Can Algorithms Predict Real Estate Prices? Multiple Regression Model Analyses

Isadora A. T. Previtalle
Faculdade de Tecnologia de São Paulo
(FATEC Adamantina)
R. Paraná,400, Jardim Brasil, Adamantina – SP - Brasil
isaprevitalle@gmail.com

Aline R. dos Santos
Faculdade de Tecnologia de São Paulo
(FATEC Adamantina)
akumonodriver243@gmail.com

Paulo R. S. Ruiz
Faculdade de Tecnologia de São Paulo
(FATEC Adamantina)
paulo.ruiz2@fatec.sp.gov.br

Rodrigo F. da Silva
Faculdade de Tecnologia de São Paulo
(FATEC Adamantina)
rodrigo.silva517@fatec.sp.gov.br

ABSTRACT

The real estate sector plays a pivotal role in economic growth, making accurate property appraisal predictions essential for informed decision-making and investments. This study aimed to evaluate and compare the performance of supervised learning algorithms - Classification and Regression Trees (CART), K-Nearest Neighbors (KNN), Multiple Linear Regression (MLR), Support Vector Machine (SVM), Random Forest (RF) in predicting property values using a dataset from Recife, Brazil, spanning 1915 to 2024. Key property characteristics were selected using attribute selection, and models were assessed using R^2, MAE, MSE, and RMSE metrics. RF emerged as the most robust model, achieving a strong balance between accuracy and generalization, while SVM exhibited poor performance with large errors and limited predictive capability. Although MLR achieved the highest R^2, it struggled with inconsistent predictions. These results underscore the importance of algorithm choice and the influence of data characteristics, such as correlations and variable distributions, on model performance. This study contributes to real estate analytics by providing insights into effective machine learning applications for property value prediction, supporting both academic research and practical decision-making in the sector.

Keywords

Machine Learning; Property Valuation; Real State Analytics; Feature Selection.

1. INTRODUCTION

The real estate sector is deeply intertwined with economic growth, establishing itself as one of the primary drivers of the national economy. This importance is evident in recent economic projections, which have shown significant adjustments in forecasts, impacting the construction sector's GDP growth. Expectations for growth have risen from 2.3% to 3%, according to assessments by the Brazilian Chamber of Construction Industry (CBI) [1].

Amid this favorable scenario, a survey conducted by Brain Intelligence Estrategica in 2021 [2] revealed an increasing number of Brazilians interested in investing in real estate, seeking new avenues for profitability. However, entering this business requires caution, detailed analysis, and prior knowledge to determine fair values and identify attractive investment opportunities [3].

Machine Learning, as a tool capable of extracting insights from diverse datasets, holds significant potential for real estate price forecasting. It can identify complex, non-linear patterns and generate more precise predictions tailored to the specifics of each real estate market. Complex datasets can be leveraged to extract critical features applicable to various contexts, supporting corporate decision-making processes [4].

In this context, the present study aims to apply Machine Learning (ML) techniques to evaluate property pricing based on market data. The research utilized a dataset from the city of Recife, encompassing construction data from 1915 to 2024, to develop a prediction tool based on regression and meta-learning. This tool is designed to serve as a benchmark and guide for potential investors in the sector.

2. BACKGROUND

ML is defined as a set of algorithms capable of extracting information from datasets without requiring a predefined mathematical model [5]. These algorithms learn from the data provided, identifying patterns autonomously, and thus automating the discovery and extraction of information [6].

However, selecting an algorithm from the wide array available can be challenging. An initial data analysis provides critical insights, including the number and types of attributes, available data, and the presence or absence of target classes. Machine learning techniques are categorized into supervised learning (requiring a target class), unsupervised learning (for clustering, without a target class), and reinforcement learning (based on rewards and penalties) [7]. The choice of algorithm depends on the objectives and data characteristics, directly affecting the model's accuracy [8].

2.1 Feature Selection

Feature selection techniques are essential for identifying the most relevant attributes for training. The goal is to understand relationships between variables, eliminating redundancies that add

no value to the classification model and unnecessarily increase computational requirements [9].

Commonly used methods include variance-based approaches, correlation-based methods, and techniques leveraging feature importance from predictive models. Variance-based methods remove attributes with insignificant variability, assuming they do not contribute meaningfully to class discrimination or predictions. Correlation-based methods assess the relationship between attributes and the target variable, as well as inter-attribute correlations, selecting only the most impactful features and reducing redundancies [10]. On the other hand, methods that use importance calculated by predictive models, such as decision trees and random forests, quantify the impact of each attribute on the model's decision-making process, enabling the empirical identification of key variables [11]. These approaches are discussed in the literature and are recommended depending on the nature of the data and the problem being analyzed. Thus, attribute selection stands out as a fundamental technique for promoting efficiency and robustness in machine learning systems.

2.2 Meta-Learning

Meta-learning is an advanced technique that systematically identifies patterns in data changes to recommend the most suitable model for a given scenario, considering dataset characteristics and project objectives [12].

The recommendation process involves analyzing dataset features alongside model performance metrics. By relating these variables, the method suggests an algorithm that aligns with the desired outcomes and performance metrics. Despite its robustness, meta-learning does not guarantee that the recommended algorithm will yield the best results [13]. Moreover, machine learning algorithms often struggle with large datasets, a challenge linked to the number of objects, attributes, or both [14].

2.3 Prediction Algorithms

Several algorithms are available for prediction tasks:

- Decision Tree Algorithms: introduced by Quinlan in 1986 with the ID3 algorithm, decision trees construct hierarchical models using rules derived from recursive data splits, optimizing metrics like information gain or impurity reduction [15]. Decision trees are interpretable but prone to overfitting, especially with noisy datasets. One of the algorithms of this nature applied to regression is Classification and Regression Trees (CART), based on the construction of binary decision trees, serving as an efficient method for segmentation and prediction [16].
- Random Forest (RF): developed by Breiman in 2001, RF addresses decision tree limitations by combining multiple trees trained on random subsets of data and features, using ensemble learning techniques to improve accuracy and generalization. RF excels in handling high-dimensional data and provides intrinsic feature importance measures [17].
- K-Nearest Neighbors (KNN): introduced by Cover and Hart in 1967, KNN is a non-parametric method that classifies samples based on their proximity to the nearest neighbors in a multidimensional space. Despite its simplicity, KNN is sensitive to noise and hyperparameter choices, such as the value of k [18].
- Support Vector Machines (SVM): proposed by Vapnik in the 1990s, SVM relies on the concept of maximum margins to identify the optimal hyperplane separating data in a high-dimensional space. It is particularly effective in non-linear problems when paired with appropriate kernels [19].
- Multiple Linear Regression (MLR): rooted in the works of Galton and Pearson in the late 19th century, MLR models linear relationships between multiple independent variables and the target variable. While simple and widely used, it struggles with non-linear relationships and multicollinearity among features [20].

In summary, decision-tree-based algorithms are versatile and effective for various applications, while methods like KNN, SVM, and MLR are better suited for specific scenarios, depending on data structure and variable relationships.

3. MATERIALS AND METHODS

3.1 Materials

A public dataset from the Recife City Hall was used in this study. It contains 3.531 records, including address details (street, number, complement, neighborhood, city), year of construction, land area, built area, finishing standard, construction type, occupancy type, appraised property value, sale transaction date, property condition, and property type.

The Python programming language was chosen for data processing due to its simplicity in implementing machine learning algorithms, supported by libraries such as Pandas [21] and Scikit-learn [22]. Development was conducted on the Google Collaboratory platform (Google Colab) [23], leveraging its cloud computing capabilities.

3.2 Methods

The proposed methodology follows the Knowledge Discovery in Databases (KDD) process (Figure 1), which consists of five key steps: selection, preprocessing, transformation, data mining, and evaluation, aiming to develop an accurate prediction tool [6]. Each step is detailed below.

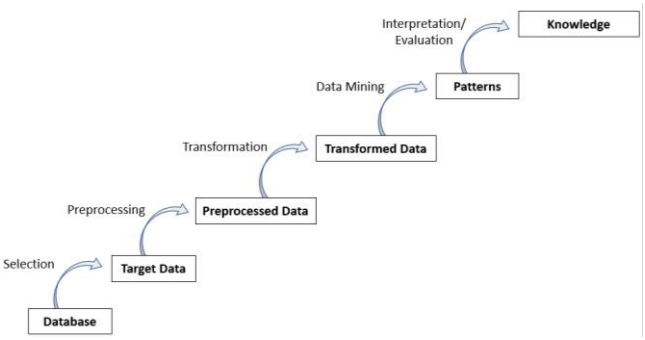

Figure 1. Methodological flowchart.

3.2.1 Selection

In this step, a preliminary analysis was conducted to evaluate property characteristics and the region where it is located, focusing on an initial classification of prices for previously sold properties. This analysis provided critical insights to support the supervised regression algorithm in predicting property values based on correlations between price and property features.

3.2.2 *Preprocessing*

Preprocessing involved handling inconsistent and null data to prevent negative impacts on the analysis. The *dropna* and *fillna* methods from the Pandas library were employed to remove rows or columns with null values and to replace null values with the mean of the attribute, respectively. Additionally, categorical attributes were converted into numerical values using *LabelEncoder* and *OneHotEncoder* from the *Scikit-learn* library.

3.2.3 *Transformation*

The transformation step focused on data standardization to ensure all attributes were on a common scale, eliminating potential biases during model training. Standardization adjusted the data to have a mean of 0 and a standard deviation of 1, as represented in Eq. (1):

$$Z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the data value to be standardized, μ is the mean of the data distribution, and σ is the standard deviation of the data distribution [24]. This transformation was performed using the *StandardScaler* method from the *Scikit-learn* library.

3.2.4 *Data Mining*

Five regression algorithms were implemented: CART, KNN, MLR, SVM and RF. Data was split into training (80%) and testing (20%) sets.

Except for MLR, all algorithms were implemented using Scikit-learn. CART utilized *DecisionTreeRegressor* with *random_state=26* for reproducibility. KNN employed *KNeighborsRegressor* with parameters: *n_neighbors=3* (three nearest neighbors), *metric='euclidean'* (Euclidean distance), *weights='uniform'* (equal neighbor weighting), and *algorithm='auto'* (automatic algorithm selection). SVM used the *SVR* method with default settings: *kernel='rbf'* (mapping data to a higher-dimensional space), *C=1.0* (controlling training error penalties), and *epsilon=0.1* (tolerance margin for ignored points). RF applied *RandomForestRegressor* with *n_estimators=100* (100 decision trees) and *random_state=26* for consistent results.

MLR was implemented using the *statsmodels* library. Predictor variables (X) included attributes like location and property characteristics, while the target variable (y) represented property prices. Categorical variables in X were converted to dummy variables with *drop_first=True* to prevent multicollinearity. An intercept was added with *sm.add_*constant, and the model was fitted using the Ordinary Least Squares (OLS) method, providing metrics such as the coefficient of determination.

3.2.5 *Evaluation*

Each algorithm was evaluated by testing the fitted model on the *X_test* dataset to generate predictions (*y_pred*), which were compared to the actual values (*y_test*). The evaluation metrics included R^2 (*r2_score*), measuring the coefficient of determination; Mean_Absolute_Error (MAE), the mean of absolute errors; Mean_Squared_Error (MSE) with squared=True, penalizing larger deviations; and Root Mean Squared Error (RMSE), the square root of MSE, maintaining consistency with the original variable's units. These metrics were computed using *sklearn.metrics* from *Scikit-learn*, selected for its simplicity and efficiency in machine learning model evaluation.

4. RESULTS

The attribute selection process reduced the dataset to 9 columns, retaining the following variables: street, neighborhood, year of construction, land area, built area, finish standard, type of construction, type of occupation, property condition, property type, and appraisal value.

Figure 2 highlights a strong positive correlation (96%) between the built area and the appraisal value, indicating that larger properties tend to have higher values. Additionally, a moderate positive correlation (20%) between finish standard and property condition suggests that higher-quality finishes are associated with better property conditions. Negative correlations were also identified, such as between street and year of construction (-26%) and street and land area (-15%). These relationships imply that property location, age, and size are interrelated factors that influence pricing. These insights are crucial for regression modeling, as they reveal how independent variables impact the dependent variable (appraisal value).

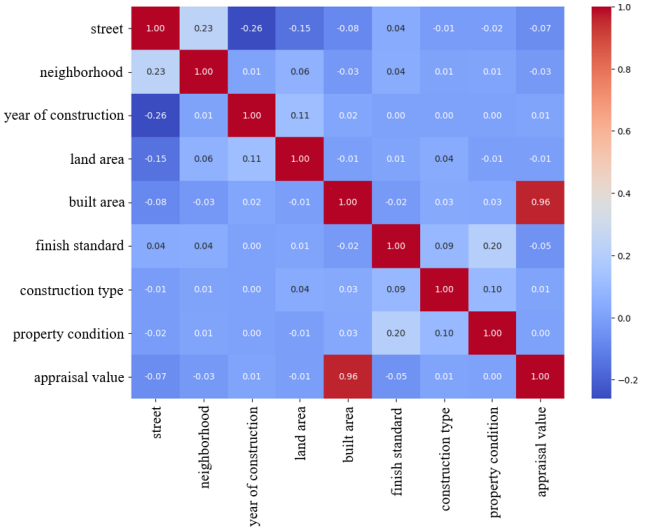

Figure 2 – Attribute correlation.

Figure 3 presents histograms of the selected attributes, showing skewed distributions across several variables. The following patterns were observed:

- Street and neighborhood: long-tailed distributions, indicating highly concentrated values in the initial categories.
- Year of construction: most properties were built after 2000.
- Land area and built area: concentration around lower values with noticeable outliers.
- Finish standard: three distinct categories, with most properties falling into low or medium finish standards.
- Type of occupation: dominated by residential properties.
- Property condition: most properties are categorized as either "good" or "regular."
- Appraisal value: highly skewed distribution, with most properties falling into lower value ranges and a few high-value outliers.

These distributions suggest the need for preprocessing steps, such as normalization or transformation, to optimize model performance.

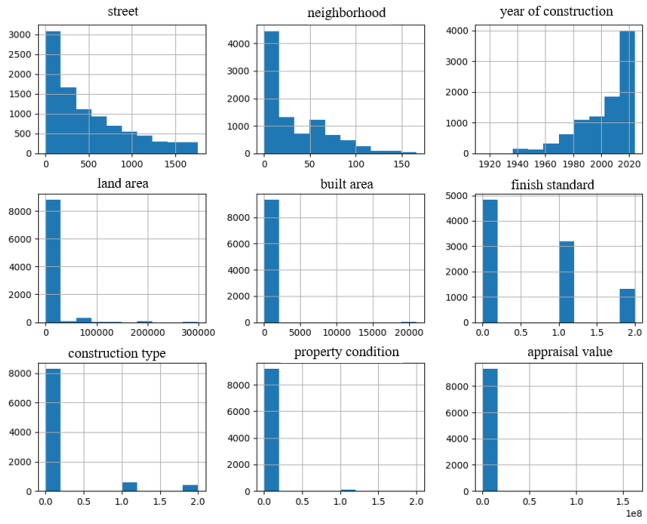

Figure 3 – Attribute histogram.

Supervised learning algorithms were applied using regression techniques on the refined dataset. Performance metrics for the models are summarized in Figure 4:

- CART: delivered improved results with an R^2 of 0.73, MAE of 5.033, MSE of 10.688, and RMSE of 5.344, outperforming KNN.
- KNN: moderate performance with an R^2 of 0.72, MAE of 5.114, MSE of 10.696, and RMSE of 5.348, indicating reasonable predictive accuracy.
- MLR: despite the highest R^2 (0.93), it exhibited a high MAE (5.411), MSE of 11.346, and RMSE of 5.673, indicating that while the model captured overall variability well, it struggled with accurate individual predictions.
- SVM: performed poorly, with a negative R^2 (-0.09) and the highest error metrics (MAE: 5.409, MSE: 10.980, RMSE: 5.646), reflecting large deviations and low precision.
- RF: achieved the best results, with an R^2 of 0.86, MAE of 4.933, MSE of 10.406, and RMSE of 5.204, demonstrating superior ability to reduce errors and capture patterns.

RF emerged as the most robust and reliable model, while SVM and MLR faced significant challenges in providing consistent results.

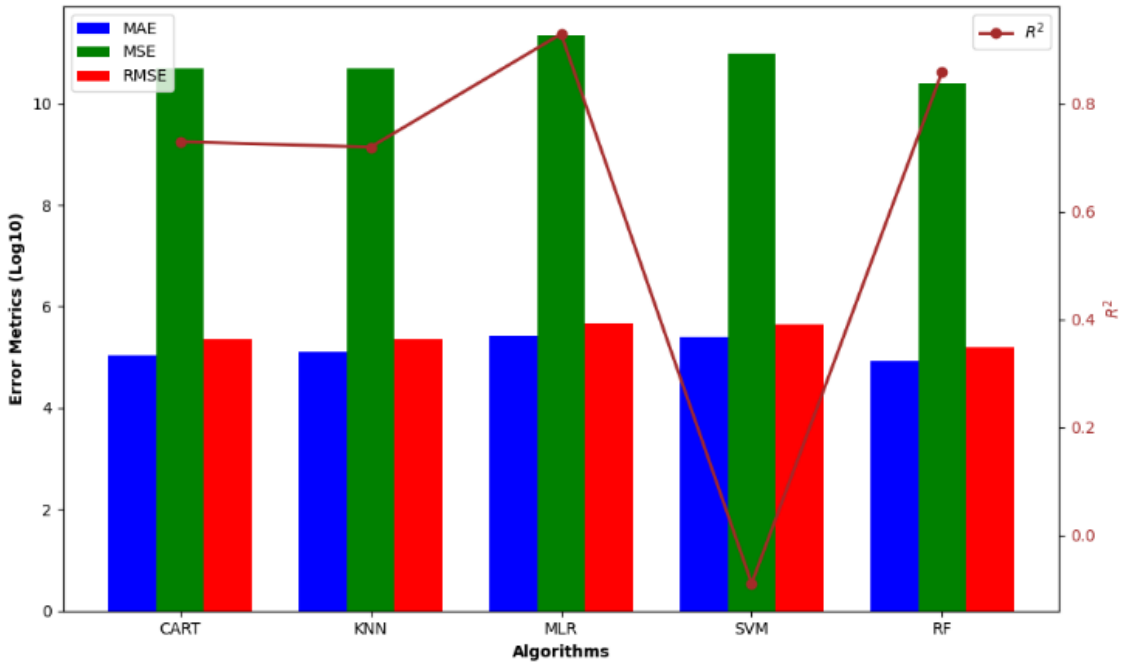

Figure 4 – Performance metrics.

Training times varied significantly among the models, as illustrated in Figure 5:

- CART and KNN: fastest training times, delivering intermediate performance with low hardware requirements.
- RF: longer training time due to the complexity of constructing multiple decision trees, but the superior results justify the resource demand.
- SVM: one of the slowest, with poor performance, making it less suitable for this dataset.
- MLR: a training time of 3 seconds, reflecting its computational simplicity, but with mixed performance outcomes.

The balance between processing time and predictive accuracy makes KNN and Decision Tree attractive options for resource-

limited scenarios. However, RF's higher resource demand is justified by its significantly better accuracy.

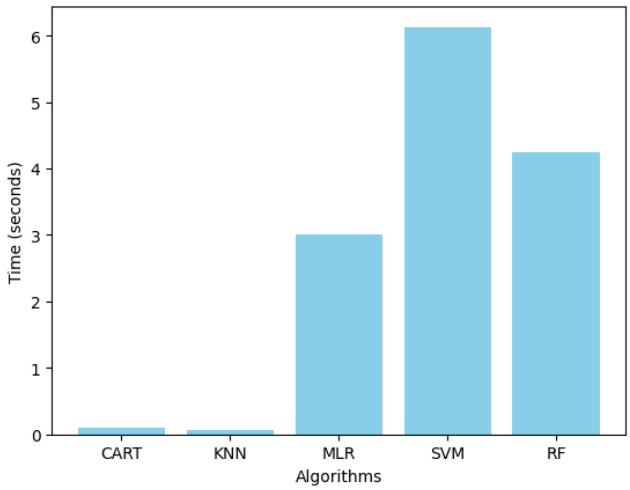

Figure 5 – Model training time.

5. CONCLUSIONS

This study evaluated and compared the performance of various supervised learning algorithms in predicting property appraisal values, utilizing a dataset comprising key variables related to property characteristics. The analysis incorporated Attribute Selection techniques and five regression algorithms: CART, KNN, MLR, SVM and RF. Model performance was assessed using standard metrics widely recognized in the literature, including R^2, MAE, MSE, and RMSE.

The findings demonstrated that RF outperformed other models, achieving a superior balance between accuracy and generalization. This result underscores its robustness in capturing complex patterns within the dataset, enabling more reliable and consistent predictions. Conversely, SVM exhibited the weakest performance, with a negative R^2 and the highest MAE, MSE, and RMSE values, indicating its inability to effectively model the relationships inherent in the data. While Multiple Regression achieved the highest R^2, its high individual prediction errors highlighted its limitations in providing consistent performance across all instances.

This study makes a valuable contribution to the literature by presenting a detailed comparative analysis of machine learning approaches for real estate price prediction. It also sheds light on the influence of variable correlations and distributions on modeling outcomes, particularly the relationships between attributes such as built area and appraisal value, and finish standard and property condition. Moreover, the comparative evaluation of performance metrics emphasizes the critical importance of selecting models that balance predictive accuracy, generalization, and computational efficiency. These insights provide a foundation for future research in real estate price prediction, demonstrating the relevance of supervised learning techniques in developing robust and reliable predictive models

6. REFERENCES

[1] CBI. *Brasileiros querem investir mais em imóveis*. Available at: <https://cbic.org.br/brasileiros-querem-investir-mais-em-imoveis/>. Acessed on: September 2024.

[2] NSCTotal. *Como o branding pode potencializar o bom momento do mercado imobiliário?* Available at: <https://www.nsctotal.com.br/noticias/como-o-branding-pode-potencializar-o-bom-momento-do-mercado-imobiliario>. Acessed on: September, 2024.

[3] Nakama, V. K., and Rufino, B. (2022). Os fundos de investimento como movimento do complexo financeiro-imobiliário no Brasil. *Revista brasileira de estudos urbanos e regionais*, 24, e202233.

[4] Kwon, O., Lee, N., and Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International journal of information management*, 34(3), 387-394.

[5] Fontana, É. (2020). Introdução aos algoritmos de aprendizagem supervisionada. *Departamento de Engenharia Química, Universidade Federal do Paraná*.

[6] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-37.

[7] Baldo, G., Cardoso, H. H. F. and da Silva Ruiz, P. R. (2024). Fact or Fake? Comparing News Classifications Using Different Artificial Neural Network Architectures. Revista de Sistemas e Computação-RSC, 14(1).

[8] Mitchell, T. M. (1997). *Machine Learning* (Vol. 1, No. 9). New York: McGraw-hill.

[9] Prado, F. F. and Digiampietri, L. A. (2020). A systematic review of automated feature engineering solutions in machine learning problems. In *Proceedings of the XVI Brazilian Symposium on Information Systems* (pp. 1-7).

[10] Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4), 491-502.

[11] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.".

[12] Melo, F. A. (2023). Meta aprendizado para detecção de mudança de conceito não supervisionada. Master's thesis — University of Brasília, Instituto de Ciências Exatas, Departamento de Ciência da Computação, 2023.

[13] Souza, B. F. et al. (2021). Meta-Aprendizado para recomendação de algoritmos. *Jornal de Inteligência Artificial*, [s.l.], v. 1, n. 1, p. 1-10.

[14] Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A. D., and Carvalho, A. C. P. D. L. F. D. (2021). *Inteligência artificial: uma abordagem de aprendizado de máquina*. 2. ed. Rio de Janeiro: LTC.

[15] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1, 81-106.

[16] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1986). *Classification and Regression Trees*. Wadsworth & Brooks/Cole.

[17] Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

[18] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*,

13(1), 21-27.

[19] Vapnik, V. (1995). *The nature of statistical learning theory*. Springer science & business media.

[20] Draper, N. R. (1998). *Applied regression analysis*. McGraw-Hill. Inc.

[21] Pandas. 2023. *Python Data Analysis Library - pandas: Python Data Analysis Library*. Available at: https://pandas.pydata.org/. Accessed on: November, 2024.

[22] Scikit-Learn. 2024. *Scikit-Learn: machine learning in Python*. Available at: https://scikit-learn.org/stable/. Accessed on: November, 2024.

[23] Google Colaboratory. 2023. *Colab*. Available at: https://colab.research.google.com. Accessed on: November 15, 2024.

[24] Spiegel, M. R., Schiller, J. J., Srinivasan, R. A and Viali, L. (2012). Probabilidade e estatística. 3. ed. Porto Alegre: Bookman