REVISTA DO DEPARTAMENTO DE GEOGRAFIA
Universidade de São Paulo
www.revistas.usp.br/rdg - ISSN 2236-2878
Volume 44 (2024), e212297
DOI: 10.11606/eISSN.2236-2878.rdg.2024.212297

Mass appraisal of apartments using Random Forest and Gradient Boosting algorithms: case study of Florianópolis, Brazil

Avaliação em massa de apartamentos com uso dos algoritmos Randon Forest e Gradient Boosting: estudo de caso de Florianópolis, Brasil

Carlos Augusto Zilli*[1]; Lia Caetano Bastos[2]

[1] Federal Institute of Santa Catarina (IFSC), Florianópolis, SC, Brazil.
[2] Federal University of Santa Catarina (UFSC), Florianópolis, SC, Brazil.
E-mail: liacbastos@gmail.com
*Email para correspondência: carloszilli@gmail.com

Recebido (*Received*): 23/05/2023
Aceito (*Accepted*): 16/11/2024

Resumo: O imposto sobre a propriedade imobiliária é um importante instrumento de política urbana e tem como base de cálculo o valor venal do imóvel, normalmente determinado por meio de avaliações em massa. Este estudo avalia o desempenho preditivo dos algoritmos de machine learning random forest e gradient boosting na avaliação em massa de imóveis urbanos, comparando-os à regressão linear clássica. Foram coletados 8.694 dados do mercado imobiliário usando técnicas de web scraping e, após um processamento inicial com critérios de inclusão, 1.572 dados de apartamentos da região central de Florianópolis, Brasil, foram selecionados para modelagem. Os resultados indicaram que o modelo gradient boosting superou todos os demais em métricas como RMSE, MAE, MAPE, COD, PRD e R², com predições até 30% mais precisas, confirmando seu potencial para estimar o valor venal de apartamentos de forma robusta e equitativa. Esses achados reforçam o gradient boosting como uma alternativa viável para a geração da base de cálculo do imposto imobiliário, possibilitando uma tributação mais justa e equânime, alcançando, assim, justiça fiscal e transparência tributária.

Keywords: Avaliação em massa de imóveis; Imposto de propriedade; Aprendizado de máquina; Web scrapping.

Abstract: *Property tax is an important tool of urban policy, and its calculation is based on the assessed market value of the property, typically determined through mass appraisals. This study evaluates the predictive performance of the machine learning algorithms random forest and gradient boosting in the mass appraisal of urban properties, comparing them to classical linear regression. A total of 8,694 real estate market data points were collected using web scraping techniques, and after initial processing with inclusion criteria, 1,572 apartment data points from the central region of Florianópolis, Brazil, were selected for modeling. The results indicated that the gradient boosting model outperformed all others across metrics such as RMSE, MAE, MAPE, COD, PRD, and R², with predictions up to 30% more accurate, confirming its potential to estimate apartment market values in a robust and equitable manner. These findings reinforce gradient boosting as a viable alternative for generating the property tax base, enabling fairer and more equitable taxation, thereby achieving fiscal justice and tax transparency.*

Keywords: *Mass appraisal of properties; Property tax; Machine learning; Web scrapping.*

1. Introdução

Accurate property valuation is essential for the calculation of municipal taxes, particularly for the Urban Property and Land Tax (Imposto Predial e Territorial Urbano, IPTU) in Brazil, which is levied based on the assessed value of urban real estate as established by the Federal Constitution of 1988. Inaccurate assessments can lead to inefficiencies in the tax base, resulting in fiscal inequalities and injustices among taxpayers.

Mass appraisal of properties plays a significant role in economic indicators and serves as a barometer for a country's development (YILMAZER *et al.* 2020). It involves the systematic valuation of properties within a jurisdiction, typically requiring automated approaches to effectively manage large datasets (IAAO, 2013). These appraisals are fundamental for determining municipal tax bases and are instrumental in compensation calculations and the implementation of urban policy instruments. In urban contexts, accurate appraisals support fiscal equity by ensuring fair property tax assessments and equitable distribution of tax responsibilities among property owners (CARRANZA *et al.* 2022).

Traditionally, multiple linear regression has been the predominant technique for property valuation due to its simplicity and interpretability (UBERTI *et al.* 2018; FARIA FILHO *et al*. 2019; BENJAMIN *et al*. 2020). However, these models often fail to capture the complexities of modern datasets, such as nonlinear relationships and spatial dependencies inherent in real estate markets. Alternative methods like geostatistics (HORNBURG; HOCHHEIM, 2017; THEODORO *et al.* 2019; DUARTE, 2019) and non-parametric regression techniques (FILHO *et al.* 2005) have been explored, but limitations persist, particularly regarding scalability and adaptability to diverse market conditions.

With the advent of big data—large and complex datasets generated in real time—there is a growing need for advanced analytical methods capable of handling high-dimensional data and uncovering intricate patterns. Machine learning has emerged as a promising tool, potentially overcoming the limitations of traditional models by effectively managing complex data structures and improving predictive accuracy (ZILLI *et al*. 2024; OLIVEIRA *et al.* 2024). Machine learning models, such as Random Forest and Gradient Boosting, are being explored for their ability to reduce biases and capture spatial dependencies in property data, which are common challenges in mass appraisal.

Despite the growing interest, there is still no consensus on the effectiveness of machine learning models in improving mass appraisal performance, particularly in specific regional markets. This gap in the literature underscores the need for empirical studies that evaluate the applicability and advantages of these models in different contexts. Understanding how these advanced algorithms perform in various market conditions is crucial for municipalities considering their adoption for property tax assessments.

Therefore, this study addresses the question: Can the use of machine learning algorithms for property data improve the performance of mass property appraisal compared to traditional regression models? Specifically, we investigate this question in the context of apartment valuation in the central region of Florianópolis, Brazil. The objective is to apply machine learning algorithms and assess their performance in predicting market values, thereby contributing to a fairer and more accurate tax base. By doing so, we aim to provide insights that could inform policy decisions and advance the field of property appraisal.

2. Literature review

In many countries, the real estate sector plays a dominant role, measured by its volume, economic contribution, and workforce. However, its relatively low turnover rate compared to markets such as stocks or bonds limits its profitability for traditional financial institutions (BREUER, 2020). Mass appraisal, a systematic approach to determining the assessed value of all properties in a municipality, is essential for establishing IPTU tax rates for residents. In this case, the higher the assessed property value, the higher the proportional tax amount.

To achieve fair tax values, the appraisal process must address the complexities and variations inherent to property attributes, including spatial location, environmental and structural factors, and community amenities (CARRANZA *et al.* 2022). Studies conducted by Delfino (2016) in the central area of Florianópolis identified various categories impacting real estate purchase prices, including environmental factors, such as noise and air pollution, and structural factors, such as sanitation, property area, and proximity to commercial, health, and security facilities.

Locational factors are also crucial, as properties near essential services tend to have higher market values, reflecting demand for these amenities. This connection between location and value underscores the need for a detailed approach to mass appraisal to ensure a fair market-based value for taxation purposes.

The literature emphasizes that mass appraisal has become increasingly important due to its role in economic measures and urban policy instruments across various countries (YILMAZER, 2020). Real estate values, driven by population growth and urban development, contribute significantly to the national GDP and play a fundamental role in municipal fiscal policies. A recent literature review on property valuation indicated that most researchers and scholars are now focusing on mass appraisals, recognizing them as more relevant and applicable than other property valuation methods (DIMOPOULOS, 2019). Mass appraisal efforts not only impact real estate tax revenue but also contribute to equitable access to housing and urban infrastructure, aligning property taxes with actual market values (CARRANZA *et al.* 2022).

Niu (2019) observed that the demand for intelligent automated appraisal systems is increasing in response to urbanization and the commercialization of real estate, which in turn creates the need for efficient, scalable, and transparent appraisal solutions. Price forecasting is essential for forming real estate policies, as it enables stakeholders such as property owners and brokers to make informed decisions (PAI, 2020). However, due to the vast volume and complexity of data involved, particularly in urban areas, big data analysis and machine learning have become central to handling this volume and variability, offering greater predictive power and accuracy over traditional appraisal methods (OLIVEIRA *et al.* 2024).

In the context of mass appraisal using machine learning, Yu *et al.* (2021) researched real estate pricing methods in China, identifying machine learning as a vital approach for data-rich environments. Machine learning models, particularly Random Forest and Gradient Boosting, excel in handling complex real estate datasets, providing more accurate and unbiased valuations by reducing common biases found in traditional methods (ZILLI *et al.* 2024). The authors highlight that these models outperform classical linear regression due to their ability to capture nonlinear and spatial relationships, a crucial advantage in real estate datasets where location and property characteristics are highly interdependent.

Machine learning methods, such as those described by Baldominos *et al.* (2018) and Rave (2019), have proven effective in identifying undervalued properties and predicting market prices, leveraging big data. Additionally, in a study on mass appraisal techniques, Dimopoulos (2019) emphasized the need for transparency in machine learning models to ensure accountability, as automated systems can mitigate human biases in valuation but require careful validation to maintain fairness. Dimopoulos (2019) asserts that while machine learning provides statistical consistency in errors, traditional human appraisal may introduce varied biases into valuations.

3. Study area, materials and methods

This study adopts a quantitative and exploratory approach to evaluate the real estate market in the central region of Florianópolis. The choice of a quantitative methodology is based on the need to create robust predictive models that enable mass property appraisal, essential for defining the assessed values used in the calculation of taxes, such as the IPTU (Urban Property Tax). Considering the complexity and volume of real estate data, supervised learning methods were applied to capture the variation in market values accurately, adapting to the particularities of the studied area.

3.1. Study area

The study covers the central region of Florianópolis, including the neighborhoods of Centro, Agronômica, Trindade, Santa Mônica, José Mendes, Pantanal, Córrego Grande, Itacorubi, Monte Verde, João Paulo, Saco Grande, and Saco dos Limões. These neighborhoods were selected for their relevance in the local real estate market, given the high concentration of apartments and diversity of socioeconomic and structural characteristics. These attributes ensure robust representativeness for mass appraisal modeling in the city. The neighborhoods are highlighted in blue within the frame in **Figure 1**.

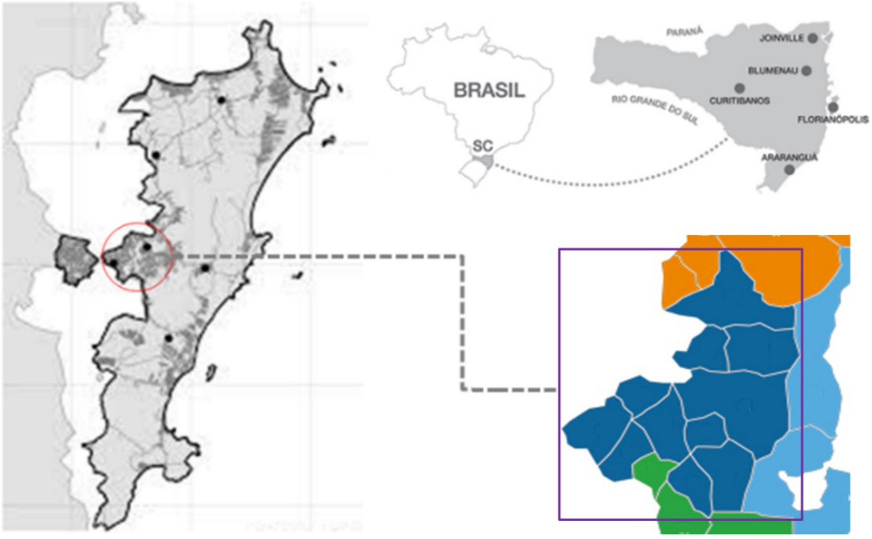

Figure 1: Neighborhoods in the central region of Florianópolis included in this study

3.2. Data collection

Data collection was conducted through web scraping on the Viva Real platform, which provides updated information on properties listed by brokers and individual owners. The process was executed in R programming language, with scripts specifically developed for this purpose. Scraping parameters were set to filter exclusively for apartment properties in Florianópolis. **Figure 2** presents the methodological flow of the study.

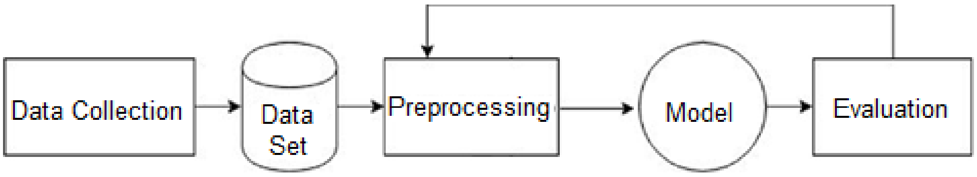

Figure 2: Complete flowchart of the study's methodological stages.

Figure 3 further details the web scraping process, including URL listings and the storage of collected data in a CSV file for subsequent analyses.

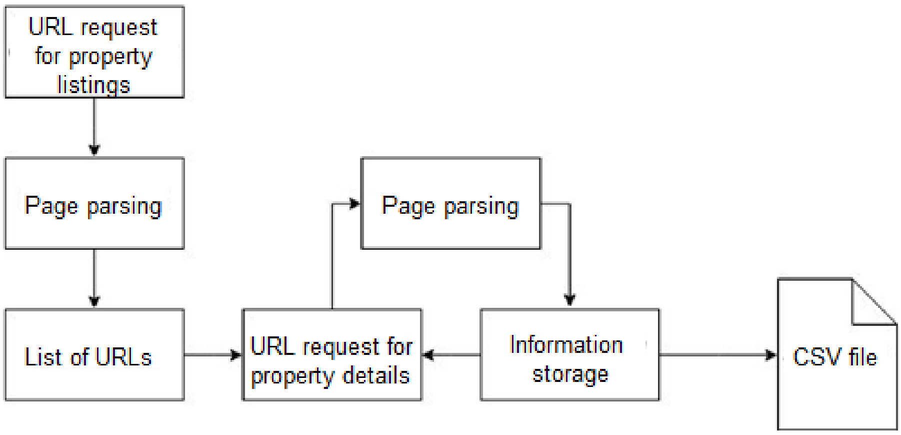
URL request for property listings
Page parsing
List of URLs
URL request for property details
Page parsing
Information storage
CSV file

Figure 3: Flowchart of the web scraping process from the Viva Real platform.

After collection, the data were georeferenced using Universal Transverse Mercator (UTM) coordinates, adjusted to time zone 22S, specific to Florianópolis's location, ensuring spatial accuracy in the analyses and forming a fundamental base for detailed modeling of real estate data.

3.3. Preprocessing

During preprocessing, a data-cleaning process was undertaken, excluding records with incomplete addresses, missing information, and data considered "outliers." In this study, "outliers" were defined as records with extreme or inconsistent values for essential variables such as total apartment area, number of rooms, and sale price, falling beyond three standard deviations from the mean. These criteria were adopted to ensure data reliability and reduce the impact of severe outliers on model results, as suggested by Carranza *et al.* (2022) in mass appraisal studies.

The dataset included the following 23 explanatory variables: total apartment area, number of rooms, number of bathrooms, number of suites, number of parking spaces, presence of a gym, distance to Avenida Beira Mar, distance to the nearest hospital, presence of a barbecue area, presence of a closet, gated community, American kitchen, presence of an elevator, gourmet space, presence of a sauna, furnished apartment, presence of a pool, playground, 24-hour security, game room, gourmet balcony, and geographic X and Y coordinates of the property. **Figure 4** illustrates the spatial distribution of the data used.

3.4. Exploratory Data Analysis

Exploratory analysis was conducted using statistical tools such as box plots, scatter plots, moments, correlations, and histograms to examine the distributions and relationships of independent and dependent variables. Additionally, a logarithmic transformation was applied to the dependent variable (total property value) to correct asymmetries and normalize the distribution, as recommended by Dantas (2014) in real estate appraisal studies. To detect influential points, Cook's Distance was used, an effective statistical method for identifying observations that may disproportionately affect modeling outcomes.

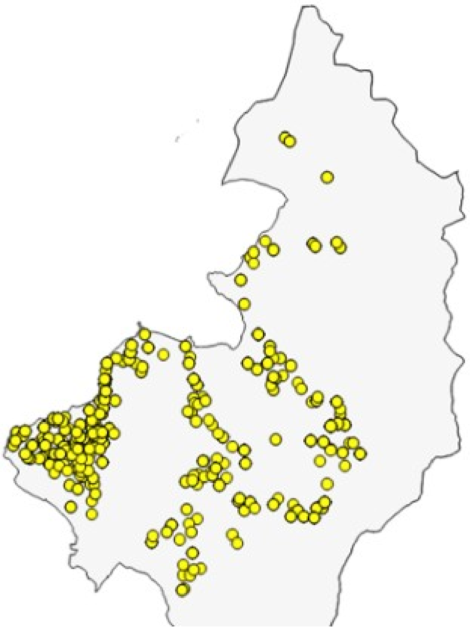

Figure 4: Spatial distribution of the 1,572 data points from the central region of Florianópolis.

3.5. Data modeling

For data modeling, three main methods were chosen: multiple linear regression (MLR), random forest (RF), and gradient boosting (GB). Multiple linear regression was selected as a baseline method due to its widespread use and ability to provide direct interpretation of variable coefficients. To obtain the best multiple linear regression model that could accurately explain the real estate market, several simulations were performed, both with and without transformations of the dependent variable.

To verify the adequacy of the regression model and ensure it meets the assumptions of classical linear regression, various tests were conducted, including: Jarque-Bera Test: assesses the normality of residuals, essential for efficient estimators; Breusch-Pagan Test: examines homoscedasticity, or the constancy of error variance; Graphical analysis of the dependent variable against explanatory variables: evaluates the linear trend in the data; Variance Inflation Factor (VIF): checks for multicollinearity among explanatory variables; RESET Specification Test: ensures the model is correctly specified, free from omissions or functional form errors; Snedecor's F-Test: evaluates the overall significance of the model, verifying if there is a significant linear relationship between the dependent variable and the set of explanatory variables.

The machine learning models, Random Forest and Gradient Boosting, were selected for their ability to handle complex and voluminous data, a crucial characteristic for the real estate sector. According to the literature (ZILLI *et al.* 2024; OLIVEIRA *et al.* 2024; CARRANZA *et al.* 2022), both models demonstrate high accuracy in contexts requiring mass appraisal and were applied in this study to explore their effectiveness in generating fair and precise estimates. Both models utilized a sample of 1,572 data points with the same explanatory variables applied in the multiple linear regression model. Ten-fold cross-validation was implemented to assess model robustness across different data subsets. The ideal parameter settings were determined through grid search optimization. Both the Random Forest and Gradient Boosting models were configured in Orange software, with specific adjustments for the optimal number of trees and depth based on the parameters that performed best in cross-validation.

In the Random Forest model, the Out-of-Bag (OOB) error was calculated, an accuracy estimator obtained by resampling training data multiple times to create decision trees, as described by Breiman (1996). The OOB error was essential for identifying the optimal number of trees, balancing predictive accuracy with computational efficiency. Additionally, Random Forest allows variable importance analysis, highlighting the most influential ones for prediction. According to Geron (2017), significant variables frequently appear in the initial nodes of trees. Importance was quantified by two indicators: the mean decrease in accuracy (%IncMSE) and the mean decrease in Gini index (IncNodePurity), which assess each variable's impact on the model's predictive ability and identify the most relevant factors in market value estimation.

For the Gradient Boosting model, the optimal number of trees was determined through simulations using the "gbm" library in R, adjusting depth levels and observing error variation according to the number of trees. This process allowed for choosing the number of trees that minimized error, optimizing precision and computational efficiency. Gradient Boosting was configured with specific shrinkage and depth parameters: the shrinkage factor adjusts each tree's contribution to the final model, enhancing stability and reducing the risk of overfitting, while depth level controls tree complexity, balancing the ability to capture variable interactions with the need to avoid overfitting.

3.6. Performance evaluation

The performance evaluation of the models was conducted using the following indicators: MAPE (Mean Absolute Percentage Error), MAE (Mean Absolute Error), RMSE (Root Mean Square Error), COD (Coefficient of Dispersion), PRD (Price-Related Differential), and R² (Coefficient of Determination). These indicators were chosen due to their widespread recommendation in the real estate appraisal literature for measuring model accuracy and consistency (CARRANZA *et al.* 2022).

R²: measures the proportion of variability explained by the model, with 1 being the ideal value; *RMSE:* provides an absolute error measure between adjusted and observed values; *MAE:* indicates the average absolute error, important for evaluating prediction accuracy; *MAPE:* represents the mean percentage error, useful for comparing models in percentage terms; *COD:* measures dispersion relative to the median, with recommended values below 15% and *PRD:* used to identify regressivity bias, with ideal values between 0.98 and 1.03.

To complement the analysis, predictive power graphs were created to compare observed and adjusted values, and an absolute relative error surface was generated in Surfer 3.2 software, allowing error pattern visualization throughout the study area.

4. Results and discussion

4.1. exploratory analysis

The initial exploratory analysis focused on understanding the distribution of the dependent variable, "total property value" (VT), in both its original and transformed forms. **Figure 5** presents frequency histograms for the VT variable, both in the original and log-transformed scales.

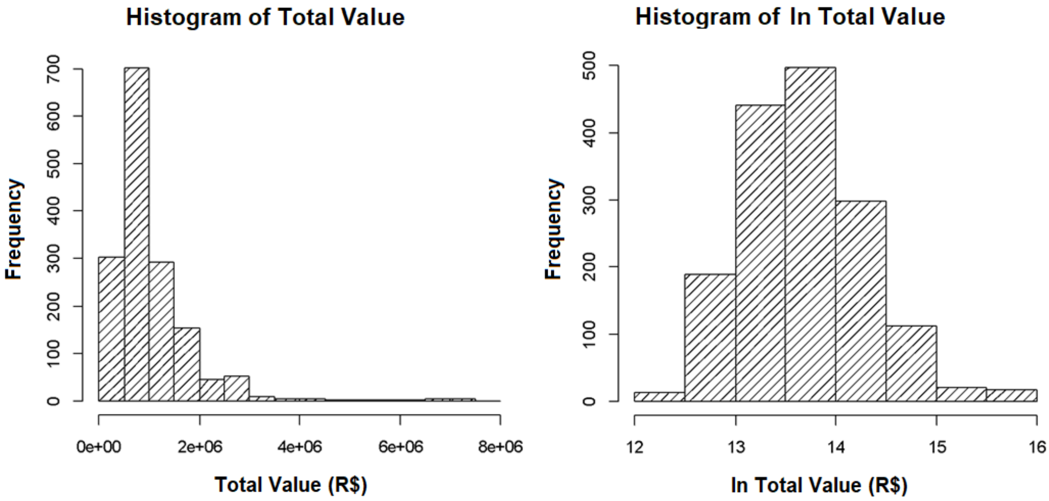

Figure 5: Frequency histograms of the total property value variable (VT and ln VT).

In the original VT scale, the data showed a slight positive skew and a platykurtic distribution, suggesting a rightward spread in property values. After applying the logarithmic transformation, the VT distribution presented a less pronounced positive skew and a more platykurtic shape, indicating a more uniform distribution of property values in the sample.

A final sample of 1,572 observations was used for subsequent stages after removing 21 data points considered influential or outliers. These included the influential points AP_208, AP_601, AP_969, AP_1091, and the outliers AP_136, AP_190, AP_371, AP_387, AP_424, AP_550, AP_678, AP_810, AP_902, AP_922, AP_1036, AP_1118, AP_1120, AP_1474, AP_1501, AP_1520, and AP_1539.

4.2. Multiple Linear Regression Model (MLR)

To identify the Multiple Linear Regression (MLR) model that best explains the real estate market, several simulations were conducted, both with and without transformations on the dependent variable. After excluding the 21 discrepant data points, a model was obtained that did not violate the basic assumptions of classical regression.

The Jarque-Bera test indicated normality of the residuals at the 5% significance level, with a p-value of 0.165. The Breusch-Pagan test indicated homoscedasticity at the same level, with a p-value of 0.1036. A graphical analysis of the dependent variable (ln VT) relative to the explanatory variables, both on the scale used in the model, showed a linear trend, and the Variance Inflation Factor (VIF) presented a maximum value in the "suites" variable (VIF_SUT = 5.738), indicating no multicollinearity. According to Gujarati *et al.* (2018, p. 348), VIF values exceeding 10 indicate high collinearity, which was not the case in this study.

The RESET specification test was performed and yielded a value of 2.7036 with a p-value of 0.5582, indicating that the model does not have specification problems. The adjusted model is presented in Eq. (1):

$$\ln(VT) = \beta_0 + 4.7163\text{x}10^{-3} \times (ARE) + 0.0533 \times (BAN) + 0.1236 \times (SUT) + 0.1849 \times (GAR) - 9.4389\text{x}10^{-5} \times (DBM) - 1.9977\text{x}10^{-5} \times (DHP) + 0.0752 \times (ACD) + 0.0727 \times (PSN) + 0.0305 \times (PGR) + 0.0925 \times (SJO) + 0.1156 \times (SAU) \quad (1)$$

Where: ARE is the property's total area in square meters; BAN is the number of bathrooms; SUT is the number of suites; GAR is the number of parking spaces; DBM is the distance to Avenida Beira-Mar (in meters); DHP is the distance to the nearest hospital (in meters); ACD indicates the presence of a gym; PSN indicates the presence of a swimming pool; PGR indicates the presence of a playground; SJO indicates the presence of a game room; and SAU indicates the presence of a sauna.

Table 1 presents the coefficients, standard errors, calculated t-values, and the significance of each regressor in the multiple linear regression model. All were considered significant at the 10% level, meeting the NBR 14.653-2 (2011) evaluation standard.

Table 1: Regression model parameter statistics

Variable	Coefficients	Error	t-Statistic	Significance
ARE	$4.7163\text{x}10^{-3}$	0.0000	39.762	0.000
BAN	0.0533	0.0102	9.948	0.000
SUT	0.1236	0.0118	18.061	0.000
GAR	0.1849	0.0117	23.032	0.000
DBM	$-9.4389\text{x}10^{-5}$	0.0000	20.39	0.000
DHP	$-1.9977\text{x}10^{-5}$	0.0000	3.220	0.001
ACD	0.0752	0.0175	6.038	0.000
PSN	0.0727	0.0160	5.869	0.000
PGR	0.0305	0.0156	2.470	0.002
SJO	0.0925	0.0190	6.159	0.000
SAL	0.1156	0.0242	5.178	0.000

The coefficient signs align with local real estate market expectations. The Snedecor F-test showed that the model was significant at the 1% level. Since Fcalc = 770.30 is greater than Fcrit = 2.26, the hypothesis of a significant regression relationship is accepted. Thus, the adopted model did not violate classical regression assumptions and passed all tests, proving statistically adequate to explain the real estate market in the study area.

4.3. Random Forest Model (RF)

For the Random Forest model, the Out-of-Bag (OOB) error was calculated to determine the optimal number of trees, as shown in **Figure 6**.

Figure 6 shows the OOB error plot obtained after running the Random Forest. The number of trees at which the error remains approximately constant is x = 150. Thus, 150 trees were adopted for this model. Beyond this number, the OOB error does not significantly change, but operational cost increases. Using the Random Forest package in Orange software, the model was configured with 150 trees and $p \approx m^{0.5} \Rightarrow p \approx 11^{0.5} \approx 3$ variables in each iteration.

A notable feature of tree-based methods is the ability to rank attributes by importance. The importance of each variable in the Random Forest model can be observed in **Figure 7**, with "distance to Avenida Beira-Mar (DBM)" and "total area (ARE)" ranking as the most important variables for explaining property values in the study region.

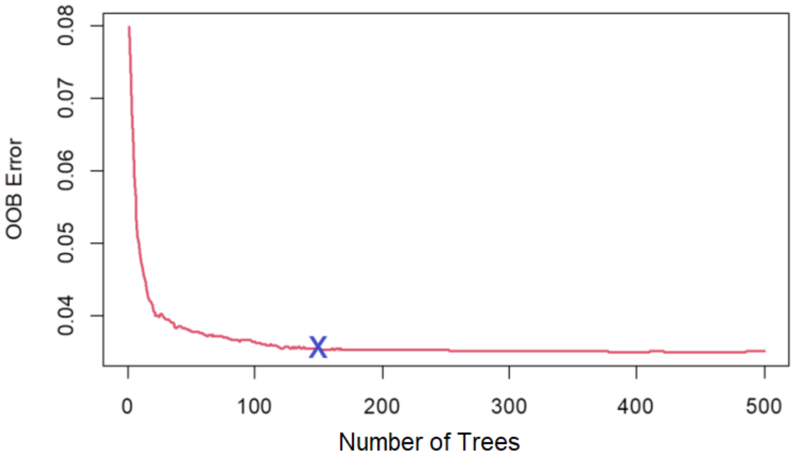

Figure 6: Out-of-Bag Error Diagram for the Random Forest Model.

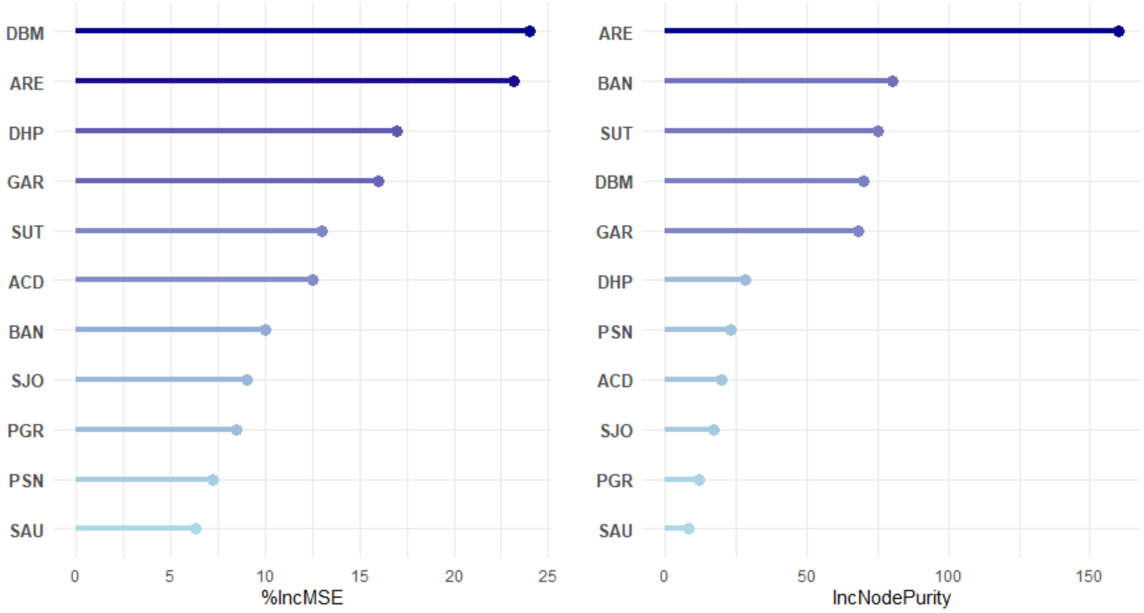

Figure 7: Importance of explanatory variables in the Random Forest Model.

The total area of a property is expected to significantly influence its market value, as it is a primary determinant in real estate valuation models (BOURASSA; HOESLI, 2023). Notably, the Random Forest model effectively captures the impact of proximity to Avenida Beira-Mar on apartment values in the central neighborhoods of Florianópolis, highlighting the importance of premium locations in the local real estate market context.

4.4. Gradient Boosting Model (GB)

To define the optimal number of trees in Gradient Boosting, simulations were conducted using the "gbm" library in R, adjusting depth and the number of trees. The error was minimized at around 2000 trees, as illustrated in **Figure 8**.

With 2000 trees, the model yielded satisfactory results. A shrinkage factor of 0.01 was used, softening the learning and helping to avoid overfitting by limiting the impact of each tree. The depth of 8 enables capturing complex interactions without overfitting, balancing precision and generalization. The analysis of variable importance indicated that "total area" and "distance to Avenida Beira-Mar" are the most relevant variables, as shown in **Figure 9**.

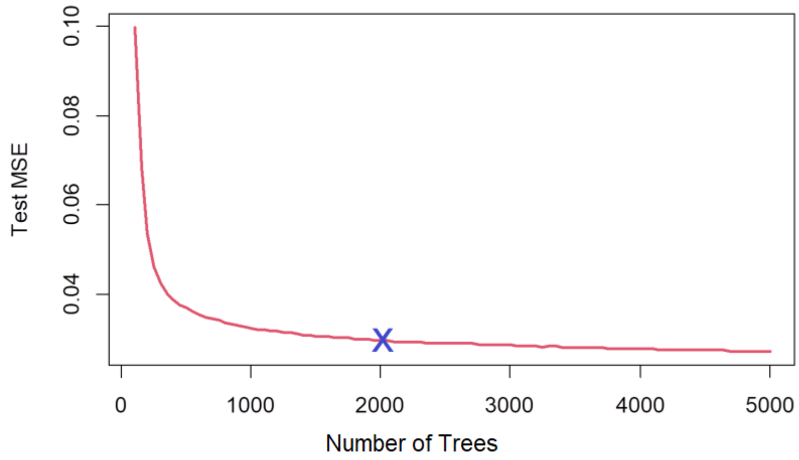

Figure 8: Error diagram for the Gradient Boosting Model.

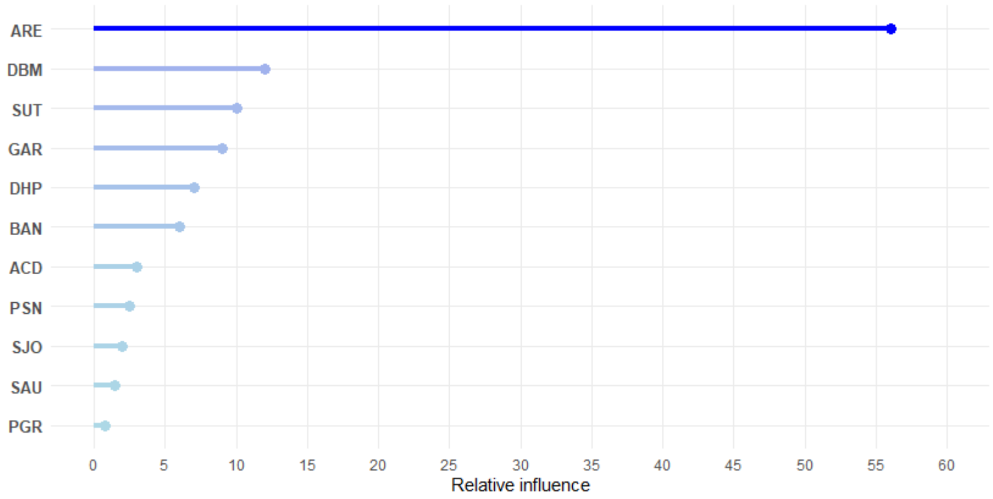

Figure 9: Importance of explanatory variables in the Gradient Boosting Model.

The Gradient Boosting algorithm also identifies the total area and distance to Avenida Beira-Mar as key variables in determining property values in the study region. However, in this model, the "total area" variable shows significantly greater importance than the other variables.

4.5. Modeling performance

The summary of key performance metrics analyzed in this study is presented in **Table 2**. These results were obtained through the models mentioned and the “test and score” function in Orange software.

Table 2: Summary of Model Performance Metrics.

Metric	Linear Regression	Randon Forest	Gradient Boosting
RMSE (R$)	360.249,02	304.990,82	249.475,92
MAE	212.415,38	166.040,65	135.913,88
MAPE (%)	18,98%	14,41%	12,25%
Ratio R	1,002	1,006	1,000
COD (%)	18,95%	14,32%	12,26%
PRD	1,062	1,060	1,020
Coeficiente R²	0,842	0,887	0,924

It can be observed that the Root Mean Square Error (RMSE) was lowest in the gradient boosting model, with a value approximately 30% lower than that of the classical linear regression model. The Mean Absolute Error (MAE) is a useful metric for evaluating prediction accuracy, and it also performed best in the gradient boosting model. The Mean Absolute Percentage Error (MAPE) remained below the 30% limit established by Ministry of Cities Ordinance 511/09 for all models; however, the gradient boosting model demonstrated the best performance on this metric.

The R level, as recommended by the IAAO (2013), needs to be close to unity, and it can be observed that the gradient boosting model achieved an R level of 1.0, precisely matching the expected value. Considering the Coefficient of Dispersion (COD), it can be seen that the gradient boosting model exhibited better performance, showing a lower value and indicating greater uniformity in the assessments. For properties in heterogeneous regions, the IAAO (2013) standard suggests dispersion coefficients below 15%. The Price-Related Differential (PRD) is a metric established by the IAAO (2013) for property assessment and indicates whether there is vertical inequality in mass assessments, recommending values between 0.98 and 1.03. For PRD, the results indicate that the gradient boosting model was the only one to maintain its values within the range specified by the standard. This is an excellent indicator for verifying so-called fiscal fairness.

Finally, in terms of the Coefficient of Determination (R²), the gradient boosting model (GB) outperformed both the classical linear regression model and the random forest (RF) model, more accurately explaining 92.4% of the market value of properties in the studied region.

To supplement the analysis, a chart was constructed to represent the predictive power of the models. All predictions were obtained using the "predictions" function in the Orange tool. In this case, a chart displaying observed and adjusted values for each model is presented. The closer the point cloud is to the bisector, the greater the predictive power of the model. This situation can be observed in the charts in **Figures 10** (A-C).

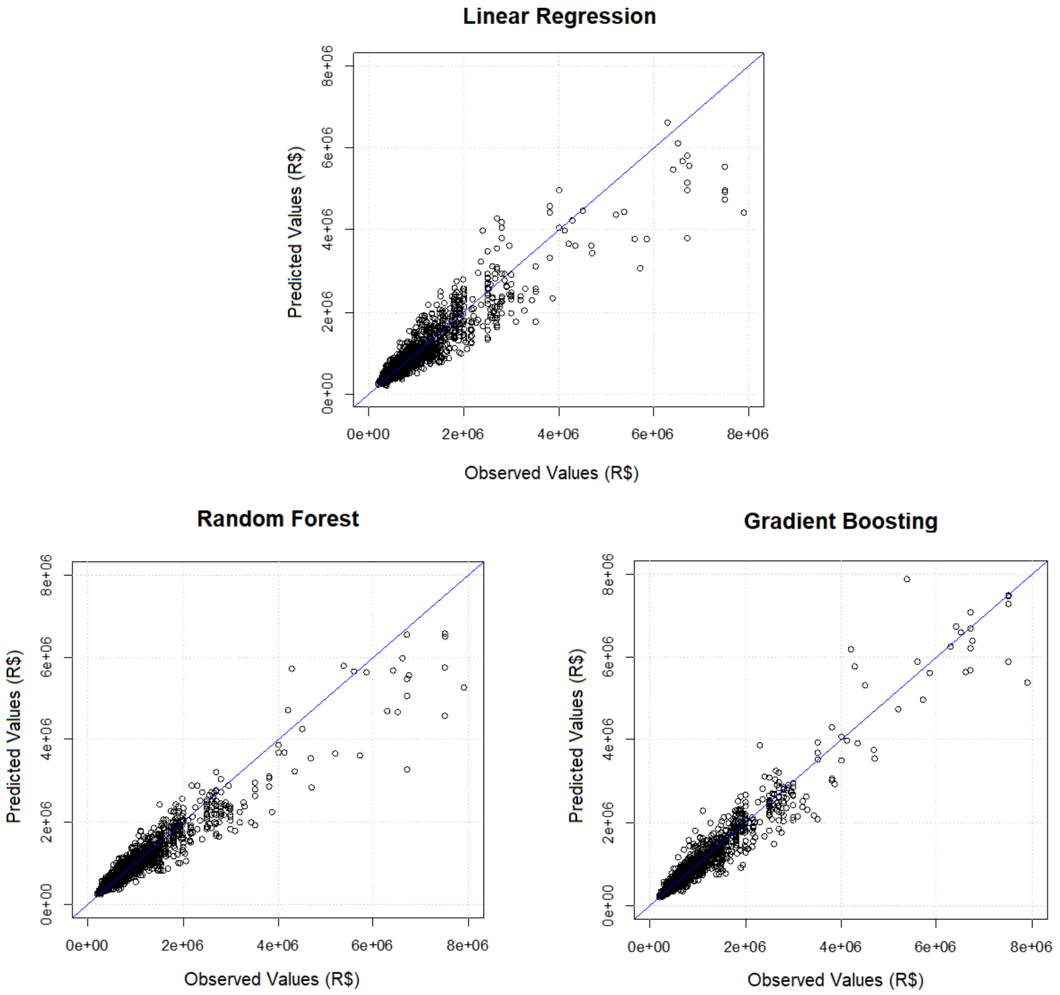

Figure 10 (A-C): Observed vs. predicted values chart for models.

It is clearly observed that the gradient boosting model was able to predict property market values with greater accuracy, with points closer to the blue-marked bisector line, which is desirable. This modeling also addressed a common problem in property assessment, where high-value properties are assessed below their market values, leading to a phenomenon of fiscal regressivity. This correction is consistent with the values presented by the PRD, indicating an improvement in fiscal equity.

Finally, the distribution of absolute relative errors of the values predicted by the models was examined. A linear interpolation was then performed to generate a gradient surface containing the estimated absolute relative errors for each of the models. The results are shown in **Figure 11**.

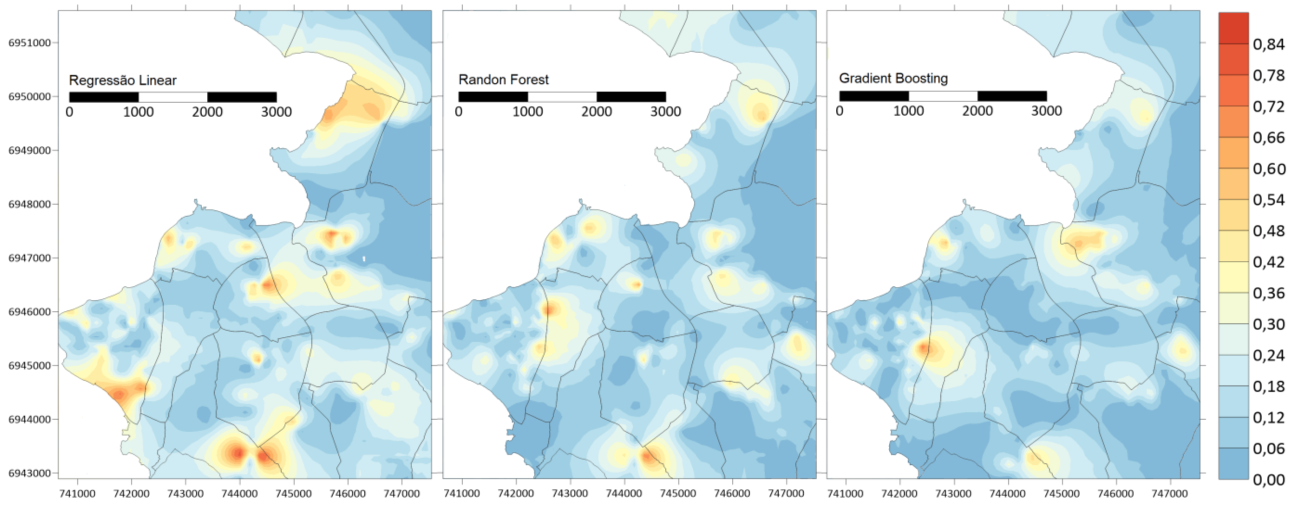

Figure 11: Estimated absolute relative errors for each model.

Based on the figures presented, it can be observed that the error surface for the gradient boosting model is much more uniform, with errors of smaller magnitude closer to zero, as indicated by the blue shading.

5. Conclusions and future perspectives

This study demonstrated the effectiveness of machine learning models, specifically Random Forest and Gradient Boosting, in comparison with traditional multiple linear regression for mass property appraisal. Using a comprehensive dataset from the central region of Florianópolis, it was observed that the machine learning models significantly outperformed linear regression in terms of accuracy, with Gradient Boosting showing the best performance among the approaches analyzed. This result suggests that using advanced techniques can improve market valuation accuracy, thus contributing to a fairer and more equitable property tax (IPTU) base.

The contribution of this work lies in the implementation of machine learning algorithms capable of more fully capturing the complexity of real estate attributes, surpassing the limitations of traditional methods, which often struggle to adequately handle highly interdependent variables. Moreover, the variable importance analysis provided by tree-based models offers valuable insights for public managers and urban planners, who can use this information to guide fiscal and urban development policies.

Despite the promising results, it is important to acknowledge that this study focused on a specific region and dataset. Therefore, the findings may not be directly generalizable to other regions or types of properties. Future research can expand this approach to other urban areas, test the application of different machine learning algorithms—such as deep neural networks or hybrid models—and incorporate additional data, such as environmental and socioeconomic characteristics. Understanding the performance of these advanced algorithms in various market conditions is crucial for municipalities considering their adoption for property tax assessments.

Ultimately, advances in data modeling can contribute to fairer tax policies, enhancing fiscal equity and efficiency in municipal revenue collection. The adoption of machine learning models by municipalities can not only improve the accuracy of assessments for IPTU purposes but also promote greater transparency and trust in the tax system, encouraging fiscal compliance and potentially increasing municipal revenue.

Acknowledgments

The present study was conducted with the support of the Coordination for the Improvement of Higher Education Personnel (CAPES), Brazil – Funding Code 001, the Support Fund for Maintenance and Development of Higher Education (FUMDES), the Federal University of Santa Catarina (UFSC), the Graduate Program in Engineering and Knowledge Management (PPGEGC), and the Federal Institute of Santa Catarina (IFSC), Brazil.

References

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 14653-2: Avaliação de Bens.** Parte 2: Imóveis Urbanos. Rio de Janeiro, 2011. 53 p.

ANTIPOV, E.A.; POKRYSHEVSKAYA, E.B. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. **Expert Systems with Applications**, v. 39, p. 1772–1778, 2012. https://doi.org/10.1016/j.eswa.2011.08.077.

BALDOMINOS, A.; BLANCO, I.; MORENO, A.J.; ITURRARTE, R.; BERNÁRDEZ, Ó.; AFONSO, C. Identifying Real Estate Opportunities Using Machine Learning. **Applied Sciences**, v. 8, p. 2321, 2018. Disponível em: https://doi.org/10.3390/app8112321

BASHA, A.M.; ANKAIAH, B.; SRIVANI, J.; DADAKALANDER, U. Real Estate Analytics With Respect To Andhra Pradesh: Machine Learning Algorithm Using R-Programming. **International Journal of Scientific & Technology Research**, v. 9, n. 4, 2020.

BRASIL. **Constituição (1988).** Constituição da República Federativa do Brasil. Brasília, DF, 1988.

BOURASSA, S.C.; HOESLI, M. Hedonic, residual, and matching methods for residential land valuation. **Journal of Housing Economics**, v. 58, Part A, p. 101870, 2022. Disponível em: https://doi.org/10.1016/j.jhe.2022.101870.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, p. 123–140, 1996.

BREIMAN, L. Random forests. **Machine Learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BREUER, W.; STEININGER, B.I. **Recent trends in real estate research: a comparison of recent working papers and publications using machine learning algorithms**. 2020. Disponível em: https://doi.org/10.1007/s11573-020-01005-w.

CARRANZA, J.P; PIUMETTO, M.A.; LUCCA, C.M.; DA SILVA, E. Mass appraisal as affordable public policy: open data and machine learning for mapping urban land values. **Land Use Policy**, v. 119, p. 106211, 2022. https://doi.org/10.1016/j.landusepol.2022.106211.

CEH, M.; KILIBARDA, M.; LISEC, A.; BAJAT, B. Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. **ISPRS International Journal of Geo-Information**, v. 7, p. 168, 2018. https://doi.org/10.3390/ijgi7050168.

CHENG, C.; CHENG, X.; YUAN, M.; CHAO, K.; ZHOU, S.; GAO, J.; XU, L.; ZHANG, T. A Novel Architecture and Machine Learning Algorithm for Real Estate. In: Sun, S., et al. (eds.), **Signal and Information Processing, Networking and Computers**, Lecture Notes in Electrical Engineering, v. 473, 2018. Disponível em: https://doi.org/10.1007/978-981-10-7521-6_60.

DANTAS, R.A. **Engenharia de avaliações: uma introdução à metodologia científica**. 3. ed. São Paulo: Pini, 2014.

DELFINO, D.; SPANIOL, E.; BUGLIONE, S. **O setor imobiliário de Florianópolis na perspectiva da nova sociologia econômica e das inserções sociais como categoria de análise.** 2016. Disponível em: https://periodicos.ufsc.br/index.php/geosul/article/download/2177-5230.2016v31n62p367/32611/156702.

DIMOPOULOS, T.; BAKAS, N. Sensitivity Analysis of Machine Learning Models for the Mass Appraisal of Real Estate. Case Study of Residential Units in Nicosia, Cyprus. **Remote Sensing**, v. 11, p. 3047, 2019. https://doi.org/10.3390/rs11243047.

DUARTE, D.C.O. **Análise multicritério e geoestatística aplicadas na avaliação em massa de imóveis urbanos.** 2019. 150 f. Tese (Doutorado em Engenharia Civil) - Universidade Federal de Viçosa, Viçosa, 2019.

FARIA FILHO, R.F.; GONÇALVES, R.M.L.; LUIZ, H.T.G. Statistical models for generating the plants of generic values: an application in a small municipality. **Urbe - Revista de Gestão Urbana**, v. 11, 2019. https://doi.org/10.1590/2175-3369.011.001.e20180192.

FILHO, C.M.; BIN, O. Estimation of hedonic price functions via additive nonparametric regression. **Empirical Economics**, v. 30, p. 93–114, 2005. https://doi.org/10.1007/s00181-004-0224-6.

FORTI, M**. Técnicas de machine learning aplicadas na recuperação de crédito do mercado brasileiro.** 74 f. Dissertação (Mestrado em Economia) – Faculdade de Economia da Fundação Getúlio Vargas, 2018.

FREUND, Y.; SCHAPIRE, R.; ABE, N. A short introduction to boosting. **Journal Japanese Society for Artificial Intelligence**, v. 14, p. 771–780, 1999.

GERON, A. **Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems**. O'Reilly Media, Inc., 2017.

GHASEMAGHAEI, M.; CALIC, G. Assessing the impact of big data on firm innovation performance: big data is not always better data. **Journal of Business Research**, v. 108, p. 147-162, 2020. Disponível em: https://doi.org/10.1016/j.jbusres.2019.09.062.

GROVER, P. **Gradient Boosting from scratch**. 2017. Disponível em: https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d.

GRUS, J. **Data Science from Scratch**. Sebastopol: O'Reilly, 2015.

GUJARATI, D.N.; PORTER, D.C. **Econometria básica**. 5. ed. Porto Alegre: AMGH Bookman, 2018.
HO, T.K. Random Decision Forests. In: **Proceedings of the 3rd International Conference on Document Analysis and Recognition**, Montreal, QC, 1995, p. 278–282.

HONG, J.; CHOI, H.; KIM, W.S. A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. **International Journal of Strategic Property Management**, v. 24, n. 3, p. 140–152, 2020. https://doi.org/10.3846/ijspm.2020.11544.

HORNBURG, R.A.; HOCHHEIM, N. Avaliação em massa de imóveis usando geoestatística e krigagem bayesiana: um estudo de em Balneário Camboriú/SC. **RECC - Revista Eletrônica de Engenharia Civil**, v. 13, n. 1, 2017. https://doi.org/10.5216/reec.v13i1.42347.

IAAO - International Association of Assessing Officers. **Standards on Ratio Studies**. Missouri: IAAO, 2013.

IBM. IBM-**Bringing Big Data to the Enterprise**. 2015.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning: with applications in R**. 2013.

JAROSZ, M.; KUTRZYŃSKI, M.; LASOTA, T.; PIWOWARCZYK, M.; TELEC, Z.; TRAWIŃSKI, B. Machine Learning Models for Real Estate Appraisal Constructed Using Spline Trend Functions. **Intelligent Information and Database Systems**. ACIIDS 2020. Lecture Notes in Computer Science, v. 12033, Springer, 2020. https://doi.org/10.1007/978-3-030-41964-6_55.

LEE, J.; PARK, S.C.; KIM, S.H. Comparison of Models to Forecast Real Estates Index Introducing Machine Learning. **Journal of the Architectural Institute of Korea**, v. 37, n. 1, p. 191, 2021. https://doi.org/10.5659/JAIK.2021.37.1.191.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R News**, v. 2, p. 18–22, 2002.

MARSLAND, S. **Machine Learning: An Algorithmic Perspective**. 2. ed. Taylor & Francis Group, 2015.

MAYRINK, V.T.M. **Avaliação do algoritmo Gradient Boosting em aplicações de previsão de carga elétrica a curto prazo.** 91 f. Dissertação (Mestrado em Modelagem Computacional) - Universidade Federal de Juiz de Fora, 2016.

MCCLUSKEY, W.J.; ANAND, S. The application of intelligent hybrid techniques for the mass appraisal of residential properties. **Journal of Property Investment and Finance**, v. 17, n. 3, p. 218–238, 1999. https://doi.org/10.1108/14635789910270495.

MITCHELL, T.M. **Machine Learning**. McGraw-Hill, 1997.

MURPHY, K.P. **Machine learning: a probabilistic perspective**. MIT Press, 2012.

NIU, J.; NIU, P. **An Intelligent Automatic Valuation System for Real Estate Based on Machine Learning.** ACM, 2019. ISBN 978-1-4503-7633-4/19/12…$15.00.

OLIVEIRA, A.A.F. **Avaliação em massa com modelos de aprendizado de máquina aplicados aos terrenos urbanos do município de Fortaleza**. 80 f. Dissertação (Mestrado em Economia) - Universidade Federal do Ceará, Fortaleza, 2020.

OLIVEIRA, A.A.F.; REYES-BUENO, F.; GONZÁLEZ, M.A.S.; DA SILVA, E. Comparing traditional and machine learning techniques in apartments mass appraisal in Fortaleza, Brazil. **Aestimum**, Just Accepted, 2024. https://doi.org/10.36253/aestim-15344.

PAI, P.-F.; WANG, W.-C. Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices. **Applied Sciences**, v. 10, p. 5832, 2020. https://doi.org/10.3390/app10175832.

PELLI NETO, A. **Redes neurais artificiais aplicadas às avaliações em massa: estudo de caso para a cidade de Belo Horizonte/MG**. 111 f. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal de Minas Gerais, Belo Horizonte, 2006.

PINTER, G.; MOSAVI, A.; FELDE, I. Artificial Intelligence for Modeling Real Estate Price Using Call Detail Records and Hybrid Machine Learning Approach. **Entropy**, v. 22, p. 1421, 2020. https://doi.org/10.3390/e22121421.

RAFIEI, M.H.S.; ADELI, H. A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units. **Journal of Construction Engineering and Management**, v. 142, n. 2, 2016. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001047.

RAVE, J.I.P.; MORALES, J.C.C.; ECHAVARRÍA, F.G. A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes. **Journal of Property Research**, 2019. https://doi.org/10.1080/09599916.2019.1587489.

SAMUEL, A.L. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development**, p. 210–229, 1959. https://doi.org/10.1147/rd.33.0210.

SELIM, H. Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. **Expert Systems with Applications**, v. 36, p. 2843–2852, 2009. https://doi.org/10.1016/j.eswa.2008.01.044.

THEODORO, L.T.C.; UBERTI, M.S.; ANTUNES, M.A.H.; DEBIASI, P. Avaliação em massa de imóveis rurais através da regressão clássica e da geoestatística. **Revista Brasileira de Cartografia**, v. 71, n. 2, p. 459-485, 2019. https://doi.org/10.14393/rbcv71n2-47458.

TRAWIŃSKI, B., et al. Comparison of expert algorithms with machine learning models for real estate appraisal. **IEEE International Conference on Innovations in Intelligent Systems and Applications**, p. 51-54, 2017. https://doi.org/10.1109/INISTA.2017.8001131.

UBERTI, M.S.; ANTUNES, M.A.H.; DEBIASI, P.; TASSINARI, W. Mass appraisal of farmland using classical econometrics and spatial modeling. **Land Use Policy**, v. 72, p. 161-170, 2018. https://doi.org/10.1016/j.landusepol.2017.12.044.

VERIKAS, A.; LIPNICKAS, A.; MALMQVIST, K. Selecting neural networks for a committee decision. **International Journal of Neural Systems**, v. 12, n. 5, p. 351–362, 2002. https://doi.org/10.1142/S0129065702001229.

YILMAZER, S.; KOCAMAN, S. A mass appraisal assessment study using machine learning based on multiple regression and random forest. **Land Use Policy**, v. 99, 2020. https://doi.org/10.1016/j.landusepol.2020.104889.

YOO, S.; IM, J.; WAGNER, J.E. Variable selection for hedonic model using machine learning approaches: a case study in Onondaga County, NY. **Landscape and Urban Planning**, v. 107, p. 293–306, 2012. https://doi.org/10.1016/j.landurbplan.2012.06.009.

YU, Y.; LU, J.; SHEN, D.; CHEN, B. Research on real estate pricing methods based on data mining and machine learning. **Springer-Verlag London Ltd.**, 2020. https://doi.org/10.1007/s00521-020-05469-3.

ZILLI, C.A.; BASTOS, L.C.; DA SILVA, L.R. Machine learning models in mass appraisal for property tax purposes: a systematic mapping study. **Aestimum**, v. 84, p. 31-52, 2024. https://doi.org/10.36253/aestim-15792.