









Feature Selection for Stock Market Prediction: A Comparison of Relief and Information Gain Methods

Humberto O. Bragança¹^a, Rafael A. Berri¹^b, Bruno L. Dalmazo¹^c, Eduardo N. Borges¹^d,
Viviane L. D. de Mattos¹^e, Richard F. Pinto¹^f, Fabian C. Cardoso²^g and Giancarlo Lucca³^h

¹Federal University of Rio Grande (FURG), Rio Grande, Brazil

²University of Rio Verde (UniRV), Rio Verde, Brazil

³Catholic University of Pelotas (UCPel), Pelotas, Brazil

humberto.obj@gmail.com, {dalmazo, rafaelberri, eduardoborges, vivianemattos, richard.pinto}@furg.br,

Keywords: Machine Learning, Feature Selection, Stocks, Technical Analysis, Financial Market.

Abstract: This study explores an approach to predictive analysis in the financial market, using a data set composed of financial information from different companies listed on the stock market, which provides a more detailed and contextualized view of the behavior of shares. Based on these indicators, feature selection methods, such as Relief and Information Gain, are applied to identify the most relevant variables for building predictive models. One of the main contributions of this work is the use of cross-validation to evaluate attribute selection, a technique that has not yet been explored in this context with this dataset. The results show that the combination of new financial indicators and cross-validation offers a solid basis for more accurate analysis, with important implications for investors, financial analysts and policymakers in the stock market. This work expands the boundaries of the literature on feature selection and opens possibilities for future research in emerging markets.

1 INTRODUCTION

The Brazilian financial market, B3¹, is a large and dynamic emerging market with unique characteristics that require adapted analytical and predictive approaches (Chen and Metghalchi, 2012). Its complexity, driven by diverse economic sectors and volatility, presents challenges and opportunities for financial analysis (Bouri et al., 2020).


In recent years, predictive analysis using machine learning has proven effective for financial decision-making. Supervised learning models help assess risks and make informed decisions, emphasizing their importance in managing corporate financial performance (Cuervo, 2023).


Feature selection, a key step in improving model efficiency and generalization, involves identifying the most relevant variables to enhance prediction accuracy and streamline the learning process (Chandrashekar and Sahin, 2014). This process is particularly important in financial markets, where the volume of data can overwhelm traditional methods, improving model performance and reducing overfitting (Htun et al., 2023).


Despite the importance of feature selection in global markets, there is limited research on its application in the Brazilian context. The country's economic and financial specificities, such as its regulatory environment and market structure, impact the behavior of financial indicators. Emerging technologies and new indicators derived from detailed data can improve financial analysis and provide deeper insights into the Brazilian market (Kohn and Moraes, 2007).


The application of machine learning methods, coupled with extensive datasets, can significantly enhance the accuracy and reliability of economic forecasts in Brazil, underscoring the importance of tailored approaches for financial analysis in emerging markets (Araujo and Gaglianone, 2023). The com-


^a <https://orcid.org/0009-0006-6610-500X>


^b <https://orcid.org/0000-0002-3812-4186>


^c <https://orcid.org/0000-0002-6996-7602>

^d <https://orcid.org/0000-0003-1595-7676>

^e <https://orcid.org/0000-0002-3512-6290>

^f <https://orcid.org/0009-0007-0176-3383>

^g <https://orcid.org/0000-0002-2842-0387>

^h <https://orcid.org/0000-0002-3776-0260>

¹<https://www.b3.com.br/>

bination of cross-validation and feature selection in Brazilian stock market data is still underexplored, highlighting a research opportunity to enhance predictive models and forecasting precision.

To address this gap, this study proposes the utilization of an dataset with financial indicators specific to Brazil, sourced from decades of detailed financial data. This dataset is used to apply advanced feature selection techniques and evaluate the predictive performance of models using cross-validation, a technique little explored in the national context.

The main objective of this work is to evaluate how different feature selection methods, applied to this set of financial data, can improve the performance of predictive models in the Brazilian financial market. Specifically, Information Gain and Relief methods are used to choose the key features.

This document is organized as follows: Section 2 covers Feature Selection Techniques and Technical Analysis Indicators, in the Literature Review. Section 3 summarizes key prior studies. Our methodology is detailed in Section 4, while Section 5 presents the study's findings. Finally, Section 6 discusses the results and future research opportunities.

2 BACKGROUND

This section aims to give insight into key concepts for the article, starting with the technical analysis indicators used in the models, followed by the feature selection methods.

2.1 Technical Analysis Indicator

Technical analysis indicators are vital instruments used to examine the price trends of various financial assets, such as stocks, currencies, and commodities. Their main goal is to predict future market movements through graphical analysis, employing mathematical formulas based on historical price and trading volume data of the assets (Shi et al., 2022).

2.1.1 Moving Average

Moving average is a statistical technique used to smooth the volatility of a time series of data (Billah et al., 2024), facilitating the identification of patterns and trends by reducing random variation. It computes the average of a set of values within a sliding window over time, offering a clearer insight into the underlying movements within the time series.

2.1.2 Standard Deviation

The standard deviation is a key metric in stock technical analysis, measuring price volatility by calculating the variability of closing prices around their moving average (Altman and Bland, 2005). It is derived from the variance, which averages the squared differences between prices and the mean, with its square root yielding the standard deviation. This measure is essential for constructing Bollinger Bands, identifying overbought and oversold levels, and assessing asset risk—where higher values indicate greater volatility and risk, while lower values suggest stability.

2.1.3 MACD

The Moving Average Convergence Divergence (MACD) is a key technical analysis tool used to identify changes in an asset's trend strength, direction, momentum, and duration. By leveraging historical data, it helps forecast price movements in financial markets. The MACD is computed using two exponential moving averages (EMAs) (Halilbegovic, 2016), which assign greater weight to recent data. Typically, these EMAs are based on 26-period and 12-period time frames. Additionally, a signal line, which is a nine-period EMA of the MACD line, is included in the dataset as a feature.

Several indicators stem from the MACD, including the MACD Slope and MACD Histogram. The MACD Slope measures the rate of change of the MACD over time, representing its angular coefficient. A rising MACD Slope suggests a strengthening uptrend, whereas a falling slope indicates a downtrend. It is computed by measuring the variance between MACD values at different time points. The MACD Histogram, another derivative indicator, represents the difference between the MACD and the signal line ($\text{MACD} - \text{Signal}$) (Kang, 2021), visually depicting momentum shifts and trend changes.

2.1.4 Relative Strenght Index (RSI)

The RSI, created by J. Welles Wilder in 1978, gauges whether a stock is overbought or oversold by analyzing recent closing prices.

It is considered as an oscillator, ranging from 0 to 100. commonly applied to identify swing points, where it is an overbought or oversold conditions of an asset, helping to predict potential trend reversals. This indicator can effectively predict market movements by identifying overbought or oversold conditions, further supporting its practical application in financial markets (Bansal, 2016).

It is also used the indicators VSDME12 and VS-

DME26, which are a variation of the moving average, it is an adaptive moving average, which incorporates volatility and speed in the calculation. The VSDME (which stands for Volatility and Speed Divergence Moving Average), utilizes α equal to 12 and τ of 26 for VSDME12, and for VSDME26 utilize α of 26 and τ equal to 52.

$$\text{VSDME} = \text{VSDME}_{\alpha} - \text{VSDME}_{\tau} \quad (1)$$

2.2 Feature Selection Methods

Feature selection is vital in model development. While more features can improve performance, too many, especially with limited training data can hinder learning and cause overfitting. The goal is to retain only essential attributes, remove redundancies, and improve model efficiency (Janecek et al., 2008).

Feature selection is vital in model development. While more features can improve performance, too many—especially with limited training data—can hinder learning and cause overfitting. The goal is to retain only essential attributes, remove redundancies, and improve model efficiency.

2.2.1 Information Gain

Information Gain is a metric used to measure the reduction in uncertainty or entropy in a set of data when a characteristic (or attribute) is chosen to divide the data. It is often used in machine learning algorithms, such as decision trees, to determine which feature should be used at each node. The central idea is that dividing the data based on a characteristic should result in purer subsets, that is, with less unpredictability.

Information Gain is calculated based on, the difference between the original entropy (before splitting) and the sum of the entropies of the subsets generated after splitting. The greater the Information Gain of a feature, the more relevant it is to predict the target and, therefore, the more useful it is in building the predictive model. The effectiveness of Information Gain in selecting relevant features in high-dimensional contexts, such as microarray data, demonstrates its applicability in different domains (Yu and Liu, 2016).

2.2.2 Relief method

The Relief an individual valuation filter method (Urbanowicz et al., 2018), that evaluates the relevance of attributes based on the proximity of instances of different classes. For each instance in the dataset, the algorithm identifies the closest instance of the same class (near neighbor) and the closest instance of a different class (far neighbor).

It then adjusts the attribute weights based on how those attributes help differentiate instances of different classes. Attributes that help distinguish between classes receive greater weight, while those that do not make a difference have reduced weight. This method is useful in problems with complex, high-dimensional data, as it selects the most informative features for the learning model.

3 RELATED WORK

With the huge amount of data generated by the financial market, more predictions are being made by Machine Learning algorithms (Jain and Vanzara, 2023). A notable example is the application of deep learning techniques, particularly Long Short-Term Memory (LSTM) networks, to the S&P 500 dataset for predicting stock price movements based on historical data (Kamalov et al., 2020). The study emphasizes the importance of daily closing values and trading volumes, analyzing data from 1990 to 2020.

The proposed model outperformed several benchmark models in predicting the directional movements of the index. For example, one study applied Support Vector Regression (SVR) to predict stock prices, focusing on preprocessing the NASDAQ (National Association of Securities Dealers Automated Quotations) dataset (Dash et al., 2023).

Technical analysis indicators like MACD, ADX, Williams, and MFI were converted into correlation tensors for enhanced processing in deep learning models, including LSTM and DNN networks. This method improved stock price predictions and buy/sell signal detection (Kamalov et al., 2019).

A recent publication in the academic literature introduces the BovDB as a benchmark dataset for research in stock market prediction (Cardoso et al., 2022). This dataset, which is publicly accessible and pre-processed, encompasses daily stock data for all companies listed on B3 from 1995 to 2020. Notably, the authors have introduced a novel metric referred to as the “factor” aimed at mitigating the influence of significant events within the dataset. Utilizing both the factor and the BovDB allows for a comprehensive analysis of the historical time series of Brazilian stock prices, tracing back to the inception of Brazil’s Real monetary plan.

This article presents an innovative approach by integrating new financial indicators, developed from an unprecedented dataset composed of detailed financial information from Brazilian companies over several decades. Unlike conventional indicators, these new indicators capture nuances of local financial behavior,

providing a more in-depth and relevant view for predictive analysis in the national context. The creation of this dataset not only fills a critical data gap, but also establishes a solid foundation for future research, allowing for more robust and contextualized analyses.

Furthermore, the use of cross-validation as part of the methodology for feature selection is an innovative approach in the context of the Brazilian stock exchange. Although cross-validation is a technique widely used in machine learning and feature selection studies, its specific application in the selection of financial attributes for analyzing the Brazilian market is still rare.

By employing this technique, we ensure that the results obtained are not only specific to the dataset used, but also generalizable, increasing the reliability and practical applicability of the conclusions. This rigorous approach raises the methodological standard of research in emerging markets, encouraging the adoption of more robust and replicable practices. Theoretically, the work enriches the feature selection literature by introducing a new perspective based on financial indicators specific to the Brazilian market, while, in practice, it offers valuable insights for investors, financial analysts and policymakers.

4 METHODOLOGY

This section presents the methodology for evaluating predictive models in the Brazilian financial market. The study constructs a dataset with daily trading data from B3 (Brazilian Stock Exchange) covering 1995 to 2020, with a focus on 2010-2020. This dataset, structured with price data, dates, and stock identifiers, enables market trend analysis and serves as the foundation for generating technical analysis indicators.

Feature selection methods, including Relief and Information Gain, are applied to identify the most relevant attributes. Sequential techniques such as Sequential Forward and Backward Selection refine the feature set further (Aha and Bankert, 1995). Features are eliminated iteratively, prioritizing model accuracy. Cross-Validation ensures robust performance evaluation by dividing data into k subsets, reducing bias and improving generalization.

Stratified K-Fold Cross-Validation is used at key feature selection points, preserving class distributions. Performance metrics such as accuracy and F1-score are averaged across folds. Visualization tools highlight critical feature contributions, and models undergo final training on the entire dataset before deployment. This approach minimizes overfitting and enhances predictive reliability for the Brazilian finan-

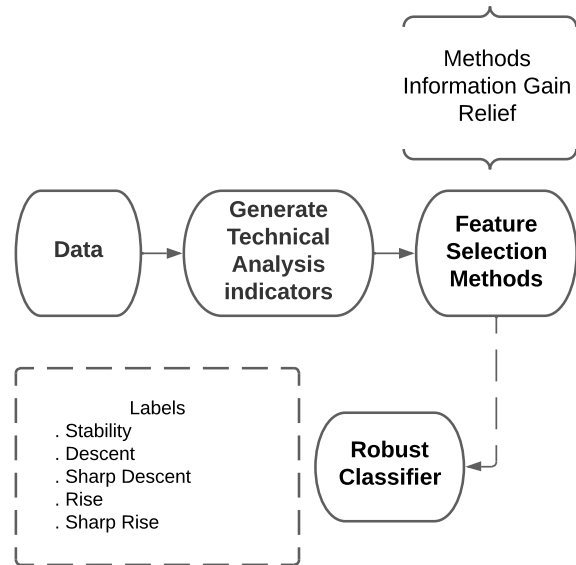


Figure 1: Diagram methodology.

cial market.

5 RESULTS

In this chapter, we present the findings and insights obtained from the research, organizing the discussion into two subsections. The first subsection 5.1 focuses on the BovDb (Cardoso et al., 2022) and (Souza et al., 2024), offering a detailed examination of the included tables and how we managed it. The second subsection outlines the results from the cross-validation process and evaluates the feature selection methods employed.

5.1 Input Data

The data of Brazilian Stocks are available to the public in text files format, organized in a raw form. The raw data is available in B3's website. This study utilizes data collected from BovDb (Cardoso et al., 2022)², which is a preprocessed dataset, from the shares in the B3, it allows a better understanding of the market and its behave. It contains data of daily exchange of all shares in B3 from 1995 to 2020, but we focused on the 7 most representative shares on the Brazilian stock market and considered only the period of 2010 to 2020. During this shorter period of time, the companies generated an ample amount of data, ensuring the relevance of the analysis. BovDb comprise five distincts tables, providing a deep view of the market landscape.

²<https://sol.sbc.org.br/index.php/dsw/article/view/17411>

First table is the Company table, this table correlates the name and identification for every company that has had a presence in B3 between the years of 1995 and 2020. It encompasses a total of 1728 companies within this database. The column "id_company" is the auto-incremented integer, serves as the unique identifier for the company, functioning as the primary key. Additionally, it is utilized as a foreign key on the Ticker to reference the aforementioned company. And the other column is the "Company" column, referring to the company's name.

The Ticker table stores the data of the stocks. It relates the code of the stock for each company, the codes are formed by a pattern of numbers and letters that helps the investor to identify each company and the type of share that corresponds with it, the table contains 2540 stocks in it. The difference between the amount of companies and stocks, is due to the fact that a single company can have more than one type of share. The first column is the "id.ticker", which is an auto incremented integer serving as the Ticker identifier, acting as a primary key. It is also utilized as a foreign key in both the EventPrice and Price tables to reference the former. The other column is the "ticker", being the company's stock symbol. The "codisi" column is the stock code in B3.

The "Price" table stores the data negotiation of the trading floor for each stock, providing us with enough information to understand the movement of the stock throughout the trading floor. The "date" column is the date of trade for a stock, serves as a crucial component in conjunction with the id_ticker, collectively forming a composite primary key for identifying a specific ticket on any given date. Within the context of EventPrice, the date, along with the id_ticker and id_event, forms a composite primary key signifying the occurrence date of a particular event.

Each one is a column of a given date, "open" represents the opening price, 'high' represents the highest price, 'low' represents the lowest price, 'average' represents the average price, 'close' represents the closing price, 'buy_offer' represents the best offering price, 'business' represents the quantity of transactions executed with the stock, 'sell_offer' represents the best selling price, 'amount_stock' represents the aggregate trading volume on the stock. The last column is the "Factor" which is the combined effect of events is considered from the most recent to the oldest until a particular date is attained, showcasing the chronological progression.

The "Event" table presents different types of events, containing 12 occurrences. The "id_event" is the auto-incremented integer serves as the unique identifier for the Event, making it the primary key.

Additionally, it functions as a foreign key on the EventPrice to reference the aforementioned Event. The "description" column is a description of the event. And the "ds_bovespa" is the abbreviated Event designation, as indicated in the documentation supplied by B3.

The last table is the "Eventprice" table, showing that over time a stock can undergo different events, this table presents the trading floor days that happened an event. It also presents if factor was applied in a stock, and its value. For example, stock split, in which the number of shares increases to provide greater liquidity without affecting the total value of the company's capital, factor is applied in this case, so it is possible to perform a better analysis of the stock over time. The "factor" column is how significant is the event on a specific stock and trade. The "applied" column represents if an event has occurred or not in a specific day.

The "Price" table was used to build the Technical Analysis Indicators, that serves as features for the Train and Test dataset. For the Moving Average and Standard Deviation calculations, we conducted a thorough analysis of various stock market trading sessions, each lasting window size of γ length. We examined the opening, maximum, minimum, average, closing, offer/buy, offer/sell, volume, and business amount data. To determine the values for MACD, MACD Histogram, MACD Slope, MACD DF, MACD VSDME12, and MACD VSDME26, we captured the opening, closing, maximum, average, and minimum. Additionally, for the MACD Signal, we utilized the prices from the other MACD indicators, considering 9 stock market trading windows. Similarly, for RSI, we performed analogous calculations using β stock market trading sessions for each data point.

The length of the stock market trading sessions are β length = 7, 14 and 21 γ length = 5, 10, 15, 20, 25, 30, 60, and 90 (representing the previous prices for the ongoing analysis).

This resulted in 194 features of Technical Analysis Indicators, being necessary to normalize it because the data was not in the same range. After analyzing the price table and considering the percentage of gain or loss in the stock market trading sessions ahead, we have identified 5 distinct labels. Descent indicates a 0.5% decrease in the stock value, while Sharp Descent signifies a 1% decrease. On the other hand, Rise denotes a 0.5% increase, and Sharp Rise indicates a 1% increase. Lastly, Stability represents a negligible fluctuation in the stock value, either up or down, by less than 0.5%.

With these data, we build the train and test dataset,

the train dataset is composed of 15237 rows in total, being 2756 rows of Stability, 3314 of Descent, 2612 of Sharp Descent, 3472 of Rise and 3083 of Sharp Rise. The test dataset is composed of 3810 rows, being 689 rows of Stability, 829 of Descent, 653 of Sharp Descent, 868 of Rise, and 771 of Sharp Rise.

5.2 Evaluation

The prediction was performed using Random Forest, a machine learning technique (Breiman, 2001). The algorithm combines multiple decision trees to improve accuracy and reduce the risk of overfitting, making it ideal for classification and regression tasks. In this study, we use Random Forest to evaluate the performance of the model based on selected attribute data. The model was configured to generate 100 decision trees, without depth restriction, allowing the trees to grow to their maximum height to capture complex interactions in the data. Data sampling was performed with 100% of the dataset in each tree, ensuring a complete view during the construction of each tree. We did not calculate the importance of attributes and all characteristics were used in the trees without restriction. Additionally, out-of-bag validation has been disabled, with a focus on other evaluation metrics to ensure model robustness.

The model performance evaluation was carried out using accuracy and F1-Score metrics. To calculate these metrics, a simple average of the results obtained in the 9 folds of the cross-validation process was applied. The final accuracy was calculated as the simple average of the accuracies of each of the k-fold cross-validation iterations.

Specifically, in the first set of experiments, the k-fold cross-validation technique with 9 folds was used. Then, in the second part of the experiment, we applied stratified k-fold cross-validation, ensuring that the distribution of classes was maintained in each of the 9 folds, which is particularly important in unbalanced data sets. During this process, performance metrics, such as accuracy and F1-Score, were calculated and the simple average of these metrics was used to evaluate the overall performance of the model. This ensures that the model is trained and tested on representative distributions of classes across all folds, avoiding any variation that may occur randomly in balanced datasets. Furthermore, the use of stratification helps to reduce variation in performance metrics, such as accuracy and F1-score, providing a more consistent evaluation of the model.

Figure 2 and 3 shows the performance of the models as features are removed according to the Information Gain and Relief method, respectively. It is worth

highlighting that, it was employed Random Forest as the classifier, in which, using all available features (194 total) achieved an average F1-Score of 0.475 and an accuracy of 0.476 in the test data. Each graph illustration presents an orange dot and a green dot, which means, the highest accuracy and a limit indicating that the removal of features from that point onwards drastically reduces the accuracy of the models. Showing the importance of the remaining features.

Our first analysis addresses the Information Gain method. Initially the accuracy increases as the features are being removed, until reaching its peak, and then declining. The order in which the features were eliminated corresponds to the reverse sequence obtained from the Information Gain feature selection approach. The accuracy results shown in the graphs are derived from Cross-Validation conducted without stratification.

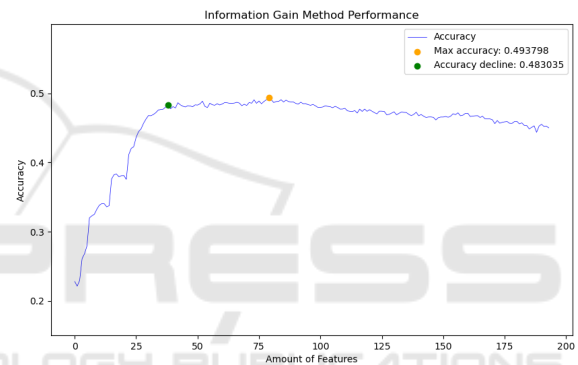


Figure 2: Information Gain features removal.

In this analysis, the OD indicates that the feature count stands at 79. Initially, the model trained achieved an accuracy of 0.493, in the GD, where the feature count was 38, the model delivered an accuracy of 0.483.

In sequence, the same approach is adopted with the Rellief method. The accuracy increases as the features are being removed, the OD and GD are closer to each other, in comparison to the Information Gain method.

In our analysis, the OD on the chart includes 73 distinct features. We initially employed Cross-Validation, which yielded an accuracy of 0.498, then we applied this methodology to the GD, which consists of 52 features, the model's first accuracy measurement was 0.490.

Note that the Relief method initially performs better, achieving higher heals as it removes the initial features. However, after obtaining these initial accuracies superior to the Information Gain method, the Relief method models were unable to maintain them over time, causing the curve of accuracies to begin

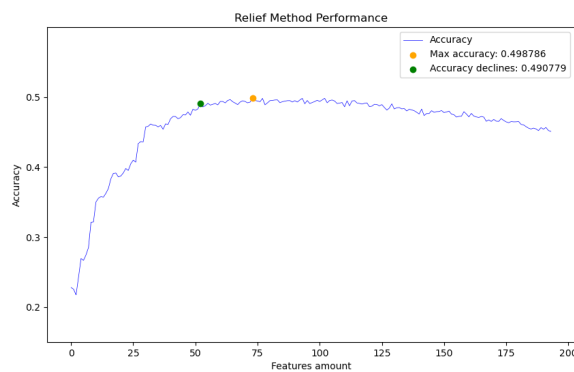


Figure 3: Relief method features removal.

earlier. This can be better observed when we compare the two GDs, where the Relief method with 52 features achieved an accuracy of 0.490 and the Information Gain method with 14 fewer features achieved an accuracy of 0.483.

The Table 1 presents the top 10 most relevant features according to Information Gain and the Relief method.

Table 1: Features selected by Information Gain and relief.

Information Gain		Relief	
Rank	Features	Rank	Features
1	sd_90_average	1	dp_90_offer/sell
2	sd_90_minimum	2	dp_90_offer/buy
3	sd_90_opening	3	dp_90_minimum
4	sd_90_closing	4	dp_90_opening
5	sd_90_maximum	5	dp_90_closing
6	sd_60_opening	6	dp_90_average
7	sd_90_offer/sell	7	dp_90_maximum
8	sd_90_offer/buy	8	dp_60_offer/buy
9	sd_60_maximum	9	dp_60_offer/sell
10	sd_60_minimum	10	rsi_21_average

The standard deviation financial indicator stands out as extremely relevant in both methods, occupying all positions in the top 10 in each of them, except the tenth position in the Relief method. In the information gain method, eight standard deviation indicators refer to the period of 90 windows and two to the period of 60 windows. In the Relief method, seven indicators correspond to the period of 90 windows, two to the period of 60 windows, while the tenth place was occupied by the RSI in 21 windows.

The Relief method with 73 features utilizing Stratified K-fold Cross-Validation achieved an accuracy of 0.493. And the model developed through Stratified K-fold Cross-Validation was evaluated using a test dataset, where it achieved an accuracy of 0.515 and an F1-Score of 0.516. Applying the same approach to the GD, representing a model with 52 features, it initially achieved an accuracy of 0.488 with Strati-

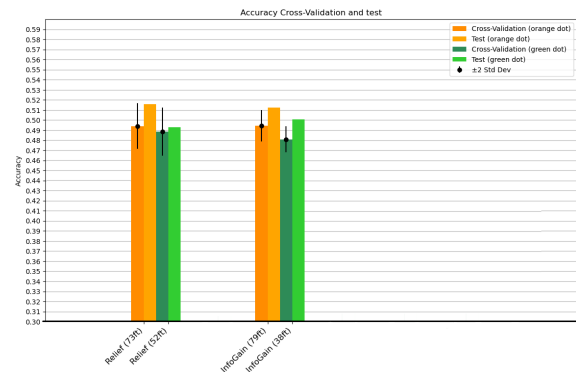


Figure 4: Graphic Information Gain and Relief methods.

fied K-fold Cross-Validation and the model developed through Stratified K-Fold Cross-Validation was evaluated using a test dataset, where it achieved an accuracy of 0.492 and an F1-Score of 0.493

In this case of Information Gain, the OD is marking where the number of features is 79. Then it is utilized Stratified K-fold Cross-Validation to build a new model, after the training it achieved an accuracy of 0.494. The model built with Stratified Cross-Validation was assessed using the test dataset, it achieved an accuracy of 0.512 and F1-Score of 0.512. The same procedure was adopted for the GD, with Stratified Cross-Validation it achieved an accuracy of 0.480, and then this model was assessed using the test dataset, it achieved an accuracy of 0.500 and F1-Score of 0.501.

The accuracy values around 50% can be partially explained by the complexity of the financial market, but also by the multiclass nature of our classification problem. The dataset was structured to predict five distinct price movement classes which inherently makes the classification task more challenging. In financial prediction, price movements are often subtle and influenced by numerous external factors, and distinguishing between similar classes, such as Stability and small rises or descents, adds complexity.

6 CONCLUSIONS

This research explored the effectiveness of Information Gain and Relief methods in improving predictive performance in the Brazilian financial market, using Random Forest models. The data set, composed of 194 technical analysis indicators, was subjected to attribute selection processes, with the methods evaluated by progressive attribute removal and validation by Cross-Validation.

The results presented provide valuable insights for the development of more efficient models in the context of the Brazilian financial market, and future studies could explore the application of other attribute selection methods or the adaptation of the methodology in different financial scenarios.

For future works, we intend to explore the potential of the selected features for new analyses, leveraging these optimized features in advanced machine learning models, such as deep learning architectures, to enhance prediction accuracy.

ACKNOWLEDGEMENTS

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

- Aha, D. W. and Bankert, R. L. (1995). A comparative evaluation of sequential feature selection algorithms. In *Pre-proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*.
- Altman, D. G. and Bland, J. M. (2005). Standard deviations and standard errors. In *Bmj*. British Medical Journal Publishing Group.
- Araujo, G. S. and Gaglianone, W. P. (2023). Machine learning methods for inflation forecasting in brazil: New contenders versus classical models. In *Latin American Journal of Central Banking*. Elsevier.
- Bansal, S. (2016). Investigating the efficacy of rsi in the nifty 50 index. In *Global journal of Business and Integral Security*.
- Billah, M. M., Sultana, A., Bhuiyan, F., and Kaosar, M. G. (2024). Stock price prediction: comparison of different moving average techniques using deep learning model. In *Neural Computing and Applications*. Springer.
- Bouri, E., Demirer, R., Gupta, R., and Sun, X. (2020). The predictability of stock market volatility in emerging economies: Relative roles of local, regional, and global business cycles. In *Journal of Forecasting*. Wiley Online Library.
- Breiman, L. (2001). Random forests. In *Machine learning*. Springer.
- Cardoso, F. C., Malska, J. A. V., Ramiro, P. J., Lucca, G., Borges, E. N., de Mattos, V. L. D., and Berri, R. A. (2022). Bovidb: a data set of stock prices of all companies in b3 from 1995 to 2020. In *Journal of Information and Data Management*.
- Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. In *Computers & electrical engineering*. Elsevier.
- Chen, C.-P. and Metghalchi, M. (2012). Weak-form market efficiency: Evidence from the brazilian stock market. In *International Journal of Economics and Finance*. Citeseer.
- Cuervo, R. (2023). Predictive ai for sme and large enterprise financial performance management. In *arXiv preprint arXiv:2311.05840*.
- Dash, R. K., Nguyen, T. N., Cengiz, K., and Sharma, A. (2023). Fine-tuned support vector regression model for stock predictions. In *Neural Computing and Applications*. Springer.
- Halilbegovic, S. (2016). Macd-analysis of weaknesses of the most powerful technical analysis tool. In *Independent Journal of Management & Production*. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo.
- Htun, H. H., Biehl, M., and Petkov, N. (2023). Survey of feature selection and extraction techniques for stock market prediction. In *Financial Innovation*. Springer.
- Jain, R. and Vanzara, R. (2023). Emerging trends in ai-based stock market prediction: A comprehensive and systematic review. In *Engineering Proceedings*. MDPI.
- Janecek, A., Gansterer, W., Demel, M., and Ecker, G. (2008). On the relationship between feature selection and classification accuracy. In *New challenges for feature selection in data mining and knowledge discovery*. PMLR.
- Kamalov, F., Smail, L., and Gurrib, I. (2019). Stock price prediction using technical indicators: a predictive model using optimal deep learning. In *Learning*.
- Kamalov, F., Smail, L., and Gurrib, I. (2020). Forecasting with deep learning: S&p 500 index. In *2020 13th International Symposium on Computational Intelligence and Design (ISCID)*. IEEE.
- Kang, B.-K. (2021). Improving macd technical analysis by optimizing parameters and modifying trading rules: evidence from the japanese nikkei 225 futures market. In *Journal of Risk and Financial Management*. MDPI.
- Kohn, K. and Moraes, C. d. (2007). O impacto das novas tecnologias na sociedade: conceitos e características da sociedade da informação e da sociedade digital. In *XXX Congresso Brasileiro de Ciências da Comunicação*.
- Shi, Y., Li, B., Long, W., and Dai, W. (2022). Method for improving the performance of technical analysis indicators by neural network models. In *Computational Economics*. Springer.
- Souza, A. S., Lucca, G., Borges, E. N., Cardoso, F. C., Dalmazo, B. L., and Berri, R. (2024). Dataset for Intraday Analysis of B3 stock prices.
- Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., and Moore, J. H. (2018). Benchmarking relief-based feature selection methods for bioinformatics data mining. In *Journal of biomedical informatics*. Elsevier.
- Yu, L. and Liu, H. (2016). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*.