Analysis of Bitcoin Trends Through the Integration of On-Chain Financial Indicators and Machine Learning

Arthur G. Bubolz[1(✉)], Giancarlo Lucca[2], Lizandro de S. Oliveira[2], Thiago Teixeira[1], Rafael A. Berri[1], Eduardo N. Borges[1], and Bruno L. Dalmazo[1]

[1] Federal University of Rio Grande, Rio Grande, Brazil
{arthurgomesbubolz,rafaelberri,eduardoborges,dalmazo}@furg.br
[2] Catholic University of Pelotas (UCPel), Pelotas, Brazil
{giancarlo.lucca,lizandro.oliveira}@ucpel.edu.br

Abstract. With the significant growth of the cryptocurrency market, predicting its short-term movements has become a significant challenge, requiring methodologies that combine both on-chain data and traditional financial indicators to improve prediction accuracy. This work proposes an integrated approach to predicting short-term movements in the cryptocurrency market by combining on-chain data with traditional financial indicators within the framework of the Rainbow Bitcoin Chart. The methodology encompasses data preparation, model development, and performance evaluation, aiming to build a robust and accurate tool for classifying market trends. By integrating blockchain structural analysis with financial signals, this approach seeks to enhance forecast accuracy and support more informed decision-making. The results show that the normalized model with 14-day windows achieved the best performance among the tested configurations, reaching an accuracy of over 75%.

Keywords: Cryptocoin · Bitcoin · On-chain · Machine Learning · Indicators

1 Introduction

In recent decades, the growing adoption of Bitcoin has demonstrated a relevant phenomenon in global financial markets [6]. The analysis of this asset presents challenges and opportunities that drive the development of predictive techniques. Among them, the combination of on-chain data and traditional financial metrics has shown promise [7]. This study proposes a model that integrates this information to improve the forecast of market movements, seeking to overcome the challenges identified in previous research [5].

Although many studies use traditional methods for price prediction [8], there is a limitation in considering the structural aspects of blockchain. Some

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
O. Gervasi et al. (Eds.): ICCSA 2025, LNCS 15649, pp. 35–50, 2025.
https://doi.org/10.1007/978-3-031-96997-3_3

approaches exploit unsupervised techniques [10] or on-chain data [12], but without integration with financial indicators [28]. Other work focuses on anomaly detection [14] or the use of deep learning [15], but does not directly address the combination of these techniques.

This research focuses on Bitcoin, applying predictive models to identify short-term trends. The challenges include choosing on-chain data, defining model parameters and selecting the ideal time interval [4]. With the increase in interest in cryptocurrencies, methodologies that assist in decision making become essential, especially in markets with high uncertainty [1]. The study seeks to evaluate the effectiveness of machine learning-based models in predicting price movements, contributing to more accurate analysis and informed strategies.

In light of this, the goal of this study is to propose a predictive model that integrates the use of a validated meta-classifier, that is, a model that serves as an additional layer to refine the predictions made by individual classifiers. In this case, the Rainbow Chart is used, a model based on logarithmic curves that segments the Bitcoin price range into different zones, providing a clear view of its historical behavior, allowing the identification of appreciation patterns and possible moments of overbought or oversold, helping to interpret market trends. On-chain data is also used, which are metrics extracted directly from the Bitcoin blockchain, such as transaction volume, active addresses and exchange inflows and outflows for respective predictions. Furthermore, the model will incorporate traditional financial indicators, such as the Relative Strength Index (RSI), an oscillator that measures the speed and magnitude of price variations to identify potential moments of overbought (when the asset may be overvalued) or oversold (when it may be undervalued), helping to make strategic decisions within the market. To accurately predict asset price movements in the short term. This model aims to increase assertiveness in negotiations, taking into account the relevance of Bitcoin as the main cryptoactive on the market and its influence on other cryptocurrencies.

The remainder of this paper is organized as follows. In Sect. 2, we present a brief overview of key aspects to understand this work better. In Sect. 3, we describe the related work, following a methodology used to select the most prominent studies on blockchain forecasting. Section 4 presents an assessment methodology to train the machine learning model. In Sect. 5, we detail the performance evaluation and discuss the obtained results. Finally, in Sect. 6, we conclude the paper with remarks.

2 Theoretical Foundation

In this chapter, fundamental concepts are covered for understanding the rest of the work.

2.1 Fundamentals of Cryptoassets

Digital assets operate in a decentralized manner, using cryptography to ensure the security and authenticity of transactions. These operations are recorded on

the blockchain, a distributed ledger that keeps all transactions transparent and immutable, eliminating the need for intermediaries [9].

In the cryptoactive market, technical analysis and financial indicators are essential for predicting trends. Technical analysis evaluates price movements based on historical data, using charts and indicators such as moving averages, RSI and Bollinger bands to identify market patterns and support strategic decisions [10].

Financial indicators organize price history into zones that reflect levels of confidence in the market trend. Through metrics such as volatility bands and logarithmic curves, these indicators simplify the interpretation of financial data, helping to detect bubbles and buying opportunities.

Additionally, on-chain analytics examine data recorded directly on the blockchain, such as transaction volume and active addresses, providing insights into the asset's supply and demand. In the case of Bitcoin, this approach allows you to evaluate market dynamics and anticipate possible price movements [11].

2.2 Machine Learning

Machine learning is an approach that allows an algorithm to learn from a set of labeled data, where the correct answers are already known. The main objective of this technique is to allow the model to map inputs to outputs, enabling predictions or classifications to be made on new, unseen data. Within supervised learning, we find two main categories of problems: regression, which predicts continuous values, and classification, which assigns labels or categories to inputs.

Both techniques are essential in building predictive models, especially in the context of the cryptoactive market, where the ability to predict trends is crucial for making informed decisions (ZEKIYE et al., 2023).

Among supervised algorithms, the choice depends on the characteristics of the data. Linear models such as linear and logistic regression [20] are efficient for simple patterns, while probabilistic algorithms such as Naive Bayes [21] offer probability-based approaches. Tree-based methods, such as decision trees and Random Forest [22], are effective for more complex and multi-variable data.

Furthermore, neural networks are inspired by the functioning of the human brain and are highly effective in detecting complex patterns and non-linear data (KIM et al., 2022). In the financial market, they are widely used to predict prices, detect fraud and perform advanced analysis.

2.3 Blockchain

Blockchain [23] is a technology that works like a distributed ledger, recording transactions in a secure and immutable way, the structure is composed of chained blocks of information where each block stores data from network transactions, along with a unique identifier (hash). This identifier references the block within the chain, ensuring data integrity and allowing historical verification of all recorded transactions.

The network has in its structure the intrinsically linked concept of decentralization, where the need for centralized intermediaries for respective transactions is eliminated. The network is maintained by multiple participants (nodes), where they validate and store transaction information through consensus protocols, in the case of Bitcoin, Proof of Work (PoW). Such mechanisms guarantee reliable and validated recording of transactions.

In addition to its essential role in cryptocurrencies, blockchain has been applied in several areas, such as smart contracts, supply chain tracking and identity authentication. Its ability to provide security, transparency and decentralization makes this technology a promising tool for transforming various industries, enabling more efficient and reliable operations.

2.4 On-Chain Data

Among the main on-chain metrics used to analyze the cryptoactive market [24], Weight Mean stands out, which measures the average weight of transactions processed on the blockchain, considering the adoption of Segregated Witness (SegWit), which reflects the efficiency in the use of block space. Difficulty Mean indicates the average level of difficulty to validate blocks within a given period, being adjusted periodically to maintain network stability; its increase suggests greater competition between miners, while a reduction may indicate a decrease in the participation of these agents.

Reward Mean represents the average of incentives received by miners for each validated block, including both the issuance of new cryptoactives and transaction fees. In the case of Transaction Sum, it corresponds to the total movement on the network in a given period, offering an overview of economic activity on the blockchain. The witness sum (Witness Sum), related to SegWit, measures the total volume of witness data included in transactions, an important factor for evaluating the optimization of block space and network security. Also regarding sums, there is the sum of fees paid (Fee Total Sum) reflects the total transaction fees paid by users within a period, indicating moments of greater congestion on the network, as periods of high demand result in higher fees.

The input sum and output sum metrics quantify, respectively, the total values entered and removed from transactions on the blockchain, which is useful for understanding capital flows and network liquidity. Finally, the total number of blocks mined (Total Blocks) measures the number of blocks validated in a given time interval, reflecting the activity of miners and the regularity in the production of new blocks. These metrics, together, offer a comprehensive view of the network's behavior, helping to identify possible overbought or oversold scenarios and allowing for more informed decisions in the cryptoasset market.

2.5 Rainbow Chart

The Rainbow Chart [25] is a popular indicator that uses logarithmic curves to categorize Bitcoin's price range over time. It is not based on conventional

technical analysis, but serves as a visual tool to identify possible market trends, helping investors to contextualize the asset's current price concerning its history.

The structure of the Rainbow Chart is made up of different color bands, which represent pricing zones, the lower layers indicate regions considered undervalued, suggesting opportune moments for buying, while the upper layers signal euphoric zones, which may indicate overbought and a possible risk of correction. These ranges are based on logarithmic models, which adjust Bitcoin's growth curve over the years, taking into account its deflationary nature and appreciation cycles (Figs. 1).

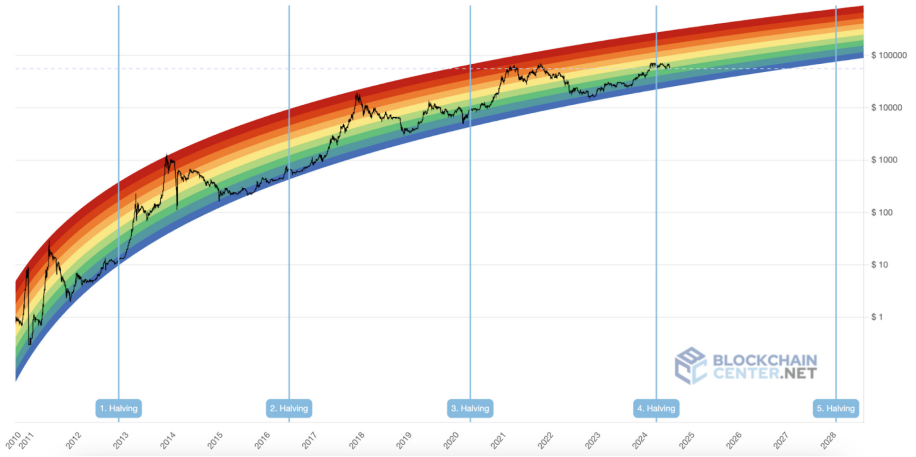

Fig. 1. Rainbow chart indicator. Source: Blockchain Center (2025)

2.6 Trend Forecasting

Trend forecasting [26] is an analytical approach that seeks to identify the future direction of a variable based on historical patterns and contextual factors. In the financial market, this technique does not necessarily focus on predicting exact values, but rather on determining whether the price of an asset is more likely to rise (uptrend), fall (downtrend) or remain relatively stable (sideways movement). The types of forecast may vary beyond these three types, covering different classifications for more significant price changes [13].

In addition to traditional methods based on mathematical and statistical analyses, more advanced approaches, such as machine learning, allow the modeling of complex relationships between multiple variables. Classification models such as decision trees and neural networks can learn non-obvious patterns in data and improve forecast accuracy by providing probabilities of trend continuation or reversal based on a large volume of historical information.

Although trend forecasting is never an exact science, as the market is influenced by unpredictable factors, the combination of different approaches and

tools can significantly increase the assertiveness of analyses, allowing for more informed and strategic decisions.

2.7 Evaluation Metrics

The confusion matrix [27] is a fundamental tool in evaluating classification models, allowing you to visualize the model's performance when comparing its predictions with real values, organizing the results into four categories: true positives (TP), when the model hits a positive prediction; false positives (FP), when it incorrectly predicts a positive case; true negatives (TN), when a negative prediction is correct; and false negatives (FN), when it incorrectly classifies a negative case.

Precision is a metric that measures the proportion of correct positive predictions, calculated as Eq. 1. It is especially important when the costs of a false positive are high, such as in anti-fraud systems, as it indicates the reliability of the positive predictions made by the model.

Recall, also called sensitivity, evaluates the model's ability to correctly identify all positive cases defined as Eq. 2. This metric is essential in scenarios where the detection of all positive cases is a priority, such as in the diagnosis of serious diseases, where false negatives can have severe consequences.

The F1-score is a metric that seeks to balance precision and recall, being the harmonic mean between the two values, calculated as Eq. 3. It is useful when there is a trade-off between minimizing false positives and false negatives, ensuring that the model performs well overall without overly favoring one of the metrics.

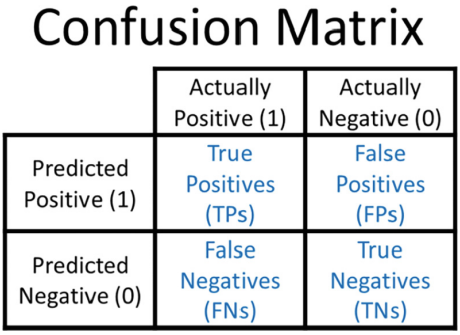

Fig. 2. Confusion matrix. Fonte: Glassbox Medicine (2025)

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

3 Related Work

Recent advances in blockchain data analysis and forecasting have been extensively explored in studies such as [9], addressing challenges and proposing innovative solutions. One key area of study involves the extraction, ingestion, and analysis of Ethereum blockchain data, tackling issues related to identifying outliers and suspicious smart contracts. Extracting transaction metadata and mapping smart contract interfaces are considered complex tasks due to the diversity and detail of information within the blockchain. Tools like Pegasus have been employed for data consumption in private blockchains, while machine learning techniques have been applied to clustering and anomaly detection, aiming to identify atypical accounts on the public Ethereum blockchain. However, scalability and data handling issues, along with the inherent complexity of smart contracts, continue to complicate the representation of relevant information.

Another significant field of study involves forecasting Bitcoin volatility through the analysis of on-chain data combined with Tweets posts, particularly from the `@whale_alert` account, as discussed in [11]. The use of reinforcement learning, specifically Q-learning, enabled the construction of a high-precision predictive model for classifying Bitcoin price variations. Nevertheless, data availability was highlighted as a challenge, especially considering Bitcoin's extensive historical records.

Machine learning has also been applied to analyzing risky cryptocurrencies using techniques like K-means++ for clustering historical data. The research [16] demonstrates that a large percentage of cryptocurrencies vanish from the market within 1,000 days, while only a small fraction have longer lifespans. Variable clustering, followed by normalization and correlation analysis, proved effective in assessing asset quality, though challenges such as algorithmic bias and the subjectivity in selecting change points remain.

Efforts to detect low-latency pump-and-dump schemes led to the proposal in [17], which analyzes 1-minute charts and uses an LSTM autoencoder to forecast cryptocurrency prices and identify anomalies in real time. Although the method showed efficiency, the limited number of samples may restrict its broader applicability.

Ethereum price forecasting has also been investigated through on-chain analysis, leveraging machine learning models that correlate various metrics, such as block size and height, with the cryptocurrency price. Studies like [12] propose evolutionary adjustments to improve prediction accuracy, despite concerns about data inconsistency and correlation.

Self-adaptive deep learning models, such as jSO-LSTM, have been explored in [10] to predict Bitcoin prices, accounting for multiple on-chain metrics. Data normalization and the selection of highly correlated variables have been crucial for model performance, which outperformed conventional methods.

Additionally, frameworks like SAM-LSTM, explored in [8], utilize attention mechanisms combined with LSTM for cryptocurrency price prediction, emphasizing the importance of detecting change points to adapt the model to previously unseen variations. Despite criticisms regarding subjectivity in selecting these points, the model achieved promising results in testing.

Forecasting crypto market phases, particularly to guide investment strategies, has also been addressed in studies like [18] using on-chain data. Predictive models such as SARIMA, LSTM, and CNN were compared, showing the advantage of deep learning methods over traditional approaches, although incorporating actual market movement predictions remains an area to explore.

Innovative approaches for security have been developed in research like [19], where machine learning algorithms are applied to Ethereum smart contracts, featuring unique characteristics like using integers instead of floating points to calculate probabilities. While this methodology proved effective and scalable, its practical applicability still depends on greater integration with the research's end goal.

Anomaly detection in DeFi transactions was explored in [14], using deep learning models like Anomaly VAE-Transformer, which combine variational autoencoders with transformers to assess anomaly probabilities in Olympus DAO data, demonstrating accuracy in correctly detecting anomalies.

Finally, cryptocurrency price forecasting using ensemble deep learning models, which combine CNNs and RNNs, was highlighted in [15] as a promising approach. These models can handle blockchain data's volatility and complex dependencies, surpassing existing methodologies and paving the way for new applications in crypto asset price modeling.

The mentioned studies address the use of on-chain data for various purposes, such as cryptocurrency price forecasting, anomaly detection, and market manipulation scheme identification. While presenting innovative and technically sophisticated approaches, they often focus on direct price predictions or isolated historical trend analysis. Even when normalizing data, they treat context uniformly and disaggregated, limiting the ability to provide a more strategic and contextual view of market movements.

This work aims to fill the gaps left by previous studies by introducing a meta-indicator that considers the market region to guide investment decisions. Instead of merely predicting the future price of Bitcoin or other cryptocurrencies, our model will assess the asset's current position within a predefined market region. Using on-chain data and advanced machine learning techniques, our goal is to determine whether the market is in a bullish or bearish phase and how this may influence buy and sell decisions. This approach will enable more robust and contextual analysis, offering investors a more precise tool for strategic decision-making.

Thus, the study stands out by combining trend forecasts with market region analysis, offering a more holistic and practical approach to cryptocurrency analysis, using Bitcoin as the basis for our experiments. By integrating these elements, we aim to provide a solution that not only indicates potential price movements

but also contextualizes these movements within the broader market scenario, allowing investors to make more informed and strategic decisions.

Table 1. Summary of the main characteristics of the related work

Main features	[9]	[11]	[16]	[17]	[12]	[10]	[8]	[18]	[19]	[14]	[15]
Uses technical data	✓		✓	✓	✓		✓	✓			✓
Uses textual data		✓									
Regressive Approach				✓	✓	✓	✓				✓
Classification Approach	✓	✓	✓	✓				✓	✓	✓	
Short Forecast	✓	✓	✓	✓	✓	✓	✓			✓	✓
Long Forecast								✓			
Traditional ML Algorithms	✓		✓						✓		
Deep Learning			✓	✓	✓	✓	✓	✓		✓	✓
Reinforcement Learning		✓									

4 Methodology

The study proposed training a machine learning model using financial and on-chain data, to capture patterns and predict market trends. The data used includes price information, such as opening, maximum, minimum, and closing (Open, High, Low, Close), as well as technical indicators, such as the Relative Strength Index (RSI). The model is guided by the analysis of "Bitcoin's Rainbow Chart", which serves as a reference for identifying the asset's macro trend. We also seek the automated extraction and verification of on-chain data, normalizing it to generate meaningful features, integrating this information into the financial context to enrich the training base and improve the predictive capacity of the algorithm.

The combination of this on-chain data with financial indicators provides a comprehensive and structured analysis, allowing for more assertive and informed information about cryptocurrency market trends. Implementation details are presented in Fig. 3.

In the first step (1 in Fig. 3), the on-chain data is extracted, which provides a detailed view of transactions and activities on the blockchain, in step two (2 in Fig. 3), the historical records of financial indicators are extracted, which offer a traditional perspective on market behavior, step three (3) involves the extraction of the historical records of the base indicator, such as the rainbow chart, which helps in the interpretation of market phases.

With all this data available, in step four 4 in Fig. 3), the data is integrated and normalized, ensuring that it is adequately prepared for the model to understand and process the information coherently. In step five (5), the model is trained, where it learns to identify patterns and trends based on the data provided. Finally, in the sixth stage (6), the model is used to make predictions about the movement of the asset, allowing market movements to be anticipated.

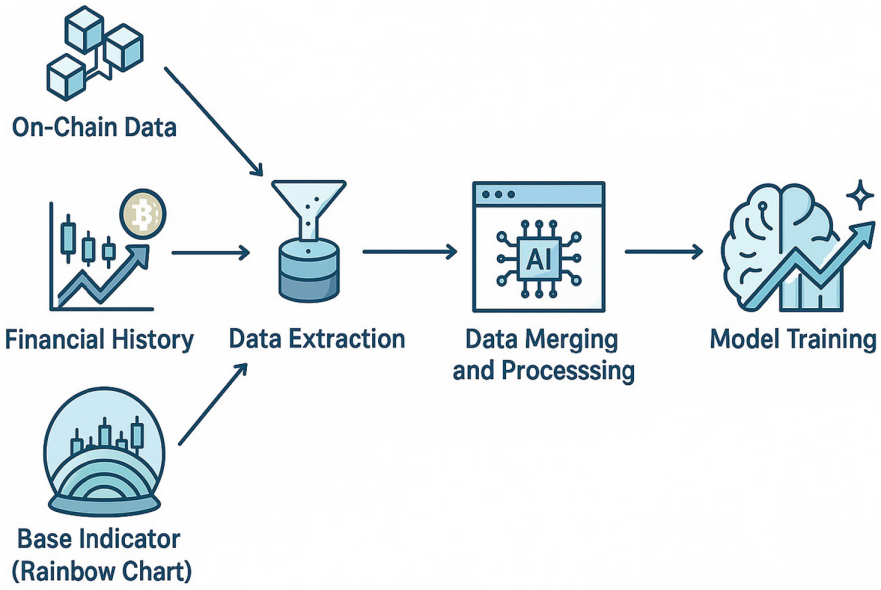

Fig. 3. Conceptual model.

4.1 Normalization

In this part of the work, the process of normalizing the data used will be discussed, detailing the stages of collection, transformation and consolidation of metrics extracted from the Bitcoin network. Normalization is essential to adjust the scale of data, facilitating trend analysis and the application of predictive models.

On-Chain Data: The notebook implements a complete process of collecting, transforming and normalizing on-chain data from the Bitcoin network, with a focus on consolidating daily metrics for trend analysis.

Initially, libraries such as pandas for data manipulation, requests for HTTP requests, BeautifulSoup for HTML analysis, and other auxiliary tools for organizing and processing the files are loaded. The Blockchair platform is used to access the Bitcoin network's block record page, through a GET request, the code captures the HTML content of the page, which is analyzed to extract links to daily block records from 2013 to 2025.

After downloading the files, each record is read and processed to calculate relevant metrics such as average block weight, average difficulty, average USD reward, transaction counts, entries, exits, witnesses, and the total sum of fees paid. The `df_to_row` function transforms each block DataFrame into a single row with normalized statistics for the respective day, consolidating the information into a format suitable for time series analysis. The normalized records are

added to a final DataFrame (`df_result`), with exception handling to deal with empty files or possible reading errors. Finally, the records are ordered by date to maintain temporal coherence.

Financial Data: Initially, Bitcoin's historical data is loaded from CSV files, which were obtained from historical records made available by CoinMarketCap. As the site allows the export of data in CSV format for only one year at a time, it was necessary to perform the removal year by year, each CSV file were representing daily records from a specific year. These files are processed iteratively, treating different delimiters and removing irrelevant columns, such as open, close, high, and low times, as well as the asset name, the date is normalized to daily format, rounding the timestamps, and the resulting DataFrames are concatenated and ordered chronologically.

To model market phases, a logarithmic curve is fitted to closing prices using the curve_fit function from the scipy library. The fitting function follows the form, and the fitted parameters define the base curve. Based on this curve, fixed multipliers determine the price ranges, classifying the asset ranging from the minimum "Fire Sal" level (price much below the predicted value [0.4x]) to the maximum "Super Bubble" level (extremely high price [5x or more]).

Subsequently, RSI is calculated based on a standard 14-day period by analyzing price changes, positive and negative differences are separated to calculate exponential moving averages (EMA) of gains and losses, values presented on the indicator above 70 indicate overbought (possible bearish reversal), while values below 30 indicate oversold (possible bullish reversal).

After adding the Rainbow Chart and RSI, null values are discarded, resulting in a normalized, enriched dataset ready for analysis or to be used as input in machine learning models.

5 Performance Tests and Comparisons

This section aims to present the additional tests carried out to identify the best conditions for forecasting, including the choice of data normalization techniques, the definition of the day intervals used in model selection, and other variables relevant to the forecasting process.

5.1 Description of the Tests

To carry out the tests, a set of data was used in which three additional columns were created based on the closing value of each record. The columns represented the price behavior after 3, 7 and 14 days, with the values classified as 1 when there was an increase in the price and 0 when there was a drop, always concerning the original closing, allowing the problem to be treated as a binary classification task, facilitating the evaluation of the performance of the algorithms.

After creating these columns, the numerical data was normalized by two different methods. The first method consisted of normalization in sliding windows,

where values were adjusted to the interval [0, 1] based on blocks of 37 days, a value that represents the median number of days that a rainbow-chart class remains active, which made it possible to capture local patterns and seasonal variations in addition to a more contextualized targeting of the market moment classification that the rainbow-chart indicated. The second method involved global normalization, applied to the entire database, adjusting the values to the same range, but considering all observations. It is important to highlight that normalization was applied exclusively to numerical attributes, since categorical attributes had already been transformed using one-hot encoding, being naturally restricted to the range [0, 1].

To evaluate the effectiveness of each normalization method, k-fold cross validation with 5 divisions was used. The metric used to measure the performance of the algorithms was the average accuracy, calculated from the results obtained in the five folds, enabling a more robust evaluation and reducing the risk of overfitting in the construction of future models, providing a more reliable view of the predictive capacity of the models, with all metrics results for each model presented in the Table 2.

The selection of algorithms used in the cross-validation stage was made with the aim of contemplating different methodological approaches to the binary classification problem, allowing a richer and more comprehensive comparison between linear and non-linear models. Among the linear models, Logistic Regression was chosen, which bases its decision on the sigmoid function to estimate the probability of classes, and Gaussian Naive Bayes, which adopts a probabilistic approach based on Bayes' theorem, assuming independence between variables and normal distribution of data, both included to verify how algorithms with simple structures behave in comparison to more complex and non-linear models.

In the set of non-linear models, three algorithms based on decision trees were selected: Decision Tree Classifier, Random Forest Classifier and Gradient Boosting Classifier, although they share the logic of recursive division of the attribute space, differ in the way they combine the trees, whether individually, in random forests (bagging) or in a sequence of improvements (boosting). In addition to these, the MLPClassifier was also tested, representing artificial neural network methods, the KNeighbors Classifier makes decisions based on the proximity of samples in the attribute space, while the Support Vector Classifier (SVC), seeks an optimal separation between classes through margin maximization in a space possibly transformed by kernels. This diversity of approaches allows us to evaluate not only the accuracy, but also the bias and generalization capacity of each model given the specific nature of the data set.

5.2 Construction of the Model and Final Tests

After carrying out tests with cross-validation, a significant improvement in the performance of the models was suggested in cases where windowed normalization was applied, when compared to normalization considering the complete set of data, providing that windowed normalization allows the models to more

Table 2. Cross-validation results

	acc mean - 3 days		acc mean - 7 days		acc mean - 14 days	
	win	comp	win	comp	win	comp
DecisionTreeClassifier	0.5452	0.5300	0.5616	0.5341	0.6284	0.4789
GaussianNB	0.5979	0.4869	0.6069	0.4876	0.6479	0.4926
GradientBoostingClassifier	0.6057	0.5210	0.6446	0.5360	0.6915	0.5685
KNeighborsClassifier	0.5753	0.5087	0.5950	0.5071	0.6555	0.5391
LogisticRegression	**0.6154**	0.5348	**0.6510**	0.5407	**0.7043**	0.5948
MLPClassifier	0.5971	0.5348	0.6230	0.5267	0.6867	0.5511
RandomForestClassifier	0.6142	0.5265	0.6502	0.5414	0.6948	0.5677
SVC	0.6128	0.5443	0.6308	0.5402	0.6943	0.5711

efficiently capture the underlying temporal relationships in the data, contributing to more robust transfers.

For the construction of the final model and performance evaluation, it was decided to carry out a planned procedure with a horizon of 14 days, since this interval presented the greatest accuracy in the preliminary experiments. Three algorithms representing different approaches were selected: linear regression to capture linear relationships, Random Forest to model nonlinear interactions through decision trees, and an MLP Classifier to explore complex patterns through neural networks.

Each model was optimized using a grid search (Grid Search) with cross-validation, with AUC-ROC as the main metric, to find the best hyperparameter configuration for class discretization. Firstly, the logistic regression model was treated, the best configuration found was with L2 penalty, C equal to 1 and "liblinear" solver, resulting in an approximate AUC-ROC of 0.7789 in the training data and 0.7759 in the test set.

The Random Forest algorithm obtained the best results with 200 estimators, maximum depth of 20, minimum of 2 examples per division and 2 leaves per node, reaching an AUC-ROC of approximately 0.8240 in training and 0.8218 in testing. The MLP Classifier showed better performance with two hidden layers of 50 neurons each, activation function "tanh", solver "adam" and alpha of 0.0001, resulting in an AUC-ROC of 0.8210 in training and 0.8148 in testing.

Finally, the adjusted models with the best hyperparameters were combined in a hard voting Fig. 2, which resulted in a final accuracy of 0.7417, consolidating the contributions of each approach, balancing their strengths to provide a more stable and generalizable solution, being demonstrated through its resulting confusion matrix Fig.5.

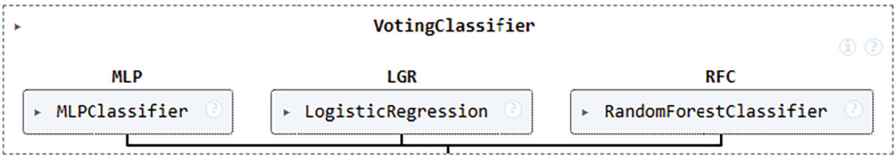

Fig. 4. Ensemble Model Architecture.

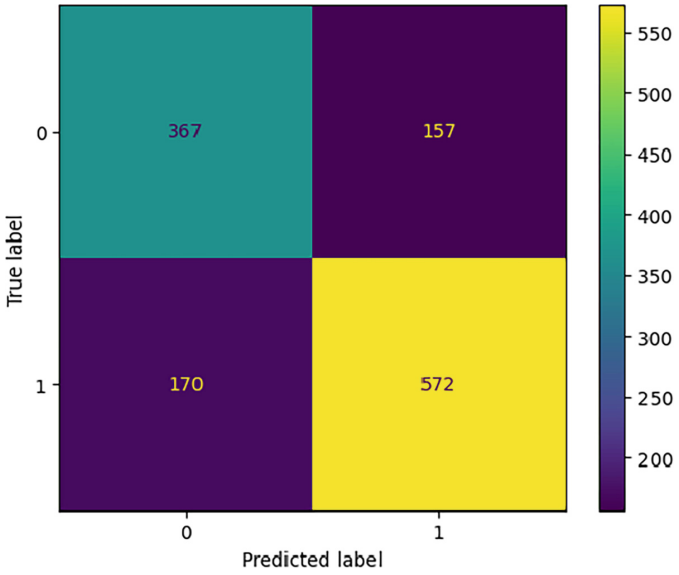

Fig. 5. Confusion Matrix of the model.

6 Conclusions

This work explores cryptocurrency price prediction, proposing an approach that integrates on-chain data, technical indicators, and the Rainbow Bitcoin Chart. Our approach achieved the best performance among all tested configurations, utilizing normalization in moving windows with the dataset and a 14-day forecast horizon. It is believed that over longer intervals, the parameters become more discretized, enhancing the model's ability to make more accurate predictions.

Furthermore, the fact of using windowed normalization strengthens the hypothesis that this method can be more effective than total data normalization. This is because windowed normalization makes the data less sensitive to abrupt variations over time, in addition to aligning more precisely with the meta-indicator used.

Based on the results obtained, it is concluded that the model demonstrates promising performance. This suggests that future research could explore other crypto assets or enhance the current model by integrating additional approaches

to refine predictions, leading to significant advancements in the field. For evaluation and transparency of results, we maintain all source codes publicly available in a git repository[1].

Acknowledgements. This work is supported by FAPERGS (24/2551-0001396-2, 23/2551-0000773-8), CNPq (305805/2021-5) and FAPERGS/CNPq (23/2551-0000126-8).

References

1. Kristoufek, L.: BitCoin meets google trends and Wikipedia: quantifying the relationship between phenomena of the Internet era. Sci. Rep. **3**(1), 3415 (2013). Nature Publishing Group UK London
2. Lahmiri, S., Bekiros, S.: Cryptocurrency forecasting with deep learning chaotic neural networks. Chaos Solitons Fractals **118**, 35–40 (2019). Elsevier
3. Dias, M.A., Henriques, M.A.A.: Estudo sobre a relação custo-benefício de mecanismos de consenso Proof-of-Stake para blockchains públicas
4. Greaves, A., Au, B.: Using the bitcoin transaction graph to predict the price of bitcoin. No data **8**, 416–443 (2015)
5. Chevallier, J., Guégan, D., Goutte, S.: Is it possible to forecast the price of bitcoin? Forecasting **3**(2), 377–420 (2021). MDPI
6. Yermack, D.: Is Bitcoin a real currency? An economic appraisal. In: Handbook of Digital Currency, pp. 29–40 (2024). Elsevier
7. Kim, H.-M., Bock, G.-W., Lee, G.: Predicting Ethereum prices with machine learning based on Blockchain information. Expert Syst. Appl. **184**, 115480 (2021). Elsevier
8. Kim, G., Shin, D.-H., Choi, J.G., Lim, S.: A deep learning-based cryptocurrency price prediction model that uses on-chain data. IEEE Access **10**, 56232–56248 (2022)
9. Starikov, V., Kozlyak, R., Opletina, P., Mityagin, S.: Forecasting cascading effects in network models as applied to urban services provision assessment. In: Gervasi, O., Murgante, B., Garau, C., Taniar, D., C. Rocha, A.M.A., Faginas Lago, M.N. (eds.) Computational Science and Its Applications - ICCSA 2024. ICCSA 2024. LNCS, vol. 14813. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-64605-8_30
10. Jagannath, N., et al.: A self-adaptive deep learning-based algorithm for predictive analysis of bitcoin price. IEEE Access **9**, 34054–34066 (2021)
11. Muminov, A., Sattarov, O., Cho, J.: Forecasting Bitcoin Volatility through On-Chain and Whale-Alert Tweet Analysis using the Q-Learning Algorithm. IEEE Access (2023)
12. Jagannath, N., et al.: An on-chain analysis-based approach to predict Ethereum prices. IEEE Access **9**, 167972–167989 (2021)
13. Souza, A.S., Lucca, G., Borges, E.N., Cardoso, F.C., Dalmazo, B.L., Berri, R.: Dataset for Intraday Analysis of B3 stock prices, Harvard Dataverse, Draft Version (2024). https://doi.org/10.7910/DVN/TMB4IG

[1] https://github.com/arthur-bubolz/implementation_codes_iccsa.

14. Song, A., Seo, E., Kim, H.: Anomaly VAE-transformer: A deep learning approach for anomaly detection in decentralized finance. IEEE Access (2023)
15. Ali, F., Ravate, S., Nimbore, S.: Ensemble model based on deep learning for forecasting crypto asset futures in markets. In: 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), pp. 1–8. IEEE (2023)
16. Zekiye, A., Amroush, F., Utku, S., Özkasap, Ö.: AI-assisted investigation of on-chain parameters: risky cryptocurrencies and price factors. In: 2023 Fifth International Conference on Blockchain Computing and Applications (BCCA), pp. 52–59. IEEE (2023)
17. Bello, A. S., Schneider, J., Di Pietro, R.: LLD: a low latency detection solution to thwart cryptocurrency pump & dumps. In: 2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), pp. 1–9. IEEE (2023)
18. Casella, B., Paletto, L.: Predicting cryptocurrencies market phases through on-chain data long-term forecasting. In: 2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), pp. 1–4. IEEE (2023)
19. Badruddoja, S., Dantu, R., He, Y., Thompson, M., Salau, A., Upadhyay, K.: Making smart contracts predict and scale. In: 2022 Fourth International Conference on Blockchain Computing and Applications (BCCA), pp. 127–134. IEEE (2022)
20. Worster, A., Fan, J., Ismaila, A.: Understanding linear and logistic regression analyses. Can. J. Emerg. Med. **9**(2), 111–113 (2007). Cambridge University Press
21. Bayes, T.: Naive Bayes classifier. Article Sources and Contributors, pp. 1–9. BIOMISA, Department of Computer and Software Engineering (1968)
22. Dabiri, H., Farhangi, V., Moradi, M.J., Zadehmohamad, M., Karakouzian, M.: Applications of decision tree and random forest as tree-based machine learning techniques for analyzing the ultimate strain of spliced and non-spliced reinforcement bars. Appl. Sci. **12**(10), 4851 (2022). MDPI
23. Di Pierro, M.: What is the blockchain? Comput. Sci. Eng. **19**(5), 92–95 (2017). IEEE
24. Stober, A., Sandner, P.: Using on-chain and market metrics to analyze the value of crypto assets. FSBC Working Paper. Frankfurt am Main: Frankfurt School of Finance (2020)
25. Bernardi, D., Bertelli, R.: Bitcoin price forecast using quantitative models. In: Proceedings of the Conference on Quantitative Finance (2021)
26. Dai, Y., Zhang, Y.: Machine learning in stock price trend forecasting. Stanford University, 2, 2023 (2013). Retrieved from http://cs229.stanford.edu/proj2013/DaiZhang-MachineLearningInStockPriceTrendForecasting.pdf
27. Yacouby, R., Axman, D.: Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. In: Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, pp. 79–91 (2020)
28. Amorim, M., et al.: Systematic review of aggregation functions applied to image edge detection. Axioms **12**(4), 330 (2023). https://www.mdpi.com/2075-1680/12/4/330. https://doi.org/10.3390/axioms12040330