Refas
Revista Fatec Zona Sul
1ª. Edição Outubro de 2014

IFLOG
Congresso Internacional de Logística e Operações do IFSP – Suzano

XI IFLOG – IFSP– Campus Suzano

USING DECISION TREES TO FORECAST REAL ESTATE RENTALS IN THE CITY OF SÃO PAULO, SP

USO DE ÁRVORES DE DECISÃO PARA PREVISÃO DO ALUGUEL DE IMÓVEIS NA CIDADE DE SÃO PAULO – SP

USO DE ÁRBOLES DE DECISIÓN PARA PRONÓSTICO DE ALQUILER DE INMUEBLES EM LA CIUDAD DE SÃO PAULO - SP

Gustavo Alberto Albino[1]
Mario Henrique Bueno Moreira Callefi[2]
Fernando José Gómez Paredes[3]

Artigo recebido em janeiro de 2026
Artigo aceito em abril de 2026

DOI: 10.26853/Refas_ISSN-2359-182X_v12n04_01

ABSTRACT

The real estate market stands out from economies for its huge importance in social and economic aspects of life. It represents an important part of growth and development of the countries, reflecting on the job market, collected taxes and banks´ role. The goal of this research is to verify the decision trees´ machine learning models to predict rent prices in São Paulo city. A regression methodology has been used in the algorithms of Decision Trees, Random Forest, and Gradient Boosting to predict the values of rent, taking as a parameter the variables on real estate company's sites, separating data in training and test. Among the models, the Random Forest algorithm produces a smaller residual error and higher r-squared metrics, which means the prediction obtained is closer to real data. The result of this research shows that ensemble models were better than Decision Tree model when using the hiperparameters tunning, providing this way a more accurate prediction about properties rental prices.

Keywords: Real Estate Market; Decision Tree; Random Forest.

RESUMO

O mercado imobiliário se destaca nas economias pela sua grande importância social e econômica, representando uma grande parcela do crescimento e desenvolvimento dos países, tanto nos papeis de geração de empregos, impostos arrecadados, e papel dos bancos. O objetivo do artigo é verificar os modelos de *Machine Learning* de árvore de decisões para prever os valores de aluguéis na cidade de São Paulo. Foi utilizada uma metodologia de regressão dos dados nos algoritmos de Árvore de Decisões, Floresta Aleatória e Gradiente *Boosting* para prever os valores de aluguéis com base nas variáveis

[1] Administração de empresas, USP Esalq. E-mail: gstgus@hotmail.com. Lattes: http://lattes.cnpq.br/8744152890847944. OrcId: 0009-0005-3321-2987.
[2] Doutor em Engenharia de Produção, Chemnitz University of Technology. E-mail: mariocallefi@gmail.com. Lattes: http://lattes.cnpq.br/7217112899449956. OrcId: 0000-0001-8698-0043.
[3] Doutor em Engenharia de Produção, UFMS CPTL. E-mail: fernando.gomez@ufms.br. Lattes: http://lattes.cnpq.br/5244252329793782. OrcId: 0000-0002-8465-943X.

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

contidas no site de imóveis, realizando a separação entre base de treino e teste. Dentre os modelos, o algoritmo de Floresta Aleatória teve o menor erro dos resíduos e maior coeficiente de determinação, ou seja, se aproximando mais dentro dos dados reais. O resultado do trabalho demonstrou que os modelos de Ensemble foram superiores ao modelo de Árvore de Decisão padrão quando realizados os *tunnings* dos hiper parâmetros, ficando bem próximos a predição dos preços de imóveis.

Palavras-chave: Mercado imobiliário; Árvore de decisões; Random Forest.

RESUMEN

El mercado inmobiliario destaca en las economías por su gran importancia social y económica, ya que representa una gran parte del crecimiento y el desarrollo de los países, tanto en lo que respecta a la generación de empleo, la recaudación de impuestos y el papel de los bancos. El objetivo del artículo es verificar los modelos de aprendizaje automático de árboles de decisión para predecir los valores de los alquileres en la ciudad de São Paulo. Se utilizó una metodología de regresión de datos en los algoritmos de árbol de decisión, bosque aleatorio y gradiente boosting para predecir los valores de los alquileres basándose en las variables contenidas en el sitio web inmobiliario, realizando la separación entre la base de entrenamiento y la de prueba. Entre los modelos, el algoritmo de bosque aleatorio tuvo el menor error de residuos y el mayor coeficiente de determinación, es decir, se acercó más a los datos reales. El resultado del trabajo demostró que los modelos de conjunto fueron superiores al modelo de árbol de decisión estándar cuando se realizaron los ajustes de los hiperparámetros, quedando muy cerca de la predicción de los precios de los inmuebles.

Palabras clave: Mercado inmobiliario; Árbol de decisión; Bosque aleatorio

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

1 INTRODUCTION

The real estate market plays a significant role in the global economy, standing out for its ability to generate jobs and income across distinct phases, from project design to the provision of products and services after property construction. This sector comprises a complex network of factors that influence the closing price of real estate, where not only supply and demand but also the property's characteristics and the consumer's perception of value are determining factors at the time of purchase (Arraes and Souza, 2008).

According to Wissenbach (2008), once completed, a real estate development continues to generate value through services such as building management, security, maintenance, and rental. This dynamic reveals how the real estate market is interconnected with several other sectors, serving as a vital component of long-term economic flows.

Additionally, there is a growing demand for rental properties, especially among young adults, students, and tourists, as pointed out by the information company PiniWeb (2024). These groups, increasingly active in the rental market, have been putting pressure on prices, especially in large cities such as São Paulo, where “micro-apartments” are emerging as a new housing standard. Rolnik (2023) highlights the impact of these properties on rental values in the city, creating growing tension for São Paulo residents, who face a significant increase in housing expenses.

Between 2000 and 2008, Brazil also experienced a boom in the real estate market, driven by credit policies that allowed many Brazilians to realize their dream of home ownership. For those who did not have the resources to make a down payment on financing, renting became an

option, as well as a means of building financial reserves (Damaso, 2008). Public policies, such as the *Minha Casa Minha Vida* program, also played a crucial role in this context.

The purchase of a property can have major implications, tying up capital and potentially leading to losses when a quick sale is necessary, such as when moving to another city. In addition, the opportunity cost of investing in real estate may be better utilized in other investments, such as starting a business, which makes renting a viable alternative. This is only possible in legally secure and developed markets (Damaso, 2008).

Real estate pricing is a key aspect, whether for residential or investment purposes. The prohibitive cost involved and the complexity of the factors that determine price, such as location and environmental characteristics, require detailed analysis. ABNT NBR 14653-2 establishes methodologies for real estate valuation in Brazil, the most common being the comparison of market data through statistical inference, using multiple linear regression (Gonçalves, 2022). In addition, other methodologies, such as hedonic methods, can be explored in greater depth (Fávero, Belfiore, Lima, 2008).

According to Rita (2018), decision tree models offer simplicity and high explanatory power in price forecasting. The model divides observations into groups with common characteristics and establishes decision rules to predict the value of a variable. Although more complex statistical models can be difficult to interpret, simpler models, such as decision trees, may not capture all relevant factors, requiring continuous adjustments according to buyer preferences.

The challenges for real estate price forecasting arise from the vast diversity of data, including size, location, custom-designed furniture, and macroeconomic factors, as well as economic and political changes, such as those observed during the pandemic. Therefore, given the limitations of traditional methods and the small number of studies that apply decision tree techniques to the real estate context, this research is justified.

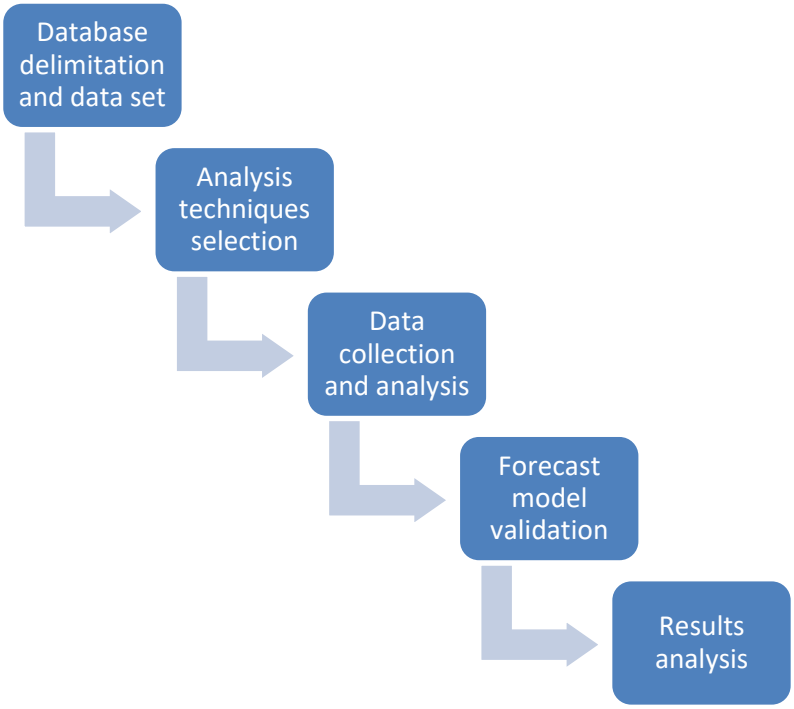
In this context, the objective of this study is to test different pricing techniques for real estate in the city of São Paulo, which accounts for 30% of the national real estate market (Charlie, 2023). São Paulo was chosen due to the market's complexity, where neighboring neighborhoods can differ by up to 300% in rental values, as observed between *Jardim Peri* and *Vila Olímpia* (Estadão, 2024).

2 MATERIALS AND METHODS

The research methodology adopts five main methodological steps, as shown in Figure 1. These are: (1) delimitation of the database and data set used; (2) choice of analysis techniques to be applied; (3) detailed analysis of the data collected; (4) analysis and validation of the forecast models generated; and, finally, (5) analysis of the results obtained. These steps were carefully structured to ensure the consistency and robustness of the research process.

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

Figure 1 - Workflow Research Method



Source: This research

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

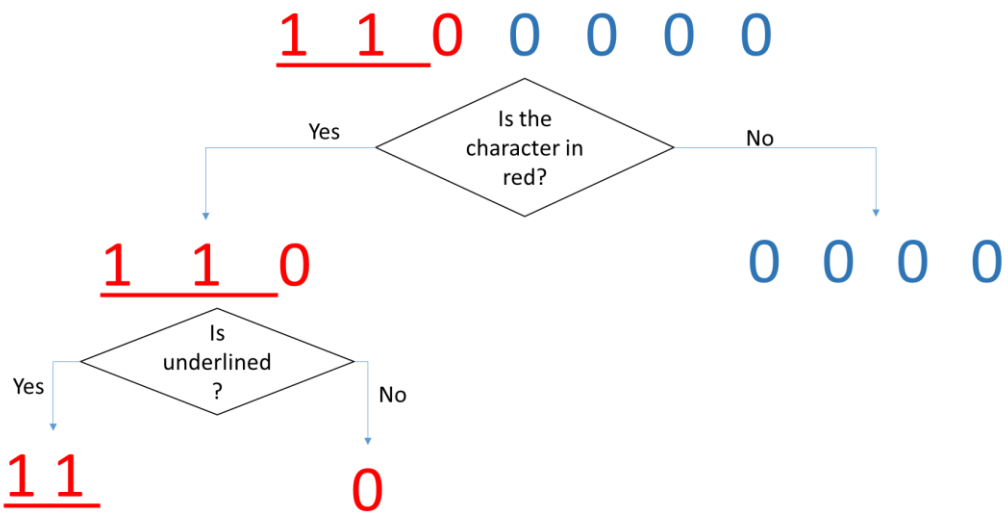
2.1 Delimitation of the database and data set used

In this study, data from the real estate company Quinto Andar, extracted from the Kaggle website, were processed. This website, widely used in both academic and professional circles, is known for its relevance in data projects and for being a frequent platform in competitions. Kaggle offers high-quality, reliable, and completely free datasets. The use of this platform ensured not only access to structured data important to the research, but also the ability to work with datasets that had already been extensively tested and discussed by the data science and machine learning communities.

2.2 Choice of analysis techniques to be applied

A decision tree is a supervised learning algorithm used for classification and regression. It partitions the feature space into progressively purer subsets by recursively selecting the split that maximizes a chosen criterion (e.g., information gain or Gini reduction). The root node contains the most informative variable for the prediction task, and subsequent internal nodes represent further splits, yielding terminal leaves that correspond to class labels or numerical predictions. Its primary advantage is interpretability: the sequence of if–then rules makes the model transparent for decision-making and enables it to reveal patterns in the data (Yiu, 2021). As an illustrative example (Figure 2), a binary target with observations labeled 0 and 1 can be separated by asking sequential questions—first on color (e.g., “Is the color red?”), then on underlining—until groups are internally homogeneous and maximally distinct from one another (Yiu, 2021).

Figure 2 - Decision Tree example



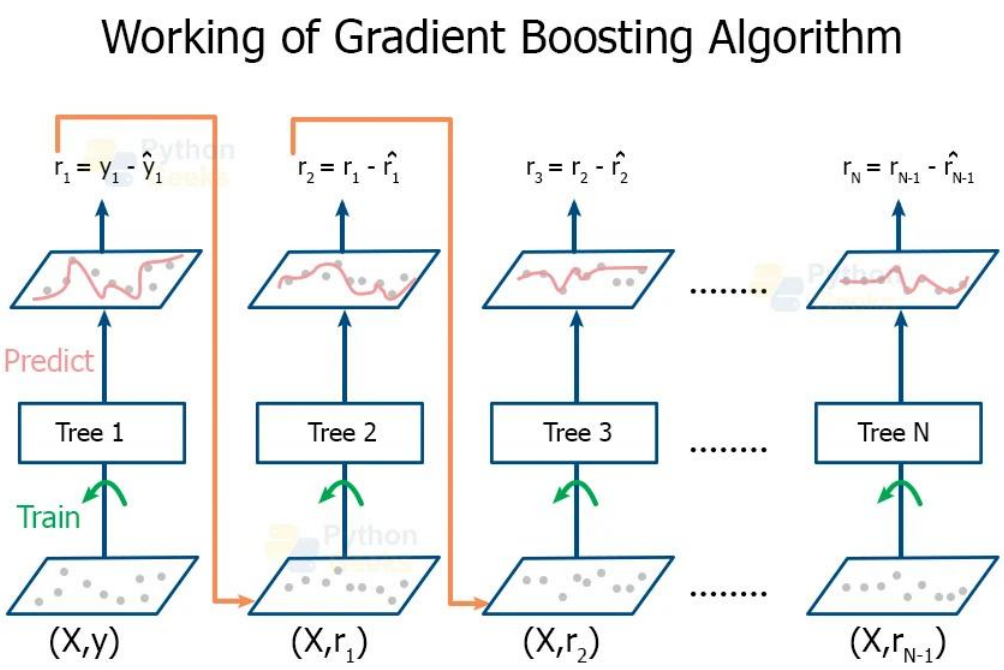
Source: Yiu (2021)

Random Forest extends decision trees through ensemble learning, combining many decorrelated trees to improve generalization and reduce variance. Developed by Leo Breiman and Adele Cutler, Random Forest can handle both classification and regression, supports mixed feature types (continuous and categorical), is robust to outliers, and tends to mitigate overfitting relative to a single deep tree (Sruthi, 2024). The algorithm employs bagging (bootstrap aggregating) as its ensemble mechanism: training data are repeatedly sampled with replacement to create bootstrap replicas; a separate decision tree is fit on each replica; and predictions are aggregated by majority vote for classification or by averaging for regression (Breiman, 1996). Random-feature subsampling at each split further reduces tree correlation, thereby enhancing ensemble performance.

Gradient Boosting is another ensemble approach that builds trees sequentially, with each learner trained to correct the residual errors of the previous ensemble. Starting from a simple base learner, the method iteratively fits shallow trees to the current residuals and updates the model as a weighted sum of learners, thereby minimizing the loss function step by step. Gradient Boosting can achieve high predictive accuracy for both classification and regression but requires careful regularization (e.g., learning rate, tree depth, number of estimators) to avoid overfitting (Medium, 2020). Conceptually, as shown in Figure 3, the process alternates between computing residuals (actual minus predicted), fitting a new weak learner to those residuals, and updating the ensemble; this continues until a stopping criterion is met (PythonGeeks Team, 2024).

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

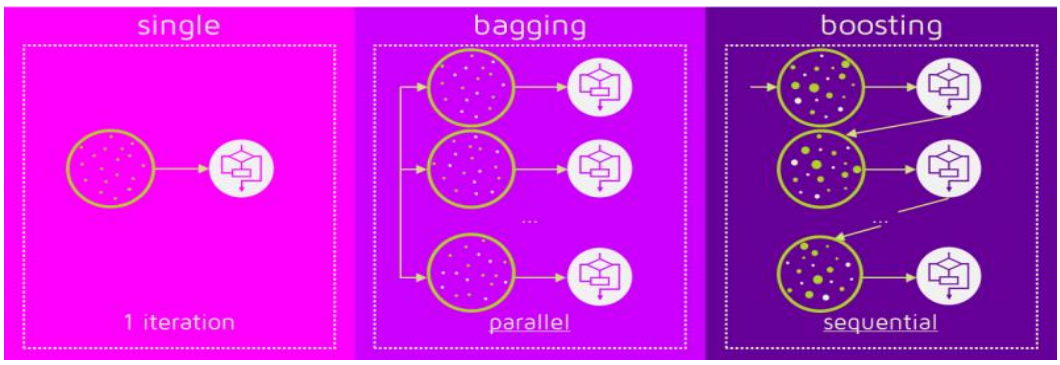
Figure 3 – How the Gradient Boosting algorithm works



Source: PythonGeeks Team (2024)

Figure 4 summarizes the mechanisms and contrasts among a single decision tree, bagging (Random Forest), and boosting (Gradient Boosting). While a single tree is highly interpretable but variance-prone, bagging reduces variance by averaging many high-variance trees built on bootstrap samples, and boosting reduces bias by sequentially adding trees that target remaining errors; together, these ensemble strategies typically deliver superior accuracy and more stable generalization than a lone tree (Aporras, 2016). In addition In Figure 4, it was possible to observe the mechanisms of the single decision tree model and the different ensembles, such as bagging, in the case of the random forest, which could be with or without replacement (bootstrapping), and the boosting model, in gradient boosting, which successively improved the accuracy of the predictions with each new observation in sequential mode.

Figure 4 - Differences of ensembles



Source: Aporras (2016)

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

2.3 Detailed analysis of the data collected

Table 1 presents the explanatory variables used in the model, all of which are considered relevant to the property valuation process. Among them, square footage is a key variable, as the average area of new residential units in São Paulo has been declining due to rising construction costs and population density. The number of bedrooms reflects a similar trend, as smaller units are increasingly common in newly constructed buildings. The number of bathrooms and the availability of parking spaces are also relevant attributes; however, parking availability has become increasingly limited due to municipal restrictions applied to new developments in central areas.

Table 1 – Explanatory variables

Variable	Description
Square footage	Size of the property in m2
Bedrooms	Number of bedrooms
Bathrooms	Number of bathrooms
Parking spaces	Number of parking spaces
Floor	Floor number of the residence
Pets allowed	Whether the landlord allows pets (Categorical variable Yes or No)
Furnished	Whether furniture is included in the lease (Categorical variable Yes or No)
Near subway	Whether there is a subway station near the property (Categorical variable: Yes or No

Source: This research

Floor level is another influential variable, as higher floors tend to offer better views and therefore command higher rental prices. The variable pets allowed was included for completeness, even though, due to recent legislation protecting tenants' rights, pet restrictions are no longer legally enforceable. Still, some rental listings maintain restrictions, so the variable was retained. The furnished condition was considered because furnished units typically have higher rental values. Finally, the proximity to subway stations was included, as reduced parking availability in São Paulo has made public transport access a relevant determinant of rental price.

Table 2 summarizes the descriptive statistics for the quantitative variables. One outlier was removed prior to modeling. The data indicate that 75% of the listed properties have rental values below R$ 3,755.00, and the mean square footage is 63.75 m², although the third quartile reaches only 76 m², confirming the trend toward smaller apartments. The median number of bedrooms is 1, while the third quartile is 2. As illustrated in Figure 5, rental prices are highly concentrated below R$ 4,000, with only a small number of properties exceeding R$ 10,000.

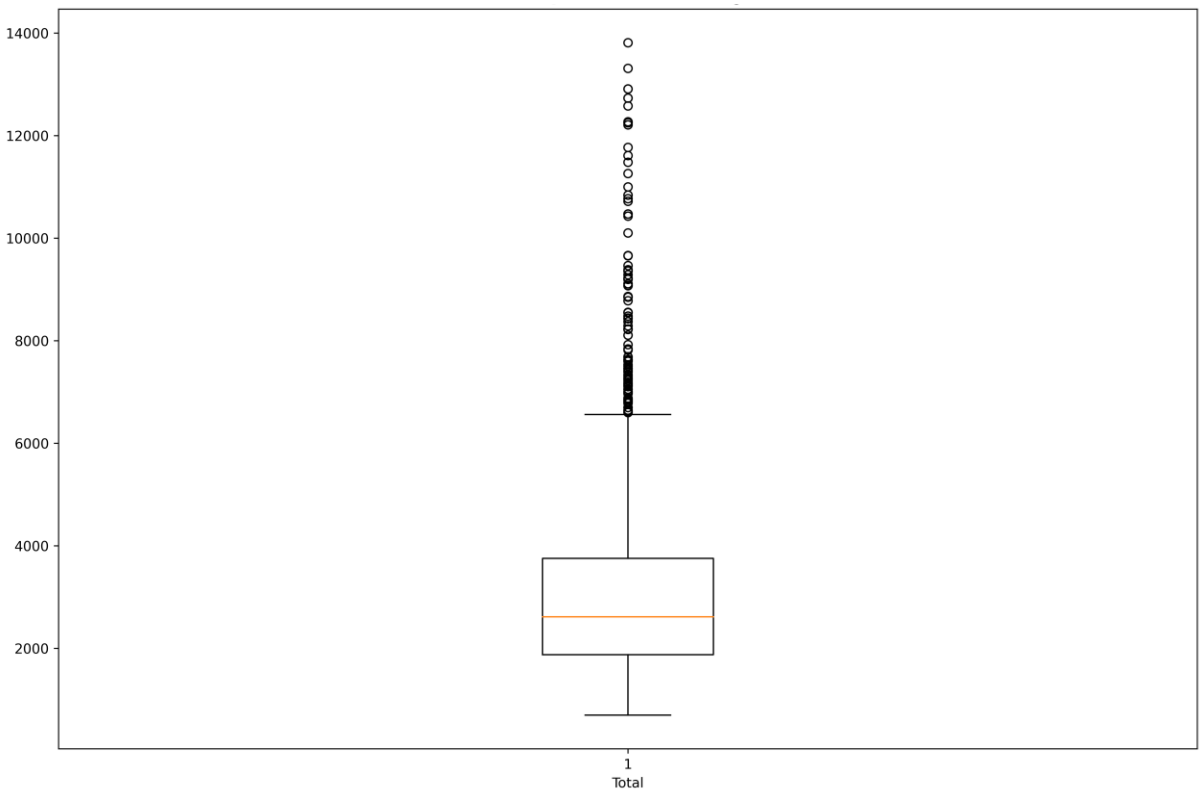
Table 2 – Description of Quantitative Variables

	Total (target)	Square ft	Bedrooms	Bathrooms	Park space	Floor
Count	2774.000	2774.000	2774.0	2774.0	2774.0	2774.0

Mean	3056.083	63.754	1.592	1.398	0.628	5.932
Standard deviation	1753.919	38.616	0.748	0.685	0.674	5.000
Min	697.000	10.0	0.0	1.0	0.0	0.0
25%	1876.500	40.0	1.0	1.0	0.0	2.0
50%	2615.500	52.0	1.0	1.0	1.0	5.0
75%	3755.000	76.0	2.0	2.0	1.0	9.0
Max	13810.000	587.0	5.0	7.0	6.0	43.0

Source: Data collected

Figure 5 – Boxplot of rent value distribution

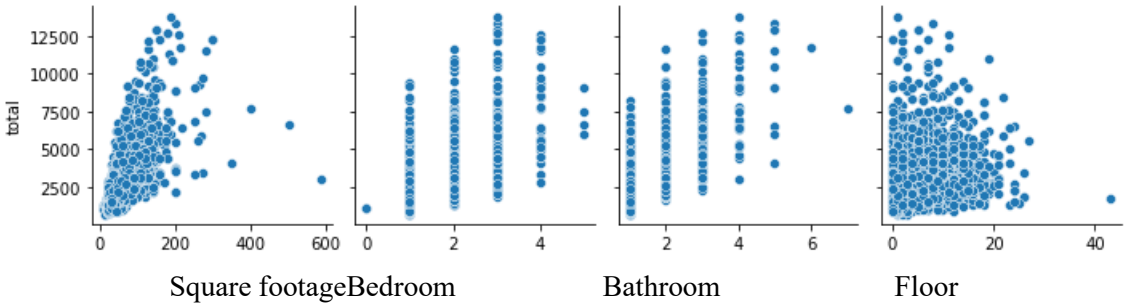


Source: This research

In Figure 6, all explanatory variables show a positive association with rental price. However, square footage displays the strongest linear pattern, with observations more closely aligned around the trend line. The variables bedrooms and bathrooms also show a positive relationship with rental value, but with less linear behavior and greater dispersion. The variable floor demonstrates a slight upward trend but with high scatter, indicating a weaker linear correlation.

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

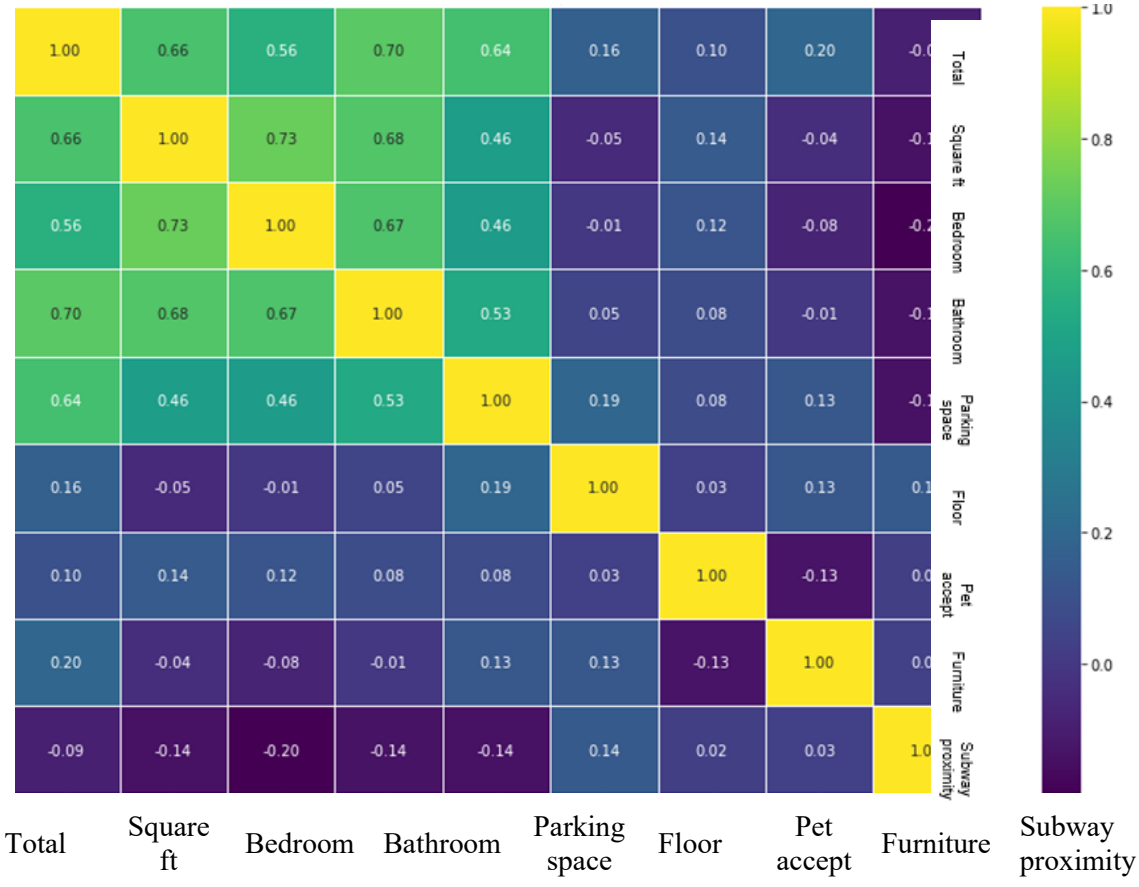
Figure 6 – Scatterplots: Predictor vs. Quantitative Variables



Source: This research

The correlation matrix in Figure 7 provides an overview of the relationships among the model variables. Pearson's correlation coefficient ranged from −1 to +1, with values close to +1 indicating strong positive correlations. The results show a strong positive correlation between rental price, square footage, number of bedrooms, number of bathrooms, and parking spaces. This suggests that higher rents are typically associated with larger units that accommodate more residents and therefore require more bathrooms. The number of parking spaces is also positively correlated with rent, reflecting the scarcity of parking in densely populated areas of São Paulo. By contrast, the correlation between rental price and subway proximity is weak, suggesting that smaller, lower-priced properties are more likely to be near public transportation hubs.

Figure 7 – Correlation Matrix for variables



Source: This research

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

2.4 Analysis and validation of the forecast models generated

The analysis was conducted using a comprehensive database containing more than 3,000 property records advertised in the city of São Paulo to assess the predictive accuracy of the model developed. After the initial training phase, the model was validated using a randomly selected sample of properties to determine its robustness. Evaluating predictive models with data not part of the training set is essential to detect necessary adjustments and verify the model's ability to generalize when exposed to new patterns of error and success. This validation stage is therefore fundamental for determining the model's effectiveness in real, heterogeneous scenarios.

To evaluate model performance, the Mean Squared Error (MSE) was used. MSE measures the average squared difference between the predicted and actual values, penalizing larger deviations more severely. Thus, the lower the MSE value, the better the model's predictive performance, as it indicates minimal discrepancy between estimated and observed outputs (Equation 1).

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{Yi} - \widehat{\mathrm{Y}}\mathrm{i})^2 \qquad (1)$$

The Root Mean Squared Error (RMSE), shown in Equation (2), is derived from the MSE and retains the same penalization properties. However, the square root is applied to return the metric to the same unit of measurement as the dependent variable, improving interpretability, whereas MSE is expressed in squared units (Kumar, 2024).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Yi} - \widehat{\mathrm{Y}}\mathrm{i})^2} \qquad (2)$$

The coefficient of determination (R-squared), presented in Equation (3), indicates how well the model explains the variability of the dependent variable. Its value ranges from 0 to 1, with higher values indicating greater explanatory power. Because it is a scale-independent metric, R^2 allows comparisons across models using the same predictors (Oliveira, 2021).

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i}^{n}(yi - yi)^2}{\sum_{i}^{n}(yi - yi)^2} \qquad (3)$$

To reduce model complexity and avoid overfitting, pruning techniques were applied to the decision trees. Pruning removes branches that offer minimal predictive value, improving model interpretability and reducing computational cost. The most widely used method is cost-complexity pruning, which balances model performance with structural simplicity by eliminating branches that do not contribute meaningfully to prediction accuracy.

The optimal hyperparameters were identified using the BayesSearchCV tool, which applies Bayesian optimization to minimize a cost function across a multidimensional search space. This approach uses Gaussian regression to estimate the behavior of the objective function and iteratively selects the most promising regions for evaluation (Michel, 2020). In this study,

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

the objective function was defined to minimize the squared error, enabling the model to converge efficiently toward the configuration that yielded the lowest prediction error.

3 RESULTS AND DISCUSSION

3.1 Model Performance Analysis

The performance of the predictive models was assessed using three standard regression metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These measures allow comparison not only between the machine learning models but also against a traditional benchmark — Ordinary Least Squares (OLS) regression. As shown in Table 3, the Random Forest model achieved the strongest performance, reaching an R^2 of 0.7902, indicating that nearly 79% of the variance in rental prices was explained by the model. This result reflects the model's ability to capture patterns that are nonlinear and influenced by interactions among predictors, a capability that classical linear models do not possess.

The Gradient Boosting model also performed well, with an R^2 of 0.7775, although it required noticeably more computational processing. The single Decision Tree obtained an R^2 of 0.7041, confirming its lower predictive capacity despite being the simplest model in terms of structure and interpretability. Finally, the OLS regression yielded the weakest results, with an R^2 of 0.6831 and the highest MSE and RMSE values, reinforcing that the pricing dynamics in the São Paulo real estate market are not linear and are better represented by ensemble learning methods than by parametric estimation.

Table 3 – Model performance

Error measurement	OLS	Decision Tree	Random Forest	Gradient Bosting
MSE	926014.4349	864486.8106	612839.6271	649942.0371
RMSE	962.2964	929.7778	782.8407	806.1898
R2	0.6831	0.7041	0.7902	0.7775

Source: This research

3.2 Importance and Interpretation of Explanatory Variables

In addition to model accuracy, it was essential to understand how each explanatory variable contributed to predicting rental values. Table 4 presents the variable importance rankings for each model. Across the Decision Tree, Random Forest, and Gradient Boosting algorithms, the most influential variable was bathroom count, followed by square footage and parking spaces. The prominence of bathroom count — a variable often overlooked in traditional pricing analyses — confirms that internal property usability and comfort features significantly affect rent pricing in São Paulo.

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

Table 3 – Variable importance

Variables	Decision Tree	Random Forest	Gradient Bosting
Pet accept	0.007568	0.012304	0.004059
Floor	0.048178	0.103940	0.067560
Bathroom	0.484455 *	0.405329 *	0.436789 *
Bedroom	0.026110	0.034634	0.030025
Square footage	0.265892	0.282474	0.279483
Furniture	0.041046	0.035920	0.043006
Parking space	0.126751	0.125400	0.139078

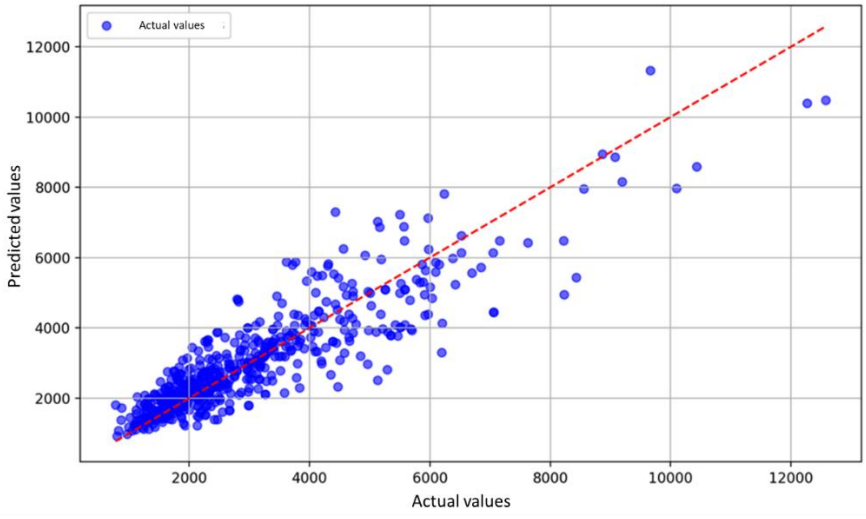
Source: This research

Another relevant result is the moderate importance of the "floor" variable, which, although less dominant than structural attributes, reflects the market's valuation of vertical housing characteristics, such as views, noise levels, and perceived safety. On the other hand, variables commonly assumed to be strong determinants — such as allowing pets or the presence of furniture — showed comparatively low influence in all models. This suggests that such amenities function more as secondary differentiators rather than primary determinants of price, especially in highly competitive rental markets.

3.3 Visualization of Predictive Accuracy

To further support the numerical results, Figure 8 compares the actual rental prices with those predicted by the Random Forest model. The closer the points are to the diagonal line, the better the prediction accuracy. The distribution reveals that the model performs well for the majority of rental values up to approximately R$7,000, where prediction density is highest. As expected, greater dispersion is observed in higher-value properties, which tend to exhibit more heterogeneous characteristics and a lower statistical frequency in the dataset.

Figure 8 – Actual value vs Predicted values



Source: This research

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

The visual evidence supports the conclusion that Random Forest effectively generalizes the price structure of most properties, particularly in the rental range where market demand is highest. Outlier behavior is present but proportionally low, especially given the dataset's size. This graphical inspection reinforces the interpretation of model robustness and confirms the superiority of nonlinear ensemble methods for real estate price modeling.

3.4 Consolidated Discussion of Findings

The comparison between models shows that the complexity of the São Paulo rental market is better captured by algorithms that model non-linearities and variable interactions. The significant gain in predictive quality from OLS to ensemble learning confirms that a single dominant variable does not determine housing prices; rather, it is a combination of structural, locational, and qualitative attributes. This aligns with the existing housing economics literature, which interprets price formation as a multivariable hedonic process rather than a linear relationship.

Another meaningful result is confirmation that physical property attributes — especially internal utility features such as the number of bathrooms and square footage — exert greater influence on pricing than external factors such as proximity to the subway or landlord restrictions. While this may initially contradict urban mobility theories, it reflects a specific characteristic of São Paulo: the rental market remains strongly segmented by property quality and size, even in a city where transportation and accessibility are highly valued.

4 FINAL CONSIDERATIONS

This study developed and evaluated a predictive pricing model for rental properties in the city of São Paulo using supervised machine learning techniques, including Decision Tree, Random Forest, and Gradient Boosting regressors. The dataset was collected from an online real estate platform and processed through exploratory statistical analysis, correlation mapping, and model training with performance comparison. The results demonstrated that ensemble-based models, particularly Random Forest, achieved the highest predictive accuracy, balancing interpretability, and performance, and outperformed both the simple decision tree and the linear regression benchmark.

Although the Random Forest model provided the strongest explanatory power for rental prices, the findings must be interpreted within the study's contextual limits. The data used refers exclusively to the metropolitan region of São Paulo, a market with high verticalization, strong segmentation, and price dynamics that may not represent cities in the countryside or other global urban areas. Therefore, external validity is restricted, and model generalization should be tested before broader application.

The study advances the literature by demonstrating how modern machine learning techniques can refine rental price predictions beyond traditional econometric models, particularly in dense, heterogeneous housing markets. The results confirm the relevance of structural attributes, such as the number of bathrooms, square footage, and parking availability, in driving price variation, while also illustrating the reduced influence of frequently assumed factors, such as pet acceptance or furnishing. The research therefore contributes both empirically and methodologically, reinforcing the usefulness of ensemble-based ML in real

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

estate analytics and providing an evidence-based framework for future academic and professional applications.

From an applied perspective, the findings are relevant for real estate market agents, property managers, and public decision-makers. Machine learning models such as Random Forests can support more accurate, data-driven rental pricing, improve investment valuation strategies, and guide regulatory planning — for instance, by identifying which attributes strongly influence affordability. Moreover, the results may assist urban policy discussions involving housing availability, zoning practices, and the economic effects of new developments in metropolitan areas.

However, the study presents important limitations. The dataset lacked property geocoordinates and neighborhood labels, which prevented a spatial econometric analysis, a major factor in real estate valuation. Information on qualitative property attributes (age, finishing level, maintenance status, or building amenities) was also absent, reducing the model's capacity to capture subjective dimensions of value. Additionally, the number of explanatory variables was limited, preventing analysis of macroeconomic and temporal factors such as inflation, supply-demand cycles, and post-pandemic shifts in housing behavior.

Future research should address these limitations by incorporating geolocated data, spatial attributes, and time-series components to capture price variation over economic cycles. The inclusion of the Human Development Index (HDI) or socioeconomic indicators could offer a deeper understanding of rental segmentation by neighborhood. Other machine learning techniques — such as XGBoost, CatBoost, deep learning regressors, or hybrid spatial-ML models — may further improve performance and interpretability. Finally, extending the analysis to price per square meter or comparing cities with different housing markets could expand the findings and strengthen generalization.

5 REFERENCES

APPORAS. **What is the difference between Bagging and Boosting?** 2016. Available at: https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/. Accessed on: Jun. 20, 2024.

ARRAES, R. A.; SOUZA, F. E. **Externalidades e formação de preços no mercado imobiliário urbano brasileiro: um estudo de caso**. Economia Aplicada, v. 12, p. 289-319, 2008.

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

CHOUDHURY. **What is Gradient Boosting? How is it different from Ada Boost?** Available at: https://medium.com/analytics-vidhya/what-is-gradient-boosting-how-is-it-different-from-ada-boost-2d5ff5767cb2. Accessed on: Jun. 10, 2024.

DAMASO, O. R. **O potencial do mercado de locação residencial no Brasil**. Conjuntura da Construção, p. 8-9, 2008.

FÁVERO, L. P. L.; BELFIORE, P. P.; LIMA, G. A. S. F. **Modelos de precificação hedônica de imóveis residenciais na Região Metropolitana de São Paulo**: uma abordagem sob as perspectivas da demanda e da oferta. Estudos Econômicos, v. 38, p. 73-96, 2008.

GONÇALVES, M. M. **Precificação de imóveis utilizando regressão linear múltipla e**

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026

árvores de decisão. Universidade de Santa Catarina, 2022. 75 p.

KUMAR, A. Available at: https://vitalflux.com/mse-vs-rmse-vs-mae-vs-mape-vs-r-squared-when-to-use/.

MICHEL, R. **Optimizing hyperparameters the right way**. Medium, 2020. Available at: https://towardsdatascience.com/optimizing-hyperparameters-the-right-way-3c9cafc279cc. Accessed on: Sep. 22, 2024.

OLIVEIRA, C. **Métricas para regressão: entendendo as métricas R², MAE, MAPE, MSE e RMSE**. 2021. Available at: https://medium.com/data-hackers/prevendo-n%C3%BAmeros-entendendo-m%C3%A9tricas-de-regress%C3%A3o-35545e011e70.

PINIEWEB. **Mercado imobiliário de aluguel está aquecido em São Paulo**. 2024. Available at: https://piniweb.com.br/mercado-imobiliario-de-aluguel-esta-aquecido-em-sao-paulo/. Accessed on: Apr. 01, 2024.

PYTHONGEEKS, Team. **Gradient boosting algorithm in machine learning**. 2024. Available at: https://pythongeeks.org/gradient-boosting-algorithm-in-machine-learning/. Accessed on: Sep. 20, 2024.

RITA, Pedro Artur Alves. **Aplicação data mining para análise e previsão das estratégias de pricing em companhias aéreas**: estudo de caso: registos das tarifas da rota SSA-LIS. 2018. 92 f. Dissertation (Master's) – Curso de Gestão de Informação com Especialização em Gestão do Conhecimento e Business Intelligence, Universidade Nova de Lisboa, Lisboa, 2018.

ROLNIK, Raquel. **Mercado de aluguel de imóveis em São Paulo está sob tensão**. Jornal da USP, 2023. Available at: https://jornal.usp.br/radio-usp/mercado-de-aluguel-de-imoveis-em-sao-paulo-esta-sob-tensao/. Accessed on: Sep. 16, 2024.

SRUTHI, E. **Understand random forest algorithm with examples**. 2024. Available at: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/. Accessed on: May 29, 2024.

WISSENBACH, Tomás Cortez. **A cidade e o mercado imobiliário: uma análise da incorporação residencial paulistana entre 1992 e 2007**. 2008. 142 f. Dissertation (Master's) – Curso de Geografia, Universidade de São Paulo, São Paulo, 2008.

YIU, T. **Understanding random forest**. Towards Data Science, 2021. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2. Accessed on: May 14, 2024.

Revista Fatec Zona Sul - Refas - Edição 54 - Abril de 2026