DOI: https://doi.org/10.14393/REE-v40n2a2025-72560
Instituto de Economia e Relações Internacionais – Universidade Federal de Uberlândia

economia
ensaios

Machine Learning Models for IBOVESPA Stock Price Trend Prediction

Métodos de Machine Learning na predição da tendência de preços do IBOVESPA

Elton Massahiro Saito Loures [a]
Lucas Santana da Cunha[b]

Abstract: Financial markets play a fundamental role in the economic organization of countries, driving investors to seek improvements in technical analysis tools to optimize returns. This study applies various Machine Learning models to predict stock price trends (upward, downward, and sideways) in the São Paulo Stock Exchange Index (IBOVESPA – *Índice da Bolsa de Valores de São Paulo*) on a weekly basis. Technical indicators are used as explanatory variables. Statistical parameters such as precision, accuracy, the ROC curve, and the AUC were evaluated. The KNN, Random Forest, and Logistic Regression models achieved the best performance. The results indicate that Machine Learning models are effective in the investment sector. They provide satisfactory outcomes for both the market and future research.

Keywords: Machine Learning; Investment; Technical Analysis; Prediction.
JEL Classification:

Resumo: Os mercados financeiros desempenham um papel fundamental na organização econômica dos países, impulsionando investidores a buscar melhorias nas ferramentas de análise técnica para otimizar os ganhos. Este estudo aplica diversas técnicas de *Machine Learning* para prever as tendências (alta, baixa e lateral) do índice IBOVESPA em um intervalo semanal entre os indicadores técnicos (variáveis explicativas). Parâmetros estatísticos como precisão, acurácia, curva ROC e a AUC foram avaliados, destacando-se o desempenho dos modelos KNN, *Random Forest* e Regressão Logística. Conclui-se que as técnicas de Machine Learning são eficazes no setor de investimentos, oferecendo resultados satisfatórios para o mercado e futuras pesquisas.

Palavras-chave: Aprendizado de Máquina; Investimento; Análise Técnica; Previsão.
Classificação JEL:

[a] Graduate Student in Machine Learning and Big Data – State University of Londrina. Email: masssahirosaito@gmail.com. ORCID: https://orcid.org/0000-0003-1067-1476.
[b] Collaborating Professor, State University of Londrina. Email: lscunha@uel.br. ORCID: https://orcid.org/0000-0001-8513-2309.

1. Introduction

Financial markets have long played a central role in shaping the social and economic organization of countries. The automation of financial markets has expanded access, allowing not only for data consultation but also for trading and decision support for economic agents. According to Madeo et al. (2012), the presence of information technology in stock exchanges has been a key instrument for developing competitive advantages during a period of sharp growth in the daily volume of transactions.

According to Khaidem, Saha, and Dey (2016), anticipating stock market price movements is a complex challenge due to the many uncertainties involved and the wide range of variables that influence market value, such as economic conditions, investor behavior, and political or natural events. The authors point out that stock market time series are typically dynamic, non-parametric, noisy, and chaotic. Therefore, the direction of stock prices is often regarded as a random process. However, some assets follow linear trends in the long run. Because of their inherent volatility and complexity, investments carry substantial risks that can be mitigated through an advanced understanding of expected price movements. The continuous growth of investments and trading activity in stock markets drives an ongoing search for more effective tools aimed at reducing risk and increasing returns. Hendershott, Riordan et al. (2009) define algorithmic trading as the automatic execution of trading decisions, including order submission and response management, through the use of computational algorithms. According to Portnoy et al. (2011), algorithmic trading involves the statistical analysis of an asset’s historical data and, based on current market conditions, seeks to predict the direction of price movements to identify profit opportunities from operations based on predictions.

Two main approaches stand out in the attempt to predict asset price behavior in capital markets: Fundamental Analysis and Technical Analysis. According to Bodie, Kane, and Marcus (2014), Fundamental Analysis is an approach that evaluates stock prices based on earnings prospects, future interest rates, dividend expectations, and the company’s risk assessment. It typically begins with an examination of the firm’s financial statements and past profits. In contrast to Fundamental Analysis, Austin (2004) emphasizes that technical analysts focus on the price of the asset itself, which is determined by supply and demand in financial markets. Goldberg, Nitzsch, and Morris (2001) note that the main objective of Technical Analysis is to identify trends. Western Technical Analysis originated from the Dow Theory (RHEA, 1993), which presented its views on price behavior and market movements. Although its techniques were long rejected by the academic finance community (Boainain, 2007), Technical Analysis and the use of technical indicators have continued to spread and are now employed by an increasing number of investors. According to SILVA et al. (2018), technical indicators allow investors to analyze relevant price patterns, construct charts alongside asset prices, and provide multiple perspectives for trend analysis, helping to minimize the risks involved in trading operations.

According to Castro (1979), analyzing observed trends makes it possible to identify more accurate trading opportunities. For technical analysts, these trends exhibit patterns that tend to repeat over time, allowing for the anticipation of some of their future movements.

This study applies different Machine Learning techniques from January 2, 2007, to January 13, 2021, to predict the stock price trend of the São Paulo Stock Exchange Index (IBOVESPA – *Índice da Bolsa de Valores de São Paulo*) as upward, downward, or sideways. According to Takamatsu, Lamounier, and Colauto (2008), this index reflects the behavior of major trading activities on the BOVESPA and serves primarily as an average indicator of market performance. In 2020, the IBOVESPA closed at slightly above 119,000 points (B3, 2021). The analysis employed several of the most widely used models, including Naive Bayes, K-Nearest Neighbors, Support Vector Machine, Decision Trees, Random Forest, and Logistic Regression, which are extensively discussed in the literature (e.g., Hastie, Tibshirani, and Friedman, 2009). The study seeks to contribute to research involving the application of Machine Learning to investment decision-making in the IBOVESPA market.

2. THEORETICAL FRAMEWORK

2.1. Stock Market and the Market Hypothesis

According to Arestis, Demetriades, and Luintel (2001), stock markets play a major role worldwide, with countless studies linking financial development to economic growth. They are capable of promoting long-term growth by encouraging the acquisition and dissemination of information and by helping reduce the costs of mobilizing savings, as they stimulate investment. According to Santos et al. (2020), the use of algorithms to predict future trends in stock market prices contradicts the Efficient Market Hypothesis (EMH) (MALKIEL; FAMA, 1970).

Malkiel and Fama (1970) define an “efficient market” as one in which the prices of financial assets provide accurate signals for resource allocation. That is, a market in which firms can make production and investment decisions, and investors can choose assets representing ownership of those firms, under the assumption that asset prices at any point in time fully reflect all available information (CARVALHO; JR; GOULART, 2008). Therefore, if an individual identified an advantage by using historical stock data, the entire market would recognize this benefit, and, as a result, prices would adjust. However, according to Malkiel (2003), in the early twenty-first century, many financial economists and statisticians began to adopt the view that stock prices are partially predictable based on historical price patterns.

Adaptive Market Hypothesis (AMH) states that stock prices can be predictable, making it possible to obtain profits through predictions. Technical Analysis and Fundamental Analysis are the two main schools of thought used by professionals to guide their decisions in the stock market (ROCKEFELLER, 2019; LO, 2019; SANTOS et al., 2020).

2.1. Machine Learning and Predictive Models

In everyday life, people are constantly faced with uncertain questions such as: “What are the profiles of clients who most frequently use this service?”, “Should I buy or sell my stocks?”, “Which factors can influence a heart attack?”, or “What is the best movie or song for this user’s profile?”. These questions illustrate the ongoing search to anticipate future events and make better-informed decisions. Machine Learning, a term popularized by engineer Samuel (1959), is a specific field within Artificial Intelligence (AI) based on the idea that systems can learn from data and interactions, identify patterns to improve their performance in specific problems, and make decisions with minimal human intervention by means of predictive modeling.

According to Machado (2018), predictive models represent the relationship between the specific performance of a unit within a sample and one or more known attributes of that unit, with the objective of assessing the likelihood that a similar unit in a different sample will exhibit the same specific performance.

Applications of Machine Learning can be found in various contexts. For instance, the AI system developed by OpenAI Five became the first to defeat world champions in the eSports game Dota 2 (BERNER et al., 2019). During the 1991 Gulf crisis, the DART system (CROSS; WALKER, 1994) was employed—an automated tool that managed up to fifty thousand vehicles, cargo airlifts, and personnel simultaneously, aiming to perform logistical planning that accounted for routes, points of departure, and conflict resolution (RUSSELL; NORVIG, 2004). Automatic translation, market research, robotics, speech recognition, and fake news detection are just a few among the many examples that exist and continue to evolve.

2.3. Preprocessing

According to Han, Kamber, and Pei (2011), preprocessing can improve data quality and, consequently, enhance the efficiency of subsequent data mining processes. This step is crucial, as databases are often large and may contain records that compromise data quality, such as missing information, inconsistent or duplicated entries, asymmetry, and outliers (SCHMITT et al., 2005).

The initial data screening, along with information gathering, source reliability assessment, evaluation of the quantity and meaning of each recorded variable, and the use of appropriate data visualization techniques, are all essential for building an adequate model.

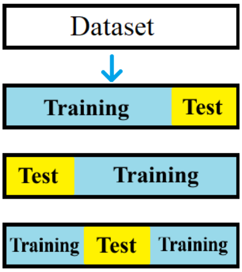
2.4. Machine Learning Techniques

2.4.1 K-Fold Cross-Validation (Grid Search)

For optimizing the parameters of each Machine Learning technique, the K-Fold Cross-Validation method is employed (BURMAN, 1989). This method consists of dividing the sample into k parts of similar size $(d1, d2, d3, \dots, dk)$. Thus, there will be K iterations, and in each iteration, the validation sample will be given by dk, while the training sample used to create the predictor will consist of the remaining $K - 1$ parts (CUNHA, 2019). Its main advantage is that all data are used in both the training and validation phases.

Here is an example of K-fold cross-validation, where $K = 3$.

Figure 1: Example of K-Fold Cross-Validation with $K = 3$.



Source: The authors.

According to Borra and Ciaccio (2010), the bias in this method decreases as the value of K increases; however, this comes at the cost of greater computational time and processing effort.

2.4.2 Naive Bayes

Naive Bayes is a classification technique based on Bayes' Theorem, which describes the probability of an event given prior knowledge that may be related to it. The conditional probability of Bayes can be expressed as (MORETTIN; BUSSAB, 2017):

$$P(A|B) = \frac{P(A).P(B|A)}{P(B)} \tag{1}$$

where $P(A|B)$ represents the probability of A occurring given that B has occurred; $P(A)$ is the probability of event A; $P(B)$ is the probability of event B; and $P(B|A)$ is the probability of B occurring given A.

Considering a set of attributes $B_1, B_2, \dots, B_n$ and $A_1, A_2, \dots, A_k$, we can rewrite:

$$P(A_k\ |B_1, \dots, B_n) = \frac{P(A_k).P(B_1, \dots, B_n|A_k)}{P(B_1, \dots, B_n)} \tag{2}$$

This method assumes that B_i is independent for all $i \in \{1,\ldots,p\}$. In other words, this classifier assumes that the presence of a particular feature in a class is not related to the presence of any other factor (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). By expressing the model as a conditional probability and assuming that each attribute B_i is conditionally independent of all other B_j for $J \neq i$ and that $(B_i|A, B_j) = p(B_i|A)$, the model can be expressed as:

$$\hat{y}\, classifier = argmax\, p(A_k) \prod_{i=1}^{n} p(B_i\, |A_k), k \in 1,\ldots,k \qquad (3)$$

where for each attribute, its probability distribution is assumed to be normal.

2.4.3 K-Nearest Neighbors (KNN)

The KNN method, or K-Nearest Neighbors (FIX; HODGES, 1951), is widely used in classification tasks due to its simplicity and ease of implementation. It is an analogy-based classifier in which the training set consists of n-dimensional vectors, and each element of this set represents a point in an n-dimensional space.

The classification of an element that does not belong to the training dataset is determined according to a distance criterion, most commonly the Euclidean distance. The KNN classifier searches for K elements that are closest to the input element of unknown class—that is, those with the smallest distances—and assigns the most frequent class among these neighbors by majority vote.

The Euclidean distance between two elements X_l and X_k, with $l \neq k$ attributes, is defined as:

$$d(X_l\, , X_k) = [(X_l - X_k)'(X_l - X_k)^{1/2} = \sqrt{\sum_{i=1}^{p} (X_{il} - X_{ik})^2} \qquad (4)$$

where X_{il} represents the value of the i-th attribute of observation l; X_{ik} is the value of the i-th attribute of observation k; and p is the total number of attributes.

Many datasets contain continuous attributes that span different ranges or vary due to their nature or measurement scales. These differences can be significant and must be accounted for (FACELI et al., 2011). To avoid issues such as discrepancies among values, data are usually normalized using the following expression (HAN; KAMBER; PEI, 2011):

$$x_{ij}^{*} = \frac{x_{ij} - min_j}{max_j - min_j} \qquad (5)$$

where x_{ij} is the value of attribute j for observation i; min_j and max_j are, respectively, the minimum and maximum values of attribute j in the dataset; and x_{ij}^{*} is the normalized value of x_{ij}.

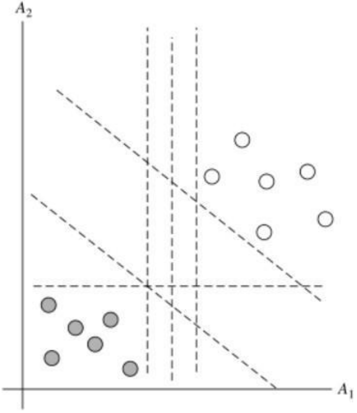
In the case of categorical attributes, it is possible to define distance by binarization—assigning a distance of 1 if the attribute values are identical across observations and 0 otherwise—or by using a scale for ordinal elements.

KNN is a method capable of modeling complex decision spaces; however, it can be computationally expensive and slow during the classification process, since it requires calculating all distances between the unknown element and the training observations. To mitigate this, some researchers propose creating a hypersphere with a defined radius *R* to select elements within it and apply majority voting. This approach greatly reduces computational cost and processing time, although it carries the drawback that no points may fall within the hypersphere.

2.4.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is based on the statistical learning theory developed by Vapnik (2013), designed to solve pattern classification problems. It was originally developed for binary classification, constructing a hyperplane as a decision surface that separates linearly separable classes. SVM can be applied to both regression and classification tasks. Although it may have a slow training time, it is highly accurate due to its ability to model complex nonlinear functions. Figure 2 represents a separation in a two-dimensional plane.

Figure 2: Separation in hyperplanes.



Source: The authors.

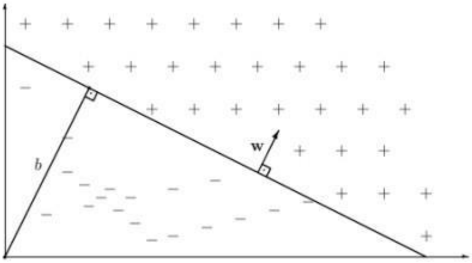
SVM has a lower probability of overfitting—when the model performs excellently but, when applied to new datasets, yields poor results; in other words, training accuracy is high, but test accuracy is low.

To construct a linear decision boundary that separates elements in the dataset belonging to different classes, the separating hyperplane method is used. For the two-class case, a value of $+1$ is assigned if $f(x) \geq 0$ and -1 if $f(x) < 0$, and it can be expressed as (LIMA, 2002):

$$\sum_{i=1}^{n} = \vec{w_i} x_i + \vec{b}, \qquad (6)$$

where $\vec{w}$ is the weight vector that defines a direction perpendicular to the hyperplane, and b is the bias which, when varied, moves the hyperplane parallel to itself. These parameters are responsible for controlling the function and the decision rule (LIMA, 2002). The values of these parameters are obtained through the learning process from the input data (GONÇALVES, 2015).

Figure 3: Geometric interpretation of $\vec{w}$ and b on a hyperplane.

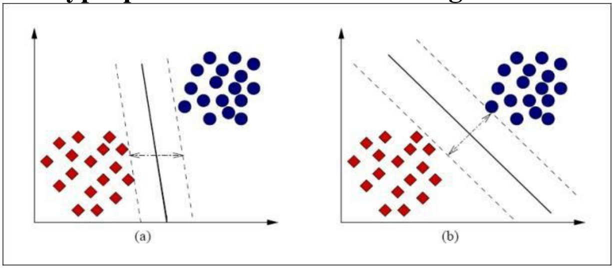


Source: Lima (2002), Gonçalves (2015).

In order for a linear SVM to find the optimal hyperplane that best separates the data from each class, the hyperplane must be considered a maximum-margin separator—that is, if the hyperplane separates a set of vectors without errors, the distance between the closest vectors of different classes and the hyperplane is as large as possible. Therefore, assuming that the dataset is linearly separable, the optimal hyperplane is the one that achieves the largest margin:

$$f(x) = \langle \vec{w} . \vec{x} \rangle + b = 0 \qquad (7)$$

Figure 4: (a) Separating Hyperplane with small margin. (b) Separating Hyperplane with maximum margin.



Source: Lima (2002), Gonçalves (2015).

To ensure that the margin is always greater than the distance between the hyperplanes $\langle \vec{w}.\vec{x}_{\square}\rangle + b = 0$ and $|\langle \vec{w}.\vec{x}_{\square}\rangle + b = 1|$, the following constraint is assumed.

$$\langle \vec{w}.\vec{x}\rangle + b \geq +1, for\ \ y_i = +1 \tag{8}$$

$$\langle \vec{w}.\vec{x}\rangle + b \leq -1, for\ y_i = -1 \tag{9}$$

Linear classifiers that separate the training dataset have a positive margin, and the combination of these equations is referred to as hard margins. To find the optimal separating hyperplane, it is necessary to determine the distance between the hyperplane and the elements closest to it, known as support vectors:

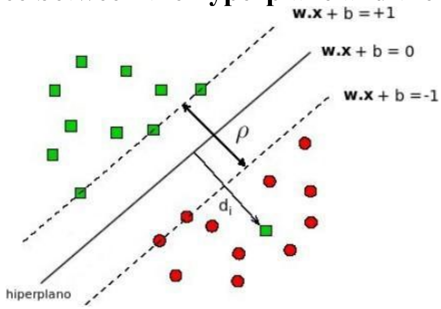
$$d_i = (\vec{w}; b; \vec{x_i}) = \frac{y_i(\langle \vec{w}.\vec{x}\rangle + b)}{||\vec{w}||}, \tag{10}$$

where $d_i = (\vec{w}; b; \vec{x_i})$ is the distance from a given $\vec{x_i}$ to the hyperplane $\overrightarrow{(w, b)}$ (LIMA, 2002). Considering the combination of the constraints from Equations 8 and 9, we have:

$$\rho = \frac{2}{||\vec{w}||}, \tag{11}$$

where ρ is the margin of a separating hyperplane.

Figure 5: Distance between the hyperplane and the support vectors.



Source: Gonçalves (2015).

Since the margin is always greater than the last instance, minimizing ||w|| leads to maximizing the margin (GONÇALVES, 2015). Therefore, the distance between a hyperplane and the support vectors can be found based on the theory of Lagrange Multipliers:

$$L = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i . x_j\rangle \tag{12}$$

where α_i are the Lagrange multipliers.

Hence, it becomes possible to obtain all necessary parameters and estimates for classification. For the case of nonlinearly separable pattern classification, SVM performs a dimensionality transformation through a kernel function, which allows the problem to be treated as a linear classification problem. The most commonly used kernel functions are the polynomial kernel, linear kernel, and sigmoidal kernel. A set of inputs from a nonlinearly separable training sample is mapped using a ϕ function to obtain a new dataset that is linearly separable in a higher-dimensional space.

With the training data mapped to the feature space, the $\phi(x)$ mapped values are used instead of x, resulting in:

$$L = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i). \phi(x_j) \rangle \qquad (13)$$

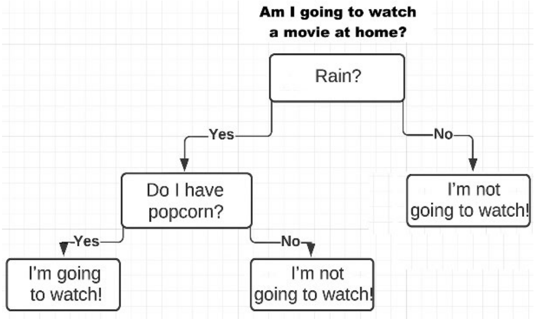
2.4.5 Decision Tree

A Decision Tree is a model represented graphically by nodes and branches, resembling a tree but inverted in direction (MEIRA; RODRIGUES; MORAES, 2008). It is commonly used in algorithms for data classification and aims to build classifiers that predict classes based on the attribute values of a dataset. This technique can be applied to categorical variables (WITTEN et al., 2005). It is an easy-to-understand model, useful for exploring and classifying data, but care must be taken to avoid overfitting. Attention is also needed to prevent information loss in the case of continuous variables.

According to Meira, Rodrigues, and Moraes (2008), the root node is the first node of the decision tree, located at the top of its structure. The internal nodes, including the root node, are the decision nodes. Each node contains a test on an independent variable, and the results of this test form the branches of the tree. The leaf nodes, located at the extremities of the tree, represent predicted values for the dependent variable or probability distributions of these values.

The following figure illustrates a decision tree representing whether a person chooses to watch a movie at home or not.

Figure 6: Decision tree: Watch or not watch a movie at home.



Source: The authors.

In this situation, at the first level (root node), it is verified whether it is a rainy day. If it is not raining, the decision "I will not watch!" is made, whereas if it is raining, it is then verified whether there is popcorn to support the decision. Table 1 presents an analysis of a person's decision, based on the decision tree, regarding whether or not to watch a movie at home.

Table 1. Decision on whether or not to watch a movie at home.

Day	Raining?	Is There Popcorn?	Will I Watch a Movie at Home?
1	Yes	Yes	Yes
2	Yes	No	No
3	Yes	Yes	Yes
4	No	Yes	No
5	No	No	No
6	No	Yes	No

Source: The authors.

After the tree is completed, its output contains organized and compacted data that are used to classify new cases. Thus, it becomes possible to generate predictions, as proposed in this study. Because decision trees tend to grow very large, algorithms usually include rules that consider the path from the root node to a leaf node. This occurs because such rules can be easily modularized and understood without the need to reference other rules (INGARGIOLA, 1996).

The decision on how to split the nodes can greatly influence the accuracy of the algorithm and, consequently, its results. To minimize impurity in the leaves, different attribute selection methods can be applied, among which the Gini Index is one of the most widely used (BREIMAN et al., 1984).

The Gini Index measures the impurity of a data partition: when two samples are randomly selected from a population, they should belong to the same class, and the sum of their probabilities equals one if the partition is perfectly pure. It is used for categorical variables applied in binary splits. The higher this index, the more homogeneous the partition will be. This measure was developed as a variance measure for categorical data (LIGHT; MARGOLIN, 1971). With K classes and p_{mk} representing the proportion of observations in class k at node m, the equation is expressed as (MOISEN, 2008):

$$I_G = \sum_{k=1}^{k} p_{mk}(1 - p_{mk}) \qquad (14)$$

2.4.6 *Random Forest*

Random Forest (BREIMAN, 2001; LIAW; WIENER et al., 2002) is a $\{h_k(X), k = 1,2,\dots,N.\}$ classifier composed of multiple classifiers designed specifically for decision trees, where T_k are independent and identically distributed random samples, and each tree decides the class for the input of X. Random vectors are generated from a fixed probability distribution over the initial input vector. The accuracy of this technique is evaluated probabilistically based on the classifier's margin, given a set of classifiers $h_1(x), h_2(x), \dots, h_k(x)$ and a random training set from the vector$Y$, X (GÓMEZ et al., 2012).

Therefore, for each tree formed based on the input and output values, a probability is obtained for each tree. Using the data from all trees, a final average is calculated—that is, a decision is made based on the frequency of decisions across the trees. For this study, a restriction of 200 trees was applied, with a maximum of 50 nodes and 2 to 4 randomly sampled variables as candidates at each split, to avoid overfitting.

2.4.7 Logistic Regression

A logistic regression function takes as input a real value Z_i, which ensures that the probability of success P_i returns values within the interval [0, 1]. It can be expressed as (GUJARATI; PORTER, 2011):

$$P_i = \frac{1}{1+e^{-Z_i}}, \qquad (15)$$

where $z_i = \beta_0 + \beta_1 X_i \in \mathbb{R}$ and $z_i = \beta_0 + \beta_1 X_i \in \mathbb{R}$.

If P_i is the probability of success, then $(1 - P_i)$ is the probability of failure. Therefore, the odds ratio is given by:

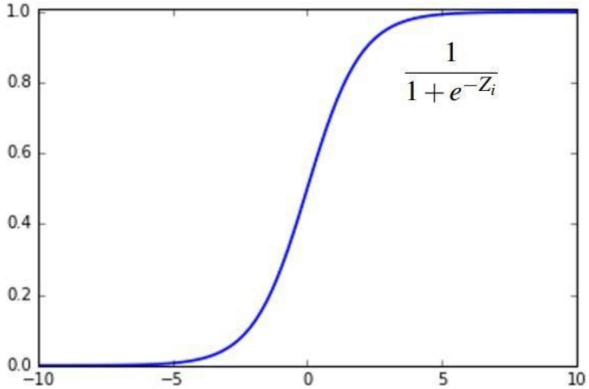
$$\frac{P_i}{1 - P_i} = \frac{1 + e^{Z_i}}{1 + e^{-Z_i}} = e^{Z_i} \qquad (16)$$

To estimate the parameters—i.e., using the Ordinary Least Squares (OLS) method—the function must be linearized (GUJARATI; PORTER, 2011):

$$L_i = ln\left(\frac{P_I}{1 - P_i}\right) = Z_i = \beta_0 + \beta_1 X_i \qquad (17)$$

The graph of the logistic function has the following behavior.

Figure 7: Logistic cumulative distribution function.



Source: (SKIENA, 2017).

2.4.8 Hold-out Cross-Validation and Confusion Matrix

Once the Machine Learning model is prepared, its performance must be measured. The algorithm could be overfitting, which would impair future predictions. The Hold-out Cross-Validation method (DEVROYE; WAGNER, 1979) proposes dividing the sample into two groups: the first group is used to train the model with data presented to the algorithm, and the second group is used to test the model. The composed predictor $p = \frac{4}{5}$ was used as the training sample, and therefore $\frac{1}{5}$ as the test sample.

To visualize an algorithm's performance, it is common to use the confusion matrix, also known as the error matrix. This matrix consists of rows representing elements in a predicted class and columns representing elements in an actual class. That is, the main diagonal of the matrix identifies the correct predictions, while the diagonal from right to left represents the prediction errors of a model.

Table 2. Confusion Matrix.

	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Source: The authors.

Representing the proportion of actual positive cases that were correctly identified, recall (true positive rate) is given by the following equation:

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

Precision represents the proportion of correctly identified positive predictions:

$$Prec = \frac{TP}{TP + FP} \tag{19}$$

Accuracy also indicates the overall performance of the model, showing how many classifications, out of all predictions, the model correctly identified. It can be expressed by Equation 20,

$$Acc = \frac{TP + TN}{N} \tag{20}$$

where $N = TP + TN + FP + FN$.

2.4.9 ROC Curve and AUC Measure

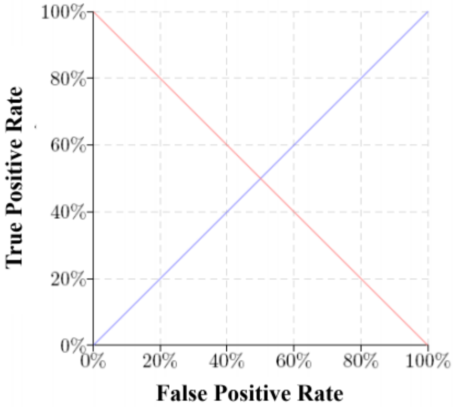
Receiver Operating Characteristics (ROC) is a graphical method used to visualize, organize, and select classifiers based on the performance of a classification model. It was originally developed in signal detection theory to evaluate the quality of signal transmission through a noisy channel (EGAN; EGAN, 1975). Today, it is applied in various fields, such as assessing income inequality (GASTWIRTH, 1971), evaluating the accuracy of weather forecasts, and even determining the quality of clinical tests (ZHOU; MCCLISH; OBUCHOWSKI, 2009; MYLNE, 2002; PRATI et al., 2008).

The ROC curve is obtained by plotting the probability of detection (recall) against the probability of false alarms, or the false positive rate (FPR), which can be expressed as:

$$FPR = \frac{FP}{(FP + TN)} \tag{21}$$

To construct the ROC graph, the FPR is plotted on the x-axis and recall on the y-axis. According to Prati et al. (2008), the point (0,0) represents the strategy of never classifying an instance as positive—models corresponding to this point have no false positives but also fail to identify any true positives. The point (100%,100%) represents a model that always classifies a new instance as positive. The point (0,100%) represents a model in which all instances are correctly classified, whereas (100%,0) represents a model that always misclassifies its predictions. As shown in Figure 8, the descending diagonal line indicates classification models that perform equally across both classes ($Recall = 1 - TFP$) . To the left of this line are models that perform better for the negative class at the expense of the positive class, while to the right are models that perform better for the positive class.

Figure 8: ROC space.

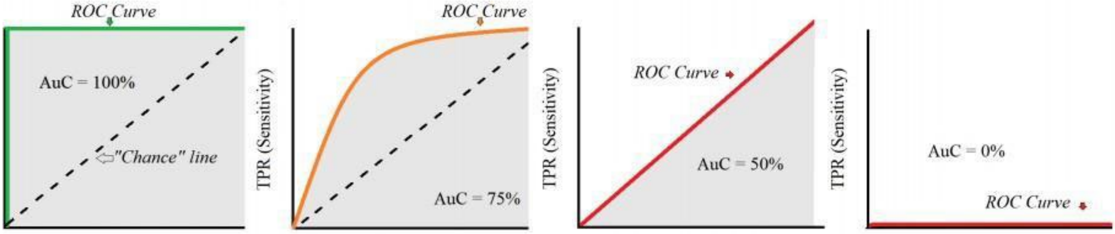


Source: (PRATI et al., 2008).

The ascending diagonal represents a stochastic behavior in which each point can be obtained by predicting the positive class with probability p and the negative class with probability $1 - p$. Points to the left of this diagonal represent models with better-than-random performance, while points to the right represent worse performance (PRATI et al., 2008).

To evaluate the ROC, the Area under the ROC Curve (AUC) is calculated as an integral measure that provides an aggregate indicator of performance across all possible classification thresholds. According to McClish (1989), it can be interpreted as the probability that the model ranks a randomly chosen positive example higher than a randomly chosen negative example, ranging from 0 (completely incorrect predictions) to 1 (100% correct predictions). This method has the advantage of measuring how well predictions are ranked and is invariant to the classification threshold, as it assesses model quality independently of the chosen decision boundary. Figure 9 presents the ROC graph and the AUC measure for different values: from left to right, as the area decreases, the AUC also decreases, and consequently, the model's performance declines.

Figure 9: AUC measure.



Source: (GLEN, 2019).

2.5. Technical Indicators

In general, technical indicators are data series derived from applying a formula based on the combination of the high, low, close, open, and volume of a given period—for example, of any asset (SILVA et al., 2018).

2.5.1 Relative Strength Index (RSI)

The Relative Strength Index (WILDER, 1978) measures the acceleration and deceleration of buying and selling pressure levels over a period of time. It can be expressed by Equation 22.

$$IFR = 100 - \frac{100}{1 + \frac{GM}{PM}} \tag{22}$$

where GM represents the average of gains over the last n periods in which the asset's closing price was higher than the previous day's, and PM corresponds to the average of losses over the periods in which the closing price was lower than the previous day's. Each average is calculated according to the periods that showed gains or losses, respectively. Wilder (1978) and Wong, Manzur, and Chew (2003) recommend a 14-day interval for calculating this indicator.

2.5.2 Moving Average Convergence and Divergence (MACD)

Moving Average Convergence and Divergence, or MACD (APPEL; APPEL, 2008), is widely used to detect changes in market trends. It consists of two lines: the fast line (MACD), generally calculated as the difference between the 12-day exponential moving average and the 26-day exponential moving average, and the slow line, also called the signal line, which is usually the 9-day moving average of the MACD (MOZZER, 2015). Thus,

$$MACD = EMA_{12}(Cl) - EMA_{26}(Cl) \tag{23}$$

and

$$Signal\ Line = EMA_{9}(MACD) \tag{24}$$

where EMA_n is the exponential moving average for n days and Cl represents the time series of closing prices.

According to Elder (2004), the fast line reflects the short-term consensus of value, while the slow line reflects the longer-term consensus. Therefore, when the two lines cross, it indicates a potential trend reversal, and the MACD lines follow market trends.

2.5.3 Price Rate of Change (PROC)

The Price Rate of Change (PROC) is an indicator developed to evaluate the rate of change in an asset's price. It is used to compare current prices with those from n previous periods, indicating divergence as well as overbought and oversold zones (MURPHY, 1999). When this oscillator varies positively, it signals an upward trend; otherwise, it indicates a downward trend (CAMPOLINA; BATISTA, 2019).

$$PROC(t) = \frac{Cl(t) - Cl(t-n)}{Cl(t-n)} \tag{25}$$

where $PROC$ is defined at time t and $Cl(t)$ is the closing price at time t. A value of $n = 14$ was adopted, since most indicators suggest a 14-day period.

2.5.4 On-Balance Volume (OBV)

The On-Balance Volume (OBV) is one of the most important analytical tools for the stock market (Debastiani, 2008). Rather than addressing the question "What is the market's direction?", this indicator focuses on "Who is driving the market where it goes?" The OBV indicator operates by applying price variation to the traded volume, based on the premise that large volumes of trades are executed by major investors. Therefore, following their movements—trading the same asset at the same time as large investors—can be an excellent strategy for small investors (Debastiani, 2008).

The OBV can be expressed as:

$$OBV(t) = \begin{cases} OBV_{t-1} + Vol(t), & se\, Cl(t) > Cl_{t-1} \\ OBV_{t-1} - Volt(t), & se\, Cl(t) < Cl_{t-1} \\ OBV_{t-1}, & se\, Cl(t) = Cl_{t-1} \end{cases} \qquad (26)$$

where OBV is defined at time t, $Vol(t)$ is the traded volume at time t, and $Cl(t)$ is the closing price at time t.

The OBV follows the same direction as the price trend. If prices tend to fall, the OBV also tends to move downward; conversely, if prices show an upward trend, the OBV line also moves upward (SANTOS et al., 2020).

2.5.5 Aroon Indicator

The Aroon indicator (CHANDE; KROLL, 1994) is used to measure trend changes in the price of a given asset, as well as the strength of that trend. According to Fernandes (2019), this indicator measures the time difference between the most recent high and the most recent low over a specified number of periods. It consists of two curves that measure upward and downward trends, defined respectively by Equations 28 and 29.

$$Aroon_{up} = \frac{N - P_{max}}{N}.100 \qquad (27)$$

$$Aroon_{down} = \frac{N - P_{min}}{N}.100 \qquad (28)$$

where P_{max} and P_{min} represent the number of periods since the last high and the last low, respectively, within the interval of N periods.

The Aroon indicator is obtained by combining Equations 28 and 29:

$$Aroon = Aroon_{up} - Aroon_{down} \qquad (29)$$

3. DATA AND METHODOLOGY

3.1. Planning

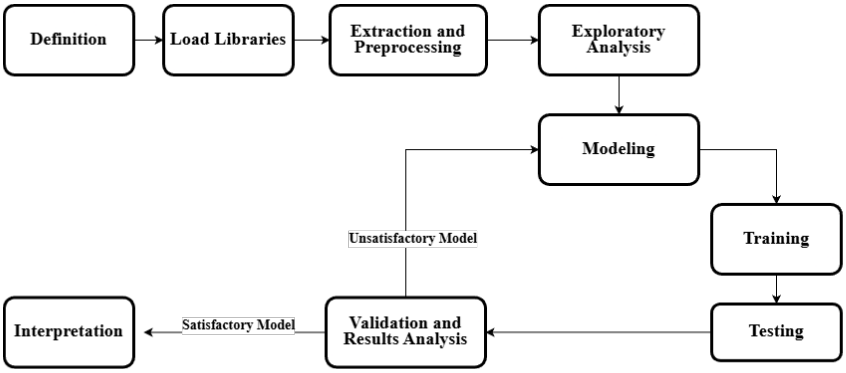
The algorithms were implemented using the R programming language (R Core Team, 2019) within the RStudio integrated development environment (ALLAIRE, 2012). For data extraction related to the index, as well as for calculating some of the indicators presented in this study, the “quantmod” (RYAN, 2025) and “TTR” (ULRICH, 2023) libraries were used. Finally, for the Machine Learning techniques and the ROC and AUC performance analyses, the "caret" (KHUN, 2008) and "pROC" (ROBIN, 2011) libraries were used, respectively.

The models were developed and evaluated on a computer with the following configuration: Intel(R) Core(TM) i7-4700MQ CPU @ 2.40 GHz, 4 cores and 8 logical processors, and 8.00 GB of physical RAM.

The research followed the workflow below:

- **Definition**: organization, identification of necessary tools and data;
- **Load Libraries**: loading the required libraries for data extraction, indicator calculation, and subsequent analyses;
- **Extraction and Preprocessing**: extraction, organization, and initial screening of the data;
- **Exploratory Analysis**: graphical analysis of the response variable and measurement of association for the explanatory variables;
- **Modeling**: application of the models and hyperparameter optimization using the Grid Search technique to test all possible parameter combinations before training and select those that produced the smallest error;
- **Training**: 80% of the total sample was used to train the model;
- **Testing**: 20% of the total sample was used to test the model trained in the previous step;
- **Validation and Results Analysis**: the results were analyzed by adjusting the estimated parameters and interpreting them through the confusion matrix to obtain satisfactory results;
- **Interpretation**: interpretation of the results obtained in the previous step.

Figure 10: Process flow



Source: The authors.

3.2. Data Collection

Using the getSymbols() function from the "quantmod" library (RYAN, 2025) to load and manage data from Yahoo Finance, it was possible to extract data for the IBOVESPA index variable for the period from January 2, 2007, to January 13, 2021. This function allows for the acquisition of historical price series, which provide the basis for calculating technical indicators as new explanatory variables for the study.

During the data processing stage for the IBOVESPA index, records with missing values were identified on certain calendar days. These values do not represent data failures but correspond to days when there was no trading session on B3 (national, state, or municipal holidays, as well as recess periods). Since these points do not contribute to modeling or temporal analysis, days without trading were removed to ensure data consistency.

The statistics calculated for the selected period were: High Price (Hi), Low Price (Lo), Trading Volume (Vol), Closing Price (Cl), and Opening Price (Op). Based on these statistics, it is possible to compute the explanatory variables, i.e., the technical indicators. According to RYAN (2025), the closing price column appears to be adjusted for stock splits twice.

The categorical response variable has three classifications:

- Upward: $price_t > price_{t-14}$;
- Downward: $price_t < price_{t-14}$;
- Sideways: $price_t = price_{t-14}$;

Therefore, if the current price is higher than the price fourteen days earlier, it will be classified as Upward in the response variable. Likewise, if it is lower, it will be classified as Downward, and if it is equal, as Sideways.

Since trend indicators will be used as explanatory variables, a time lag of seven days was established between them and the response variable. In other words, the explanatory variables correspond to day $t - 7$, while the response variable corresponds to day t.

To avoid inconsistent results regarding the IBOVESPA price trend, it is important that the explanatory variables be aligned. Thus, a correlation analysis was performed to verify whether there is a positive relationship among the technical indicators. Correlation analysis, also known as association, aims to assess the degree of relationship between two or more variables and establishes a value summarizing the strength of this relationship, which ranges from -1 to +1 (GUJARATI; PORTER, 2011).

For the K-Fold Cross-Validation, the value of K was set to K = 10, as suggested by researchers such as Borra and Ciaccio (2010) and Cunha (2019), who reported good performance.

4. RESULTS AND DISCUSSION

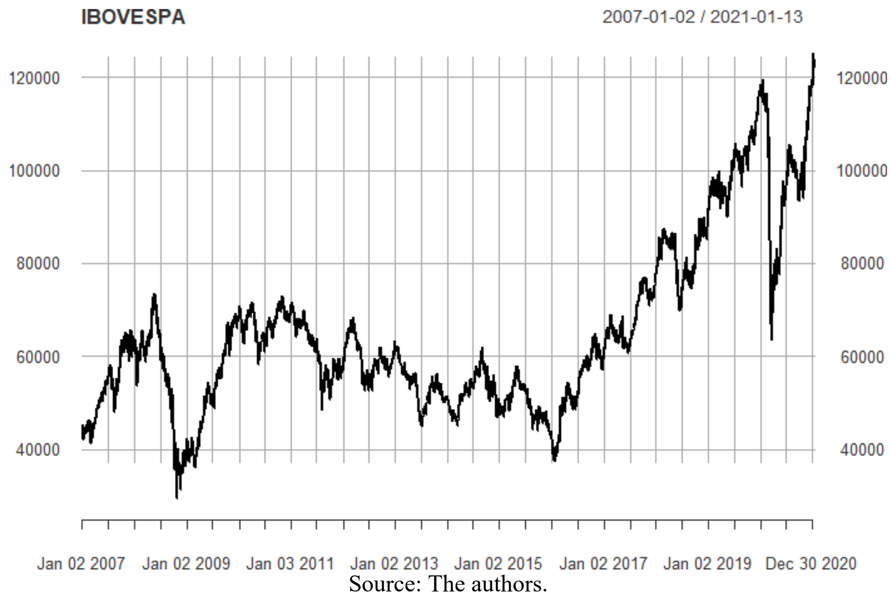
4.1. Exploratory Analysis

The dataset covers the IBOVESPA index over a period of approximately 14 years (from January 2, 2007, to January 13, 2021). As shown in Figure 11, the index exhibited an upward trend in 2007 but followed a downward trend from early 2008 to 2016. In the following years, the index returned to an upward trajectory.

Some of the sharpest declines were observed in 2008 and at the beginning of 2020. According to Schneider (2020), the 2008 subprime financial crisis was considered the most severe since the Great Depression of 1929. Its origin, following the collapse of the housing bubble, was driven by the rapid expansion of bank credit and reinforced by the use of new financial instruments. The crisis deepened with the bankruptcy of the investment bank Lehman Brothers in September 2008, after the Federal Reserve Board (FED) refused to provide financial assistance to the institution. The author points out that the main effects of the crisis on the Brazilian economy were the drop in stock prices and the increase in the exchange rate of the U.S. dollar. Subsequently, credit availability decreased, international investments declined, and economic growth expectations weakened, leading to less optimistic GDP predictions.

As for the beginning of 2020, this period corresponds to the global recession caused by the coronavirus (SARS-CoV-2), combined with the context of the oil price war (DANTAS, 2020). According to the FIA Business School (FIA - *Fundação Instituto de Administração*, 2020), when the pandemic began in Brazil in March 2020, the coronavirus crisis had major impacts on the economy. The institution notes that both industrial activity and consumption experienced a progressive slowdown, producing almost immediate consequences for the stock markets.

Figure 11: Time series of adjusted daily closing prices.

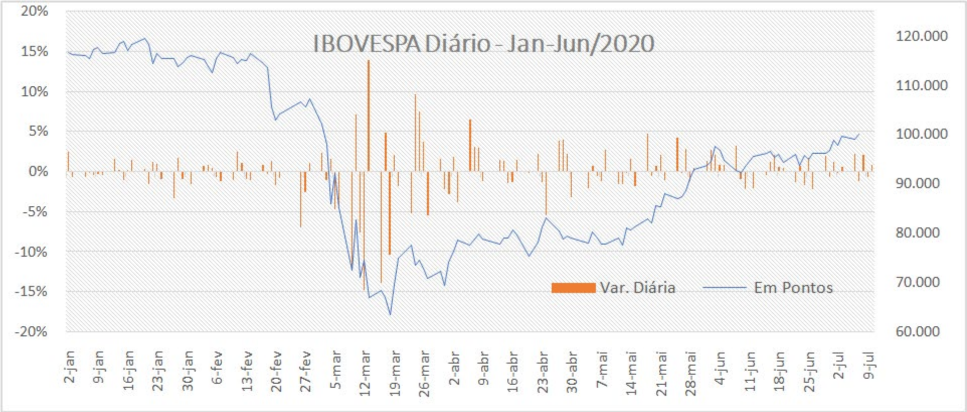


Source: The authors.

According to Martins (2020), the historical series of several Brazilian economic indicators began to form a "V" shape in 2020—an abrupt decline followed by a rapid recovery—indicating the effectiveness of the fiscal and monetary stimulus measures implemented, which can be observed in four historical series.

The first series is that of the IBOVESPA index. Due to a sharp drop in the Selic rate, combined with fiscal stimulus, the quick adaptation of many sectors of the economy, and the depreciation of stock prices, demand for equities increased.

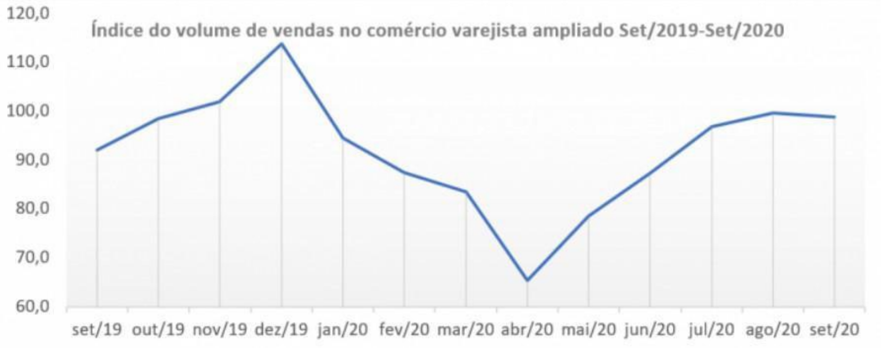
Figure 12: Daily IBOVESPA variation and points. Expanded from January 2020 to June 2020.



Source: Economatica. Prepared by Martins (2020).

The second series covers retail sales performance. Combined with the effects of the emergency aid program, stimulus to the construction sector through lower interest rates, the growth of e-commerce, and incentives for housing finance, these factors contributed to higher retail sales.

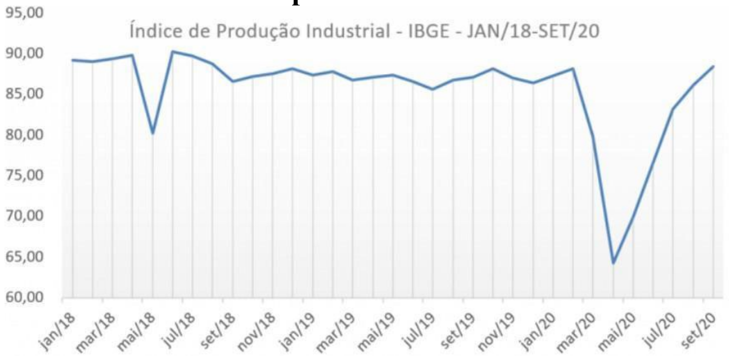
Figure 13: Volume of retail sales, expanded from September 2019 to September 2020.



Source: IBGE. Prepared by Martins (2020).

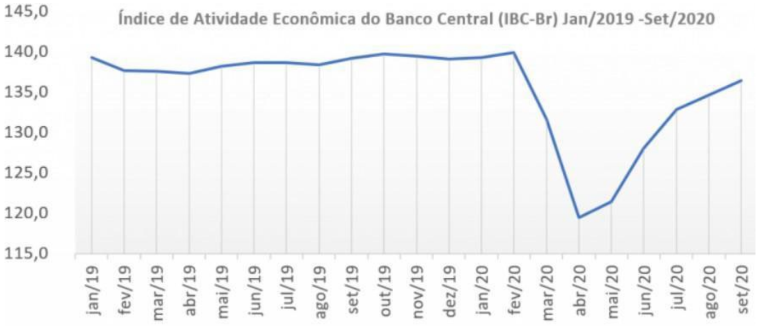
Finally, the last two historical series reflect the recovery in demand, the increase in exports, the substitution of some imports in industrial production, and the rebound in the services sector, represented by the IBC-BR index.

Figure 14: Industrial Production Index, expanded from January 2018 to September 2020.



Source: IBGE. Prepared by Martins (2020).

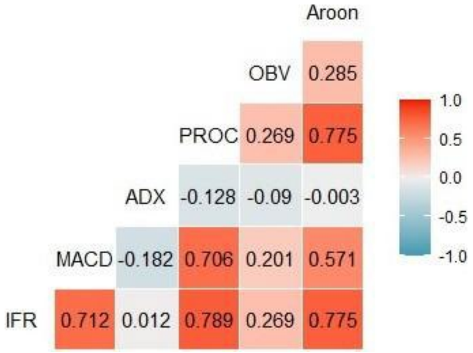
Figure 15: Central Bank Economic Activity Index (IBC-BR), expanded from January 2019 to September 2020.



Source: IBGE. Prepared by Martins (2020).

Subsequently, the Pearson correlation analysis was applied to the explanatory variables. Based on the correlation matrix (Figure 16), out of the 15 results, the variables that showed a correlation greater than 0.70 were: PROC and Aroon (0.775), PROC and MACD (0.706), PROC and RSI (0.789), RSI and Aroon (0.775), and MACD and RSI (0.712). The ADX variable showed the lowest correlation values (absolute values below 0.182) and was not substantially positive; therefore, it was not included in the models.

Figure 16: Correlation matrix of the indicators.



Source: The authors.

4.2. Implementation of Machine Learning Models

After the exploratory analysis, the proposed models were trained and tested, and a statistical analysis was performed to assess the reliability, computational efficiency, and accuracy of the applied Machine Learning techniques. Since no Sideways trend was identified in the response variable, the confusion matrix presents values only for the Downward and Upward trends.

4.2.1 Results Analysis

After completing the model training process, confusion matrices were constructed to identify the values of TP, FP, TN, and FN associated with each model, allowing the calculation of precision and accuracy. In addition, the ROC curve was plotted, and the AUC was determined, providing a broader assessment of the performance of the developed models. The same procedure was applied for the testing phase.

Table 3. Confusion matrix for the training of the Naive Bayes, KNN, and SVM models.

		Naive Bayes		KNN		SVM	
		Prediction					
		Downward	Upward	Downward	Upward	Downward	Upward
Real	Downward	901	468	797	433	822	277
	Upward	364	1006	468	1041	443	1197

Source: The authors.

Table 4. Confusion matrix for the training of the Decision Tree, Random Forest, and Logistic Regression models.

		Decision Tree		*Random Forest*		Logistic Regression	
		Prediction					
		Downward	Upward	Downward	Upward	Downward	Upward
Real	Downward	1044	276	947	228	822	277
	Upward	221	1198	318	1246	443	1197

Source: The authors.

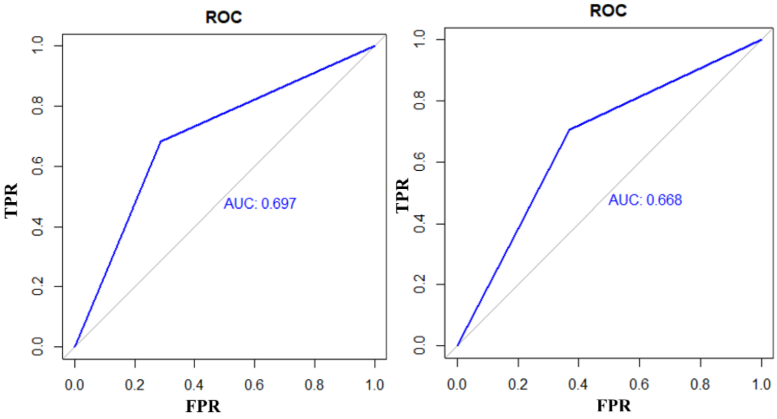
Table 5. Statistical summary of the training phase for the Machine Learning models.

Model	**Accuracy**	**Precision (Downward; Upward)**	**AUC**
Naive Bayes	0.69	0.71 0.68	0.697
KNN	0.67	0.63 0.71	0.668

SVM	0.74	0.65 0.81	0.731
Decision Tree	0.82	0.83 0.81	0.819
Random Forest	0.80	0.75 0.85	0.797
Logistic Regression	0.73	0.71 0.73	0.722

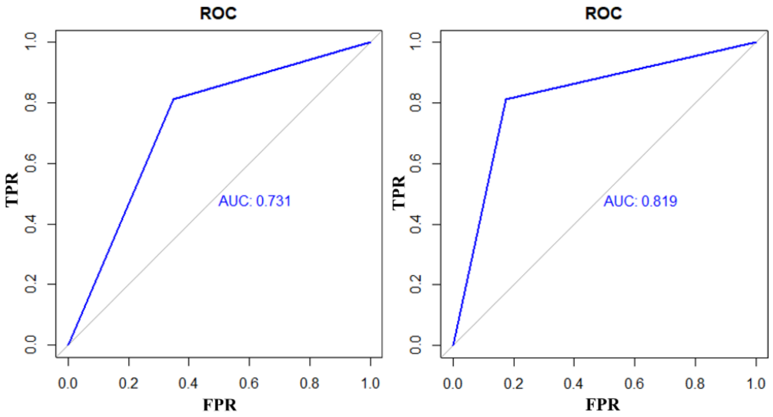
Source: The authors.

Figure 17: Training ROC and AUC Curves for the Naive Bayes and KNN Models.



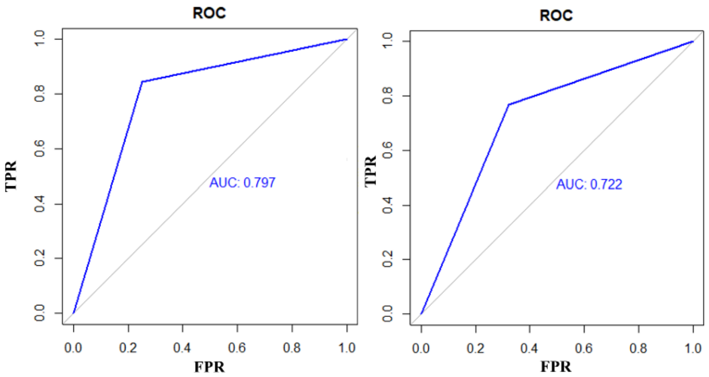
Source: The authors.

Figure 18: Training ROC and AUC Curves for the SVM and Decision Tree Models.



Source: The authors.

Figure 19: Training ROC and AUC Curves for the Random Forest and Logistic Regression Models.



Source: The authors.

Table 6. Confusion matrix for the testing phase of the Naive Bayes, KNN, and SVM models.

		Naive Bayes		KNN		SVM	
		Prediction					
		Downward	Upward	Downward	Upward	Downward	Upward
Real	Downward	166	122	178	123	152	77
	Upward	121	275	110	273	136	319

Source: The authors.

Table 7. Confusion matrix for the testing of the Decision Tree, Random Forest, and Logistic Regression models.

		Decision Tree		*Random Forest*		Logistic Regression	
		Prediction					
		Downward	Upward	Downward	Upward	Downward	Upward
Real	Downward	180	108	153	135	170	118
	Upward	125	271	83	313	82	314

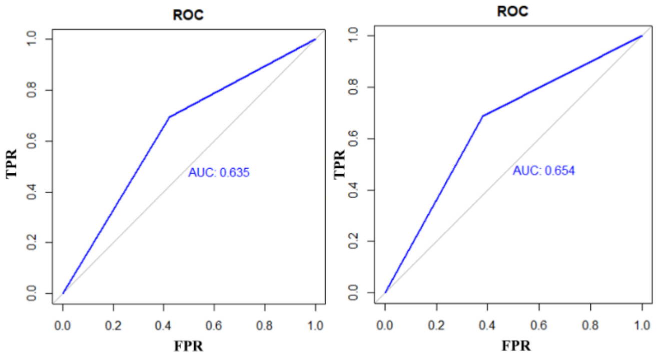
Source: The authors.

Table 8. Statistical summary of the testing phase for the Machine Learning models.

Model	Accuracy	Precision (Downward Upward)	AUC
Naive Bayes	0.64	0.69 0.69	0.635
KNN	0.66	0.62 0.69	0.654
SVM	0.69	0.53 0.81	0.667
Decision Tree	0.66	0.59 0.71	0.655
Random Forest	0.68	0.65 0.70	0.661
Logistic Regression	0.71	0.67 0.72	0.692

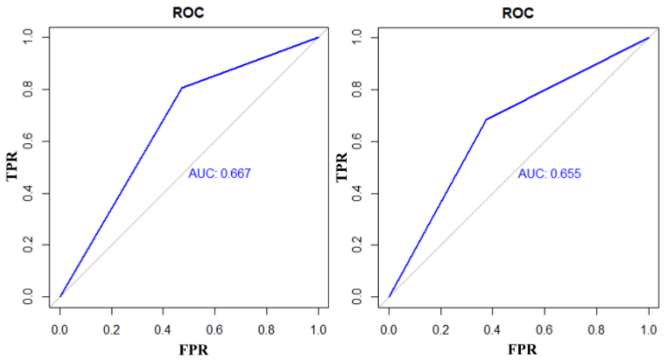
Source: The authors.

Figure 20: Testing ROC and AUC Curves for the Naive Bayes and KNN Models.



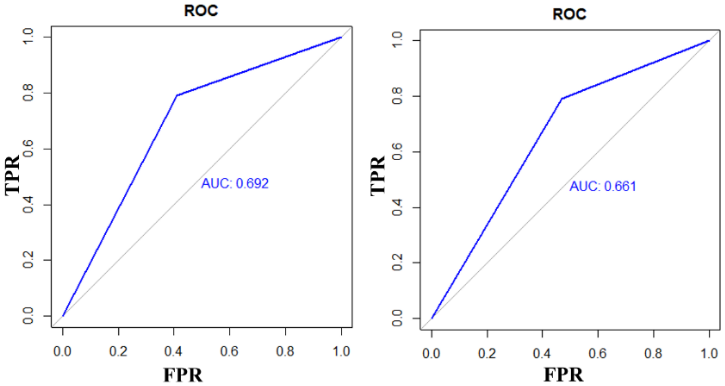
Source: The authors.

Figure 21: Testing ROC and AUC Curves for the SVM and Decision Tree Models.



Source: The authors.

Figure 22: Testing ROC and AUC Curves for the Logistic Regression and Random Forest Models.



Source: The authors.

Table 9. Summary of processing time for the Machine Learning models.

Model	Time (s)
Naive Bayes	3.92
KNN	2.86
SVM	31.55
Decision Tree	1.21
Random Forest	10.28
Logistic Regression	2.06

Source: The authors.

Based on Tables 5 and 8, it is possible to observe the accuracy, precision, and AUC values summarized for each model applied during the training and testing phases. The processing time of the program, considering both the training and testing stages for each applied method, is presented in Table 9. This analysis makes it possible to identify the differences in time and computational cost among the models.

Results show that the KNN model achieved an Accuracy and AUC below 0.70, with values of 0.67 and 0.668, respectively. Similarly, the Naive Bayes model recorded an Accuracy of 0.69 and an AUC of 0.697. Similarly, the Naive Bayes model recorded an Accuracy of 0.69 and an AUC of 0.697. The remaining models achieved results between 0.70 and 0.80 for both estimates.

Although all models showed relatively balanced precision values for Downward and Upward trends—meaning that the precision for Downward was not far from that for

Upward—Table 16 shows that some models did not perform well. The Decision Tree model experienced a significant drop compared with the training phase, with a reduction of 0.24 in Downward precision and 0.10 in Upward precision. The SVM model also showed a 0.12 decrease in Downward precision relative to training, reaching 0.53 for Downward and 0.81 for Upward in the testing phase. Although it performed well in predicting upward trends, the model was not able to adequately predict downward movements of the index.

In addition, the Random Forest model presented decreases of 0.10 and 0.15 in Downward and Upward precision, respectively, but still maintained results above 0.60, with a final accuracy of 0.68. The Naive Bayes, KNN, and Logistic Regression models did not show substantial declines in their precision values.

Regarding Accuracy in the testing phase, all models achieved values above 0.63. As for the AUC, which provides an aggregate measure of overall model performance, all models exceeded 0.630. However, in terms of computational processing time, models such as Random Forest (10.28 seconds) and SVM (31.55 seconds) required more time for training. On the other hand, Naive Bayes (3.92 seconds), KNN (2.86 seconds), Logistic Regression (2.06 seconds), and Decision Tree (1.21 seconds) presented shorter processing times.

In both academic research and practical applications, the goal is to achieve lower computational cost and faster processing. Therefore, the complexity of the algorithms and the range of parameters used in optimization should be considered—particularly in the case of Random Forest and SVM.

The analyses suggest that an investor could achieve, on average, a 64.5% accuracy rate in predicting the trend of the IBOVESPA index using the predictor variables employed in this study. These results demonstrate the applicability of the tested models in predicting the behavior of the IBOVESPA and indicate their potential for use with other assets in the Brazilian stock market, contributing to risk mitigation in investment operations.

The findings are consistent with Khaidem, Saha, and Dey (2016), who achieved accuracy levels ranging from 84.5% to 94.5% with the Random Forest model, thereby contradicting the Efficient Market Hypothesis (EMH).

Dingli and Fournier (2017) developed a binary variable indicating the direction of price change in the next period (up or down) and achieved 65% precision in predicting the behavior of currencies such as EUR/USD, GBP/USD, and BITCOIN/USD using a Convolutional Neural Network (CNN) model. Although many consider stock market fluctuations to be random, and despite its inherent volatility, the results suggest that market movements are not entirely random. Giacomel (2016) showed that neural networks correctly predicted 55% of upward movements in stocks listed on the U.S. S&P 500, and 60% of downward movements were also correctly predicted for the same market. In addition, Santos et al. (2020) highlight the effectiveness of the Random Forest model in predicting stock market trends. According to the authors, the model achieved 96% accuracy in predicting the stock performance of Vale and 92% for Petrobras.

4. CONCLUSION

The results obtained throughout this study indicate that the Machine Learning algorithms applied to predicted IBOVESPA index trends showed promising performance, despite notable variations among models. Overall accuracy—above 0.63 in all cases—demonstrates the feasibility of using these techniques both in financial market applications and in academic research.

Among the evaluated models, Random Forest (0.661), SVM (0.667), and Logistic Regression (0.692) stood out as the most effective in terms of accuracy. However, although the SVM model achieved good overall performance, it showed limited ability to correctly predict downward trends, with a precision of only 0.53 for that class. Models such as KNN, Naive Bayes, and Logistic Regression produced consistent results, particularly when considering the relationship between performance and computational processing time—an important factor for real-time applications or automated trading systems.

This study aligns with the Adaptive Market Hypothesis (AMH), which acknowledges the existence of temporarily exploitable patterns in financial markets and argues that stock prices can be predictable, allowing for profit through prediction. The superior performance of certain models during specific periods suggests that the market, although generally efficient, may display predictable anomalies, thereby relativizing the Efficient Market Hypothesis (EMH) in its strong form, which assumes absolute unpredictability of prices even when all available information is known.

Despite the achieved advances, this study has limitations. The main one concerns the use of technical indicators based solely on historical price data, which may not fully capture the complexity of the market—especially during exogenous shocks or structural breaks. Furthermore, the analyzed period may contain specific structural biases, requiring caution in generalizing the findings.

Given these results, future research should include macroeconomic and market sentiment variables, as well as additional technical indicators commonly used in Technical Analysis. The use of more advanced algorithms, such as Recurrent Neural Networks (RNNs), is also recommended. Another relevant direction for further investigation involves price reversal analysis, that is, the ability to predict turning points in trends, which could contribute to improved entry and exit strategies in financial markets.

Machine Learning techniques have shown promising results, supporting their continued use in future studies in the investment field. Although they require caution and awareness of their limitations, predictive algorithms can serve as valuable allies to investors.

References

ALLAIRE, J. **Rstudio: integrated development environment for r**. Citeseer, Boston, MA, v. 770, p. 394, 2012.

APPEL, G.; APPEL, M. **A quick tutorial in macd: Basic concepts**. Working Paper, 2008.

ARESTIS, P.; DEMETRIADES, P. O.; LUINTEL, K. B. Financial development and economic growth: the role of stock markets. **Journal of money, credit and banking**, JSTOR, p. 16–41, 2001.
DOI: https://doi.org/10.2307/2673870

AUSTIN, M. Candlesticks 101: Doji. **TECHNICAL ANALYSIS OF STOCKS AND COMMODITIES-MAGAZINE EDITION-, TECHNICAL ANALYSIS, INC.**, v. 22, p. 36–41, 2004.

B3. Índice Bovespa (Ibovespa B3) - **Evolução Diária**. 2021. Disponível em: <http://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/indice-ibovespa-ibovespa-estatisticas-historicas.htm>.

BERNER, C. et al. **Dota 2 with large scale deep reinforcement learning**. arXiv preprint arXiv:1912.06680, 2019.
DOI: https://doi.org/10.48550/arXiv.1912.06680

BOAINAIN, P. G. **“Ombro-cabeça-ombro”:** Testando a lucratividade do padrão gráfico de análise técnica no mercado de ações brasileiro. Insper. 2007.

BODIE, Z.; KANE, A.; MARCUS, A. **Fundamentos de investimentos**. [S.l.]: AMGH Editora, 2014.

BORRA, S.; CIACCIO, A. D. Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods. **Computational statistics & data analysis**, Elsevier, v. 54, n. 12, p. 2976–2989, 2010.
DOI: https://doi.org/10.1016/j.csda.2010.03.004

BREIMAN, L. **Random forests. Machine learning**. Springer, v. 45, n. 1, p. 5–32, 2001.
DOI: http://dx.doi.org/10.1023/A:1010933404324

BREIMAN, L. et al. **Classification and regression trees**. New York: Chapman and Hall/CRC, 2007.
DOI: https://doi.org/10.1201/9781315139470

BURMAN, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. **Biometrika**, Oxford University Press, v. 76, n. 3, p. 503–514, 1989.
DOI: https://doi.org/10.1093/biomet/76.3.503

CAMPOLINA, P.; BATISTA, L. Uma estratégia automatizada de day-trade por meio de comitê de indicadores técnicos. SBIC, **XIV CONGRESSO BRASILEIRO DE INTELIGÊNCIA COMPUTACIONAL**. 2019.

DOI: https://www.doi.org/10.21528/CBIC2019-53

CARVALHO, L. G. P. de; JR, N. C. da C.; GOULART, M. A. de O. Análise técnica versus hipótese dos mercados eficientes: um estudo utilizando o indicador macd. **Revista Alcance**, Universidade do Vale do Itajaí, v. 15, n. 3, p. 398–416, 2008.

CASTRO, H. O. P. **Introdução ao mercado de capitais**. Rio de Janeiro: IBMEC, 1979.

CHANDE, T. S.; KROLL, S. **The new technical trader**: boost your profit by plugging into the latest indicators. [S.l.]: John Wiley & Sons Incorporated, 1994. v. 44.

CROSS, S. E.; WALKER, E. Dart: applying knowledge-based planning and scheduling to crisis action planning. **Intelligent Scheduling**. [S.l.]: Morgan-Kaufmann Publishers, Inc, 1994. 711–729 p.
DOI: https://doi.org/10.1145/201977.1065560

CUNHA, J. P. Z. **Um estudo comparativo das técnicas de validação cruzada aplicadas a Emodelos mistos**. Tese (Doutorado) - Universidade de São Paulo, 2019.
DOI: https://doi.org/10.11606/D.45.2019.tde-26082019-220647

DANTAS, M. **Comportamento da bolsa de valores no brasil diante das crises globais de 2008 e 2020**. Pontifícia Universidade Católica de Goiás, 2020.

DEBASTIANI, C. A. **Análise Técnica de Ações: identificando oportunidades de compra e venda**. [S.l.]: Novatec Editora, 2008.

DEVROYE, L.; WAGNER, T. Distribution-free performance bounds for potential function rules. IEEE **Transactions on Information Theory**, IEEE, v. 25, n. 5, p. 601–604, 1979. DOI: https://doi.org/10.1109/TIT.1979.1056087

EGAN, J. P.; EGAN, J. P. **Signal detection theory and ROC-analysis**. [S.l.]: Academic press, 1975.

ELDER, A. **Como se transformar em um operador e investidor de sucesso**: Entenda a psicologia do mercado financeiro, técnicas poderosas de negociação, gestão lucrativa de investimentos. [S.l.]: Elsevier, 2004.

FACELI, K. et al. **Inteligência artificial:** Uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, v. 2, p. 192, 2011.

FERNANDES, L. S. de B. **Técnicas de aprendizado de máquina aplicadas a Algotrading no mercado de ações**. Universidade Federal do Rio de Janeiro, 2019.

FIX, E.; HODGES, J. **Discriminatory analysis:** nonparametric discrimination, consistency properties. [S.l.]: USAF school of Aviation Medicine, 1951.

Fundação Instituto de Administração. **Mercado financeiro e o coronavírus**: histórico, impactos e projeções. 2020. Disponível em: https://fia.com.br/blog/mercado-financeiro-e-ocoronavirus.

GASTWIRTH, J. L. A general definition of the lorenz curve. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 1037–1039, 1971.
DOI: https://doi.org/10.2307/1909675

GIACOMEL, F. d. S. **Um método algorítmico para operações na bolsa de valores baseado em ensembles de redes neurais para modelar e prever os movimentos dos mercados de ações.** 2016.

GLEN, S. **Roc curve explained in one picture**. Data Science Central [online] Disponível em:https://www.datasciencecentral.com/profiles/blogs/roc-curve-explained-in-one-picture, 2019.

GOLDBERG, J.; NITZSCH, R. V.; MORRIS, **A. Behavioral finance**. [S.l.]: John Wiley Chichester, 2001.

GÓMEZ, S. N. et al. **Random forests estocástico**. Pontifícia Universidade Católica do Rio Grande do Sul, 2012.

GONÇALVES, A. R. **Máquina de vetores suporte**. Universidade Estadual de Campinas, 2015.

GUJARATI, D. N.; PORTER, D. C. **Econometria básica**. 5. Ed. Porto Alegre Amgh Editora, 2011.

HAN, J.; KAMBER, M.; PEI, J. **Data mining:** concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, v. 5, n. 4, p. 83–124, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning:** data mining, inference, and prediction. 2. Ed.[S.l.]: Springer Science & Business Media, 2009.

HENDERSHOTT, T.; RIORDAN, R. et al. **Algorithmic trading and information**. Manuscript, University of California, Berkeley, 2009.

INGARGIOLA, G. **Building classification models:** Id3 and c4. 5. Disponível em: https://cis.temple.edu/giorgio/cis587/readings/id3c45.html, 1996.

KHAIDEM, L.; SAHA, S.; DEY, S. R. Predicting the direction of stock market prices using random forest. arXiv. Disponível em: https://arxiv.org/pdf/1605.00003.pdf, 2016.

KUHN, Max. Building predictive models in R using the caret package. **Journal of Statistical Software**, v. 28, n. 5, p. 1–26, 2008.

DOI: https://doi.org/10.18637/jss.v028.i05. Disponível em: https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

LIAW, A.; WIENER, M. et al. Classification and regression by randomforest. **R news**, v. 2, n. 3, p. 18–22, 2002.

LIGHT, R. J.; MARGOLIN, B. H. An analysis of variance for categorical data. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 66, n. 335, p. 534–544, 1971.
DOI: https://doi.org/10.1080/01621459.1971.10482297

LIMA, A. R. G. **Máquinas de vetores suporte na classificação de impressões digitais**. Universidade Federal do Ceará, Departamento de Computação, Fortaleza-Ceará, 2002.

LO, A. W. **Adaptive markets**: Financial evolution at the speed of thought. [S.l.]: Princeton University Press, 2019.

MACHADO, F. N. R. **Big Data O Futuro dos Dados e Aplicações**. [S.l.]: Saraiva Educação SA, 2018.

MADEO, R. C. B. et al. Papel estratégico e impacto dos sistemas de informação no mercado de ações: Um estudo envolvendo brasil e estados unidos. **Revista Eletrônica de Sistemas de Informação**, v. 11, n. 2, 2012.
DOI: https://doi.org/10.5329/RESI.2012.1102006

MALKIEL, B. G. The efficient market hypothesis and its critics. **Journal of economic perspectives**, v. 17, n. 1, p. 59–82, 2003.
DOI: https://doi.org/10.1257/089533003321164958

MALKIEL, B. G.; FAMA, E. F. Efficient capital markets: A review of theory and empirical work. **The journal of Finance**, Wiley Online Library, v. 25, n. 2, p. 383–417, 1970.
DOI: https://doi.org/10.2307/2325486

MARTINS, M. A. d. S. A sonhada letra"v". **Análise: conjuntura nacional e Coronavírus**. FCE/UFRGS. Porto Alegre. 25 nov., 2020.

MCCLISH, D. K. Analyzing a portion of the roc curve. Medical Decision Making, **Sage Journals**. Thousand Oaks, CA, v. 9, n. 3, p. 190–195, 1989.
DOI: https://doi.org/10.1177/0272989X8900900307

MEIRA, C. A.; RODRIGUES, L. H.; MORAES, S. de. Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. Embrapa Informática Agropecuária, **Tropical Plant Pathology**, Brasília, DF, v. 33, n. 2, p. 2-13, Mar./Apr. 2008.
DOI: https://doi.org/10.1590/S1982-56762008000200005

MELONI, R. B. da S. **Classificação de Imagens de Sensoriamento Remoto usando SVM**. Tese (Doutorado) — PUC-Rio, 2009.

MOISEN, G. Classification and regression trees. In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). **Encyclopedia of Ecology**, volume 1. Oxford, UK: Elsevier. p. 582-588., p. 582–588, 2008.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. [S.l.]: Saraiva Educação SA, 2017.

MOZZER, M. **Moving Average Convergence Divergence: funcionamento matemático e comportamento em tendências**. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2015.

MURPHY, J. J. **Technical analysis of the financial markets**: A comprehensive guide to trading methods and applications. New York, 1999.

MYLNE, K. R. Decision-making from probability forecasts based on forecast value. **Meteorological Applications**, Wiley Online Library, v. 9, n. 3, p. 307–315, 2002. DOI: https://doi.org/10.1017/S1350482702003043

PORTNOY, K. et al. **High frequency trading and the stock market**: A look at the impact of trade volume on stock price changes. 2011.

PRATI, R. et al. Curvas roc para avaliação de classificadores. **Revista IEEE América Latina**, v. 6, n. 2, p. 215–222, 2008.

RHEA, R. **The Dow theory**: An explanation of its development and an attempt to define its usefulness as an aid in speculation. [S.l.]: Fraser Publishing Company, 1993.

ROBIN, Xavier et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, v. 12, p. 77, 2011.
DOI: https://doi.org/10.1186/1471-2105-12-77

ROCKEFELLER, B. **Technical analysis for dummies**. [S.l.]: John Wiley & Sons, 2019.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. [S.l.]: Elsevier, 2004.

RYAN, Jeffrey A.; ULRICH, Joshua M. quantmod: Quantitative Financial Modelling Framework. 2025. R package version 0.4.27.
Disponível em: https://CRAN.R-project.org/package=quantmod.
DOI: https://doi.org/10.32614/CRAN.package.quantmod.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of research and development**, IBM, v. 3, n. 3, p. 210–229, 1959.
DOI: https://doi.org/10.1147/rd.33.0210

SANTOS, G. C. et al. **Algoritmos de machine learning para previsão de ações da b3**. Universidade Federal de Uberlândia, 2020.

SCHMITT, J. et al. **Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo**. Florianópolis, SC, 2005.

SCHNEIDER, F. **Variação do índice ibovespa no ano de 2008**. Pontifícia Universidade Católica de Goiás, 2020.

SILVA, I. G. A. et al. **Análise técnica**: aplicação do setup operacional de debastiani nas ações do banco do brasil e vale sa. Universidade Federal de Campina Grande, 2018.

SKIENA, S. S. **The data science design manual**. [S.l.]: Springer, 2017.
DOI: https://doi.org/10.1007/978-3-319-55444-0

TAKAMATSU, R. T.; LAMOUNIER, W. M.; COLAUTO, R. D. Impactos da divulgação de prejuízos nos retornos de ações de companhias participantes do ibovespa. **Revista Universo Contábil**, v. 4, n. 1, p. 46–63, 2008.

TEAM, R. C. R core team. **r: A language and environment for statistical computing**. Foundation for Statistical Computing, 2019. Disponível em: https://www.R-project.org/.

ULRICH, Joshua. *TTR:* Technical Trading Rules. 2023. R package version 0.24.4. Disponível em: https://CRAN.R-project.org/package=TTR.
DOI: https://doi.org/10.32614/CRAN.package.TTR.

VAPNIK, V. The nature of statistical learning theory. Berlin: **Springer science & business media**, 2013.

WILDER, J. W. **New concepts in technical trading systems**. [S.l.]: Trend Research, 1978.

WITTEN, I. H. et al. **Data Mining**: Practical machine learning tools and techniques. Morgan Kaufmann, p. 578, 2005.

WONG, W.-K.; MANZUR, M.; CHEW, B.-K. How rewarding is technical analysis? evidence from singapore stock market. **Applied Financial Economics**, Taylor & Francis, v. 13, n. 7, p. 543–551, 2003.
DOI: https://doi.org/10.1080/0960310022000020906

ZHOU, X.-H.; MCCLISH, D. K.; OBUCHOWSKI, N. A. Statistical methods in diagnostic medicine. Journal of Biopharmaceutical Statistics, 22:3, 612-614. 2012.
DOI: https://doi.org/10.1080/10543406.2012.646580