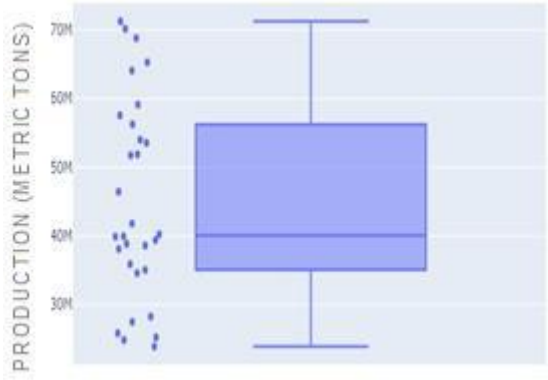
Graphs 1 and 2 show ST boxplots of annual cement production and ICC GDP, respectively.

Graph 1: Boxplot of annual cement production between 2003 and 2018.



Source: Authors (2021).

Graph 2: Boxplot of ICC GDP between 2003 and 2018.



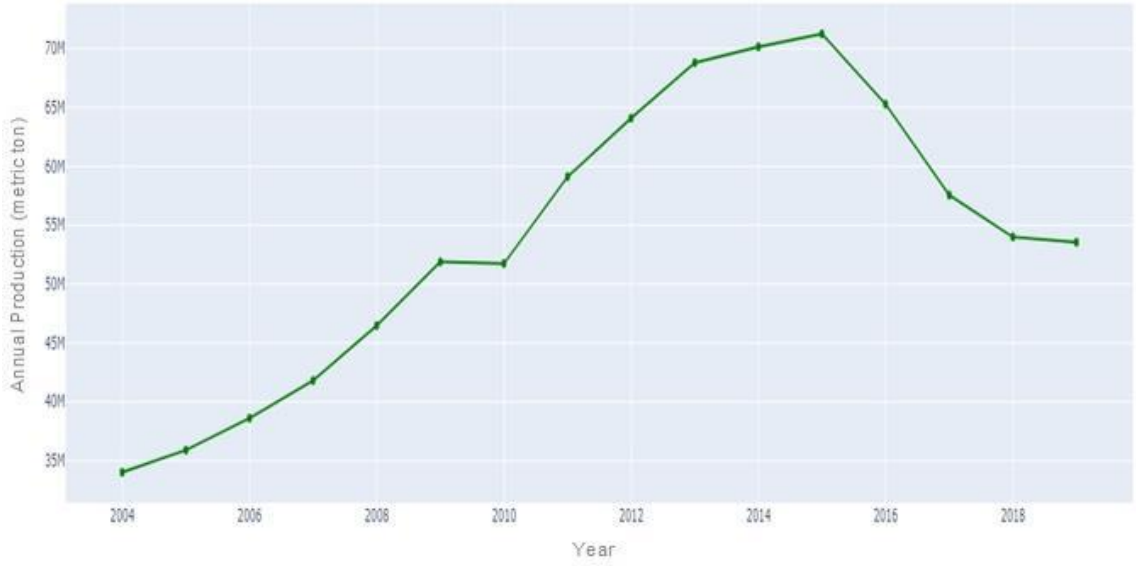
Source: Authors (2021).

It is observed in Graphs 1 and 2 that there is no presence of outliers in the STs. Therefore, to carry out the analyses, samples of size 192, that is, 16 full years, were obtained.

4.1 Exploratory Data Analysis of the ST of Cement Production in Brazil

Graph 3 shows the evolution of annual cement production, in metric tones, between 2003 and 2018.

Graph 3: Annual evolution of cement production - 2003 to 2018.



Source: Authors (2021).

According to Graph 3, 2003 was the year with the lowest cement production within the historical series, with 32,500 metric tons produced. According to Cunha (2012), this is justified by the fact that, in 2003, Brazil went through a period of low growth in the ICC sector, with no significant investments in infrastructure expansion or improvements. In the years that followed 2004 there was an increase in cement production due to public policies that earmarked investments for the infrastructure sector and also for housing programs. In the years from 2010 onwards, there has been a significant increase in

cement production in Brazil. In 2009, the "Minha Casa, Minha Vida" program was launched, which aimed to expand the number of homes for the lower social classes. According to Monteiro and Veras (2017), between 2009 and 2014, 1.7 million houses and apartments were delivered, benefiting a total of 6.8 million people. According to Leão *et al.* (2016), sporting events played an important role in heating up the ICC sector. In 2007 the Pan American Games took place, held in the city of Rio de Janeiro. In 2014 the International Federation of Association Football – FIFA – World Cup took place, held in 12 Brazilian capitals, which required large investments by the government for the renovation and construction of the stadiums, as well as the improvement of infrastructure in the cities that hosted the games. Later, in 2016, the Olympic Games took place in the city of Rio de Janeiro. And finally, it is observed that from 2015 onwards a year-by-year reduction in cement consumption begins.

A year-by-year analysis of the cement production ST was also carried out. Graphs 4, 5, 6 and 7 show the evolution of the ST, in metric tons, for the years 2003 to 2018.

.

Graph 4: Monthly cement production in the years 2003 to 2006.

Source: Authors (2021)

Graph 5: Monthly cement production in the years 2007 to 2010.

Source: Authors (2021).

Graph 6: Monthly cement production in the years 20011 to 2014.

Source: Authors (2021).

Graph 7: Monthly cement production in the years 2015 to 2018.

Source: Authors (2021).

Analyzing Graphs 4, 5, 6, and 7 it is possible to identify that, between the months of July and October, there is an increase in cement production, configuring a seasonal period. In field research, in which interviews were conducted with civil engineers, it was concluded that this characteristic is due to the dry period in most regions of Brazil. The period from December to April is composed of rainy seasons, which ends up resulting in stoppages and delays in the progress of constructions, leading to a consequent reduction in cement consumption. It is also noted that the month of May 2018 had the worst performance of the year, contrary to what happened in other years. This reduction is strongly related to the truck drivers' strike that took place in the same month, a fact that reduced the supply of cement in stores.

According to Gujarati and Porter (2011), one more analysis needs to be carried out. According to the authors, if the computed value for the test statistic exceeds the critical values, the ST is considered non-stationary. Table 1 shows the results of the DF test applied to the cement production ST in Brazil.

Table 1: DF results for the cement production ST.

Test Statistics	-2.057.684
p-value	0,261867
Critical Value (1%)	-3.468.062
Critical Value (5%)	-2.878.106
Critical Value (10%)	-2.575.602

Source: Authors (2021).

As it can be seen in Table 1, the values found by the DF for the p-value, the test statistic and the critical values were 0.261867, -2,057,684, -3,468,062, -2,878,106 and -2,575,602, respectively. It is verified, therefore, that p-value > 0.05 and that the test statistic exceeded the critical values, that is, there are statistical evidences that the cement production ST between 2003 and 2018 is non-stationary.

To be taken as a reference for a maximum allowed RMSE, a naive model (Baseline) was proposed. In the Baseline estimation, the value for period *t* is determined as equal to the real value for period *t* - 1. In the Moving Average model, the 3-year period was considered for the calculation of the estimates. For the ARIMA model, values 3, 1 and 1 were determined for its parameters (ARIMA (3,1,1)). These values were obtained using the *auto_arima* library of the Python language. Table 2 presents RMSE values for each model.

Table 2: RMSE of prediction models.

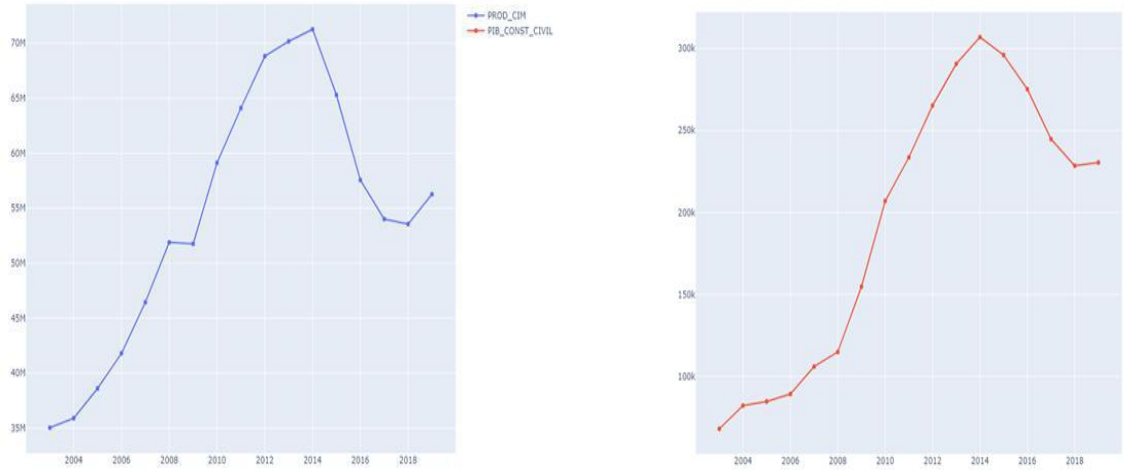
Model	***RMSE***
Baseline	303,243.52
Moving Average	421,182.67
ARIMA	327,361.63
PROPHET	169,565.83

Source: Authors (2021).

Based on the values presented in Table 2, PROPHET was the model that presented the lowest RMSE compared to the others. This result is mainly due to the ability of PROPHET to adjust to data sets that present non-stationary behavior. The Moving Average and ARIMA generated predictions with RMSE greater than the Baseline, which is justified by the fact that the models are not able to deal well with an ST that presents a trend and seasonality.

The values obtained for ρ_p e ρ_s, for the ST of the annual cement production and the GDP of the ICC in Brazil, were $\rho_p = \rho_s = 0.96$. Based on these values, it can be concluded that the ST have a strong positive and increasing linear correlation. Graph 8 shows the annual evolution of the ST in the period from 2003 to 2018.

Graph 8: Evolution of cement production ST and ICC GDP in Brazil.



Source: Authors (2021).

As can be seen in Graph 8, the ST show similar behavior, increasing and decreasing in the same periods, proving the strong positive correlation between them.

The values for R^2 and RMSE obtained by the models are shown in Table 3.

Table 3: Comparison of the accuracy of ML models.

Model	R^2	RMSE
Linear Regression	84%	36,374.85
Decision Tree	85%	35,107.20
Random Forest	98%	11,499.92
Gradient Boosting	88%	31,476.17

Source: Authors (2021).

According to the values described in Table 3, all models had the intended minimum accuracy, that is, $R^2 > 80\%$. Random Forest showed the best performance, with $R^2 = 98\%$, that is, the model explains 98% of the variability of the ICC GDP from cement production in Brazil. Regarding the RMSE, Random Forest also had the best result, 11,499.92. In second place was Gradient Boosting, with $R^2 = 88\%$ and RMSE = 31,467.17. Next was the Decision Tree model and, at last, the Linear Regression model.

Based on the analyzes carried out, it was found that there is a strong positive correlation between cement production and ICC GDP in Brazil, which allowed the use of prediction models. Analyzing the values for R^2 and RMSE obtained by the models, it was possible to conclude that the ones from Ensemble Learning were better adapted to the data sets, generating the predictions closer to the real values. Therefore, it is possible to make the extrapolation, estimating with high accuracy the ICC GDP based on cement production for future periods.

5. Final Considerations

In this article, descriptive, diagnostic and predictive analyses of cement production in Brazil were performed, using data from 2003 to 2018. The relationship between cement production and the ICC GDP in the country was also analyzed, with the intention of proposing prediction models. The ICC has great value in the economic scenario in Brazil, being one of the sectors that employ the most and which, when at an accelerated pace, moves several other sectors. The ICC has a significant participation in the Brazilian GDP, but it needs more investments in technology, which will improve the strategic and operational planning of both public and private organizations.

After extraction from the IPEA and CBIC databases, a pre-processing was carried out, in which the data cleaning and transformation process was carried out. There were no missing values or outliers in the ST. With the data prepared, descriptive and diagnostic analyses were carried out, identifying patterns and trends, as well as possible causes of variations in the behavior of cement production ST. Using the DF test, it was verified that the ST presents a non-stationary behavior. At last, the Moving Average, ARIMA and PROPHET models were tested and compared to estimate future values for annual cement production in Brazil. Based on the values obtained by the models for the RMSE, it was concluded that PROPHET generated the best estimates, proving its ability to adjust to data sets that present non-stationary behavior. The Moving Average and ARIMA were not able to deal well with the presence of trend and seasonality in the cement production ST.

To analyze the influence of annual cement production on the ICC GDP in Brazil, first the correlation coefficients ρ_p and ρ_s were calculated. $\rho_p = \rho_s = 0.96$ was obtained, indicating that there is a strong positive and increasing linear correlation

between the ST. Then, Linear Regression, Decision Tree, Random Forest and Gradient Boosting models were proposed to estimate the ICC GDP based on the annual cement production in Brazil. The models were validated considering a minimum accuracy, that is, $R^2 > 80\%$. All models reached the minimum R^2 and were then compared using the RMSE. Random Forest showed the best performance, with $R^2 = 98\%$ and RMSE = 11,499.92. Second place was Gradient Boosting, followed by Decision Tree and Linear Regression. Therefore, it is concluded that the Ensemble Learning models were better adapted to the ST, generating predictions closer to the real values. It is then possible to make extrapolations for future periods, estimating with high accuracy the ICC GDP based on cement production in Brazil.

The predictions made by the ML models can help ICC managers to implement strategies, ensuring competitive advantage in the market. Based on the analysis of future scenarios, it is possible for corporations to establish more assertive plans, as well as to assist the public administration in creating policies that promote the heating of the sector. When there are government policies encouraging the construction and improvements of infrastructure, there is an increase in cement production, which leads to increase in the ICC GDP and, consequently, in the economy in general.

References

Ajayi, S. O., & Oyedele, L. O. (2018) Waste-efficient materials procurement for construction projects: A structural equation modelling of critical success factors, *Waste Management*, 75, 60-69.

Antunes, J. L. F.& Cardoso, M. R. A. (2015) Uso da análise de séries temporais em estudos epidemiológicos. *Epidemiologia e Serviços de Saúde*, 24, 565–576.

Araujo, G. J. F. D. (Março 2020) O coprocessamento na indústria de cimento: definição, oportunidades e vantagem competitiva. *Revista Nacional de Gerenciamento de Cidades*, 8(57), 52–61.

Box, G. E. P., Jenkins, G. M.& Reinsel, G. C.& Ljung, G. M. (2015) *Time Series Analysis: Forecasting and Control.* (5a ed.), John Wiley & Sons.

Cao, L. (2016) Data science and analytics: a new era. Int J Data Sci Anal,. https://link.springer.com/article/10.1007/s41060-016-0006-1.

Cao, L.. (2017) Data science: A comprehensive overview. ACM Computing Surveys, 50(3).

Cao, L.. (2020) Data science: challenges and directions, 60(8), 59–68. https://cacm.acm.org/magazines/2017/8/219605-data-science/fulltext..

CBIC- Câmara Brasileira da Indústria da Construção.(2020) Banco de Dados CBIC. http://www.cbicdados.com.br/menu/pib-e-investimento/pib-brasil-e-construcao-civil.

Chatfield, C. (2013) *The analysis of time series: An Introduction* (6a ed.), Chapman & Hall/CRC.

CNI - Confederação Nacional da Indústria. (2019) Fato Econômico: Razões e condições da crise à recuperação do setor de construção. 1–3.

Cowpertwait, P. S. P. & Metcalfe, A. V. (2009) Introductory Times Series with R. Springer,.

Cunha, G. D. C. (2012) Importância do setor de Construção Civil para o desenvolvimento da Economia Brasileira e as alternativas complementares para o Funding do Crédito Imobiliário no Brasil. 79 f. Trabalho de Conclusão de Curso (Instituto de Economia) Universidade Federal do Rio de Janeiro. Rio de Janeiro.

De Melo, I. M., de Melo, M. M., & de Sá Morais, K. M. (2013). Direito ambiental como base para a gestão e responsabilidade ambiental da votorantim cimento de Sobral. *Essentia-Revista de Cultura, Ciência e Tecnologia da UVA*, 15(1).

Dua, S.& Du, X. (2016) *Data Mining and Machine Learning in Cybersecurity***.** (3a ed.), Imprensa CRC,.

Ehlers, R. S. (2009) Análise de séries temporais. Universidade Federal do Paraná. Curitiba,. p. 90.

Espíndola, A. M. S. de , Roth, L. , Camargo, M. E. & Fachinelli, A. C. (2016) Big Data e Inteligência Estratégica: Um Estudo de Caso Sobre a Mineração de Dados como Alternativa. *Revista Espacios*, 37, 16-35.

EVCHENKO, Mikhail et al. (2021) Frugal machine learning. arXiv preprint arXiv:2111.03731.

FARAH, M. F. S. (1996) Processo de Trabalho na Construção Habitacional: Tradicional e Mudança. Annablume.

Ferrari, D. G. & Silva, L. N. de C. (2017) Introdução a mineração de dados. Saraiva Educação SA.

Fu, T. C. (2011) A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24, 164–181,.

Gaspar, I. D. A., Gonçalves, M. R. & Matias, I. D. O. (2018) Time Series Prediction: Case study using artificial neural network techiniques for forecasting National Petroleum Production. *Interdisciplinary Scientific Journal.*, 5(1), 138-152.

Gujarati, D. N. & Porter, D. C. (2011) *Econometria*. (5a ed.), AMGH Editora,.

Gupta, C. B.& Guttman, I. (2018) *Estatística e probabilidade com aplicações para engenheiros e cientistas.* LTC ,.

Han, J., Kamber, M.& Pei, J. (2011) *Data Mining: Concepts and Techniques*. (3a ed.), Elsevier.

Hussain, K., Salleha, M. N. M., Talpur, S. & Talpura, N. (2018) Big Data and Machine Learning in Construction: A Review, *International Journal of Soft Computing and Metaheuristics.*

Ishizaki, M. Y. (2018) Reconhecimento automático de palavras 43 f. Trabalho de Conclusão de curso (Graduação em Engenharia de Controle e Automação) - Universidade Tecnológica Federal do Paraná, Cornélio Procópio,

Jacobs, W., Zanini, R. R. & Costa, M. (2014) Estudo comparativo de séries temporais para a previsão de vendas de um produto. Revista Iberoamericana de Engenharia Industrial, 6(12), 112-133,.

Jijo, B. T. & Abdulazeez, A. M. (2021) Classification Based on Decision Tree Algorithm for Machine Learning. evaluation, 6, 7.

Kantardzic, M.(2011) *Data Mining: Concepts, Models, Methods, and Algorithms*. (2a ed.), John Wiley & Sons.

Kirchgässner, G. & Wolters**,** J. (2007) Introduction to Modern Time Series Analysis. Springer.

Kumar, T. S.(2014) *Introduction to Data Mining.* (5a ed.), Pearson Education Limited.

Larose, D. T. (2005) *Discovering Knowledge in Data: An Introduction to Data Mining.* John Wiley and Sons, Inc.

Leão, A. L. M.de S., Ferreira, B. R. T. & Gomes, V. P. de M. (2016) Um "elefante branco" nas dunas de Natal? Uma análise pós-desenvolvimentista dos discursos acerca da construção da Arena das Dunas. *Revista de Administração Pública,* 50(4).

Mayrink, V. T. D. M. (2016) .Avaliação do algoritmo Gradient Boosting em aplicações de previsão de carga elétrica a curto prazo. 91f. Dissertação (Mestrado em Modelagem Computacional) – ICE/Engenharia.Universidade Federeal de Juiz de Fora. Juiz de Fora.

Mccue, C. (2007) Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis. Elsevier,

Monteiro, A. R. & Veras, A. T. de R. (2017) A Questão habitacional no Brasil. Mercator, Fortaleza, 16. https://www.scielo.br/pdf/mercator/v16/1984-2201-mercator-16-e16015.pdf.

Montgomery, D. C.& Peck, E. A. & Geoffrey, G.(2012*) Introduction to linear regression analysis*. (5a ed.), John Wiley & Sons

Moretin, P. A. C.& Toloi, C. M. (2018) *Análise de séries temporais: modelos lineares univariados*. Blucher, v. 3.

Nabavi-Pelesaraei, A., Rafiee, S., Hosseini-Fashami, F. & Chau, K. (2021) Predictive Modelling for Energy Management and Power Systems Engineering. In: Artificial neural networks and adaptive neuro-fuzzy inference system in energy modeling of agricultural products. Editora: Elsevier. Cap.11. p. 299-334.

Olson, D. L.& Delen, D. (2008) *Advanced data mining techniques*. Springer Science & Business Media.

Origuela, L. A. (2018) Estudo da influência de eventos sobre a estrutura do mercado brasileiro de ações a partir de redes ponderadas por correlações de Pearson, Spearman e Kendall. 85 f. Dissertação (Mestrado em Administração de Organizações) - Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, Ribeirão Preto.

Pacheco, C. A. R. & Pereira, N. S. (2018) Deep learning conceitos e utilização nas diversas Áreas do conhecimento. *Revista Ada Lovelace,* 2, 34-49.

Raschka, S.& Mirjalili, V. (2017) *Python Machine Learning*. (2a ed.), Packt,.

Rezende, D. A.,& Abreu, A. F. D. , (2013) Tecnologia da informação aplicada a sistemas de informação empresariais: o papel estratégico da informação e dos sistemas de informação nas empresas. (9a ed.), Atlas.

Sammut, C.& Webb, G. I. (2011) *Encyclopedia of Machine*. Springer,.

Santovena, A. Z. (2013) Big Data : Evolution , Components , Challenges and Opportunities. *Escola de Gestão Mit Sloan.* 126.

Santos, H. G. dos. (2019) Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. *Cadernos de Saúde Pública*, 35.

SNIC- Sindicato nacional da Indústria de Cimento (2019). Vendas de cimento encerram em queda de 1,2%. http://snic.org.br/assets/pdf/resultados-preliminares/1547058910.pdf.

Tan, P., Steinbach, M., Karpatne, A. & Kumar, V.(2019) *Introduction to Data Mining*. (2a ed.), Pearson.

Taylor, S. J.& Letham, B. (2017) Forecasting at Scale. PeerJ Preprints, Setembo.

Teixeira, L. P.& Carvalho, F. M. A. (Julho - Dezembro 2005) A construção civil como instrumento do desenvolvimento da economia brasileira. *Revista Paranaense de Desenvolvimento*, 109, 9-26.

Tukey, John. W. (1977) Exploratory Data Analysis Biometrics.

Vicario, G. & Coleman, S. (2020) Uma revisão da ciência de dados nos negócios e na indústria e uma visão futura. *Appl Stochastic Models Bus Ind*, 36, 6 - 18.

Zhang, Y.& Haghani, A. (2015) A gradient boosting method to improve travel time prediction. *Transportation Research Part C*, 58, 308-324.

Zhu, X., & Lafferty, J. (2005, August). Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In Proceedings of the 22nd international conference on Machine learning (pp. 1052-1059).