Enhancing Stock Market Predictions: The Role of Feature Selection Techniques in Financial Modeling

Humberto O. Bragança[1(✉)], Richard F. Pinto[1], Bruno L. Dalmazo[1], Eduardo N. Borges[1], Giancarlo Lucca[2], Viviane L. D. de Mattos[1], and Rafael A. Berri[1]

[1] Federal University of Rio Grande (FURG), Rio Grande, Brazil
{humberto.braganca,richard_pinto,dalmazo,eduardoborges, vivianemattos,rafaelberri}@furg.br
[2] Catholic University of Pelotas (UCPel), Pelotas, Brazil
giancarlo.lucca@ucpel.edu.br

Abstract. Feature selection plays a crucial role in improving the performance of predictive models, especially in complex and dynamic environments such as the Brazilian financial market. This study evaluates the impact of different feature selection methods—OneR, Symmetrical Uncertain, Gain Ratio, Pearson Correlation, mRMR, and Correlation-based Feature Subset Selection—on financial market prediction using Random Forest models. A dataset consisting of 194 technical analysis indicators was analyzed, and progressive attribute removal was applied, with model performance assessed through cross-validation. The results demonstrate that feature selection significantly enhances model efficiency by reducing dimensionality while maintaining predictive accuracy. Moreover, our findings provide valuable insights into the relevance of different financial indicators, offering a methodological contribution to financial market analysis. Future work includes leveraging the selected features in advanced machine learning architectures, such as deep learning, to further refine prediction accuracy and develop more reliable models for financial forecasting.

Keywords: Feature Selection · Financial Market · Machine Learning

1 Introduction

The Brazilian financial market, Brasil Bolsa Balcão (B3)[1], stands out as a dynamic and complex emerging market, demanding specialized analytical and predictive approaches [1]. Its volatility and diverse economic sectors pose challenges and opportunities for financial modeling, requiring strategies to enhance predictive accuracy [2].

[1] https://www.b3.com.br/.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15649, pp. 386–400, 2025.
https://doi.org/10.1007/978-3-031-96997-3_24

In recent years, machine learning has emerged as an effective tool for financial decision-making, offering data-driven insights to mitigate risks and optimize investment strategies. Supervised learning models play a crucial role in evaluating financial risks and supporting corporate decision-making, reinforcing their value in performance management [3].

A fundamental step in improving model efficiency and accuracy is feature selection, which involves identifying the most relevant variables while eliminating redundant or less informative ones [4]. This process is particularly crucial in financial markets, where the sheer volume of data can overwhelm traditional analytical methods. By refining the input features, models achieve greater predictive power while reducing overfitting and computational complexity [28].

Although feature selection has been widely explored in global financial markets, its application in Brazil remains under-researched. Unique characteristics such as regulatory frameworks, market structure, and economic particularities influence the behavior of financial indicators, requiring tailored methodologies. Recent advancements in financial data collection and emerging technologies have opened new possibilities for more precise and detailed market analysis [5].

Leveraging extensive historical datasets with machine learning techniques can significantly improve the accuracy of economic forecasts in Brazil. However, the integration of feature selection and cross-validation for Brazilian stock market data remains a largely unexplored area, which presents an opportunity to improve financial predictive models [6].

This study addresses the gap in research by utilizing a dataset of Brazilian financial indicators to evaluate the effect of advanced feature selection methods on predictive model performance. By employing Information Gain and Relief, we pinpoint key features to enhance model efficiency. Additionally, we evaluate cross-validation's role in improving predictions, leading to a stronger approach for financial analysis in emerging markets.

These methods are known for their ability to handle large datasets and identify key features that contribute most significantly to the precision of the prediction, as discussed by [7], who emphasize the practical application of feature selection to improve model efficiency. Additionally, the importance of feature selection in high-dimensional data, demonstrating how correlation-based filters can efficiently identify the most relevant features, which can significantly enhance the generalization capacity of predictive models [8].

The models are then evaluated through cross-validation, allowing analysis of these methods. The contribution of this study lies in the introduction of a new perspective for feature selection analysis in emerging markets, using new financial indicators and cross-validation to ensure applicability of the results.

This document includes: Sect. 2 on Feature Selection Techniques and Technical Analysis, followed by Sect. 3 summarizing key previous research. Section 4 details the methodology, Sect. 5 presents the results, and Sect. 6 discusses findings and future work directions.

2 Background

In this section, our objective is to provide the essential background to comprehend the research. To begin, we will introduce the technical analysis indicators that are used as input for the models, followed by an exploration of the feature selection techniques employed.

2.1 Technical Analysis Indicators

Technical analysis indicators are essential tools utilized to scrutinize the price behavior of financial assets, including stocks, currencies, commodities, and others. Their primary objective is to forecast future market movements using graphical analysis by leveraging mathematical calculations applied to historical price and trading volume data of the assets [9].

Moving Average. The moving average is a statistical method employed to reduce the volatility in a time series of data [10], making it easier to identify trends and patterns by minimizing random fluctuations.

Standard Deviation. Standard deviation measures price volatility, with higher values indicating riskier assets [11]. This metric is crucial for constructing Bollinger Bands and identifying overbought and oversold conditions in the market.

Moving Average Convergence Divergence. The Moving Average Convergence Divergence (MACD) identifies changes in trend strength, direction, and momentum using two Exponential Moving Averages [12]. It is calculated by subtracting the 26-period EMA from the 12-period EMA. The resulting MACD line is complemented by the MACD Signal, which is a 9-period Exponential Moving Average of the MACD itself. This signal line serves as a trigger for buy or sell decisions: when the MACD crosses above the signal line, it indicates a potential buying opportunity; when it crosses below, it may indicate a selling opportunity. Its derivatives include the MACD Slope, which measures the rate of change, and the MACD Histogram, which visualizes the difference between the MACD and its signal line [13].

RSI. The Relative Strength Index (RSI), developed by J. Welles Wilder, evaluates whether a stock is overbought or oversold by analyzing recent price movements [14]. This momentum oscillator helps traders identify potential reversal points in the market.

Others. The Volatility and Speed Divergence Moving Averages (VSDME) indicators enhance traditional moving averages by incorporating both volatility and speed parameters. These adaptive indicators adjust their sensitivity to market conditions, providing more dynamic insights than conventional moving averages.

2.2 Feature Selection

Feature selection plays a fundamental role in machine learning by reducing dimensionality, enhancing model performance, and minimizing overfitting, especially when working with limited data [15]. The objective is to retain only the most relevant attributes, eliminating those that are redundant or irrelevant, which leads to more efficient and generalizable models [16].

Information Gain. Information Gain (IG) is based on entropy, a measure of uncertainty in a dataset. It evaluates how much a feature reduces entropy when the dataset is split based on its values. A higher IG implies that the feature contributes significantly to class discrimination, making it useful for building predictive models [17].

Symmetrical Uncertainty. Symmetrical Uncertainty (SU) is an information-theoretic metric that quantifies the mutual dependence between a feature and the class [18]. It compensates for IG's bias toward features with many values and is effective in identifying informative attributes for classification.

Minimum Redundancy Maximum Relevance (mRMR). The mRMR method aims to select features that are highly relevant to the target class while minimizing redundancy among them. It balances relevance and redundancy by discarding features that convey information already captured by previously selected attributes [19].

Relief. Relief estimates feature relevance by comparing each instance to its nearest neighbors from the same and opposite classes. It adjusts feature weights based on their ability to distinguish between these classes, making it suitable for detecting conditional dependencies and interactions [20].

OneR. OneR is a simple yet effective feature selection method that creates a one-rule classifier for each feature. It selects the feature whose rule results in the fewest classification errors, ranking attributes based on predictive accuracy [21].

Gain Ratio. Gain Ratio [22] adjusts Information Gain by penalizing features with many distinct values, which could otherwise skew the selection process. This normalization encourages the choice of features that are informative without overfitting due to excessive specificity [27].

Pearson Correlation. The Pearson Correlation coefficient measures the linear relationship between two variables. Ranging from -1 to +1, it indicates the strength and direction of the association, with values near ± 1 representing strong correlations [23].

Correlation-Based Feature Subset Selection (CfsSubset). CfsSubset evaluates groups of features rather than individual attributes. It favors subsets that are highly correlated with the class but have low intercorrelation, aiming to capture complementary information while avoiding redundancy [24].

3 Related Works

In stock market prediction, feature selection plays a crucial role by improving model accuracy, interpretability, and computational efficiency. Given the vast volume of financial indicators and the highly dynamic nature of market data, selecting the most informative features helps reduce noise and mitigate overfitting [28].

Park and Irwin [29] demonstrated the predictive value of technical indicators, such as Moving Averages and the Relative Strength Index (RSI), across global markets. These tools are essential for trend identification and momentum analysis, supporting trading decisions based on historical price patterns. Extensions of these indicators, like the MACD Slope, further enhance responsiveness to market changes and improve entry/exit signal quality [30]. Standard deviation remains a key metric for capturing volatility, directly impacting the robustness of predictive models.

Combining technical indicators with machine learning has been shown to boost predictive performance. However, due to the high dimensionality of financial datasets, feature selection becomes necessary to isolate relevant inputs. Htun et al. [28] emphasized the challenge of handling non-linear relationships and frequent fluctuations in stock prices, where traditional models may overfit. Feature selection techniques address these challenges by retaining only the most relevant indicators and removing redundancy.

In a comparative study, Krauss et al. [31] showed that ensemble methods, such as gradient-boosted trees and random forests, outperformed deep neural networks in stock prediction tasks, offering better profitability and risk-adjusted returns. These models are well-suited for handling complex patterns in noisy data and benefit significantly from optimized feature subsets.

Our study builds on these findings by applying and comparing multiple feature selection methods, highlighting their impact on model performance. Based on the results, we recommend the OneR method due to its ability to maintain performance with a minimal set of attributes, offering a balance between accuracy, efficiency, and interpretability.

4 Methodology

This study analyzes stock market data from the seven largest companies listed on B3 between 2010 and 2020, a period selected for its data richness and market relevance. The dataset includes daily trading session details, such as opening, closing, high, low, average prices, buy/sell offers, trading volume, and number of

transactions. These price attributes served as the basis for generating technical analysis indicators widely used in financial forecasting.

From this raw price data, we developed 194 technical indicators—including Moving Averages, Standard Deviations, RSI, MACD (and its derivatives), and the price variation (DF)—using sliding windows of length γ or β trading sessions. These indicators were used as features in machine learning models aimed at classifying stock behavior into five categories: *Stability*, *Descent*, *Sharp Descent*, *Rise*, and *Sharp Rise*, determined by the price variation over the following five trading days.

Due to the high dimensionality of the feature space, feature selection was applied to identify the most relevant indicators. We employed multiple feature selection methods (described in Sect. 2.2), followed by a stepwise feature elimination strategy. In this process, features were removed iteratively in reverse order of importance, and classification performance was measured at each step using k-fold cross-validation.

For each method, we plotted the cross-validated accuracy across the feature removal iterations. We identified two key points on each curve: the model with the maximum accuracy (marked with an **Orange Dot – OD**) and the point where performance began to consistently decline (marked with a **Green Dot – GD**), indicating that the most critical features were being removed beyond this point.

To validate the reliability of the models at these points, we performed *Stratified k-Fold Cross-Validation*—preserving class distributions within each fold—and tested the resulting models on an unseen test set. This ensured generalization and mitigated overfitting.

The label definitions were as follows:

- **Stability:** price variation between -0.5% and $+0.5\%$,
- **Descent:** decrease between 0.5% and 1%,
- **Sharp Descent:** decrease $\geq 1\%$,
- **Rise:** increase between 0.5% and 1%,
- **Sharp Rise:** increase $\geq 1\%$.

The overall methodology is summarized in Fig. 1.

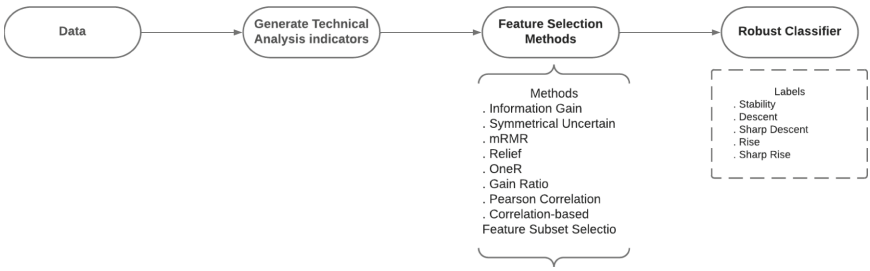


Fig. 1. Methodology flow diagram.

5 Result

This chapter we provide the results and insights of this research, it will be dissected in 2 subsections. The first subsection (reference section number) we delve into the BovDb [25], and explain in depth the tables it contains. The second subsection yields the findings from the cross-validation process and evaluation of the feature selection methods

5.1 Input Data

Brazilian stock data is accessible in raw text files on the B3 website. This study utilizes the BovDb dataset [25], which offers deeper insight into market behavior for B3 stocks from 2010 to 2020. This period features substantial data generation, enhancing the analysis's relevance. BovDb includes five tables that capture the market's dynamics.

Cardoso et al. [25] present a schematic illustration of the database that effectively highlights the interrelationships between the various tables and their respective columns.

The "price" table provided the data for constructing Technical Analysis Indicators for the Train and Test datasets. We processed the data to generate financial indicators based on different price metrics and various trading session windows, which were then used as features.

We derived 194 features from Technical Analysis Indicators and normalized the data. After reviewing the price table, we defined 5 labels:"Descent" for a 0.5% drop, "Sharp Descent" for a 1% drop, "Rise" for a 0.5% increase, "Sharp Rise" for a 1% increase, and "Stability" for fluctuations under 0.5%.

Using these features, we built the train and test datasets. The train dataset consists of 15,237 rows in total, with 2,756 rows for Stability, 3,314 for Descent, 2,612 for Sharp Descent, 3,472 for Rise, and 3,083 for Sharp Rise. The test dataset contains 3,810 rows, with 689 rows for Stability, 829 for Descent, 653 for Sharp Descent, 868 for Rise, and 771 for Sharp Rise.

5.2 Evaluation

The prediction utilized Random Forest [26], a machine learning technique that combines multiple decision trees to improve accuracy and reduce overfitting, making it ideal for classification and regression tasks. In this study, we built a model with 100 decision trees, allowing them to reach their maximum depth to capture complex data interactions. We used the entire dataset for each tree without limiting feature importance calculations and did not perform out-of-bag validation, focusing instead on other metrics to evaluate the model's effectiveness.

The model's performance was evaluated using accuracy and F1-Score metrics. To calculate these metrics, we applied a simple average of the results obtained from the 9 folds of the cross-validation process. The final accuracy was computed as the average of the accuracies from each k-fold cross-validation iteration.

In the first phase, we used 9-fold cross-validation, followed by stratified k-fold cross-validation to maintain class distribution in unbalanced datasets. We calculated performance metrics like accuracy and F1-Score, averaging them for overall model evaluation. This method helps ensure representative class distributions and reduces performance variation, yielding a more reliable model assessment.

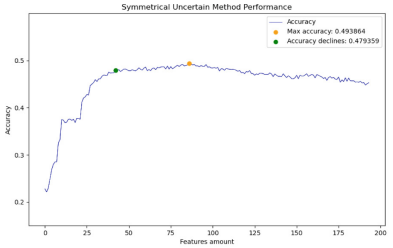
When all 194 available features were used, the model achieved an average F1-Score of 0.475 and an accuracy of 0.476 on the test data. Each graph includes an orange dot and a green dot. The orange dot marks the highest accuracy, while the green dot indicates a point which removing more features would significantly decrease the model's accuracy.

Figure 2 illustrates the accuracy patterns as features are progressively removed for each selection method. These graphs reveal critical insights into feature importance and method behavior. All methods demonstrate an initial steep accuracy increase with the first few features, followed by a plateau and eventual decline as essential features are removed. This pattern confirms the effectiveness of feature selection in identifying the most informative indicators. Notably, OneR (Fig. 2b) achieves near-optimal performance with remarkably few features, demonstrating its efficiency in identifying highly discriminative attributes. In contrast, mRMR (Fig. 2g) maintains higher accuracy with more features, suggesting its ability to capture complex feature interactions. Relief (Fig. 2f) shows a balanced profile with steady performance across a moderate feature set. The orange dots mark peak accuracy points, while green dots indicate where performance begins to decline significantly, helping identify optimal feature subsets for each method. These visual patterns complement the numerical results in Table 1 and 2, also it provides intuitive evidence of each method's theoretical strengths.

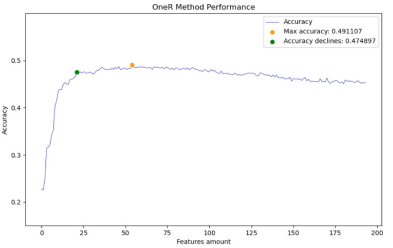
In Fig. 3 we can see a comparison of how all the Feature Selection Methods works. It is noticeable that there is a pattern in the variation of accuracy throughout the features insertion, initially the accuracy increases to a certain point, where it is the maximum accuracy, showing that these initial features removed are getting in the way of the model. Then the accuracy starts to decrease slowly until reaches GD, where from there, it starts to decrease drastically. The OneR method stands out for being able to delay the GD, managing to maintain accuracy at a level similar to the initial one using only 21 features.

To further analyze the feature selection methods, we constructed a comparative bar chart that evaluates the performance of the methods at the two critical points of each accuracy line: the OD (represented by orange bars) and the GD (represented by green bars).

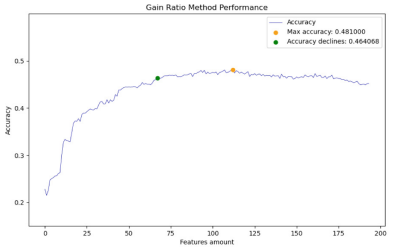
Tables 1 and 2 present the performance results obtained at the points of greatest accuracy (OD) and at the initial point where the performance drops significantly (GD) for each feature selection method evaluated. The number of features used, the accuracy averages obtained by cross-validation (simple and stratified), as well as the accuracy and F1-score values in the test using the model generated from the Stratified Cross-Validation, are displayed, allowing the comparison of the robustness and effectiveness limit of each method.



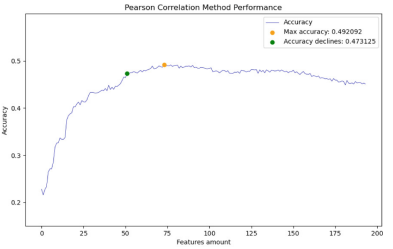
(a) Symmetrical Uncertain



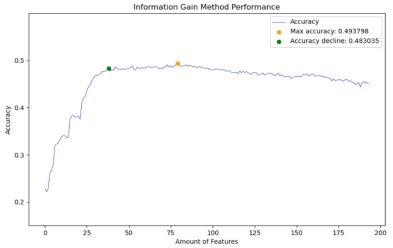
(b) OneR



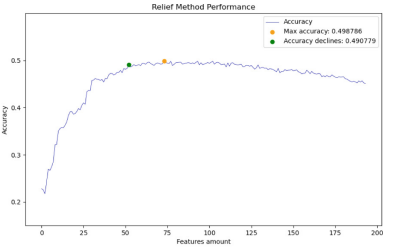
(c) Gain Ratio



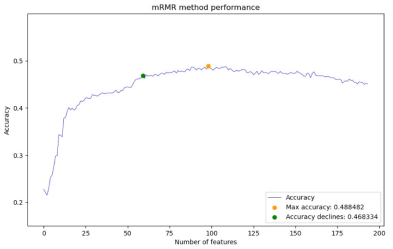
(d) Pearson Correlation



(e) Information Gain



(f) Relief



(g) mRMR

Fig. 2. Feature removal experiments

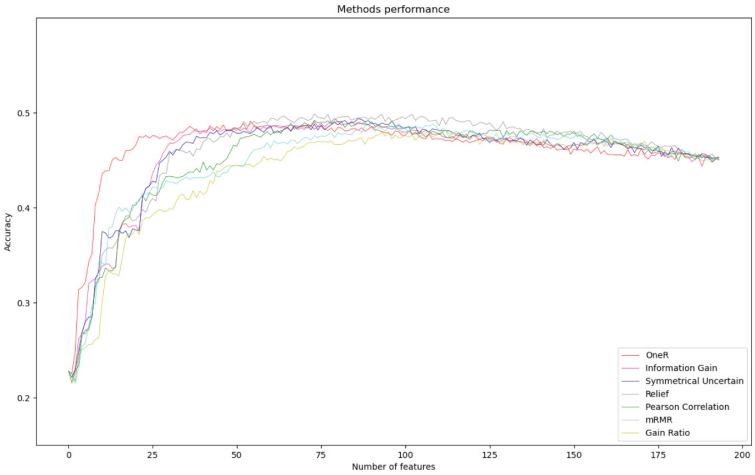


Fig. 3. All Feature Selections performance

Table 1. Performance at the points of greatest accuracy (OD) for each feature selection method.

Method	Feat.	C-Validat	Strat C-Validat	Acc. Test	F1-Score Test
OneR	54	0.491	0.484	0.497	0.498
Gain Ratio	112	0.481	0.474	0.494	0.495
Pearson Correlation	73	0.492	0.490	0.512	0.513
Information Gain	79	0.493	0.494	0.512	0.512
Relief	73	0.498	0.493	0.515	0.516
mRMR	98	0.488	0.487	0.519	0.520

Table 2. Performance at starting decay (GD) points for each feature selection method.

Method	Feat.	C-Validat	Strat C-Validat	Acc. Test	F1-Score Test
OneR	21	0.474	0.473	0.477	0.479
Gain Ratio	67	0.464	0.468	0.486	0.487
Pearson Correlation	51	0.473	0.471	0.491	0.492
Information Gain	38	0.483	0.480	0.500	0.501
Relief	52	0.490	0.488	0.492	0.493
mRMR	59	0.468	0.469	0.482	0.483

The Fig. 4 displays four bars for each method: two corresponding to the OD, representing the average accuracy from stratified cross-validation and the test set result, and two for the GD, showing the same metrics. Additionally, above

each bar representing the average accuracy from stratified cross-validation, a black line illustrates two standard deviations, both above and below, indicating the variability of accuracy across the cross-validation folds.

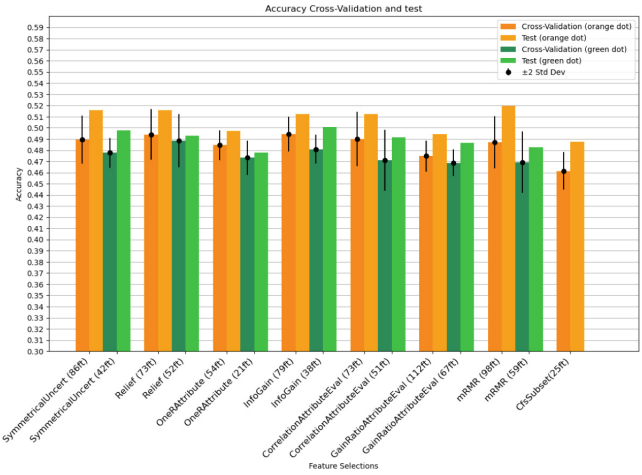


Fig. 4. Standard Deviation Graph Features Selections

Analyzing the graph, it is evident that all methods demonstrated generalization capability. The higher accuracy observed in the test set compared to the training set indicates that the combination of cross-validation and feature selection contributes positively to generalization. Notably, some methods achieved test set accuracy exceeding two standard deviations above the mean of Stratified Cross-Validation, highlighting their effectiveness in classifying unseen data.

The mRMR method stood out by achieving an accuracy of 0.520 with 98 features, making it the highest-performing method on the test data while using less than half of the original feature set. Another noteworthy result came from the OneR method, which, despite using only 21 features—approximately 10% of the initial set—achieved an accuracy of 0.477, slightly improving upon the initial accuracy. Overall, all methods demonstrated superior accuracy on the test data compared to the initial model.

The challenge of obtaining higher accuracy, around 50%, stems not only from the inherent complexity of the financial market but also from the nature of the multiclass classification task. The dataset was designed to predict five different price movement categories, which increases the classification difficulty. Additionally, financial markets are influenced by numerous external factors, and price variations tend to be subtle. This makes it particularly difficult to differentiate between closely related classes, such as Stability and minor price fluctuations.

Table 3 presents the top 10 most relevant features according to all feature selection methods, based on their average ranking across the methods, except for CfsSubset.

Table 3. Top 10 features

Rank	Feature	Average Position
1	sd_90_offer_sell	11.142857
2	sd_90_minimum	11.142857
3	sd_90_offer_buy	11.4285771
4	sd_90_average	12.142857
5	sd_90_closing	12.428571
6	sd_90_opening	12.857142
7	sd_90_maximum	15.714285
8	sd_60_offer_sell	21.142857
9	sd_60_opening	21.142857
10	sd_60_offer_buy	23.142857

It can be seen that the standard deviation in the 90 and 60-day trading sessions stands out as the most relevant, with the variables related to 90-day periods occupying the first seven positions. This indicates that the most relevant financial indicator, according to the table, is the Standard Deviation. Moreover, price volatility over longer periods plays a crucial role in predictive modeling and market behavior.

6 Conclusion

This study investigated how different feature selection techniques influence predictive performance in the Brazilian financial market using Random Forest models. Our analysis of 194 technical analysis indicators through systematic feature reduction revealed that the mRMR method delivered the best performance, achieving an accuracy of 0.520 with only 98 features—less than half of the original feature set.

The superior performance of mRMR can be attributed to its theoretical foundation, which optimizes both relevance and non-redundancy simultaneously. Unlike methods that evaluate features individually, mRMR considers feature interactions, selecting attributes that collectively provide maximum information about the target variable while minimizing redundancy among selected features. This approach is particularly effective in financial markets where indicators often exhibit complex interdependencies.

Relief also performed exceptionally well (0.515 accuracy with 73 features), likely due to its ability to detect local dependencies between features that might be missed by correlation-based methods. This capability is valuable in financial data where relationships between indicators may vary across different market conditions.

For applications where computational efficiency is critical, we recommend the OneR method, which achieved 0.477 accuracy with just 21 features (approx-

imately 10% of the initial set). OneR's effectiveness stems from its ability to identify simple yet powerful discriminative rules, making it ideal for real-time financial analysis where processing speed is essential.

Our analysis consistently identified standard deviation indicators, particularly those calculated over 90-day periods, as the most relevant features across all selection methods. This finding suggests that long-term volatility metrics capture fundamental market behaviors that are highly predictive of future price movements.

Based on our comprehensive analysis, we recommend:

- **For maximum accuracy:** the mRMR method, which balances feature relevance and non-redundancy effectively;
- **For computational efficiency:** the OneR method, which provides substantial dimensionality reduction with minimal performance loss;
- **For balanced performance:** the Relief method, which offers strong accuracy with moderate feature reduction.

These recommendations are particularly relevant for financial analysts and algorithmic traders developing predictive models for emerging markets like Brazil, where data complexity and computational constraints often present significant challenges.

Future research should focus on applying these selected features in more sophisticated machine learning approaches, such as deep learning models, and investigating how feature selection performance varies across different market conditions and economic cycles.

Acknowledgments. The authors would like to thank FAPERGS (24/2551-0001396-2, 23/2551-0000773-8), CNPq (305805/2021-5) and FAPERGS/CNPq (23/ 2551-0000126-8). Fabian thanks to Fesurv-UniRV for the pay leave, which helped to collaborate in this work.

References

1. Chen, C.P., Metghalchi, M.: Weak-form market efficiency: evidence from the Brazilian stock market. Int. J. Econ. Financ. (2012)
2. Bouri, E., Demirer, R., Gupta, R., Sun, X.: The predictability of stock market volatility in emerging economies: relative roles of local, regional, and global business cycles. J. Forecast. (2020)
3. Cuervo, R.: Predictive AI for SME and large enterprise financial performance management. arXiv preprint arXiv:2311.05840 (2023)
4. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. (2014)
5. Kohn, K., Moraes, C.H.: O impacto das novas tecnologias na sociedade: conceitos e características da Sociedade da Informação e da Sociedade Digital. XXX Congresso Brasileiro de Ciências da Comunicação (2007)

6. Araujo, G. S., Gaglianone, W. P.: Machine learning methods for inflation forecasting in Brazil: new contenders versus classical models. Lat. Am. J. Cent. Bank. (2023)
7. Htun, H.H., Biehl, M., Petkov, N.: Survey of feature selection and extraction techniques for stock market prediction. Financ. Innov. (2023)
8. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Machine Learning Proceedings 1992, pp. 249–256. Elsevier (1992)
9. Chuang, L.Y., Ke, C.H., Yang, C.H.: A hybrid both filter and wrapper feature selection method for microarray classification. arXiv preprint arXiv:1612.08669 (2003)
10. Shi, Y., Li, B., Long, W., Dai, W.: Method for improving the performance of technical analysis indicators by neural network models. Comput. Econ. (2022)
11. Billah, M.M., Sultana, A., Bhuiyan, F., Kaosar, M.G.: Stock price prediction: comparison of different moving average techniques using deep learning model. Neural Comput. Appl. **36**(11), 5861–5871 (2024)
12. Altman, D.G., Bland, J.M.: Standard deviations and standard errors. BMJ **331**(7521), 903 (2005)
13. Halilbegovic, S.: MACD-Analysis of weaknesses of the most powerful technical analysis tool. Independent J. Manag. Prod. **7**(2), 367–379 (2016)
14. Kang, B.K.: Improving MACD technical analysis by optimizing parameters and modifying trading rules: evidence from the Japanese Nikkei 225 futures market. J. Risk Financ. Manag. (2021)
15. Bansal, S.: Investigating the efficacy of RSI IN the nifty 50 index. Global J. Bus. Integr. Secur. (2016)
16. Janecek, A., Gansterer, W., Demel, M., Ecker, G.: On the relationship between feature selection and classification accuracy. In: New Challenges for Feature Selection in data Mining and Knowledge Discovery. PMLR (2008)
17. Janecek, A.G.K., Gansterer, W.N., Demel, M.A., Ecker, G.F.: On the relationship between feature selection and classification accuracy. In: Proceedings of the 2008 International Conference on New Challenges for Feature Selection in Data Mining and Knowledge Discovery, FSDM 2008, vol. 4, pp. 90–105. JMLR.org, Antwerp (2008)
18. Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning (ICML-03) (2016)
19. Lin, X., Li, C., Ren, W., Luo, X., Qi, Y.: A new feature selection method based on symmetrical uncertainty and interaction gain. Comput. Biol. Chem. **83**, 107149 (2019)
20. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
21. Urbanowicz, R.J., Olson, R.S., Schmitt, P., Meeker, M., Moore, J.H.: Benchmarking relief-based feature selection methods for bioinformatics data mining. J. Biomed. Inform. (2018)
22. Kumar, K., Kumar, G., Kumar, Y.: Feature selection approach for intrusion detection system. Int. J. Adv. Trends Comput. Sci. Eng. (IJATCSE) **2**(5), 47–53 (2013)
23. Leroux, A., Boussard, M., Dés, R.: Information gain ratio correction: improving prediction with more balanced decision tree splits. arXiv preprint arXiv:1801.08310 (2018)

24. Armya, R., Abdulazeez, A.M., Sallow, A.B., Zeebaree, D.Q.: Leukemia diagnosis using machine learning classifiers based on correlation attribute eval feature selection. Asian J. Res. Comput. Sci. **9**(3), 52–65 (2021)
25. Hall, M.A.: Correlation-based feature selection for machine learning. Ph.D. dissertation, The University of Waikato (1999)
26. Cardoso, F.C., et al.: BovDB: a data set of stock prices of all companies in B3 from 1995 to 2020. J. Inf. Data Manag. **13**(1) (2022)
27. Breiman, L.: Random forests. In: Machine Learning. Springer (2001)
28. Karegowda, A.G., Manjunath, A.S., Jayaram, M.A.: Comparative study of attribute selection using gain ratio and correlation based feature selection. Int. J. Inf. Technol. Knowl. Manage. **2**(2), 271–277 (2010)
29. Park, C.H., Irwin, S.H.: What do we know about the profitability of technical analysis? J. Econ. Surv. **21**(4), 786–826 (2007)
30. Pardeshi, Y.K., Kale, P.: Technical analysis indicators in stock market using machine learning: a comparative analysis. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6. IEEE (2021)
31. Krauss, C., Do, X.A., Huck, N.: Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. Eur. J. Oper. Res. **259**(2), 689–702 (2017)