Cristiane de Andrade Coutinho

Predictive Modeling of the Closing Price of Real Estate Investment Funds Using Random Forest

SÃO PAULO
2025

Cristiane de Andrade Coutinho

Predictive Modeling of the Closing Price of Real Estate Investment Funds Using Random Forest

Final Course Project submitted to the Institute of Technology and Leadership (INTELI), to obtain a bachelor's degree in Computer Science

Advisor: Prof. Geraldo Magela Severino Vasconcelos

Coadvisor: Prof. Adriana Vieira Coelho

SÃO PAULO
2025

Cataloging in Publication
Library and Documentation Service
Instituto de Tecnologia e Liderança (INTELI)
Data entered by the author.

Sobrenome, Nome
Título do trabalho: subtítulo / Nome Sobrenome do autor; Nome e Sobrenome do orientador. – São Paulo, 2025.
nº de páginas : il.

Trabalho de Conclusão de Curso (Graduação) – Curso de [Ciência da Computação] [Engenharia de Software] [Engenharia de Hardware] [Sistema de Informação] / Instituto de Tecnologia e Liderança.
Bibliografia

1. [Assunto A]. 2. [Assunto B]. 3. [Assunto C].

CDD. 23. ed.

Resumo

[Coutinho, Cristiane. **Modelagem Preditiva do Preço de Fechamento de Fundos Imobiliários Utilizando Random Forest**. 2025. nº de folhas. TCC (Graduação) – Curso Ciência da Computação, Instituto de Tecnologia e Liderança, São Paulo, 2025.]

O mercado de Fundos de Investimento Imobiliário (FIIs) tem crescido de forma significativa no Brasil, impulsionado pela democratização do acesso a produtos financeiros e pela busca por diversificação em portfólios de renda variável. Nesse contexto, prever o comportamento dos preços de fechamento das cotas torna-se relevante para investidores e gestores. Este estudo apresenta uma modelagem preditiva utilizando dados históricos de FIIs negociados na B3 entre 2021 e 2024, com foco na aplicação dos algoritmos Random Forest Regressor. Foram utilizadas variáveis derivadas, como retorno diário, volatilidade móvel e delta do volume, além de variáveis originais relacionadas à precificação. A validação temporal foi realizada por meio de TimeSeriesSplit, com cinco folds. O modelo Random Forest apresentou desempenho satisfatório, com MAE médio de 19,86, RMSE de 57,78 e R^2 médio de 0,7504.

Palavras-Chave: Fundos de Investimento Imobiliário; Machine Learning; Random Forest;

Abstract

[Coutinho, Cristiane. **Predictive Modeling of the Closing Price of Real Estate Investment Funds Using Random Forest**. 2025. nº of pages. Final course project (Bachelor) – Course Computer Science, Institute of Technology and Leadership, São Paulo, 2025.]

The Brazilian Real Estate Investment Funds (FIIs) market has expanded significantly over the past years, driven by increased access to financial products and the search for diversified investment portfolios. In this context, forecasting the closing prices of fund shares becomes essential for investors and portfolio managers. This study presents a predictive modeling approach based on historical data from FIIs traded on B3 between 2021 and 2024, applying Random Forest Regressor and Linear Regression algorithms. Derived variables such as daily returns, rolling volatility, and delta volume were incorporated, alongside original pricing and liquidity features. Temporal validation was performed using TimeSeriesSplit with five folds. The Random Forest model showed solid performance, achieving a mean MAE of 19.86, RMSE of 57.78, and an R^2 of 0.7504.

Key words: Real Estate Investment Funds; Machine Learning; Random Forest; Linear Regression; Volatility; Predictive Modeling Predictive Modeling.

List of Illustrations

List of Tables

Summary

1. Introduction

The pursuit of understanding the factors that influence the profitability of financial assets has been the subject of research for several decades. Classical studies analyzed the behavior of the U.S. stock market over nearly half a century, investigating the relationship between average stock returns and variables such as market risk, market value, fundamental indicators, and leverage. These studies laid the foundation for contemporary models of risk and return assessment in financial markets (MIRANDA, 2013).

In the Brazilian context, the growth in the number of investors reflects the advancement of investment culture in the country. In 2022, the market recorded more than 4.2 million individual investor accounts at brokerage firms, driven by the democratization of access to financial products, asset diversification, and the historically low interest rate environment observed between 2020 and 2021 (GOMES, 2022). Equity investments thus became widely used both for wealth accumulation and for retirement and financial protection strategies.

Among the various investment instruments, Real Estate Investment Funds (REIFs) stand out. These funds have been regulated in Brazil since Law No. 8,668/1993 and later refined through regulations issued by the Brazilian Securities and Exchange Commission (CVM). REIFs pool resources from multiple investors, enabling direct or indirect participation in real estate ventures. Fund shares are book-entry and registered, representing ideal fractions of the fund's equity, and are primarily distributed through public offerings, although recent regulations have expanded flexibility for private issuances (FIGUEIREDO, 2019).

The real estate sector, closely linked to the construction industry, plays a crucial role in economic growth, particularly during periods of crisis, when public policies and sectoral investments become relevant mechanisms for stimulating GDP and mitigating socioeconomic impacts. In scenarios of instability, such as the COVID-19 pandemic, the need for strategic planning and segmentation in the development of new projects became even more evident, as identifying consumer

profiles and their specific demands is essential for the feasibility of real estate ventures (SIEBRA, 2024).

Within this framework, REIFs have consolidated themselves as central instruments in the capital markets, contributing to the financing of sale-leaseback operations, the development of logistics warehouses, the expansion of shopping centers, and the provision of long-term capital for developers and companies in the real estate sector. The breadth of this market reinforces its relevance for portfolio diversification and risk reduction in financial operations (FIGUEIREDO, 2019).

Thus, understanding the factors that influence REIF performance becomes essential for investors, managers, and researchers alike. As the market expands and becomes more accessible, the need for models capable of analyzing financial indicators, forecasting prices, and assessing risks also increases, contributing to more informed and sustainable investment decisions.

1.1. Study Objective

The objective of this study is to develop and evaluate a predictive model to estimate the daily closing price of Real Estate Investment Funds traded on B3, using supervised learning algorithms, with a focus on the Random Forest Regressor. Specifically, the study aims to:

- Analyze the influence of price-related variables on model performance;
- Evaluate error metrics and explanatory power;
- Investigate the relevance of derived variables.

1.2. Random Forest

Decision trees are widely used in market segmentation, risk classification, and investment analysis. The Random Forest algorithm extends this concept by creating multiple decision trees with variations in the sets of independent variables and combining their results to produce a more robust prediction. This method employs

ensemble learning and bagging techniques to reduce variance and increase predictive accuracy (GOMES, 2022).

1.3. Linear Correlation Coefficient

Pearson's linear correlation coefficient (r) is used to measure the strength and direction of the relationship between two variables, ranging from −1 to 1. The coefficient of determination (R^2) indicates the proportion of the variance of the dependent variable explained by the model and is a fundamental measure for assessing goodness of fit (MARTINS, 2008).

1.4. Evaluation Metrics

To measure model accuracy, metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are used. These metrics quantify the deviation of predictions from observed values and are essential for evaluating predictive models, especially in highly volatile contexts (MIRANDA, 2013).

1.5. Volatility

Volatility is a statistical measure of risk that reflects the dispersion of returns of a financial asset. It can be calculated as the square root of the variance or as the standard deviation of the return series. Volatility is used to estimate uncertainty and price instability and is particularly relevant for short-term forecasting models (KONDO, 2008). To address sensitivity in calculations, a rolling window is commonly used, allowing more recent observations to be emphasized while balancing historical information.

2 Development

The methodology of this study was organized into sequential stages involving data collection, preparation, exploratory analysis, and predictive modeling based on machine learning algorithms, aiming to forecast the daily closing price of REIFs traded on B3. A supervised nonlinear model, the Random Forest Regressor, was employed due to its ability to capture complex relationships among financial variables.

2.1 Methodology

2.1.1 Data Used

Historical data for several REIFs listed on B3 were used, covering the period from 2021 to 2024, including:

Variable Name	Variable Type	Description
date	Date	Trading date
ticker	Nominal	Asset identifier
opening_price	Numeric	First traded price of the day
high_price	Numeric	Highest traded price of the day
low_price	Numeric	Lowest traded price of the day
closing_price	Numeric	Closing price of the day
volume	Numeric	Total traded value (number of trades × prices)

Table 1: *Description of the variables used.*

2.1.2 Additional Variables

Historical data for several REIFs listed on B3 were used, covering the period from 2021 to 2024, including:

- **Volume Delta**

$$\Delta V = V_t - V_{t-1}$$

V_t = Volume on the current day.

V_{t-1} = Volume do dia anterior.

Defined as the absolute difference in trading volume relative to the previous day, capturing abrupt or gradual changes in market activity.

- **Rolling Volatility**

$$\sqrt{\frac{1}{n-1}\sum_{i=0}^{n-1}(r_{t-i}-\bar{r}_t)^2},$$

n = Size of the rolling window.

r_{t-i} = Return of asset i at time t.

$\bar{r}_t$ = Mean return of asset i over the 20-period rolling window.

Defined as the rolling standard deviation of asset returns, representing historical volatility calculated using a rolling window. Higher values indicate greater uncertainty and price instability.

- **Return**

$$\frac{P_t - P_{t-1}}{P_{t-1}}.$$

$$\frac{P_t - P_{t-1}}{P_{t-1}}.$$

P_t = Return at time *t*.

P_{t-1} = Return on the previous day (*t − 1*).

Refers to the daily return of the asset, calculated as the percentage change in price relative to the previous day. This metric captures price appreciation or depreciation dynamics over time.

2.1.3 Descriptive Analysis

	preco_abertura	preco_max	preco_min	preco_medio	preco_fechamento	qtd_titulos	volume	delta_volume	retorno	volatilidade	volatilidade_movel
count	176586.000000	176586.000000	176586.000000	176586.000000	176586.000000	1.765860e+05	1.765860e+05	1.765860e+05	176586.000000	176586.000000	176586.000000
mean	118.528019	119.566034	117.386671	118.363698	118.396364	2.145898e+04	1.220627e+08	-2.129661e+05	0.000143	0.023309	0.015337
std	237.654779	240.172955	235.175707	237.410206	237.273012	7.309009e+04	2.373543e+08	1.747614e+08	0.136325	0.129019	0.135523
min	0.380000	0.400000	0.370000	0.380000	0.380000	1.000000e+00	5.740000e+02	-1.429267e+10	-0.946278	0.004789	0.000567
25%	66.520000	67.100000	65.820000	66.410000	66.500000	9.100000e+02	6.589592e+06	-1.196183e+07	-0.005435	0.008573	0.006817
50%	87.100000	87.850000	86.275000	87.000000	87.000000	4.773500e+03	3.364835e+07	-1.708600e+04	0.000000	0.010866	0.009527
75%	101.500000	102.000000	100.887500	101.400000	101.460000	1.984175e+04	1.449650e+08	1.118931e+07	0.004589	0.014964	0.013635
max	2943.990000	2950.000000	2911.020000	2928.790000	2925.090000	5.673655e+06	1.532104e+10	1.328176e+10	46.841924	2.986604	10.475008

Table 2: *Descriptive analysis of the data.*

The full dataset contains 176,586 records. Opening, high, low, average, and closing prices present means close to BRL 118 and relatively symmetric distributions but with a high standard deviation (approximately BRL 237), indicating strong heterogeneity among assets and periods. Trading volume and number of shares show high dispersion, reflecting REIFs with very different liquidity levels. Volume delta exhibits extremely wide variation, with minimum and maximum values in the order of billions, highlighting abrupt liquidity fluctuations. Daily returns have a mean close to zero, as expected for financial series, with a standard deviation of 0.13, indicating periods of strong price movements. Volatility and rolling volatility show low average values but significant peaks, suggesting generally stable behavior interrupted by episodes of intense fluctuation. These results highlight the statistical diversity of the dataset and reinforce the need for models capable of handling significant variability across REIFs.

2.1.4 Data Cleaning

- **Null, Empty, and Zero Values**

  Rows containing at least one null, empty, or zero value in any column were identified. Entire ticker series meeting these criteria were removed from the dataset.

- **Outlier Treatment**

  Outliers were not treated, as the variables of interest exhibit right-skewed distributions. Removing or altering extreme values could distort the natural variability of the data and negatively impact predictive performance, since outliers reflect significant financial movements that the model must learn to capture.

- **Removal of Assets with Fewer Than 250 Observations**

To input data into the model, each ticker was required to have at least 230 observations, corresponding to approximately one trading year. Tickers with fewer observations were removed.

2.1.5 Distribution Analysis

The distributions of financial variables were visualized using histograms to identify skewness, dispersion, and outliers. The x-axis range was limited to 0–200 to better visualize distribution shapes. Different observation ranges were applied for volume, volume delta, return, volatility, rolling volatility, and number of shares.

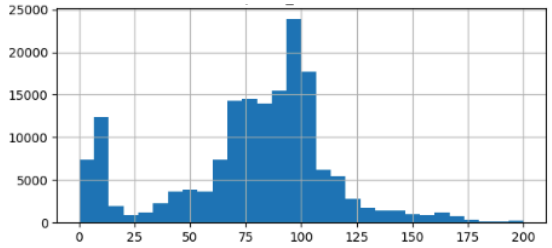

Figure 1: *Distribution of opening price*

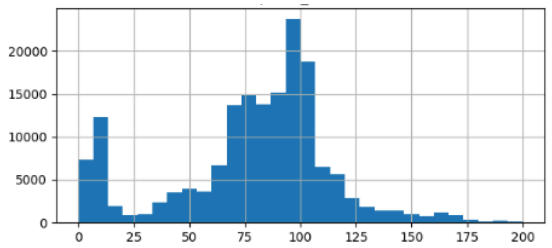

Figure 2: *Distribution of maximum price*

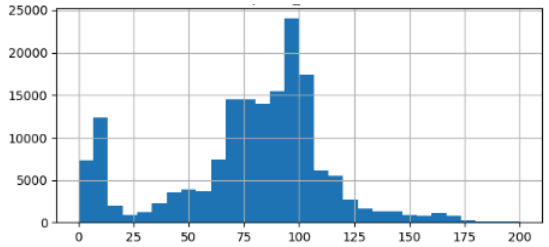

Figure 3: *Distribution of average price*

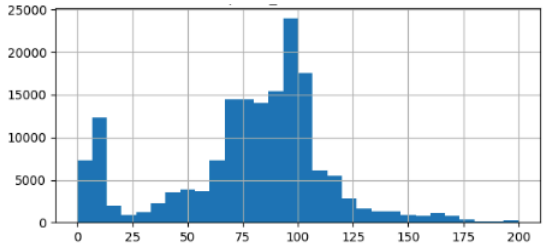

Figure 4 : *Distribution of closing price*

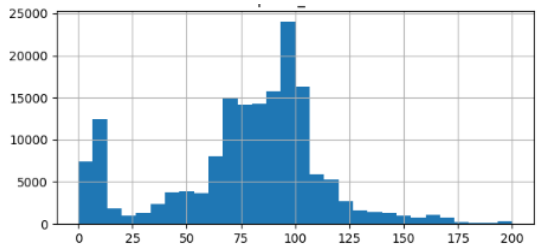

Figure 5: *Distribution of number of shares*

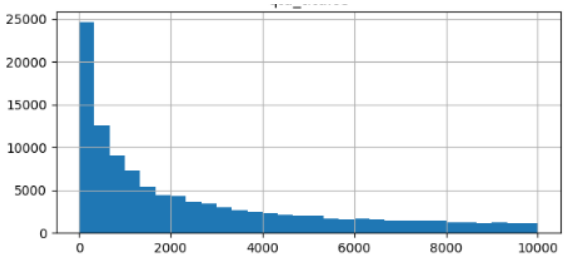

Figure 6: *Distribution of minimum price*

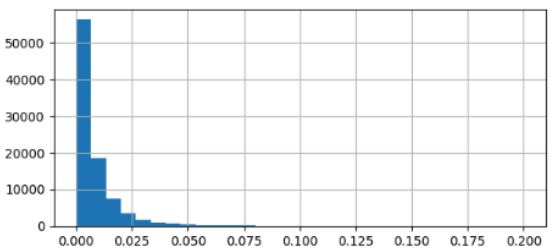

Figure 7: *Distribution of return*

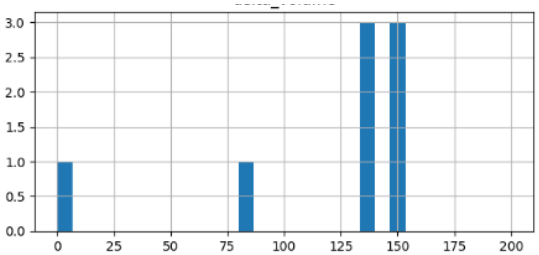

Figure 8: *Distribution of volume delta*

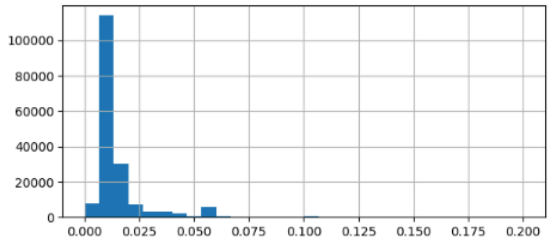

Figure 9: *Distribution of volatility*

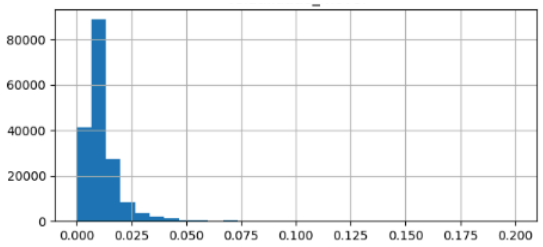

Figure 10: *Distribution of rolling volatility*

2.1.6 Relationship Between Variables

Scatter plots were constructed to visually assess relationships between pairs of variables, especially:

- **Volume delta × Closing price**

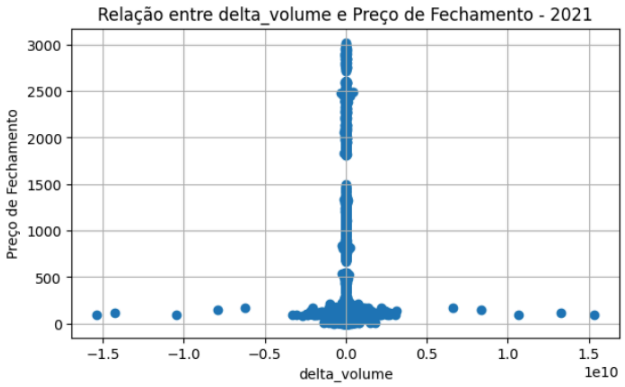

Figure 11: *Relationship between volume delta and closing price in 2021*

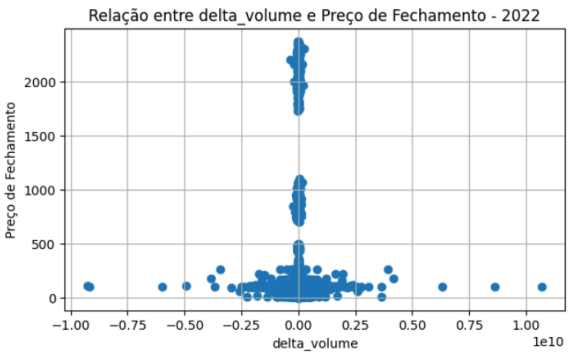

Figure 12: *Relationship between volume delta and closing price in 2022*

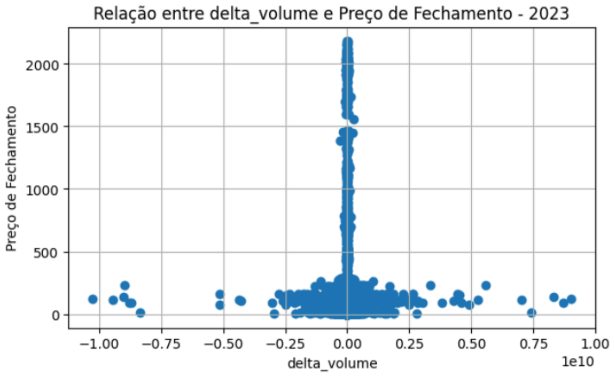

Figure 13: *Relationship between volume delta and closing price in 2023*

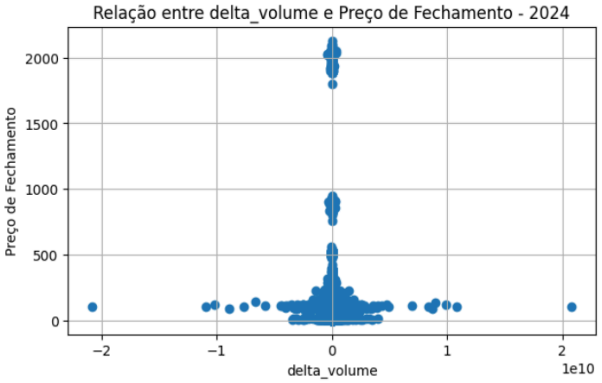

Figure 14: *Relationship between volume delta and closing price in 2024*

- **Return × Closing price**

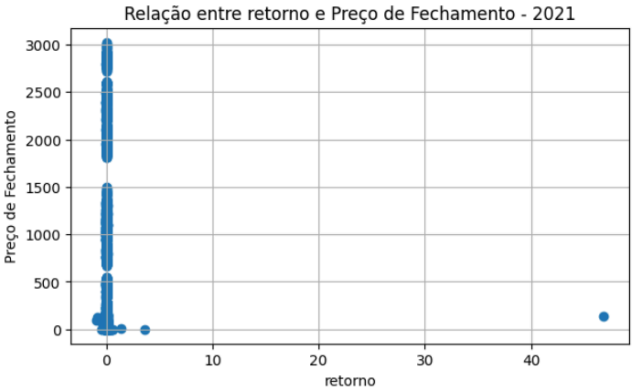

Figure 15: *Relationship between return and closing price in 2021*

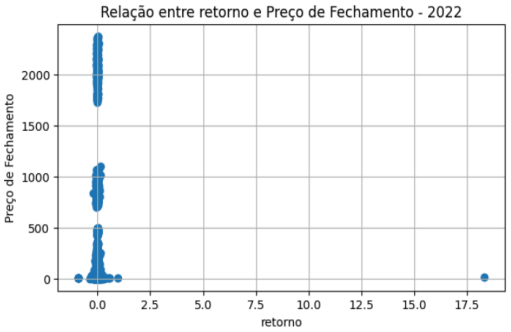

Figure 16: *Relationship between return and closing price in 2022*

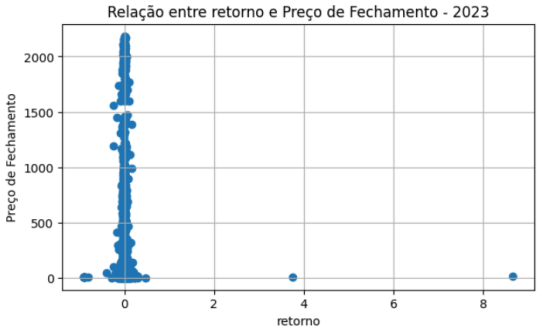

Figure 17: *Relationship between return and closing price in 2023*

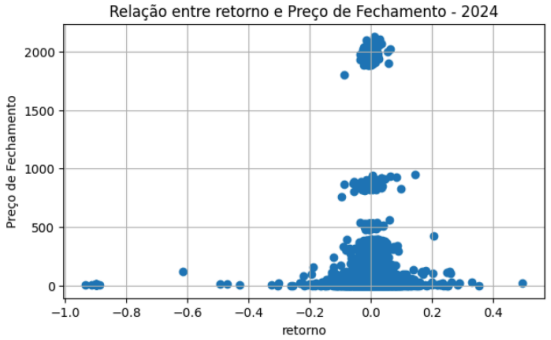

Figure 18: *Relationship between return and closing price in 2024*

- **Rolling volatility × Closing price**

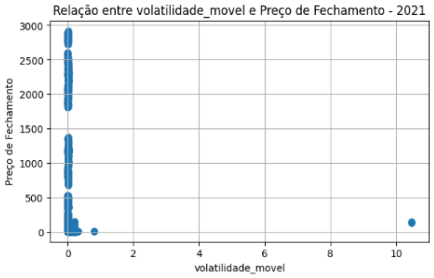

Figure 19: *Relationship between rolling volatility and closing price in 2021*

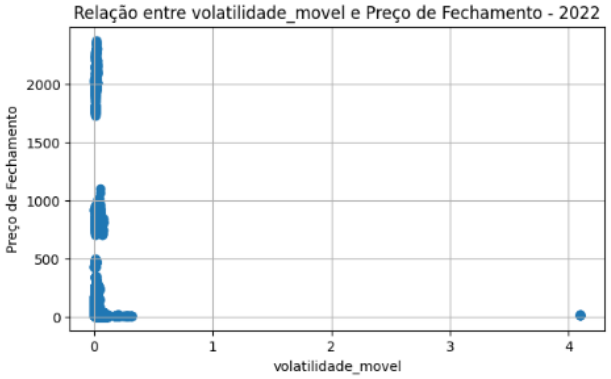

Figure 20: *Relationship between rolling volatility and closing price in 2022*

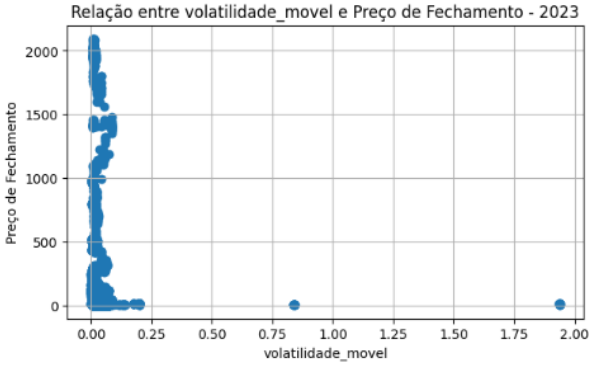

Figure 21: *Relationship between rolling volatility and closing price in 2023*

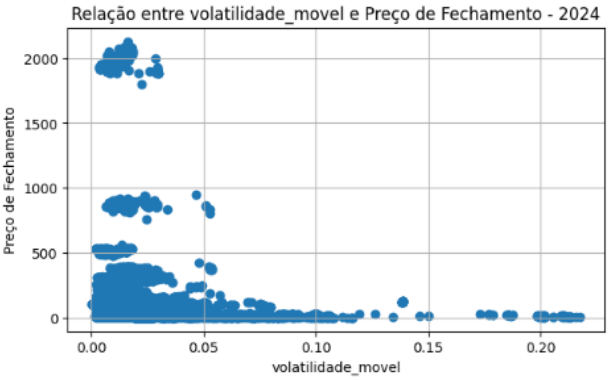

Figure 22: *Relationship between rolling volatility and closing price in 2024*

- **Number of shares × Closing price**

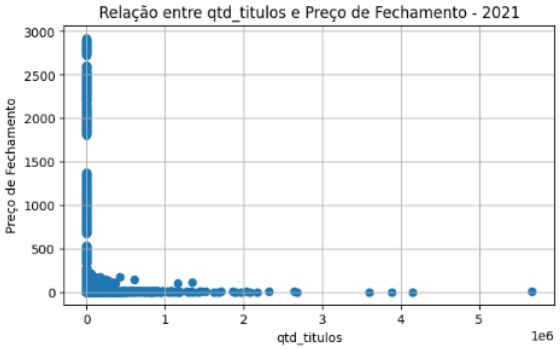

Figure 23: *Relationship between number of shares and closing price in 2021*

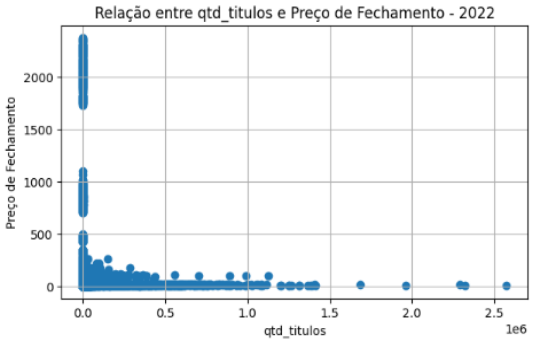

Figure 24: *Relationship between number of shares and closing price in 2022*

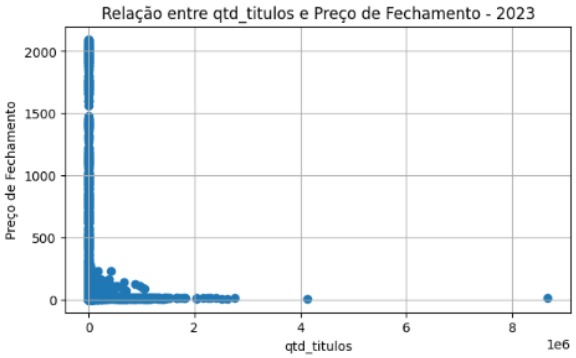

Figure 25: *Relationship between number of shares and closing price in 2023*

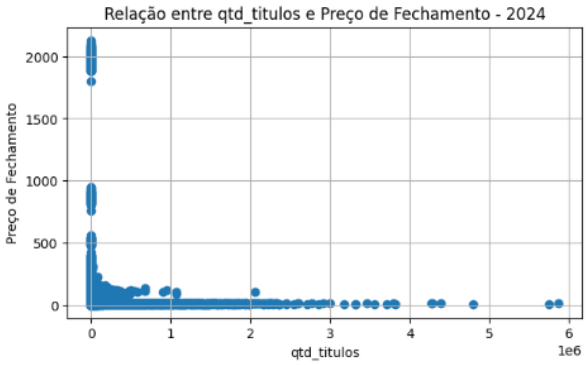

Figure 26: *Relationship between number of shares and closing price in 2024*

- **Volatility × Closing price**

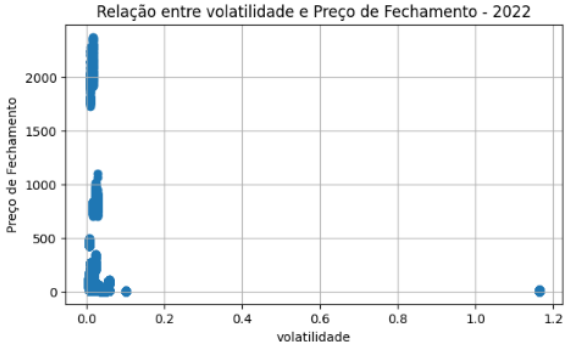

Figure 27: *Relationship between volatility and closing price in 2021*

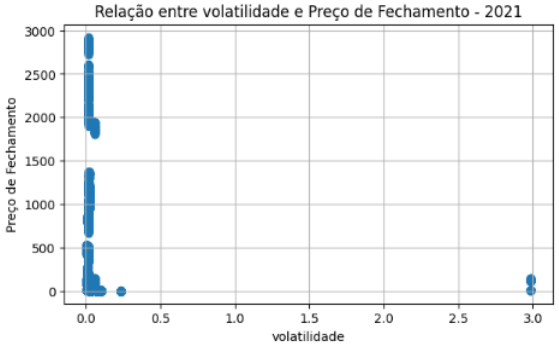

Figure 28: *Relationship between volatility and closing price in 2022*

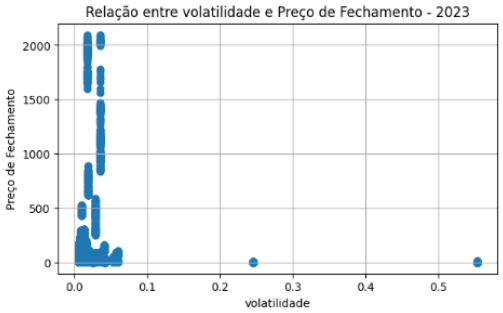

Figure 29: *Relationship between volatility and closing price in 2023*

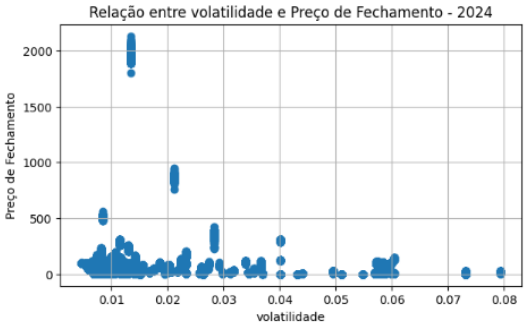

Figure 30: *Relationship between volatility and closing price in 2024*

2.1.7 Correlation Analysis

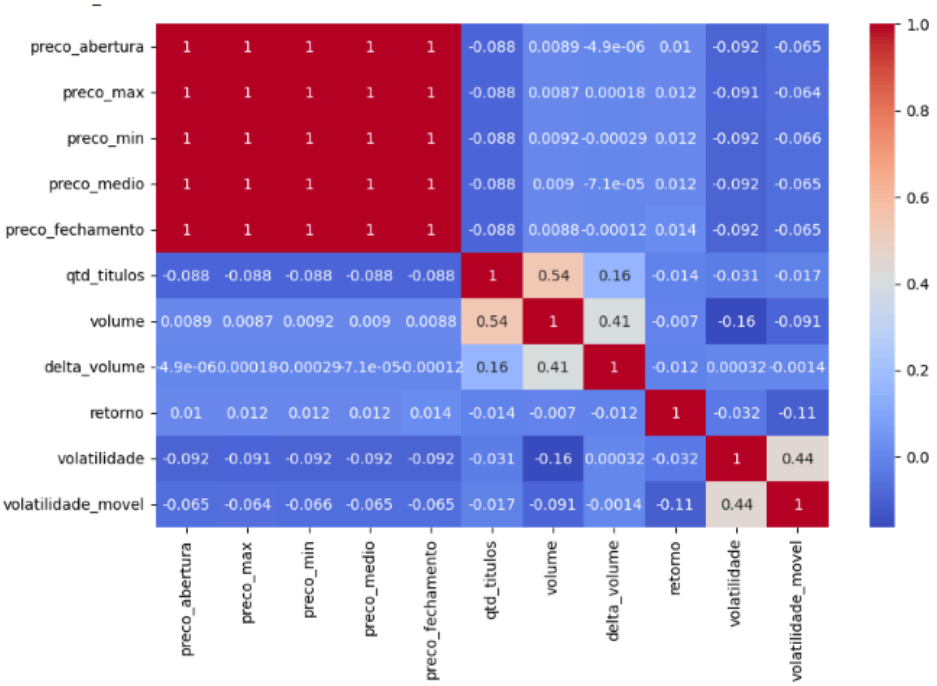

Figure 31: *Correlation plot between the studied variables*

Opening, high, low, average, and closing prices exhibit near-perfect correlation (≈ 0.999), indicating strong collinearity. Volume and number of shares show moderate correlation with each other (r ≈ 0.54) and with volume delta (r ≈ 0.40), reflecting expected liquidity dynamics. Volatility and rolling volatility present moderate positive correlation (r ≈ 0.43) but weak negative correlation with prices (−0.06 to −0.09). Daily returns show very low correlation with all other variables, reinforcing their erratic behavior.

2.1.8 Dataset Splitting

The entire dataset was used as input for the model without splitting by ticker, allowing the model to generalize across different REIFs.

2.1.9 Predictive Model Application

Random Forest is a machine learning model based on decision trees that combines multiple trees trained on random subsets of data and features. Each tree generates a prediction, and the final output is the average of all trees, reducing overfitting.

Parameters used:

- **n_estimators = 200** Number of trees in the forest.
- **max_depth = 10** Maximum depth of each tree.
- **random_state = 42** Ensures reproducibility.
- **n_jobs = -1** Uses all CPU cores to speed up training.

2.1.10 Temporal Validation

To robustly evaluate model performance over time, the TimeSeriesSplit technique was used. In this approach:

- The model is trained only on data prior to the test period;

- The training window expands with each fold, while the test window represents the subsequent period.
- Five temporal folds were used, reducing the risk of data leakage.

2.2 Results

The models were evaluated using five temporal folds (TimeSeriesSplit), ensuring that training in each fold used only data prior to the test period, thereby avoiding any temporal leakage.

The average metrics computed across the five folds were:

- **Mean MAE:** 19.86
- **Mean RMSE:** 57.78
- **Mean R^2:** 0.7504

These values indicate that, on average, the model is able to predict the closing price with a mean absolute error of approximately 20 units and a good explanatory power ($R^2 = 0.75$), meaning that about 75% of the price variance is explained by the model.

Fold	MAE	RMSE	R^2
1	41.88	122.78	0.2764
2	14.43	27.39	0.7037
3	21.55	102.94	0.9191
4	13.13	21.95	0.9729
5	8.33	13.84	0.8799

Table 2: *MAE, RMSE, and R^2 results for each training fold*

Random Forest allows the evaluation of the contribution of each variable to the prediction. The resulting feature importance was as follows:

Variable	Importance
volume	0.6876
qtd_titulos	0.3092
volatilidade_movel	0.0015
delta_volume	0.0014
retorno	0.0003

Table 3: *Variable importance in the model*

Trading volume was the most influential variable, accounting for approximately 69% of the predictive capacity. The number of traded units also showed considerable relevance (≈31%). The remaining variables had a nearly negligible impact, suggesting that, for the analyzed period and assets, the closing price is strongly related to trading volume and the number of units.

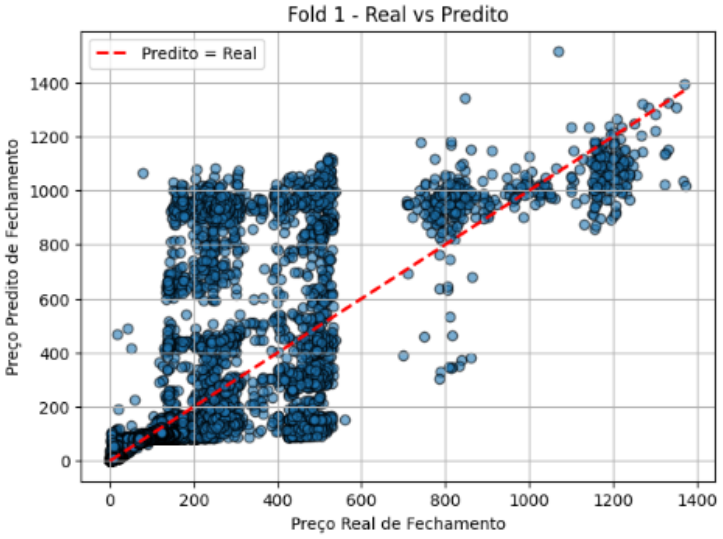

Figure 32: *Comparison between actual and predicted values for Fold 1*

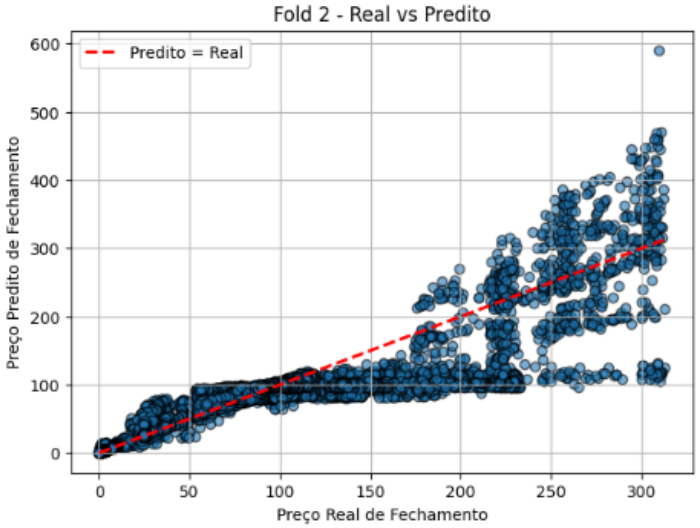

Figure 33: *Comparison between actual and predicted values for Fold 2*

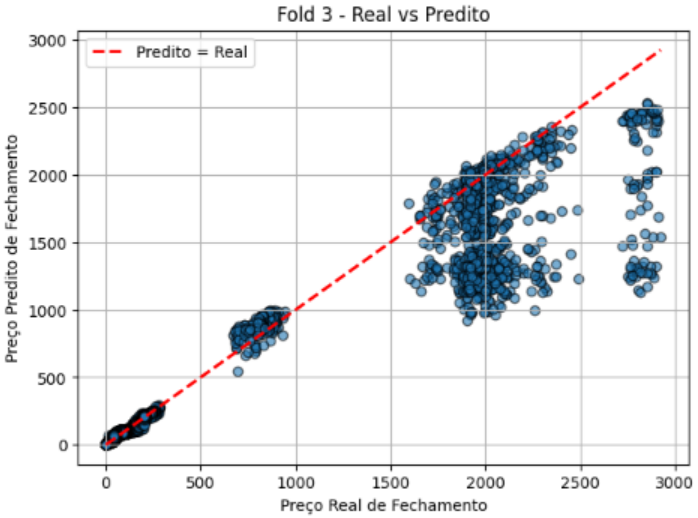

Figure 34: *Comparison between actual and predicted values for Fold 3*

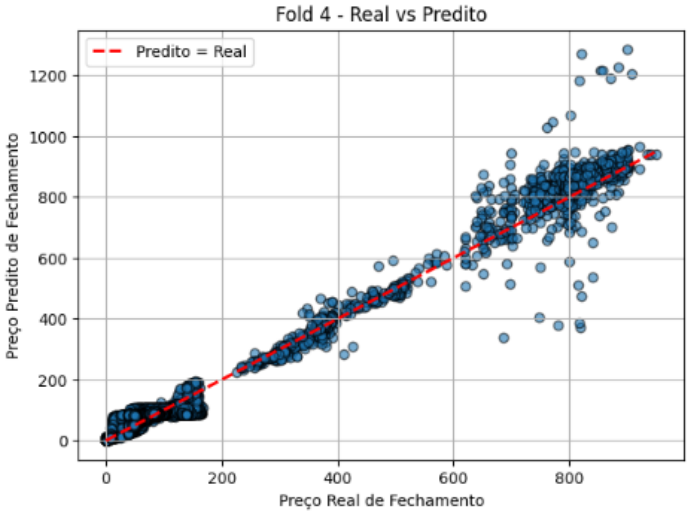

Figure 35: *Comparison between actual and predicted values for Fold 4*

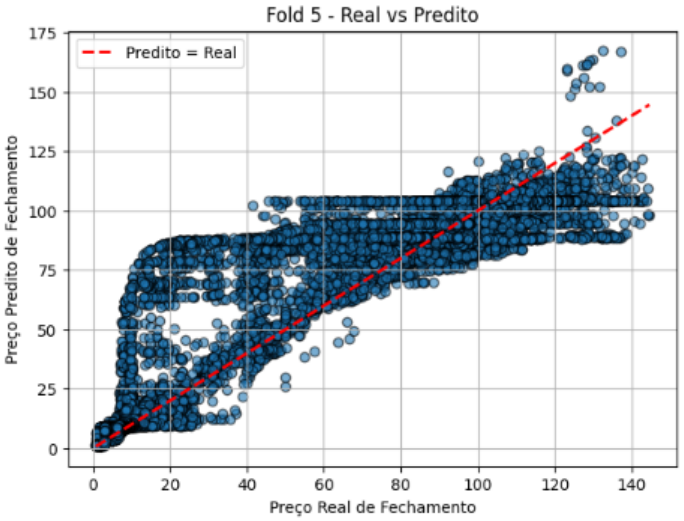

Figure 36: *Comparison between actual and predicted values for Fold 5*

2.3 Analysis or Discussion of Results

This section presents an analysis and discussion of the results obtained from the application of the Random Forest Regressor to predict the daily closing price of Real Estate Investment Funds (REIFs) traded on B3. The discussion focuses on model performance, temporal stability, variable relevance, and the financial interpretation of the findings.

- **Model Performance and Predictive Accuracy**

The Random Forest model demonstrated overall satisfactory predictive performance across the five temporal folds generated using the TimeSeriesSplit validation strategy. The average error metrics—MAE of 19.86 and RMSE of 57.78—indicate that the model was able to estimate closing prices with a relatively low absolute deviation when compared to the average price level observed in the dataset. Additionally, the average R^2 value of 0.7504 suggests that approximately 75% of the variance in closing prices was explained by the model.

These results indicate a strong in-sample explanatory capacity, especially considering the heterogeneity of the dataset, which includes multiple REIFs with distinct price levels, liquidity profiles, and trading behaviors. The ability of the model to capture nonlinear relationships between predictors and the target variable is

consistent with the known strengths of ensemble tree-based methods in financial forecasting contexts.

However, despite the favorable average performance, a closer inspection of individual folds reveals substantial variability in predictive accuracy over time. The R^2 values ranged from approximately 0.27 to 0.97, indicating that the model performed very well in certain periods while exhibiting considerably weaker explanatory power in others. This instability suggests that the model's effectiveness is highly sensitive to market regimes, particularly trends and volatility conditions present in specific temporal windows.

- **Temporal Stability and Market Regimes**

The observed variation across folds highlights a critical challenge in financial time series modeling: non-stationarity. REIF prices are influenced by macroeconomic conditions, interest rate cycles, sector-specific dynamics, and investor sentiment, all of which may change significantly over time. Periods characterized by strong price trends or stable liquidity conditions tend to favor higher predictive performance, as patterns learned during training remain valid in the test window.

Conversely, abrupt structural changes—such as shifts in monetary policy, economic shocks, or changes in real estate market expectations—can reduce the model's generalization capacity. The lower R^2 values observed in certain folds likely reflect such regime changes, reinforcing the idea that high average performance metrics may mask temporal fragility. This finding is consistent with the literature on financial forecasting, which emphasizes that strong historical fit does not necessarily translate into robust out-of-sample predictability.

- **Variable Importance and Financial Interpretation**

The analysis of feature importance revealed that traded volume and number of shares accounted for more than 98% of the model's predictive power. Traded volume alone contributed approximately 69%, while the number of shares represented roughly 31%. This dominance underscores the central role of liquidity-related variables in explaining closing price behavior within the analyzed sample.

From a financial perspective, this result is intuitive, as higher liquidity often coincides with increased price efficiency, stronger demand, and more active participation by market agents. Large trading volumes may reflect heightened investor interest, information flow, or institutional activity, all of which can influence price formation mechanisms in REIF markets.

In contrast, derived variables such as daily return, volatility, rolling volatility, and volume delta exhibited minimal contribution to the model. This suggests that short-term price fluctuations and risk measures, at least in their contemporaneous form, provide limited incremental information beyond what is already captured by liquidity variables. Additionally, the weak correlation between returns and prices observed in the exploratory analysis reinforces the erratic and noisy nature of return series in financial markets.

- **Methodological Considerations and Look-Ahead Bias**

While the strong reliance on volume and number of shares enhances predictive accuracy, it also raises important methodological concerns. Both variables reflect information consolidated at the end of the trading session, which introduces the risk of look-ahead bias when used to predict the same day's closing price. In practical forecasting scenarios, such information would not be fully available at the time investment decisions are made.

Therefore, although the model performs well from a retrospective and exploratory standpoint, its direct applicability to real-time prediction or trading strategies is limited. This limitation highlights the importance of incorporating temporal lags or using strictly ex-ante variables when the objective is decision support rather than descriptive modeling.

- **Implications for Financial Modeling of REIFs**

Overall, the results indicate that Random Forest models are effective tools for capturing nonlinear relationships and identifying dominant drivers of REIF price behavior, particularly those related to liquidity. The findings reinforce the notion that price formation in REIF markets is strongly associated with trading activity rather than short-term risk or return measures.

At the same time, the temporal instability observed across folds emphasizes the need for caution when interpreting high explanatory metrics in financial applications. Robust validation strategies, regime-aware modeling, and careful feature engineering are essential to mitigate overfitting and improve generalization.

Future research could enhance the robustness of the analysis by modeling returns instead of nominal prices, introducing lagged explanatory variables, segmenting funds by type or liquidity level, and integrating macroeconomic indicators. Such extensions may provide deeper insights into the dynamics of REIF pricing and improve the practical relevance of predictive models in the Brazilian real estate investment market.

3 Conclusion

This study aimed to investigate whether it is possible to predict the daily closing price of Real Estate Investment Funds (REIFs) traded on B3 using machine learning techniques, with an emphasis on the Random Forest Regressor, as well as to identify which variables exert the greatest influence in this predictive process. By applying the model to a large and heterogeneous dataset comprising more than 176,000 observations from 2021 to 2024, it was possible to objectively address the research question.

The results indicate that it is possible to model and explain a substantial portion of the variation in REIF closing prices using the Random Forest algorithm, as the model achieved consistent average performance, with an MAE of 19.86, RMSE of 57.78, and an average R^2 of 0.7504. These metrics demonstrate that the algorithm was able to capture relevant patterns in historical data and provide predictions with strong explanatory power from a statistical standpoint.

However, the temporal analysis revealed significant instability in performance across different folds, with considerable variation in R^2 values, indicating that the model's predictive capability is highly dependent on prevailing market conditions in each period. This behavior reinforces the non-stationary nature of financial time series and suggests that, despite good historical fit, the model's ability to generalize over time is limited.

Regarding variable relevance, the findings show that traded volume and number of shares are the primary determinants of closing prices, accounting for more than 98% of the total feature importance in the model. Derived variables such as daily return, volatility, rolling volatility, and volume delta exhibited virtually no impact. This result suggests that, within the analyzed context, REIF price formation is strongly associated with liquidity conditions, rather than with short-term risk or return measures.

Despite the model's strong explanatory performance, the study also identified important methodological limitations, particularly related to the use of contemporaneous end-of-day variables, which may introduce look-ahead bias and limit the model's applicability in real-world decision-making scenarios. Therefore, while Random Forest proves to be an effective tool for exploratory analysis and understanding the factors influencing REIF prices, methodological adjustments are required for its use in operational predictive settings.

Finally, this work contributes to the literature by demonstrating both the potential and the limitations of applying machine learning algorithms to REIF price modeling in the Brazilian market. Future studies may extend this research by using return-based transformations instead of nominal prices, incorporating lagged variables, segmenting funds by specific characteristics, and including macroeconomic factors, thereby enhancing the robustness and practical applicability of the proposed models.

References

GOMES, Gleidson Willer et al. Prediction of the closing price of a stock listed on the B3 stock exchange (Brasil, Bolsa, Balcão) using predictive analysis and the R language. **Universidade Federal de Minas Gerais.** 2022. Available at: https://repositorio.ufmg.br/server/api/core/bitstreams/a7fc8a2c-d14c-4252-8855-acb966ae1dfe/content

SOUZA, André Nunes. An analysis of the composition of the IMOB B3 real estate index. **Universidade Federal do Ceará.** 2023. https://repositorio.ufc.br/handle/riufc/74064

SIEBRA, Nélio Vitor Alves. Improvement of a linear regression model for real estate pricing and investment in Fortaleza.**Universidade Federal do Ceará.** 2024. Available at: https://repositorio.ufc.br/handle/riufc/79066

MIRANDA, Raul Figueira. Development of a linear regression model for analyzing stock price variation of construction companies based on the return on equity indicator. **Universidade Federal do Ceará.** 2013. Available at: https://repositorio.ufc.br/handle/riufc/35745

GUIMARÃES, Roberta Valente. Use of logistic regression to predict the closing of financial operations: currency forwards. **EPUSP**. Available at: https://bdta.abcd.usp.br/directbitstream/30a8e6f4-52c3-43c7-81b8-aa67333ad99f/RobertaValenteGuimaraes%20TCC-PRO06.pdf

FIGUEIREDO FILHO, Fabio. Behavior of the trading value of real estate investment fund shares in Brazil from 2013 to 2018. 2019. Monografia (MBA em Real Estate) – **Escola Politécnica, Universidade de São Paulo**, São Paulo, 2019. Available at: https://bdta.abcd.usp.br/item/003137015

KONDO, Daniel Yudi Sasahara.Volatility estimation models and their impact on the calculation of value at risk of a portfolio of financial assets. Final Project (Production Engineering) – **Escola Politécnica, Universidade de São Paulo**, São Paulo, 2008. Available at: https://bdta.abcd.usp.br/directbitstream/996737b2-1f3d-4470-a3cb-3e99bdac39da/DanielYudiSasaharaKondo%20TCC-PRO08.pdf

GONÇALVES, Gabriel Alexandre. Comparison of machine learning models for predicting the closing price of a banking sector stock listed on B3.**UNIVERSIDADE FEDERAL DE UBERLÂNDIA.** 2022. Available at: https://repositorio.ufu.br/bitstream/123456789/35505/1/Compara%c3%a7%c3%a3oDeModelos.pdf