doi:10.12662/2359-618xregea.v12i1.p121-137.2023

ARTIGOS

INVESTMENTS IN TIMES OF UNCERTAINTY: FORMATION OF PORTFOLIOS USING RANDOM FOREST

INVESTIMENTOS EM TEMPOS DE INCERTEZA: FORMAÇÃO DE PORTEFÓLIOS USANDO FLORESTA ALEATÓRIA

José Erasmo Silva
jose.erasmo@natelcontact.com.br
Doutor em administração pela Universidade Metodista de Piracicaba. Professor no MBA USP/Esalq. Piracicaba - SP - BR.

Maria Imaculada de Lima Montebello
milmonte50@gmail.com
Doutora em Agronomia pela Escola Superior de Agricultura Luiz de Queiroz. Professora dos Programas de Pós-Graduação da Universidade Metodista de Piracicaba. Piracicaba - SP - BR.

Jorge Luiz dos Santos Silva
jluiztc@gmail.com
Doutor em Administração na Universidade Metodista de Piracicaba (UNIMEP). Professor da Universidade Vale do Rio Verde (UNINCOR). Três Corações - MG - BR.

ABSTRACT

Objective: based on a systematic approach using machine learning, this research aims to propose a model of selection and allocation of assets that allows for building profitable and safe portfolios, even in times of insecurity and low predictability.

Methodology: we used the machine learning algorithm called random forest to associate the independent variables with a dependent one and learn the probability of positive returns in the month following the data collection. According to the probabilities, the stocks were allocated into long, short, or non-allocated portfolios. Finally, we allocated a share of gold, which is a protection asset much used in times of crisis and uncertainty.

Results and contributions: the study reached its goal and demonstrated being possible to build profitable and safe investment portfolios, even in times of greater uncertainty and volatility, as in 2020 due to the Covid-19 pandemic. We found that the model is effective in moments of crisis and also of greater predictability, as in the period from 2016 to 2019 when the stock exchange has an uptrend.

Relevance: the relevance of this study points to an unprecedented historical context in Brazil, where uncertainties regarding both the local and world economy have demanded advanced studies of prediction to minimize risks and contribute to results for investors. In addition, we highlight that following a short period of low Selic (2019 to 2021), the Central Bank increased the rate again, raising the interest more in profitable and safer assets than the investment in stocks.

Impact on the area: the study has a positive impact on the finance area since studies in this field promote greater stability to investors

and thus better capital flow for companies, which, in turn, contribute to society and the growth of the country.

Keywords: portfolios; machine learning; random forest.

RESUMO

Objetivo: com base em uma abordagem sistemática utilizando o aprendizado de máquinas, esta pesquisa visa propor um modelo de seleção e alocação de ativos que permita construir portefólios lucrativos e seguros, mesmo em tempos de insegurança e baixa previsibilidade.

Metodologia: utilizamos o algoritmo de aprendizagem de máquinas chamado floresta aleatória para associar as variáveis independentes a uma dependente e aprender a probabilidade de retornos positivos no mês seguinte à coleta de dados. De acordo com as probabilidades, os estoques foram alocados em portefólios longos, curtos, ou não alocados. Finalmente, alocamos uma parte de ouro, que é um ativo de proteção muito utilizado em tempos de crise e incerteza.

Resultados e contribuições: o estudo atingiu seu objetivo e demonstrou ser possível construir portefólios de investimentos rentáveis e seguros, mesmo em tempos de maior incerteza e volatilidade, como em 2020, devido à pandemia de Covid-19. Constatamos que o modelo é eficaz em momentos de crise e também de maior previsibilidade, como no período de 2016 a 2019, quando a bolsa tem uma tendência de alta.

Relevância: a relevância deste estudo aponta para um contexto histórico sem precedentes no Brasil, onde as incertezas tanto da economia local como mundial exigiram estudos avançados de previsão para minimizar os riscos e contribuir para os resultados para os investidores. Além disso, destacamos que após um curto período de baixa Selic (2019 a 2021), o Banco Central aumentou novamente a taxa, elevando o interesse em ativos mais rentáveis e seguros do que o investimento em ações.

Impacto na área: o estudo tem um impacto positivo na área financeira, uma vez que os estudos neste campo promovem maior estabilidade aos investidores e, portanto, melhor fluxo de capital para as empresas, que, por sua vez, contribuem para a sociedade e o crescimento do país.

Palavras-chave: portefólios; aprendizado de máquinas; floresta aleatória.

1 INTRODUCTION

The number of investors in the Brazilian stock exchange has increased briskly, reaching 5 million registered CPFs, according to data by the B3 (2022). At the same time, there has been an increase in publications in the finance area using approaches of machine learning and artificial intelligence (KOLANI, 2022). Such types of approaches allow for capturing linear and non-linear behaviors (AHMED *et al.*, 2022).

Currently, the stock market has become one of the main fields of investment. In this context, tools of machine learning and artificial intelligence have raised the interest of both researchers and investors. Accurate predictions for the fluctuation of stocks in the market may reduce risks and generate abundant returns (JI *et al.*, 2022)

Many factors influence the price of stocks, including supply and demand, market trends, local and global economy, companies outcomes, historical prices, news in general (positive or negative), confidential financial information, and company popularity, which may lead to whether an increase or decrease of traders (HO; DARMAN; MUSA, 2021; SRIVINAY *et al*., 2022).

Despite the growth of studies seeking to establish market predictions, some researchers

argue that the market is unpredictable. Such an idea of the unpredictability of the stock market is supported by the hypothesis of efficient markets, which is one of the pillars of the classical theory of finances, proposed by Fama in 1970 (FAMA, 1998) This hypothesis assumes that the prices of a stock incorporate all information available and are based on the rational expectations of investors who seek to enhance their profits. Fama (1995) also argues that the price of stocks follows a random way. Therefore, price changes are independent and historical prices cannot be used to predict the future price of stocks (JI *et al.*, 2022).

However, some studies have revealed that emotions are often part of the decision-making process and that such decisions are irrational and may cause disasters in the financial market. In addition, not all actors in the market are fully informed at the same time in such a way that decisions are taken from different perspectives and at different moments (KAPOOR; PROSAD, 2017) (FELIZARDO *et al*., 2022).

As shown, predicting market fluctuations is not an easy task, and investing in the stock market without a defined strategy may generate great losses for the investor. Paiva *et al*. (2019) the model was developed using a fusion approach of a classifier based on machine learning, with the support vector machine (SVM describe the stock market as dynamic, complex, evolutionary, non-linear, fuzzy, non-parametric, and chaotic by nature. In addition, according to these authors, it is extremely sensitive to political factors, microeconomic and macroeconomic conditions, as well as the expectations and insecurities of investors.

The stock market plays a fundamental role in the economic system of a country and allows its actors, like investors and traders, to enhance their wealth by investing in stocks. However, preventive measures must be taken to cope with the risks of this environment, where the value of investments may increase or decrease according to its conditions (ISMAIL *et al.*, 2020).

Based on a systematic approach using machine learning, our goal is to propose a model of selection and allocation of assets that allows for building profitable and safe portfolios, even in times of insecurity and low predictability.

To reach our goal, we used the random forest algorithm in a monthly selection of stocks. In the context of long and short, the stocks with greater probabilities of positive returns were bought to the long position, and the stocks with lower probabilities were sold to the short position of the portfolio. This method is aimed at protecting the portfolio from changes in market trends. Finally, also as a protection strategy, we allocated a percentage of gold in the portfolio aiming to, especially in moments of crisis, protect the investor's capital. The results proved satisfactory, especially for the first semester of 2020, when investors and people, in general, experienced the uncertainties caused by the Covid-19 pandemic.

In addition to this introduction, this paper presents four other sections. Section 2 introduces a review of the theoretical background and summarizes recent studies on machine learning and the prediction of a stock price. Section 3 describes the methodology, sample, and dependent and independent variables, as well as the metrics used to assess the portfolios and the proposed model. Section 4 presents our results and suggestions for further studies. Finally, section 5 introduces our final remarks.

2 THEORETICAL BACKGROUND

The increasing complexity and the dynamic nature of the stock markets are the main challenges of the financial sector, leading to the projection of inflexible strategies by experienced financial professionals who cannot achieve satisfactory performance in all market conditions (WU *et al.*, 2020).

At least from two perspectives, the way of investing has changed over the last decades. Firstly, financial information in real-time has been much more accessible, especially due

to new technologies that promote not only wide and fast access to the internet but also allow for greater efficiency in data analysis. Secondly, artificial intelligence (IA), supported by technological advances, has guided investors toward a new way of investing, where decisions are taken based on the analysis of a large volume of quantitative information (CERVELLÓ-ROYO; GUIJARRO, 2020).

With the development of AI, big data, and the technology of cloud computing, people have become gradually familiarized with quantitative investment, which issues negotiation instructions through quantitative methods and computer programming (WANG *et al*., 2020a). Over the last decades, quantitative investment has become a hot spot in the market development in capitals in Europe and the United States. The market scale for quantitative investment reached 70% of the investment market in the USA, showing stable performance and becoming a new method for investors (WANG *et al.,* 2020a).

Despite the increasing popularization of the topic, developing a consistent and accurate method of prediction for the stock market is not an easy task since many factors, such as wars, and politics, among others, affect the feelings of investors, therefore affecting the market behavior (CHEN, 2020). The stock market has a dynamic, complex, evolutionary, non-linear, cloudy, non-parametric, and chaotic nature, in addition to being extremely sensitive to political factors, microeconomic and macroeconomic conditions, as well as the expectations and insecurities of investors (PAIVA *et al*., 2019)the model was developed using a fusion approach of a classifier based on machine learning, with the support vector machine (SVM. To better cope with such a noisy environment, recently more and more researchers have focused on applications and approaches that use AI (JIANG *et al*., 2020).

From the perspective of such an increasing evolution and accessibility of new investment models, we understand that old strategies should be reviewed and evolve, as well as new ones should be created. The diversity of strategies positively contributes to greater market stability, including in moments of crisis and recession. When many investors group around the same strategy, returns, up to a point, can be together with positive, however, such a behavior can be devastating in moments of crisis (AVRAMOV *et al*., 2015).

Currently, performing predictions based on methods of machine learning is regarded as a much more effective solution to such a challenge (NIU *et al.,* 2020) Machine learning is a branch of AI that includes many models (or algorithms) that have evolved to meet different demands.

An advanced model of machine learning that has reached good results in the financial market and called the attention of investors and researchers is known as random forest, which derives from decision trees and is aimed at improving their precision and overcoming the high sensitivity to small alterations in data (FISCHER; KRAUSS, 2018;yet inherently suitable for this domain. We deploy LSTM networks for predicting out-of-sample directional movements for the constituent stocks of the S&P 500 from 1992 until 2015. With daily returns of 0.46 percent and a Sharpe ratio of 5.8 prior to transaction costs, we find LSTM networks to outperform memory-free classification methods, i.e., a random forest (RAF RAMASUBRAMANIAN; SINGH, 2017).

Ho (1995) developed the first proposal to expand the complexity of models of decision trees and improve the accuracy not only for the training period but especially for the test period. The Random Forest model was then created, whose essence is to build multiple trees in randomly selected subspaces from a larger set. In other words, the Random Forest is a supervised learning algorithm that uses combined decision trees so that each tree depends on the values of a random independent vector and with equal distribution to all trees in the forest (BREIMAN, 2001).

Such a powerful machine learning algorithm of the ensemble type can be applied to regression problems but stands out, especially in ranking issues. It is an update of the concept of

decision trees and emerged to solve the problem of overfitting, much more common when using tree isolate (RAYA *et al*., 2022).

The results from the predictive process of this algorithm mostly derive from trees in ranking issues, as well as an average prediction in regression problems, according to Tratkowski (2020). The author also highlights that the random forest algorithm can be used in complex economic environments, considered useful in the formulation of investment strategies.

The random forest combines versatility and power in a single machine-learning approach. As the set uses only a small and random part of the full set of resources, the algorithm can cope with extremely large datasets (LANTZ, 2019).

Wang *et al.* (2020a) combined a random forest algorithm with a model of stock selection with multiple factors, which generated a more universal and adaptable strategy for stock selection. They collected financial information from 2012 to 2020 from the companies listed in the CSI 300, one of the main stock indices in China. From the model, tests were run with 100, 500, and 1,000 decision trees, generating, respectively, yearly returns of 17.80%, 29.40%, and 24.50%. Thereby, the best return was obtained by using 500 decision trees, however, all of them proved rather superior to the CSI 300, which had a yearly return of 5.80%. The results were also superior concerning some quantitative funds in the market.

Raya *et al. (*2022) tested the models of support vector regression (SVR), long short-term memory network (LSTM), XGBoost, autoregressive integrated moving average model (ARIMA), and random forest (RF) to predict the closing prices of the Vanguard Total Stock Market Index Fund between 2015 and 2020. The best result was generated using the LSTM with a root mean squared error (RMSE) of 1,6149. From the perspective of mean absolute error (MAE), the RF produced the best result, with 1,344.

Zhu (2020) used the XGBoost and RF models to predict the monthly returns of 300 stocks listed in the stock exchanges of Shanghai and Shenzhen and reached 78% of accuracy for the former and 72% for the latter.

Abraham *et al.* (2022)based on the Genetic Algorithm (GA proposed a combined study of two models: one with the selection of features using a genetic algorithm, and another of ranking using RF. The authors reached 80% of accuracy in the prediction of stock trends listed in the S&P 500 and the CAC40.

Levantesi and Piscopo (2020) compared the performance of the RF model with a generalized linear model (GLM) for predictions in the real estate market in the city of London. The results were more accurate when using the RF.

Yin *et al.* (2021) used decision trees and RF. Despite seeming redundant to use both together, the approach of the authors consisted of using the decision trees to select features that better represent the sample for further establishing the predictions using the RF. The results proved promising when the model was tested in four stocks listed in the American market.

Ma, Han and Wang (2021) tested the models of RF, SVR, LSTM, neural network (NN), deep multilayer perceptron (DMLP), and convolutional neural network (CNN) to form portfolios. Therefore, the algorithms of machine learning and deep learning were used to select the stocks, and their outputs were inserted into two optimization models of portfolios: mean-variance (MV) and omega. The strategy was applied to the Chinese market between 2007 and 2015 and the best results were obtained from the RF-MV set.

Wang *et al.* (2020b) tested the long models of LSTM, SVR, RF, deep neural networks (DNN), and ARIMA. The authors found that with a wide advantage margin, the LSTM presented a better performance in predicting the stocks listed on the stock exchange of the United Kingdom from 1994 to 2019.

Ballings *et al.* (2015) tested the algorithms RF, AdaBoost, Kernel Factory (KF), NN, Logistic Regression (LR), SVM, and K-Nearest Neighbor (KNN) to predict the direction of the prices of 5767 European

companies of open capital. The authors used the AUC as an assessment metric and concluded that the RF algorithm showed the best performance in this function, followed by SVM, KF, AdaBoost, NN, KNN, and LR.

As mentioned in the previous paragraphs, many machine learning-based approaches have been used to predict the return of stocks. However, to reach our goals, we sought to go beyond such a prediction and also apply a technique of resource allocation aiming at the profitability and protection of the invested capital, according to Wang *et al.* (2020b), and Harvey *et al.* (2019)a popular question arises: What steps can an investor take to mitigate the impact of the inevitable large equity correction? However, hedging equity portfolios is notoriously difficult and expensive. We analyze the performance of different tools that investors could deploy. For example, continuously holding short-dated S&P 500 put options is the most reliable defensive method but also the most costly strategy. Holding 'safe-haven' US Treasury bonds produces a positive carry, but may be an unreliable crisis-hedge strategy, as the post-2000 negative bond-equity correlation is a historical rarity. Long gold and long credit protection portfolios sit in between puts and bonds in terms of both cost and reliability. Dynamic strategies that performed well during past drawdowns include: futures time-series momentum (which benefits from extended equity sell-offs. This model is known in the market as a hybrid.

Harvey *et al.* (2019)a popular question arises: What steps can an investor take to mitigate the impact of the inevitable large equity correction? However, hedging equity portfolios is notoriously difficult and expensive. We analyze the performance of different tools that investors could deploy. For example, continuously holding short-dated S&P 500 put options is the most reliable defensive method but also the most costly strategy. Holding 'safe-haven' US Treasury bonds produces a positive carry, but may be an unreliable crisis-hedge strategy, as the post-2000 negative bond-equity correlation is a historical rarity. Long gold and long credit protection portfolios sit in between puts and bonds in terms of both cost and reliability. Dynamic strategies that performed well during past drawdowns include: futures time-series momentum (which benefits from extended equity sell-offs tested some important strategies in the American market aiming to learn which was more efficient in mitigating losses during times of crisis or recession. In addition to testing classical strategies of protection, which are effective but erode a substantial part of the gains, they also sought to test two different strategies aimed at reducing the protection costs. The first of those is the *momentum* strategy of time series, which presented good performance both in moments of crisis and recession. The second strategy is the 'long and short' in stocks, which used different quality metrics to classify the companies cross-sectionally. This latter strategy achieved the best results, which proved satisfactory throughout the period studied (1985-2018), especially in moments of crisis and recession.

The investment strategies known as 'long and short' have been increasingly popularizing for allowing investors to work with a large number of assets more efficiently than with the 'long only' models – portfolios that allocate only bought stocks (GRINOLD; KAHN, 2000). Lewin and Campani (2020) tested the result of these strategies based on the formation of portfolios using a model introduced by Hamilton (1989), called Markov chains. The results showed that the model that allowed no short sales (long-only) generated an award of yearly risk of 4.66%, with a volatility of 7.62%, and a Sharpe index of 0.61%. The model with short sales (long and short) presented a significantly greater yearly result, with an award of 36.90%, volatility of 33.33%, and Sharpe index of 1.11%. Despite the large award of yearly risk in the portfolio, there is also some large volatility.

For Jiao, Massa, and Zhang (2016), these models also contribute to reducing the investment cost since the short position of the portfolio is a resource to cover other positions.

Beaver, McNichols and Price (2016) pointed out that despite the long and short strategy per se do not contribute to a good performance in the portfolio, the protection against an eventually unfavorable direction in the market, the reduction in the investment cost, and diversification of such strategy makes it much interesting and adopted by investors and managers of investment funds.

Finally, also protection measures for the portfolio simulated herein, especially considering the turbulences experienced in the market in 2020 due to the Covid-19 pandemic, we allocated a share of gold in the referred portfolio. It is worth highlighting that gold is used by many investors to protect investment portfolios, thus justifying its influence on our portfolios (SHABBIR; KOUSAR; BATOOL, 2020). Whether physical or electronic negotiation, gold is one of the most used assets for protection (ALSHAMMARI *et al.*, 2020; HARVEY *et al.*, 2019;a popular question arises: What steps can an investor take to mitigate the impact of the inevitable large equity correction? However, hedging equity portfolios is notoriously difficult and expensive. We analyze the performance of different tools that investors could deploy. For example, continuously holding short-dated S&P 500 put options is the most reliable defensive method but also the most costly strategy. Holding 'safe-haven' US Treasury bonds produces a positive carry, but may be an unreliable crisis-hedge strategy, as the post-2000 negative bond-equity correlation is a historical rarity. Long gold and long credit protection portfolios sit in between puts and bonds in terms of both cost and reliability. Dynamic strategies that performed well during past drawdowns include: futures time-series momentum (which benefits from extended equity sell-offs NAWAZ; AZAM; ASLAM 2020; SAMUEL, 2020).

The combination of different assets in different economic moments provides, as largely studied in finances, positive results from diversification (ALSHAMMARI *et al.*, 2020; HARVEY *et al.*, 2019;a popular question arises: What steps can an investor take to mitigate the impact of the inevitable large equity correction? However, hedging equity portfolios is notoriously difficult and expensive. We analyze the performance of different tools that investors could deploy. For example, continuously holding short-dated S&P 500 put options is the most reliable defensive method but also the most costly strategy. Holding 'safe-haven' US Treasury bonds produces a positive carry, but may be an unreliable crisis-hedge strategy, as the post-2000 negative bond-equity correlation is a historical rarity. Long gold and long credit protection portfolios sit in between puts and bonds in terms of both cost and reliability. Dynamic strategies that performed well during past drawdowns include: futures time-series momentum (which benefits from extended equity sell-offs NAWAZ; AZAM; ASLAM, 2020; SAMUEL, 2020; SHABBIR; KOUSAR; BATOOL, 2019). Even though many studies have confirmed the role of gold as a capital protection asset, some researchers are yet to find the same results, including Maghyereh and Abdoh (2020).

3 METHODOLOGY

To reach our goal, as aforementioned, we used the machine learning algorithm called random forest to associate the independent variables with the dependent one and obtain the probability of positive returns in the month following the data collection. Subsequently, the stocks were allocated based on a method known as long and short. Finally, we allocated in the portfolio a share of gold, which is a protection asset much used in times of crisis and uncertainty.

In the context of predictions, a great challenge is to learn the relationship between past and future information. Generally,

the prediction of stock prices uses models of statistics, mathematics, economy, and machine learning in general (JAYAPALAN; SOMASUNDARAM, 2020). There has been strong evidence that identifying a behavior pattern allows for designing a prediction model (PAIVA *et al.*, 2019) the model was developed using a fusion approach of a classifier based on machine learning, with the support vector machine (SVM.

Monthly information from the companies belonging to the Bovespa index – the main benchmark of variable income in Brazil, was collected from the Economatica platform, referring to the period from April 2002 to April 2022. The index was created in 1968 and gathers the most important companies in the market of Brazilian B3 capitals (2022). According to data on the Economatica platform (2022), based on a historical period since 1997, it is currently composed of 91 stocks, one of the largest numbers in the portfolio, reaching 93 stocks in the same year.

Researchers have reported that the larger the volume of data collected, the more accurate the machine learning-based models are (CERVELLÓ-ROYO; GUIJARRO, 2020; FISCHER; KRAUSS, 2018). Thereby, we created a scheme of train and testing ("cross-validation") as follows: 120 months of a train and the model was tested in the following month. Subsequently, we added a month to the training period and tested the following month, always maintaining 120 months as a train and one month as a test. Table 1 exemplifies the scheme of the periods of train and testing.

Table 1 - Periods of train and test of the algorithm

Model	Train	Test
1	04/2002 to 03/2012	04/2012
2	05/2002 to 04/2012	05/2012
3	06/2002 to 05/2012	06/2012
4	07/2002 to 06/2012	07/2012
...	...	...
138	04/2012 to 03/2022	04/2022

Source: elaborated by the author.

Based on the probabilities extracted from the models and following the logic proposed by Krauss, Do and Huck (2017)machine learning research has gained momentum: new developments in the field of deep learning allow for multiple levels of abstraction and are starting to supersede well-known and powerful tree-based techniques mainly operating on the original feature space. All these methods can be applied to various fields, including finance. This paper implements and analyzes the effectiveness of deep neural networks (DNN, long and short portfolios were formed. The long position of the portfolio allocated the 10 greatest probabilities of positive returns provided that these reached at least 50%. The short position of the portfolio allocated the 10 lowest probabilities provided that these reached up to 49.9%. It is worth emphasizing that such a procedure was repeated every month during the 120 months of the period studied, in an automated manner, on the Python software.

The allocation of the stocks into the long and short format aims to protect the portfolio from the changes in market trends, especially in moments of crisis with high volatility. In addition, such a format reduces investment costs by promoting resources with short-sold stocks (BEAVER; MCNICHOLS; PRICE, 2016; GRINOLD; KAHN, 2000; LEWIN; CAMPANI, 2020). We also resorted to another protection technique, the allocation of a share of the portfolio in gold, which is largely known as a protection asset (ALSHAMMARI *et al*., 2020; HARVEY *et al.*, 2019; NAWAZ; AZAM; ASLAM, 2020; SAMUEL, 2020).

The variables used in the study were raised from the aforementioned studies. Endogenous variables were used in addition to exogenous variables aiming to detect, through the predictions, fluctuations caused by changes in the companies and the market in general. The use of macroeconomic variables has been highlighted in studies of portfolio allocation, as well as indices representing the market and the price of commodities, which usually reflect the world economic behavior (SETIADI;

MASDUPI, 2018). The dependent variable used herein study is a dummy that was attributed with 1 if the return following the data dissemination was positive and 0 if the return following the data dissemination was negative. Table 2 presents the independent variables.

Table 2 - Independent variables

Variable	Description
Ano	Year referent to the data collected
CDI 252 DIAS	Certificate of Interbank Deposit
CDI CUMULATIVE	Monthly Certificate of Interbank Deposit
Dolar	Monthly Dollar versus Real Return
IBOV	Monthly Ibovespa Return
IBXX	Monthly IBRX Brazil Return
INCC DI TX MES	National Index of construction cost
INCC_DI	National Index of construction cost
IPC	Index of prices to the consumer
LFT	Monthly LFT return
LP	Profit on the closing price of the last stock exchange trading floor in the month
Month	Month referent to the data collected
OTNBTN	Monetary correction
OZ1D	Monthly gold return
PBV	Price on equity
PE	Price of the last stock exchange trading floor in the month over the profit of the last trimester
PEBITDA	Price on the EBITDA
PFCF	Price on free cash flow
POUP_TAXA_MES	Rate of savings account
POUPACUM	Monthly Return of savings account
Presence	Presence in the stock exchange
Return	Stock return lagged by 1 to 12 months

Note: Variables used herein that appear in previous studies.

Source: elaborated by the author.

The indicators in Table 2 are mostly monthly, except for the indicators that depend on the trimestral dissemination of the companies, such as the EBITDA. In this case, the P/EBITDA was calculated based on the price of the last train day of the model divided by the last EBITDA provided by the company.

In addition to selecting adequate variables, Sun *et al*. (2020) highlight that the accuracy of the models strongly depends on the hyperparameters selected for each of them. Therefore, before executing an algorithm, the configurations of hyperparameters that most contribute to the results must be established. Herein, we used most of the default values, except those presented in Table 3, which were defined from a grid search. This allowed for obtaining better accuracy levels.

Table 3 - Hyperparameters used with the Random Forest model

Param	Value	Description
bagging_fraction	0.9	Fraction of data to be used for each iteration
bagging_freq	1	Bagging after estimating the number of trees
boost_rounds	500	Number of trees to build
boosting_type	rf	Boosting scheme
feature_fraction	0.3	Fraction of features to be taken for each iteration
learning_rate	0.1	Boosting learning rate
min_data_in_leaf	500	Minimum amount of data per leaf
num_leaves	31	Parameter to control the complexity of the tree model
objective	binary	Learning objective
test_length	1	Test length
train_length	120	Train length

Source(s): elaborated by the author.

The models were assessed through the metrics of roc_auc, accuracy, precision, recall, and F1. The portfolios were assessed based on the mean monthly return, cumulative return, risk (standard deviation), and Sharpe index, which is much used to assess the risk-return of portfolios. It is calculated by dividing the mean of the excess return by the standard deviation of the portfolio (PAI; ILANGO, 2020).

After running the models, the stocks were ranked. Thus, those classified as positive returns with actual positive values are called true positives (TP). The stocks classified with positive returns but with negative returns are called false positives (FP). The stocks classified as negative returns with positive returns are called false negatives (FN). Finally, the stocks classified as negative returns with actual negative results are called true negative (TN). Such denominations allowed for building the following indicators:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP+FN} \quad (7)$$

$$\text{F1 Score} = 2\frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

Accuracy shows the percentage of accurate predictions among all predictions, that is, events and non-events. In turn, precision reveals the predictions of accurate events among the precisions performed as events. Both measures are very interesting to be used in the stock market, the former being used to assess both long and short positions and the latter only to assess long positions. The recall is the indicator that shows the percentage of hits when considering the sum of true positives and false negatives. Finally, the F1 score is a harmonic mean established from the precision and recall metrics. In addition, we also assessed the "Area Under the ROC Curve" (AUC).

To understand the concepts of the ROC and AUC curves, we must understand the concept of specificity. According to Fávero and Belfiore (2017), specificity refers to the percentage of hits for a given cutoff by considering only the observations that are not events, as follows:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (9)$$

The ROC curve shows the behavior per se of the tradeoff between sensitivity and specificity and the convex shape concerning the point (0.1) by bringing the values of (1-"specificity") on the abscissa axis. Therefore, a given model with a larger area below the ROC curve presents greater overall efficiency of prediction (FÁVERO; BELFIORE, 2017).

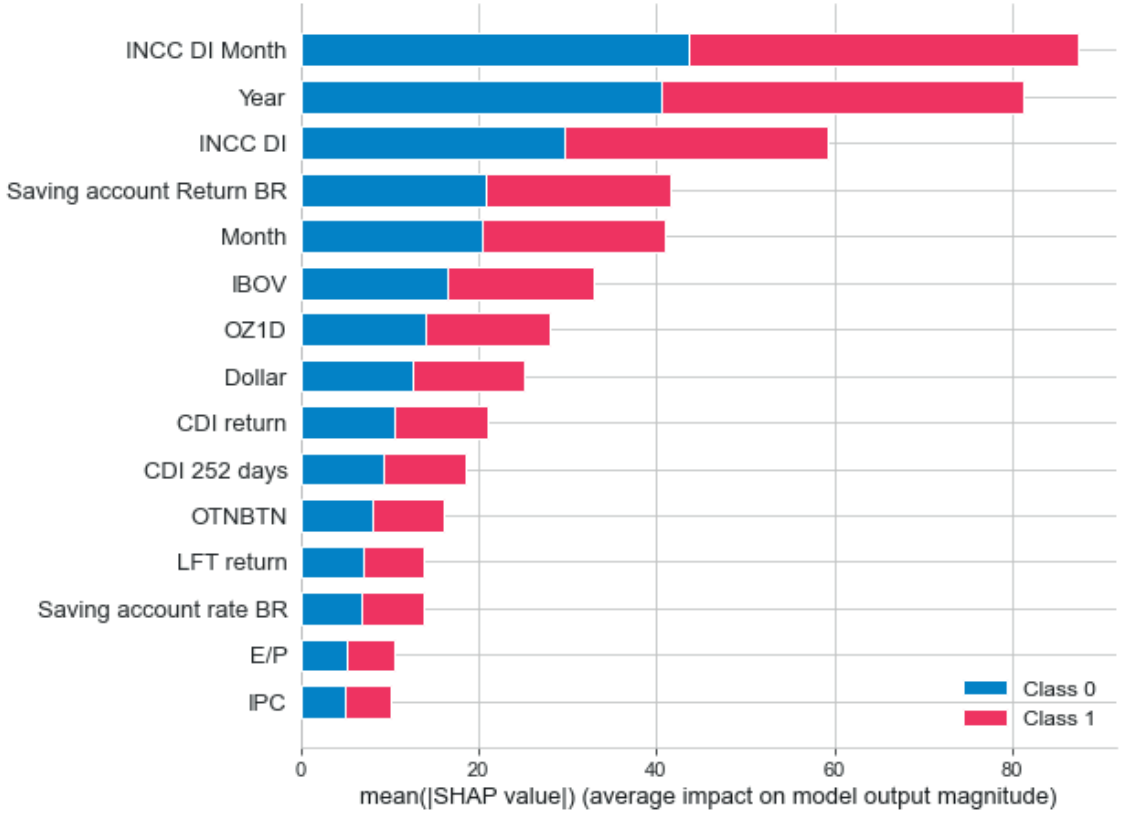
We do not intend to create a fixed model with long-term predictive capacity, that is, with the model crashing at a given moment, our goal is to build a dynamic model that is updated monthly based on the market information for long-term application. Considering the dynamics of the Brazilian stock market, it is important to support the most recent updates to preserve the model's accuracy.

4 RESULTS

Our results are presented in the next paragraphs. Figure 1 shows the variables with relevance above 5. Although machine learning models work well with many variables, it is suggested that variables of lower relevance are disregarded since they can generate more noise than results (BREIMAN, 2001).

The INCC DI stands out as the most important variable and generally has the strongest impact on stocks in construction industry companies. Still, only one variable on the company level (profit over price) presented relevance above 5. The prices of the Brazilian companies of open capital are more guided by external than internal factors (BRESSER-PEREIRA, 2012; PIMENTA; HIGUCHI, 2008; BERNARDELLI; BERNARDELLI; CASTRO, 2017; BALZANA FILHO; BORDEAUX-REGO, 2013)

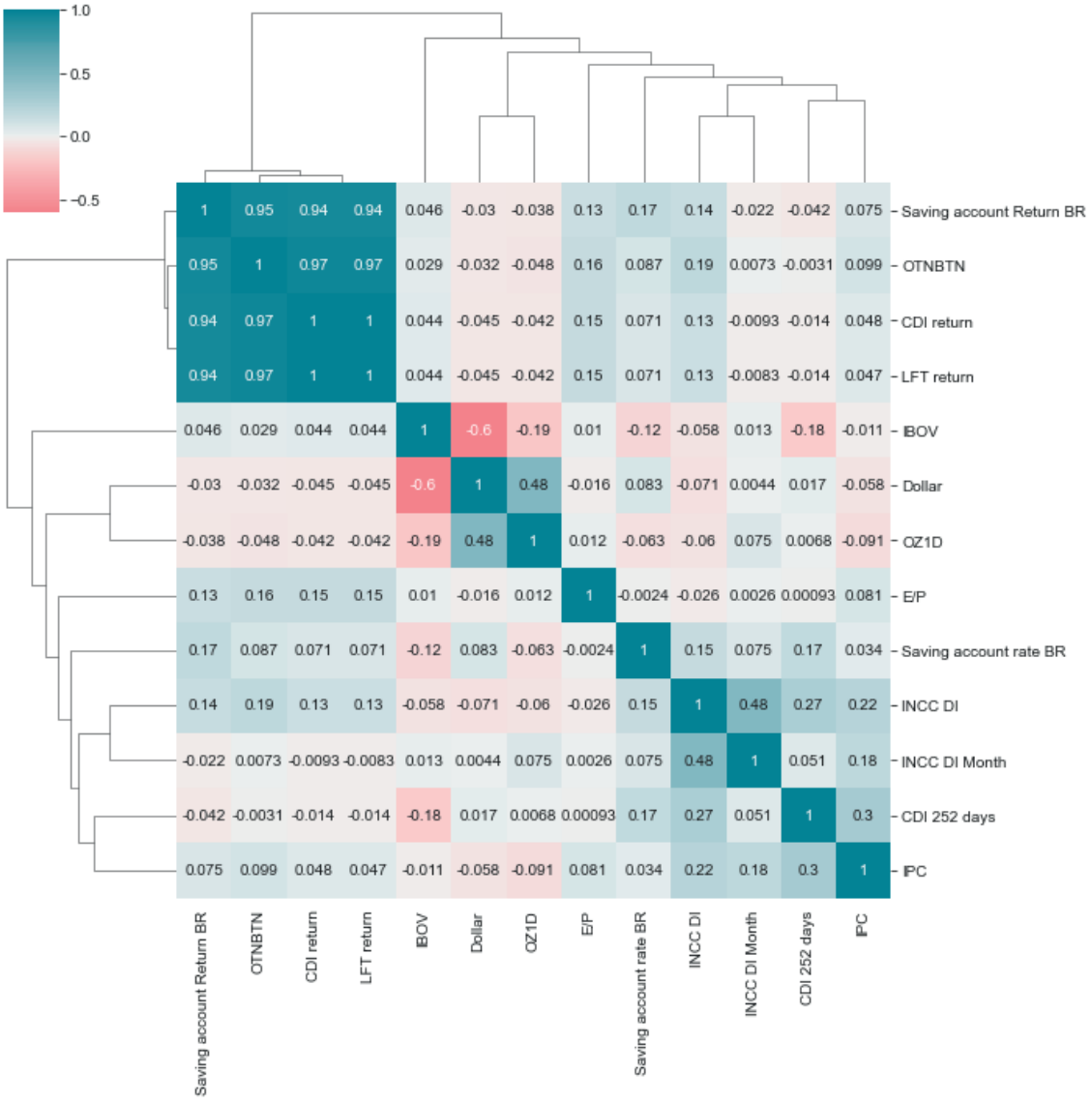
Figure 1 – Shap summary plot – importance



Source(s): elaborated by the author.

Figure 2 illustrates the heatmap with the correlations of variables in the final sample. There is a strong correlation and the formation of a group among the return variables from the savings account, CDI, LFT, and OTNBTN, which is only natural since the variables share some factors in their formation. The dollar variable presented a strong negative correlation with the IBOV.

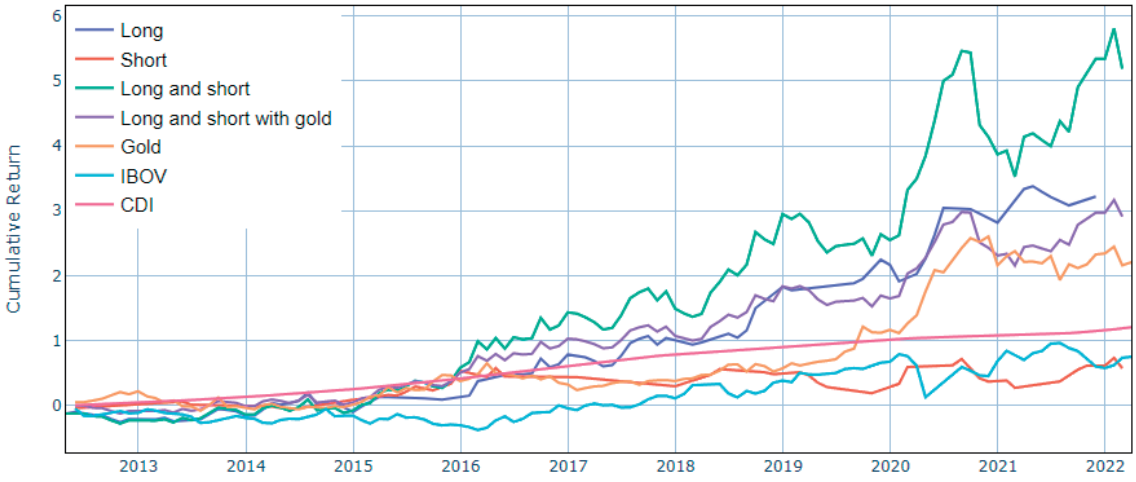
Figure 2 - Heatmap with dendrogram



Source(s): elaborated by the author.

Figure 3, which can be interpreted based on Table 4, shows the cumulative returns of the portfolios proposed over the analysis period. Still, the best cumulative result was reached by the long and short portfolios and the worst result was generated by the short portfolio isolate. In addition, the pink line is much more stable in the CDI, in this case representing a title of fixed income. Within the period analyzed, it provided a better return than the IBOV, however, worse than the remaining portfolios.

Figure 3 – Cumulative return from 2012 to 2022



Source(s): elaborated by the author.

Table 4 validates Figure 3. We highlight that despite the long-only portfolio having a greater monthly mean, it does not have a better cumulative return than the long and short portfolios. This occurs due to the volume of business, which is larger in the long and short portfolio, in addition to the risk, which is lower in the long and short portfolio. We also highlight that in some months the stocks met the criteria to enter one of the portfolios, that is, in some months the stock presented a probability above 50% to form a long-only portfolio, and in other months the stock had a probability below or equal to 49.9% to form a short portfolio. Finally, is worth emphasizing that inserting 1/3 of gold into the long and short portfolio with gold significantly reduced the risk, although compromising the cumulative return. It is believed that the power of the gold and the long and short allocation provides better results, especially in moments of great uncertainty, as demonstrated by *Harvey et al*. (2019)a popular question arises: What steps can an investor take to mitigate the impact of the inevitable large equity correction? However, hedging equity portfolios is notoriously difficult and expensive. We analyze the performance of different tools that investors could deploy. For example, continuously holding short-dated S&P 500 put options is the most reliable defensive method but also the most costly strategy. Holding 'safe-haven' US Treasury bonds produces a positive carry, but may be an unreliable crisis-hedge strategy, as the post-2000 negative bond-equity correlation is a historical rarity. Long gold and long credit protection portfolios sit in between puts and bonds in terms of both cost and reliability. Dynamic strategies that performed well during past drawdowns include: futures time-series momentum (which benefits from extended equity sell-offs.

Table 4 - descriptive analysis of portfolios

Portfolio	Cumulative	Transactions	Average (M)	Std.	Sharp
Long	321%	795	2.05%	6.98%	19.80%
Short	58%	440	1.38%	8.04%	8.80%
Long and short	519%	1235	1.79%	7.12%	15.70%
Long and short with gold	291%	1235	1.27%	4.69%	12.80%
Gold	221%		1.12%	5.29%	8.60%
IBOV	75%		0.69%	6.51%	0.30%
CDI	120%		1.27%	0.29%	

Source(s): elaborated by the author

An important factor to be considered in the studied portfolios is that the costs of transactions were not inserted and that, in general, the custodial costs for gold contracts and leases for short portfolios surpass the maintenance costs of the long-only portfolio. Herein, we decided to not include the costs of transactions due to the recent strong changes in the Brazilian market, where many brokers have proposed aggressive offers.

Table 5 - Metrics results

Metrics	Value
Roc_auc	54%
Accuracy	55%
Precision	70%
Recall	53%
F1	60%

Source(s): elaborated by the author.

Table 5 presents the assessment metrics of the machine learning model used, that is, the random forest. Despite the values being slightly above 50%, such results are common in this type of study. The precision metric stood out at 70%, which is explained especially by the strong fluctuation of the stock market uptrend starting in 2016.

5 FINAL REMARKS

We used the machine learning algorithm random forest to associate the independent variables with a dependent one and learn the probability of positive returns in the month following the data collection. The stocks were allocated into long, short, or non-allocated portfolios. Finally, we allocated a share of gold, as a protection asset.

The study reached its goal and confirmed the possibility of building profitable and safe investment portfolios, even in times of greater uncertainty and volatility, as in 2020, when the Covid-19 pandemic started. The model proved effective in moments of crisis, as well as in moments of greater predictability, such as between 2016 and 2019, when the stock exchange experienced an uptrend.

As for the performance of the portfolios throughout our study period, the best result considering the cumulative return, monthly mean, and Sharpe was attributed to the long and short portfolios. The lowest standard deviation and, therefore, the lowest risk from such perspective, was attributed to the long and short portfolio with gold allocation.

In the scenario of allocation of stocks for portfolios, we suggest that further studies apply some criterium of stop loss. We suppose that such a criterium should increase even more the results, especially regarding the losses caused to the short position of the portfolio. Still, in the allocation scenario, we find it interesting to consider different weights according to the moments of the market and the economy, following the method applied by Lewin and Campani (2020). In the context of machine learning, we suggest testing other algorithms, such as those based on boosting.

REFERENCES

ABRAHAM, R. *et al*. Forecasting a Stock Trend Using Genetic Algorithm and Random Forest. **Journal of Risk and Financial Management**, v. 15, n. 5, p. 1-18, 2022. https://doi.org/10.3390/jrfm15050188

AHMED, S. *et al.* Artificial Intelligence and Machine Learning in Finance: a Bibliometric Review. **Research in International Business and Finance**, v. 3, 101646, 2022. https://doi.org/10.1016/j.ribaf.2022.101646

B3. **5 milhões de contas de investidores**. Disponível em: https://www.b3.com.br/pt_br/noticias/5-milhoes-de-contas-de-investidores.htm. Acesso em: 13 abr. 2022.

B3. Índices amplos. Disponível em: https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm. Acesso em: 13 abr. 2022.

BALLINGS, M. *et al.* Evaluating multiple classifiers for stock price direction prediction. **Expert Systems with Applications**, v. 42, n. 20, p. 7046-7056, 2015. https://doi.org/10.1016/j.eswa.2015.05.013

BALZANA FILHO, M. D. L.; BORDEAUX-REGO, R. Uma análise da relação entre o retorno das ações do setor de construção civil brasileiro e indicadores macroeconômicos. **Engevista**, v. 16, n. 2, p. 137, 2013. https://doi.org/10.22409/engevista.v16i2.469

BEAVER, W.; MCNICHOLS, M.; PRICE, R. The costs and benefits of long-short investing: A perspective on the market efficiency literature. **Journal of Accounting Literature**, v. 37, p. 1-18, 2016. https://doi.org/10.1016/j.acclit.2016.07.001

BERNARDELLI, L. V.; BERNARDELLI, A. G.; CASTRO, G. H. L. A Influência das Variáveis Macroeconômicas e do Índice de Expectativas no Mercado Acionário Brasileiro: Uma Análise Empírica para os Anos de 1995 a 2015. **Revista de Gestão, Finanças e Contabilidade**, v. 7, n. 1, p. 78-96, 2017. https://doi.org/10.18028/2238-5320/rgfc.v7n1p78-96.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001. https://doi.org/10.1023/A:1010933404324

BRESSER-PEREIRA, L. C. A taxa de câmbio no centro da teoria do desenvolvimento. **Estudos Avançados**, v. 26, n. 75, p. 5-28, 2012.

CERVELLÓ-ROYO, R.; GUIJARRO, F. Forecasting stock market trend: a comparison of machine learning algorithms. **Finance, Markets and Valuation**, n. 1, p. 37-49, 2020. https://doi.org/10.46503/nluf8557

CHEN, L. Using Machine Learning Algorithms on Prediction of Stock Price. **Journal of Modeling and Optimization**, v. 12, n. 2, p. 84-99, 2020. https://doi.org/10.32732/jmo.2020.12.2.84

ECONOMATICA. Disponível em: www.economatica.com.br. Acesso em: 13 abr. 2022.

FAMA, E. F. Random Walks in Stock Market Prices. **Financial Analysts Journal**, v. 51, n. 1, p. 75-80, 1995. https://doi.org/10.2469/faj.v51.n1.1861

FAMA, E. F. Market efficiency, long-term returns, and behavioral finance. **Journal of Financial Economics**, v. 49, n. 3, p. 283-306, 1998. https://doi.org/10.1016/s0304-405x(98)00026-9

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados**: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Rio de Janeiro: LTC, 2017.

FELIZARDO, L. K. *et al*. Outperforming algorithmic trading reinforcement learning systems: A supervised approach to the cryptocurrency market. **Expert Systems with Applications**, 117259, 2022. https://doi.org/10.1016/j.eswa.2022.117259

FISCHER, T.; KRAUSS, C. Deep learning with long short-term memory networks for financial market predictions. **European Journal of Operational Research**, v. 270, n. 2, p. 654-669, 2018. https://doi.org/10.1016/j.ejor.2017.11.054

GRINOLD, R. C.; KAHN, R. N. The Efficiency Gains of Long-Short Investing. **Financial Analysts Journal**, v. 56, n. 6, p. 40-53, 2000. https://doi.org/10.2469/faj.v56.n6.2402

HARVEY, C. R. *et al*. The Best of Strategies for the Worst of Times: Can Portfolios be Crisis Proofed? **SSRN Electronic Journal**, 2019. https://doi.org/10.2139/ssrn.3383173

HO, M. K.; DARMAN, H.; MUSA, S. Stock Price Prediction Using ARIMA, Neural Network and LSTM Models. **Journal of Physics: Conference Series**, v. 1, p. 1723-1728, 2021. https://doi.org/10.1088/1742-6596/1988/1/012041

HO, T. K. Random decision forests. **Proceedings of 3rd International Conference on Document Analysis and Recognition**, v. 1, p. 278-282, 1995. https://doi.org/10.1109/ICDAR.1995.598994

ISMAIL, M. S. *et al.* Predicting next day direction of stock price movement using machine learning methods with persistent homology: Evidence from Kuala Lumpur Stock Exchange. **Applied Soft Computing**, 93, 106422, 2020. https://doi.org/10.1016/j.asoc.2020.106422

JAYAPALAN, V.; SOMASUNDARAM, K. Machine Learning based comparison of financial forecasting methods. **International Journal of Advanced Science and Technology**, v. l, p. 8902-8907, 2020. http://sersc.org/journals/index.php/IJAST/article/view/25617/13751

JI, G. *et al.* An adaptive feature selection schema using improved technical indicators for predicting stock price movements. **Expert Systems with Applications**, 116941, 2022. https://doi.org/10.1016/j.eswa.2022.116941

JIANG, M. *et al*. The two-stage machine learning ensemble models for stock price prediction by combining mode decomposition, extreme learning machine and improved harmony search algorithm. **Annals of Operations Research**, 2020. https://doi.org/10.1007/s10479-020-03690-w

JIAO, Y.; MASSA, M.; ZHANG, H. Short selling meets hedge fund 13F: An anatomy of informed demand. **Journal of Financial Economics**, v. 122, n. 3, p. 544-567, 2016. https://doi.org/10.1016/j.jfineco.2016.09.001

KAPOOR, S.; PROSAD, J. M. Behavioural Finance: A Review. **Procedia Computer Science**, 122, 50-54, 2017. https://doi.org/10.1016/j.procs.2017.11.340

KOLANI, D. Portfolio Selection Using Random Forest Algorithm. **International Journal of Computer Engineering and Data Science**, v. 2, n. 1, 28-36, 2022. http://www.ijceds.com/ijceds/article/view/32

KRAUSS, C.; DO, X. A.; HUCK, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. **European Journal of Operational Research**, v. 259, n. 2, p. 689-702, 2017. https://doi.org/10.1016/j.ejor.2016.10.031

LANTZ, B. **Machine Learning with R**. 2019. Packt Publishing Ltd. https://www.packtpub.com/product/machine-learning-with-r-third-edition/9781788295864

LEVANTESI, S.; PISCOPO, G. The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach. **Risks**, v. 8, n. 4, p. 1-17, 2020. https://doi.org/10.3390/risks8040112

LEWIN, M.; CAMPANI, C. H. Gestão de Carteiras sob Múltiplos Regimes: Estratégias que Performam Acima do Mercado. **Revista de Administração Contemporânea**, v. 24, n. 4, p. 300-316, 2020. https://doi.org/10.1590/1982-7849rac2020190161

MA, Y.; HAN, R.; WANG, W. Portfolio optimization with return prediction using deep learning and machine learning. **Expert Systems with Applications**, v. 165, p. 1-15, 2021. https://doi.org/10.1016/j.eswa.2020.113973

MAGHYEREH, A.; ABDOH, H. Tail dependence between gold and Islamic securities. **Finance Research Letters**, 101503, 2020. https://doi.org/10.1016/j.frl.2020.101503

NAWAZ, M. S.; AZAM, M.; ASLAM, M. Probable daily return on investments in gold. **Gold Bulletin**, v. 53, n. 1, p. 47-54, 2020. https://doi.org/10.1007/s13404-020-00273-2

NIU, T. *et al.* Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting. **Expert Systems with Applications**, v. 148, 2020. 113237. https://doi.org/10.1016/j.eswa.2020.113237

PAI, N.; ILANGO, V. Neural Network Model for Efficient portfolio Management and Time Series Forecasting. **Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020, Iciccs**, p. 150-155, 2020. https://doi.org/10.1109/ICICCS48265.2020.9121049

PAIVA, F. D. *et al.* Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. **Expert Systems with Applications**, v. 115, p. 635-655, 2019. https://doi.org/10.1016/j.eswa.2018.08.003

PIMENTA, T.; HIGUCHI, R. H. Variáveis macroeconômicas eo Ibovespa: um estudo da relação de causalidade. **Revista Eletrônica de Administração**, v. 14, n. 2, 2008. https://www.redalyc.org/articulo.oa?id=401137460003

RAMASUBRAMANIAN, K.; SINGH, A. **Machine Learning Using R**, 2017. Apress. https://doi.org/10.1007/978-1-4842-2334-5

RAYA, M. *et al.* Visualizing, Comparing and Forecasting Stock Market Prediction. **2022 IEEE Delhi Section Conference (DELCON)**, p. 1-7, 2022. https://doi.org/10.1109/DELCON54057.2022.9753359

SAMUEL, S. S. A study on growth of gold etfs : as an effective investment tool over physical gold. **UGC Care Journal**, v. 40, n. 40, p. 2328-2331, 2020. https://api.semanticscholar.org/CorpusID:216342242

SETIADI, J.; MASDUPI, E. The Effect of Macroeconomic Variables on Banking Stock Price Index in Indonesia Stock Exchange. **RJOAS**, v. 1, n. 73, 2018. https://doi.org/10.18551/rjoas.2018-01.20

SHABBIR, A.; KOUSAR, S.; BATOOL, S. A. Impact of gold and oil prices on the stock market in Pakistan. **Journal of Economics, Finance and Administrative Science, ahead-of-p**(ahead-of-print), p. 47-63, 2020. https://doi.org/10.1108/JEFAS-04-2019-0053

SRIVINAY *et al*. A Hybrid Stock Price Prediction Model Based on PRE and Deep Neural Network. **Data**, v. 7, n. 5, p. 1-11, 2022. https://doi.org/10.3390/data7050051

TRATKOWSKI, G. Construction of Investment Strategies for WIG20, DAX and Stoxx600 with Random Forest Algorithm. **Contemporary Trends and Challenges in Finance, Springer Proceedings in Business and Economics**, p. 179-188, 2020. Springer Nature Switzerland. https://doi.org/10.1007/978-3-030-43078-8_15

WANG, S. *et al*. Stock selection strategy of A-share market based on rotation effect and random forest. **AIMS Mathematics**, v. 5, n. 5, p. 4563-4580, 2020a. https://doi.org/10.3934/math.2020293

WANG, W. *et al.* Portfolio formation with preselection using deep learning from long-term financial data. **Expert Systems with Applications**, v. 143, 113042, 2020b. https://doi.org/10.1016/j.eswa.2019.113042

WU, X. *et al.* Adaptive stock trading strategies with deep reinforcement learning methods. **Information Sciences**, v. 538, p. 142-158, 2020. https://doi.org/10.1016/j.ins.2020.05.066

YIN, L. *et al.* Research on stock trend prediction method based on optimized random forest. **CAAI Transactions on Intelligence Technology**, p. 1-11, 2021. https://doi.org/10.1049/cit2.12067

ZHU, M. Construction of Quantization Strategy Based on Random Forest and XGBoost. **Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare**, p. 5-9, 2020. https://doi.org/10.1145/3433996.3433998