1 Introduction

In an evolving world of stock markets, the demand for sophisticated tools for financial forecasting is growing exponentially (Tellez Gaytan et al. 2022), forecast trends for a particular day and the behavior of the stock for following days is even more difficult with the uncertainty and variability in stock markets. Additionally, the political and social situations are external factors that can impact on the stock prediction (Botunac et al. 2024) and will be something that the past data is unable to reveal.

Traditional methods for predicting market behavior present great challenges in price predictions due to their limitations in handling large amounts of data or capturing rapid market changes. Additionally, many factors can influence stock market volatility (Chau et al. 2014). Traditional strategies based on fundamental or quantitative analysis fail to fully capture the nature of the market, generating inaccurate decisions when investing (Di Liu et al. 2019).

Despite the inherent uncertainty of the market, historically, Technical Indicators have been used for predicting the trend of the assets, this kind of indicator uses the Open/Close prices and formulas rooted in statistics to provide values that depending on the context, can provide valuable results (Botunac et al. 2024), though not always accurate.

For the reasons mentioned above, it is required a combination of some forecasting or trending abilities that a machine learning model can provide, with some expert knowledge that historically has helped investors make better decisions for the asset's predictions such as Technical Indicators. This investigation will use historical market data to validate the impact of these combinations. Conclusions from this research could lead to more data-driven decisions, better strategies, and probably better returns on investments.

This research aims to contribute to the literature and will analyze the impact of integrating machine learning classifiers with traditional technical indicators for predicting stock market trading decisions, in comparison to different studies that seek to improve the precision of the model, this investigation seeks to create a robust model that allows the integration of different indicators and is reliable in its predictions. The target group of this research are financial analysts or common people who want to invest or learn more about investment.

Research question

The research question for this thesis is "How to create a robust and replicable machine learning model to different stock markets using technical indicators?".

As a result, of this study, a model that combines the strength of the machine learning models and technical indicators will be generated. Previous investigations suggest that technical indicators work well with machine learning models (Tan et al. 2019). ensemble learning techniques improve the results of the accuracy and reduce variance, which can be helpful for trading decisions (Nti et al. 2020).

This research has the following structure, as shown in the next figure. Chapter 1 will explore the definition of stock markets, the traditional methods used for the predictions, and more recent methods such as machine learning, as well as the current investigation in this area.

In the second chapter, the Crisp DM methodology adopted for the machine learning model will be explained. The third chapter is a combination of business understanding and data understanding. this chapter will discuss the stocks selected, the data source used, and the initial description of the data, as well as the indicators that will be used. for this research, data was taken from 3 stocks with the daily opening and closing prices available in Yahoo Finance for a period of 5 years. The three stocks were selected to represent different sectors of the industry: Meta (technology), WTI Oil (energy), and iShares MSCI World Index ETF (diversified). These datasets served as the basis for the daily calculation of the financial indicators that will be the data for the model.

Chapter 4 is about data preparation and preprocessing, which includes, dimensionality reduction, outlier identification & treatment, and handling the missing values. After that, feature reduction and balancing are employed.

Chapter 5 is about modeling, here different models were employed to verify which one is the best for the classification task: models such as random forest, decision tree, XGBoost, KNN, AdaBoost, and Ensemble Learning were trained. Finally, in chapter 6 empirical analysis is performed, here evaluation metrics were employed to evaluate this classification problem. These metrics include Accuracy, Precision, Recall, F1 Score, AUC curve, the Confusion Matrix, Backtesting, and comparison to the base model.

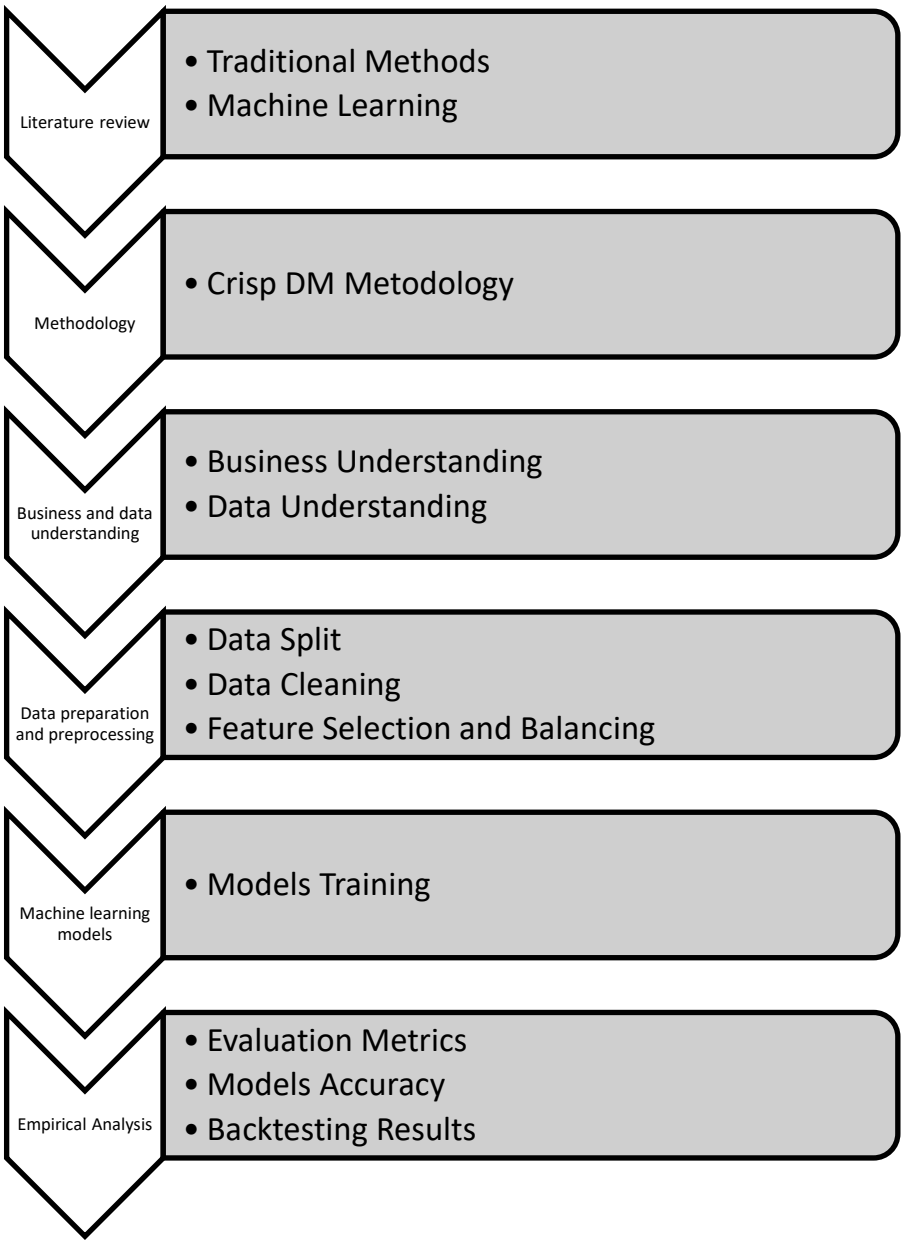

Figure 1.Structure of thesis

2 Literature review

In this chapter, the stock markets will be defined, their history, factors that influence prices as well as the participants in the market will be discussed. It also discusses traditional methods for price predictions, and then machine learning algorithms will be addressed.

The stock market, also known as the equity market, works as a platform for trading and investing in stocks. In simple terms, it will allow selling or buying fractions of companies, these fractions represent units, which in the end means that the buyer owns part of a company (Tairu 2024). The value of the stock may be affected by economic and market conditions, company performance, launch of new products, or negative publicity (Botunac et al. 2024).

The stock market as we know it today has its origins in the medieval era in Europe (Petronchak and Sidorets 2021). In cities like Bruges - Belgium, commodity traders came together, creating the first formal exchange system (Rjumohan 2019). However, these transactions were carried out without supervision and in a disorderly manner. Only until 1602, with the founding of the Amsterdam Stock Exchange, was a more organized form of stock exchange established (Rjumohan 2019).

Later, at the end of the 18th century, the London and New York Stock Exchanges were created, which managed to achieve the greatest representation in the market internationally (Rjumohan 2019). Unfortunately, there was also a whole history of economic crises from the mid-19th century to the present with deep repercussions in the stock market world and throughout society, the most important one was in 1929.

The Great Depression of 1929

Large fluctuations occurred at the beginning of the 1920s, a strong demand generated by the reopening of commerce and a high level of currency issuance had significantly influenced the rise in prices. Raw materials such as wool, products such as coffee, sugar, wool, meat and cereals increased more of 200%. Compensation imposed on the defeated countries and their allies generated hyperinflation and a very high devaluation of some European countries' currency (Marichal 2010).

In the middle of that decade the economy recovered, and almost all currencies were once again governed by the gold standard (a system that guarantees the value of circulating money in a certain number of grams of gold). Industrial production grew in the United States and the main European countries, especially England, France and Germany, which activated the rise of the stock markets (Investment Funds went from 40 in 1921 to 750 in 1929 in USA) and unfortunately also credit for disproportionate stock market speculation operations (Marichal 2010).

On October 24, 1929 – Black Thursday – the New York Stock Exchange collapsed, and was later followed by a wave of failed banks, a 60% reduction in world trade and the loss of tens of millions of dollars (Marichal 2010).

Historians and economists who analyze this phenomenon find among its main causes the absence of regulations and controls on the financial and stock market speculation, which led to the issuance of legislation aimed to controlling these activities. (Emergency Banking Act – 1933, Glass-Steagall Act-1933,Securities Exchange Act 1934 and Banking Act – 1935), with the designation of their corresponding control authorities, issued in USA in the administration of Franklin Delano Roosevelt (Marichal 2010).

From the middle of the 19th century to the present, the world economy has experienced different crises. (Marichal 2010, 36) makes a careful sequence of them. The period between the Great Depression of 1929 analyzed above and the other great crisis of 2008-2009 can be observed in the following table:

Year	Description
1929 crisis	the New York stock market collapse in October triggered a period of international financial imbalance, bank panics between 1931 and 1933 and the Great Depression.
1945 recession	At the end of World War II, severe recessions were recorded in several belligerent countries.
1971*	There was no financial crisis, but the dollar-gold parity system was over.
1982 crisis	Latin American debt crisis.
1987 crisis	Stock market crisis.
1989/1990 crisis	Financial crisis in Japan.
1997/1998 crisis	Asian crises.
2001 crisis**	Dot.com stock market crisis in the United States.
2008 crisis	The financial markets in New York and London collapsed in September and October 2008, followed by the global recession of 2008-2009.

Table 1. Major financial crises of the 20th and 21st centuries adapted from (Marichal 2010, 36)

However, from the Second World War to1982, no economic crises occurred. but from then on, A financial Tsunami began that culminated in the 2008-2009 crisis (Marichal 2010).

The Great Recession 2008-2009

Almost a century after the world was shaken by the enormous Depression of 1929, the world was once again struck by an almost similar phenomenon with serious economic and social repercussions everywhere. After the crisis of technological companies, mainly in the United States, known as the "dot/com crisis" in 2001(due to a stock market bubble). North American economy began to recover between 2004 and 2005 based on an enormous flow of capital from abroad (especially from Asia) and a reduction in interest rates in favor of public debt in the United States. These phenomena, together with the complete disassembly of the control regulations issued in the 1930s, lead to speculation in the real estate, mortgage and stock market sectors, and caused the collapse of 2008-2009 (Marichal 2010).

(Marichal 2010) describes the sociological phenomenology of these latest bubbles, especially the one related to mortgages. The new home buyers were confident that the prices were not going to decrease, many people bought, trusting that the majority was not wrong. However, when conditions change due to an increase in mortgage payments, everyone rushes to sell and the property values decrease. The consequence in United States between 2006 and 2008 was the suspension of mortgages payments and the collapse of many banks or stock market funds that were based on these values, which caused a bankruptcy situation widespread.

The effects of this crisis on the world economy were devastating. In 2009, world GDP suffered a contraction of 2.1%, the most significant since 1945, world trade fell by 12% in that year, the unemployment rate grew to 10% in the countries of the European Union, as well as in the USA, In contrast to its 2007 rate of 4.6%. OECD countries had around 15 million more unemployed people in 2009. Although the crisis was in the USA, EU also had serious consequences in the poorest countries, which saw a decrease in the level of trade in their products. As a result, it is estimated that poverty was increase for 64 million people (Keeley and Love 2010).

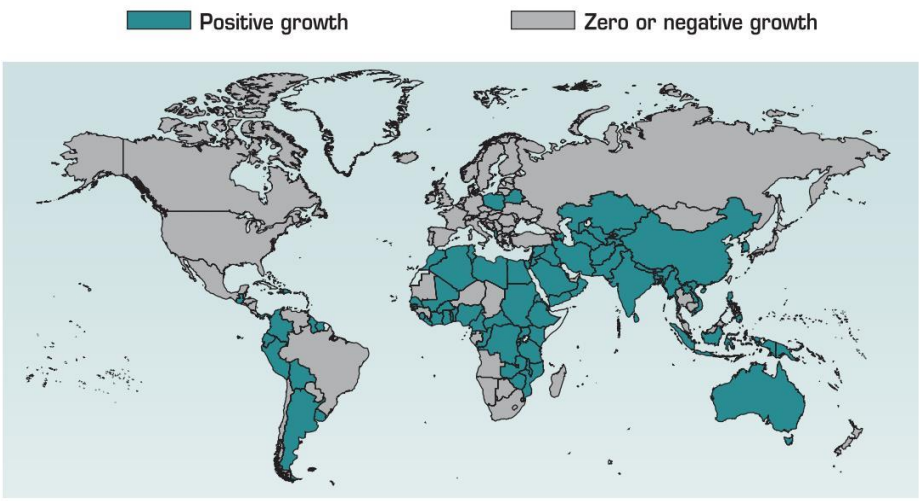

Figure 2: Change in real GDP in 2009 (Keeley and Love 2010, 40)

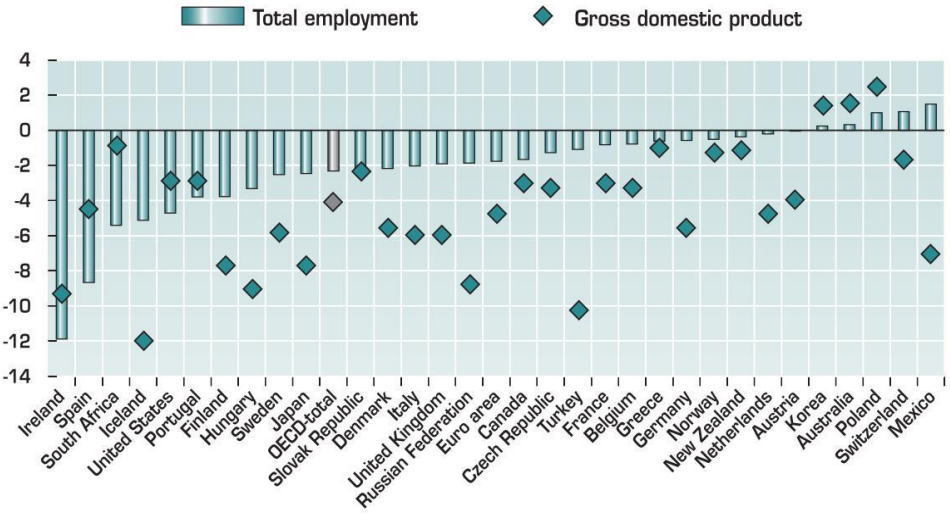

Figure 3: Falling growth and rising unemployment in the recession (Keeley and Love 2010, 38)

In the 20th century, with the digital era, the way of doing things was revolutionized, online trading and digital platforms have in some way replaced the traditional approach (Mirdamad and Mammadova Gulsum 2024), allowing more people to access the market. Having all the information available on the web has allowed for more conscious and informed transactions. Additionally, platforms have appeared that allow operations to be executed with preestablished logic (Rühr et al. 2019).

Legal Regulations

The present and future of the global economic situation in general and the stock market in particular is closely related to the so-called Data Economy. as quoted by the European commission,

"The data economy measures the overall impacts of the data market – i.e. the marketplace where digital data is exchanged as products or services derived from raw data – on the economy as a whole. It involves the generation, collection, storage, processing, distribution, analysis, elaboration, delivery, and exploitation of data enabled by digital technologies" (EUROPEAN COMMISSION 2017, 2)

The emergence of this economy implies strong competition between major participants: United States, China, and the European Union. The first two are ahead in terms of technological advancement, with developments that generate concern, especially in terms of Artificial Intelligence (AI) and Deep Learning (DL), due to the lack of regulation. In the United States and China, the tendency has prevailed not to issue strict regulations to not lag the technological race. The truth is that there are no major regulatory developments in these matters at an international level. UNESCO took a first step by issuing recommendations in 2021 for an ethical use of Artificial Intelligence and in 2024, the European Union issued the first law regulating AI on an international level.

Characteristics of stock markets

The stock markets are important in the economy growth, since issuing shares purchased represents capital for the company that issued the share, which can represent growth of the company, new products, and more employment. Additionally, it has advantages for investors because it is possible to obtain higher returns (Hanif and Alifiah 2021). This seems very attractive to new investors who want to get profits quick; however, this is not an easy task for the newcomers and can be considered as difficult to understand. However, there are some principles that can help to understand how the market works.

Market participant in trading:

Market participant	**Definition**
Issuers	These are companies that issue shares to increase their capital and are sold as parts or fractions of the company
Investors	People or companies that buy the shares.
Stock Exchanged	Markets that facilitate the business of buy and sell shares between the issuer and the investor.

Brokers	Institutions that act as investors' agents, helping with trading operations.
Regulatory Entities	Entities such as the SEC in the United States are responsible for supervising stock purchase and sale operations and preventing cases of fraud.

Table 2. Market participant

The price for the shares is not fixed, it varies even during the day. There are various factors that influence the price, such as the following:

- Supply and demand: when there are more sellers than buyers the price goes down, on the contrary, if there are more buyers than sellers then the price goes up (Almashaqbeh et al. 2021).
- Company performance: how well the company is doing in the market, whether it is making investments, brand recognition, positive and negative news can also affect prices (Almashaqbeh et al. 2021).
- Market sentiment: the optimism and pessimism generated by investors in the market can cause prices to vary (Maknickiene et al. 2018).
- Economic conditions: inflation, interest rates, economic growth.
- World events: geopolitical events, wars, pandemics, natural disasters can also affect prices (Botunac et al. 2024).

Price prediction has been of popular interest for a long time, the possibility of obtaining a return on the money invested has attracted many, which is why various strategies have been designed to obtain higher returns, some of them below:

2.1 Traditional Methods

Fundamental analysis:

In a changing market as the stock market, informed decisions need to be made when buying and selling stocks. Fundamental analysis uses financial statements, economic conditions, and publicly available information to establish the value of a company and determine the closest value to the real value of the shares (Almeida and Vieira 2023) (J. Kumaran, G. Ravi, T. Mugilan 2013).

Fundamental analysis analyzes all the factors that can affect an asset such as macroeconomic and microeconomic factors to establish the value of the asset (Badruzaman 2018).

- Microeconomics and financial conditions: the balance sheet, income statement, and cash flow are used to review the economic solvency and financial stability of a company.

- Macroeconomics and industry conditions: market trends, the launch of new products, and competitors are evaluated to establish the growth of the company.

Fundamental analysis turns out to be very useful for future earning predictions, as it can predict the intrinsic value of the company (Budiman et al. 2022). Additionally, by knowing the financial health of a company, it is possible to reduce investment risks and reduce losses in the future.

Financial analysis gives visibility to undervalued companies that have good financial health and have capacity to grow over time (Saputro and Ariyasa Qadri 2024). On the other hand, it is necessary to be cautious with short-term predictions. This analysis does not see the fluctuations and rapid price changes inherent to the market (Almashaqbeh et al. 2021). Another disadvantage is based on limited company data or not available to the public (Fikriyah and Suhartini 2023), which leads to inaccurate predictions.

In the current context, this methodology becomes important because with new technologies, data is available to the general population, which facilitates the development of more detailed analyses. With the arrival of artificial intelligence and big data, it is possible to process large amounts of financial information, which allows identifying patterns or trends for decision-making (Johnson et al. 2021). Therefore, more advances are expected shortly.

Fundamental analysis has been the subject of several papers that evaluate the performance of this methodology in diverse study fields, some recent studies will be shown below:

- (Fairfield and Whisenant 2000) This study provide evidence of the positive results of using fundamental analysis to predict negative company returns. This research used 373 companies classified by the Center for Financial Research and Analysis (CFRA) with operational problems and unusual accounting practices over a 4-year period. This organization uses fundamental analysis techniques and relies on public information available from companies, such as documents submitted to the SEC.

  Among the most notable conclusions of this research is that negative results are confirmed for up to two years after the results provided by CFRA, which uses fundamental analysis to establish its forecast. The study does not indicate how these forecasts are made through fundamental analysis, which leaves the door open for other research to explore the details. Additionally, the study indicates that financial statements provide tools to predict operational problems, but these signals are not addressed by the market in time.

- (Muhammad and Ali 2018) This study indicates that fundamental analysis facilitates the forecasting of future returns in emerging economies such as Pakistan. This study used

data from 2007 to 2017 for 115 companies listed on the Karachi Stock Exchange (KSE). Financial indicators, liquidity ratios, profitability ratios and market-based ratios were used. The result of the study indicates that it is possible to predict returns in these companies, and that it is possible to use fundamental analysis for this.

This study indicates that market-based ratios and profitability have a greater impact on stock markets than the other indicators. In accordance with other previous studies mentioned in the literature of the paper.

Technical analysis:

As indicated in (Almeida and Vieira 2023) this analysis focuses on the study of historical prices to identify trends. Technical indicators use graphical and mathematical tools to make the respective predictions. This technique assumes that it can be predicted with historical data and that trends can be repeated.

Unlike fundamental analysis, this can be translated into specific strategies for trading, which helps in decision-making to adapt quickly to the market. Despite its detractors, it is widely used in different sectors (Almeida and Vieira 2023) it is possible to detect intrinsic market patterns, which allows predicting future prices.

Technical analysis also contributes to risk management, identifying and quantifying potential risks (Rani et al. 2024). Price fluctuations can be measured by volatility, more specifically the Bollinger bands indicator (Liu 2023). Additionally, it is possible to integrate sudden price changes into the predictions using the RSI indicator (Panigrahi et al. 2021).

Other technical indicators, such as moving averages, serve as trend indicators to reveal the direction of prices (Agusta et al. 2022). RSI also gives signals of oversold or overbought (Panigrahi et al. 2021), which could be used to increase profits or reduce risk exposure.

Technical analysis has characteristics that make it very attractive to the market, but a single indicator may not have the desired precision in the predictions. as indicated in (Agusta et al. 2022), it is advisable to use a combination of indicators to have predictions more adjusted to the real market.

Technical analysis presents the advantage of carrying out a short-term strategy responding to the changing needs of the market (Almeida and Vieira 2023). It is also a great ally if it is used as validation of the fundamental analysis, by comparing the results of the financial statements with certain indicators it is possible to obtain certainty in the decision to be made.

The disadvantage of this analysis is that it does not consider external, geopolitical, or environmental factors that can significantly affect the movement of stocks. Additionally, it is observed that some of these indicators detect fluctuations and create signals after the moment has passed, which makes the strategy no longer valid for the moment. This is not a method with the exact solutions to all trading predicaments, as it also makes mistakes, but its long use in different sectors has made it the ally of many investors (Huang et al. 2019).

Among the recent advances in technical analysis are the new tools that have been created to facilitate calculations and identify trends (Huang et al. 2019). Additionally, different studies have combined these indicators with machine learning and deep learning. Several studies have evaluated the performance of technical analysis in diverse study fields. Below are some recent studies:

- (Hartono and Sulistiawan 2015) This research indicates that better returns are obtained in declining markets with the use of technical analysis. The study used data from 21 countries focusing on the year 2011 that presented significant falls in the global market and 2010 was taken as the control year. For the analysis, 11.044 technical analysis signals from Yahoo Finance were used and indicators such as simple moving average SMA and weighted moving averages were used.

  The study indicated that technical indicators perform better in declining markets than buy and hold strategies, but that their performance is lower for non-declining. The results indicate that understanding the market is important to choose the necessary strategy and that technical indicators can be affected by market conditions.

- (Souza et al. 2018) This research conducted a profitability study using technical and fundamental analysis in stock markets of BRICKS member countries. The study evaluated assets traded in the BRICKS countries. The results indicate that for some countries such as Russia and India the returns significantly exceed the investment, and the other countries on average there is a return that exceeds the value invested.

  In some market segments, strategies using technical analysis can surpass the basic buy and hold strategy. However, it is indicated that few combinations of moving averages maintain returns when compared to a long-term buy and hold strategy. The research also indicates that although some member nations share similar market characteristics, the results obtained are very heterogeneous when compared to each other.

Quantitative analysis:

This analysis contemplates the use of mathematical and statistical methods to analyze data and make financial decisions (Duan 2024). Unlike fundamental analysis, which uses financial statements for the financial analysis of a company, this focuses on establishing prediction of price movements.

These techniques are characterized by generating strategies to manage risk-making decisions based on data (shangchen 2020), using statistical and mathematical models to predict market trends based on historical price data. Among its types of analysis, the following stand out:

Regression analysis consists of understanding the relationship between variables to make predictions. This methodology uses a dependent variable that is normally the target variable and independent variables that are the features with which the target variable will be predicted. The simplest form of this analysis is linear regression, it identifies the linear relationship between two variables and the influence of one on the other (Ahmed 2015).

Time series: It is a statistical technique used in data in chronological order to analyze and seek patterns or trends in the data (Diggs-McGee et al. 2019). It is widely used in the analysis of financial markets to recognize changing trends in data (Qiangwei et al. 2022).This analysis takes specific periods with different data points. An example of this is the ARIMA model, which gives accurate predictions in short-term periods (Hou 2023).

This method allows the identification of price trends, measuring the volatility of the data, and generating signals based on statistical patterns found. However, this analysis assumes linearity in some cases as linear regression (Krivoshchekov et al. 2022), data with different distributions can lead to wrong predictions. Additionally, this method requires statistical knowledge which can make its use difficult, and it is also not capable of capturing the complexity of the market.

The effectiveness of quantitative analysis has been evaluated in numerous research areas. Below are some recent studies:

- (Ariyo et al. 2014) This study uses autoregressive integrated moving average ARIMA for stock market predictions. ARIMA is applied to stocks listed on the New York Stock Exchange (NYSE) and the Nigeria Stock Exchange (NSE). The study confirms that this methodology can accurately forecast prices for short periods and can compete with other forecasting techniques.

- (Adebayo et al. 2014) This study selects the best ARIMA model for forecasting stock markets in Nigeria and Botswana using model selection criteria such as BIC, RMSE, MAE, AIC and HQC. These criteria allowed to choose the best model for each country.

2.2 Machine learning

This discipline uses machine learning algorithms to make decisions based on data (Zhaofeng and Banghao 2024), this historical data is used to understand patterns and make future predictions. Machine learning has changed the way of doing things because it can process large amounts of information in less time, hand in hand with the development of technology, it has been possible to improve the precision of the predictions even surpassing traditional methods (El Hajj and Hammoud 2023).

The process consists of the model learning about patterns and trends in large amounts of data in the initial phase known as training, and from that learning making decisions with data that it has not yet seen in the phase known as testing or validation (Duan 2024).

The models can be classified based on the type of training into 3 main classes: supervised learning, unsupervised learning and reinforcement learning (Géron 2019). Supervised learning consists of learning the relationship between a set of data and an objective value to make a prediction. (Goodfellow et al. 2016; Dixon et al. 2020) Basically the characteristics learn from that relationship to make predictions on new data (Goodfellow et al. 2016).

Unlike supervised learning, unsupervised learning does not have an objective value or a correct value to learn from; it seeks patterns and relationships in the data (Géron 2019). Finally, reinforcement learning consists of a system of rewards or penalties in which the algorithm learns how to maximize the reward over time.

Machine learning algorithms can have different learning tasks. In case numerical results are required it is known as a regression task, in cases where the task requires classifying between classes it is known as classification. Some models work well for both classification and regression tasks, such as logistic regression (Géron 2019).

Another type of task that the model may have is clustering, in which the algorithm learns to make connections in the data and to group them by those connections (Géron 2019). In cases where clustering algorithms are required, it is possible to use DBSCAN, K-means, and Gaussian mixture models.

Machine learning can be very valuable in the financial sector as it can establish data-driven strategies based on the decisions of previously trained models (Guzman et al. 2015). Predictive models can also help reduce financial risk by forecasting volatility and credit risk. In combination with other tools such as NLP and sentiment analysis, it is possible to analyze news, tweets, or any pertinent information that helps understand market sentiment to improve predictions (Johnson 2023).

Among the limitations of machine learning, it is observed that incomplete data, with biases or missing values can lead to inaccurate predictions. In complex models, it can be difficult to interpret what the decision-making process was, which can lead to biases, lack of transparency, and ethical problems with data manipulation.

Another challenge that this methodology faces is that not all market information is present in the data (Htun et al. 2023), many factors cannot be predicted and affect future predictions and their accuracy.

Machine learning and technical indicators

Other approaches are derived from the combinations of the previous methods to reinforce their use. It is possible to combine technical analysis with fundamental analysis. As an example, financial statements can be analyzed and technical indicators can detect trends or patterns to reinforce the decision. On the other hand, it is possible to carry out a combination of technical analysis and machine learning to automate those trends that can be obtained from technical analysis and have more accurate predictions.

This combination of technical indicators with machine learning has been the subject of many scientific investigations with good results. Recent studies with their precision measurements will be shown below:

- (Agusta et al. 2022) The study presents a strategy for predicting stocks in Indonesia in which it combines technical analysis such as moving averages, simple moving averages, relative strength index, stochastic oscillator, Bollinger bands, Aroon oscillator, and SVM with an accuracy of 77.8%.

- (Shynkevich et al. 2017) This study shows a classification of 75.4% for a two-class target variable using machine learning and technical analysis, Windows Functions with a 15-day forecast window.

- (Sadorsky 2021) This study uses random forests and decision tree bagging with accuracy percentages of 85% to 90% to predict the price direction of the clean energy ETF.

- (Yong Ming et al. 2022) This study uses KNN and random forest for e-commerce prediction using technical indicators such as Relative Strength Index, moving average, Moving Average Convergence Divergence, and Stochastic Oscillator as model characteristics. The results showed that using Ma the errors in the predictions are

reduced, and daily returns are obtained that range on average between -0.0261% and 0.1940%. Additionally, high standard deviations are observed, showing high volatility.

This chapter reviewed the historical development of stock markets and the methodologies used to predict prices. It begins with the history of stock markets and their chaotic start until present day, the entities that regulate them such as the SEC in the United States, legal regulations and some possible factors that influence prices such as demand and world events.

Then the methodologies for predictions are addressed, starting with traditional method such as fundamental analysis that analyzes the financial health of companies, followed by technical analysis that employs indicators to forecast future predictions. Additionally, was explained quantitative analysis that uses mathematical and statistical models for predictions.

The chapter ends describing machine learning and its implementation for trading due to its abilities to process large amounts of information. This chapter also covers studies that combine machine learning with technical analysis. In these studies, it was observed that the combination of these two techniques can have good results in price predictions.

3 Methodology

The methodology employed to develop the machine learning model for the stock recommended action is described in this chapter, following the CRISP-DM approach. this iterative methodology definition is described, and the 6 steps to implement it.

CRISP DM

To succeed in any data mining project, it is required to have an already established benchmark and methodology, thus CRISP-DM was selected for this research. As indicated in (Schröer et al. 2021) and (Mora et al. 2024) CRISP-DM is an iterative methodology for data mining, which can be applied to any area of knowledge and different types of data. This iterative process incentivizes continuous improvement of the project or, as in this case, the model. Additionally, it provides a clear and standard guide that reduces complexity in projects and guarantees quality in the process.

CISP DM approach consists of 6 phases:

Business understanding: in this phase, the objective for the project should be set, this requires understanding the project, the stakeholders, and the project definition. Additionally, it is necessary to get a sense of the resources available.

Data understanding: data should be selected and analyzed. This involves understanding the available data, evaluating its quality and data type, it is also suggested to produce a statistical analysis for the data.

Data preparation: in this phase, it is required to clean the data and select the feature that gives better insights to the model. For cleaning includes activities for eliminating columns with constant values, imputation of missing values , and identification and treatment of outliers. In the case of an unbalanced target variable, it should be recommended to apply a balancing method. Finally, if new features are needed for the model, this step should be performed.

Modelling: in this phase, different models are selected and trained. It is recommended to choose a model that adapts to the characteristics of the data and the type of model (supervised, unsupervised, or reinforcement learning). For this thesis, a supervised classification model will be used.

Evaluation: evaluation metrics are applied for the machine learning models, this step is required to analysis the results according to the project objective, compare the performance of each model, and evaluate how well the model works with data not seen before (test set).

Deployment: This phase is not always applicable to all projects, its objective is to have a practical application in the real world, also it depends more on the stakeholders and the project objective, it can be a report or software. In this phase, future improvements, doubts, or concerns about its practical use are raised, as well as privacy issues.

In this chapter CRISP-DM was explained, the initial steps are business understanding and data understanding, both will be combined into a chapter. Next steps described were data preparation, modelling, and evaluation, all of them will be covered in a separate chapter. For this investigation, there will be no deployment, so the last chapter will be evaluation as shown in the following figure adapted from CRISP-DM methodology.

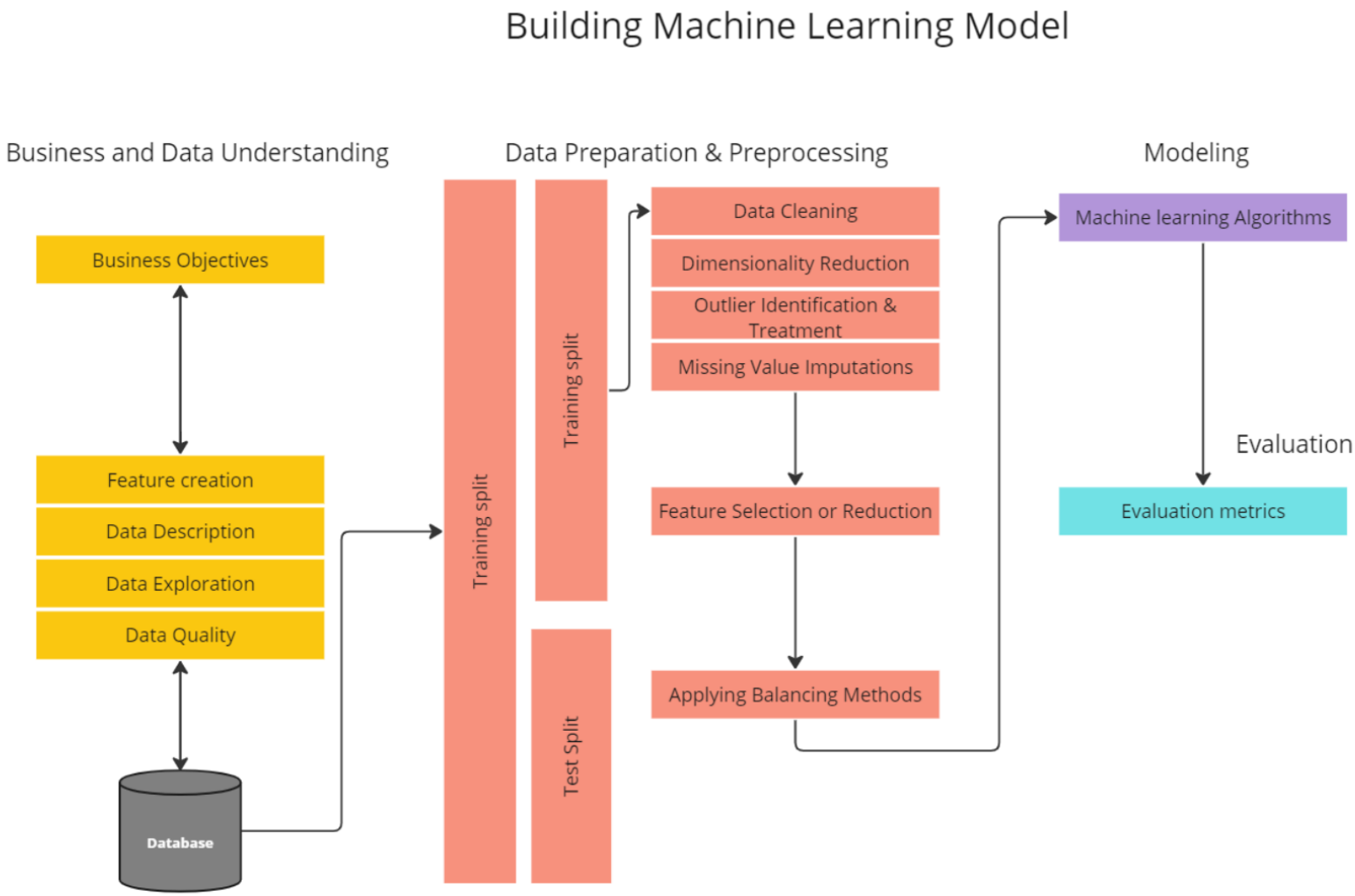

Figure 4. Crisp Dm diagram

4 Business and data understanding

This chapter will combine two important points of the CRISP-DM process, such as business understanding and data understanding. In the first, the problem that arises in this research will be reviewed, a research question is posed and the objectives that respond to that question will also be discussed.

For the second point of business understanding, the data source that will be worked on and the original data set will be considered. The creation of new features that will be used in the model, which are the technical indicators are shown, the explanation of these indicators will be made, and how the clean dataset will look like.

The chapter will end with a description of the stocks that will be studied, their behavior in the studied period, as well as the target variable for each stock and its proportion for the 3 classes considered. It will be finalized with the analysis of the quality of the data, which is important for next stages.

4.1 Business understanding

Technical indicators have been present for many years and have been used in sectors such as banking to predict the future values of stocks, but these indicators cannot always make accurate predictions (Huang et al. 2019), which is why the market is searching for more accurate options. Machine learning allows us to capture patterns that could help make more accurate predictions. This research will integrate technical indicators with machine learning and review how they interact together.

This research seeks to reach the following objectives:

Objective of this thesis

- Create a machine learning model capable of predicting the trends and classify in the 3 classes called "buy," "hold," or "sell" for a given stock. This model will have as input data technical indicators calculated from the stock opening price to identify market movements.
- Select the model that better predicts the opening prices according to the evaluation metrics.
- Compare the winning model to a base model.

Stock selection

To fulfill the objectives mentioned above and build a robust machine-learning model, 3 different stocks were selected to determine the effectiveness of the model. The assets that will be working with are the following:

- **West Texas Intermediate (WTI oil)**

  This stock, also known as Texas light sweet, is a benchmark for crude oil globally, it is used as a reference for oil prices and trading activities around the world for its quality and strategic location. It is known for its low density and lower sulfur content, which make the oil easy to process (Chen 2024).

  This stock is followed by investment because it can influence the market. However, stock market in general can be affected by different economic factors and geopolitical events as is cited in (Botunac et al. 2024).

- **Meta**

  Also known as Facebook, it is a recognized company in the technology sector, it serves as a reference to review the behavior of other companies in the technology sector, its focus is technological, which is why they invest in innovation, virtual and augmented reality.

  It is highly influenced by market news and sentiments, which represents interesting patterns for learning the model. Meta has high-growth potential and its influence on social networks guarantees its position in the market for a long time.

- **ETF iShares MSCI World Index**

  This stock incorporates an exchange-traded fund (ETF). This ETF seeks to track the performance of the MSCI World Index, which measures the performance of different companies in developed countries.

  It groups different types of stocks that provide information on the general behavior of the market. It is very attractive due to its stable behavior in the market, with more generalized and attractive movements for machine learning. Its understanding can facilitate the comprehension of global dynamics since it covers a wide range of actions around the world.

The selection of these assets provides a balanced combination that covers different types of sectors and market dynamics, which will allow the accuracy of the model to be verified with different characteristics, trends, patterns and guarantees its use for different commercial sectors.

4.2 Data understanding

The data was extracted from Yahoo Finance as the principal data source for the stock prices. This website was chosen because it presents different benefits compared to other pages that present the same data. It is free and accessible with historical data of multiple actions and indexes, it has a graphical interface that allows the visualization of stocks in different periods, it also allows the calculation of multiple indicators with the desired parameters. Regarding data extraction, it is possible to import the data into a spreadsheet for easy manipulation. On the other hand, if it is required to use the data in Python, it can be extracted by connecting to the Yahoo Finance API and through the Yahoo Finance Python SDK. This option allows extracting large amounts of data.

For this research, Yahoo Finance API through Python was selected, the historical period covered is 5 years and the frequency of the data is daily. the raw data is structured as shown in the table.

Columns Yahoo Finance dataset						
Date	Open Price	High	Low	Close	Volume	Adj Close

Table 3.Yahoo Finance data

The data extracted from Yahoo Finance was used to create the financial indicators to run the model, the information is organized chronologically, it has some missing data due to weekends or holidays. For this investigation, it will be assumed that the stock market was not active on those specific dates.

	Number of rows per stock						
Stock	Date	Open Price	High	Low	Close	Volume	Adj Close
WTI	1260	1260	1260	1260	1260	1260	1260
Meta	1258	1258	1258	1258	1258	1258	1258
iShares	1255	1255	1255	1255	1255	1255	1255

Table 4.Number of rows per stock

The difference between the number of WTI, Meta and iShares records may be because they are traded on different exchanges. Each exchange has its own schedules and holidays, as well as different data registration practices and unexpected closures.

Features for the model

Regarding the features generated for modeling in this investigation, it was decided to work with technical indicators, these indicators are calculated from Yahoo Finance data prices and can identify market trends, fluctuations and add additional value to decision-making. The most used indicators according to various sources are described in this section.

Relative Strength Index (RSI):

The indicator measures both the rate of change and the change in price, as well as determining whether the stock is oversold or overbought. The system generates numbers ranging from 0 to 100, and it uses the last 14 days to determine this. Typically, a reading above 70 will be deemed overbought, while a reading below 30 will be deemed oversold (Parth Sanghvi 2023) (Pat Tong Chio 2022). Mathematical formula is shown below from (Pat Tong Chio 2022, 10):

$$RSI = 100 - \frac{100}{1 + RS}$$

$$\text{RS}_{\text{t}} = \frac{\text{Average Gain}_{t}}{\text{Average Loss}_{t}}$$

The equation for average gain and loss is the following:

$$\text{Average Gain}_{0} = \frac{\sum_{1}^{14} \text{Gain}_{i}}{14}$$

$$\text{Average Loss}_{0} = \frac{\sum_{1}^{14} \text{Loss}_{i}}{14}$$

$$\text{Average Gain}_{t} = \frac{\text{Average Gain}_{t-1} * 13 + Gain}{14}$$

A stock is overbought when the price rises due to demand, the stock is understood to be overvalued and the price is expected to fall. Oversold is when the stock falls below its real value due to excessive market supply (Khaidem et al. 2016).

MACD (Moving Average Convergence Divergence)

This indicator is a trend momentum indicator, and it is the combination of 2 lines; MACD line and signal line (Parth Sanghvi 2023). Normally, the parameters are set for 12, 26, 9 days (Pat Tong Chio 2022) and for this research will also be used the same. The calculation for this indicator starts with MACD line, which is the difference between two exponential moving averages

(EMA) from days 12 and 26. Then this previous difference is used to calculate the signal line which is the 9th day ema from the difference (Pat Tong Chio 2022).

Mathematical formula is shown below from (Pat Tong Chio 2022, 7):

$$EMA_{12} = \frac{2}{(12+1)} * \text{ closing price }_{12} + \left(1 - \frac{2}{(12+1)}\right) * EMA_{12-1}$$
$$EMA_{26} = \frac{2}{(26+1)} * \text{ closing price }_{12} + \left(1 - \frac{2}{(26+1)}\right) * EMA_{26-1}$$

a) MACD Line:

$$MACD = EMA_{12} - EMA_{26}$$

b) Signal Line:

$$\text{Signal } = EMA_9$$

There are several ways to identify buy and sell signals for this indicator, one of them indicates that if the MACD line crosses above the signal line it is a buy signal and if the MACD line crosses below it is a sell signal (Pat Tong Chio 2022).

Moving Averages (MA)

This indicator is one of the most basic ones, basically, it will be an average of the closing price for N periods divided on the n periods. It is used for predicting trends (Chaddha and Yadav 2022).

Mathematical formula is shown below from (Paspanthong et al. 2018, 6):

$$SMA_n = \frac{P_t + P_{t-1} + \cdots + Pt - n}{n}$$ n: lookback window

For this indicator, all periods have the same weight, which could be considered a disadvantage if it is assumed that the closest prices are more relevant. Another limitation with this indicator is that it does not predict the current movement of the market, but rather goes behind the current value. A very popular strategy used is to buy when the SMA crosses the price from below and sells when it crosses the top value (Ellis and Parbery 2005).

Bollinger Bands

This technical indicator calculates standard deviation of the prices to measure volatility (Di Liu et al. 2019). It consists of 3 bands and the process is the following: the middle band is a simple moving average (SMA) of 14 days closing price, the upper band is the middle band plus (+)

standard deviation and the lower band is the middle band minus (-) standard deviation (Pat Tong Chio 2022).

The width between the bands reflects volatility of the price of the stock, small bands suggest low volatility, while bigger bands represent high volatility (Liu 2023).

Mathematical formula is shown below from (Pat Tong Chio 2022, 9):

Middle Band =SMA(14)

Upper Band = Middle Band + Std

Lower Band = Middle Band- Std

$$SMA = \frac{1}{14}\sum_{t=1}^{14} P_t$$

Stochastic Oscillator

The Stochastic Oscillator is a momentum indicator. It analyses the closing price's relation between the higher and lower price in a time frame. It is composed of two lines K% and D%, Normally, it will have values between 0 and 100 (Paik et al. 2024). If K% value is bigger than D% value it is recommended to buy the stock, in the opposite case it is recommended to sell (Karki et al. 2023)

Mathematical formula is shown below from (Karki et al. 2023, 77):

$$\%K = \frac{C_t - L_{t,t-n+1}}{H_{t,t-n+1} - L_{t,t-n+1}} x100$$

$$\%D = EMA_{3\text{ day}} \text{ of } \%K$$

Where:

t= latest time of observation

n= number of time-period

C_t Closing price at time t

$H_{(t,t-n+1)}$ = Lowest closing price from time t to time $(t-n+1)$

$H_{(t,t-n+1)}$ = Highest closing price from time t to time $(t-n+1)$

Exponential Moving Average (EMA or EWMA)

This technical indicator is a type of moving average that gives more weight to the most recent prices, giving more emphasis on the current data compared for example to the simple moving average that gives the same weight to all the prices (Liu 2023). As an advantage, it is helpful to predict trends in the data using the average price in a specific time (Botunac et al. 2024) and it is good for reducing noise in the data (Liu 2023).

Its formula used the closing price and a smoothing factor, this factor determines the importance or weight of the most recent data. The election of this smoothing factor is important because it can affect the final result of the indicators (Liu 2023).

Mathematical formula is shown below from (Liu 2023, 162):

$$\text{EWMA}_t = \begin{cases} S_0, t = 0 \\ \alpha \text{S}_t + (1-\alpha)\text{EWMA}_{t-1}, t > 0 \end{cases}$$

EMA provides more accurate results than simple moving averages but more false predictions as well because it reacts faster to price fluctuations (Liu 2023).

Ichimoku cloud

This indicator is a combination of different moving averages that allows to check market trends. This indicator provides insights into how the market changes quickly (Almeida and Vieira 2023). Here are the components and formulas of this indicator according to (Che-Ngoc et al. 2022, 1784–1785) for this project the parameters selected are 9, 26 and 56:

Tenkan-sen (Conversion Line): the average of the highest and lowest prices from the last 9 periods. Additionally, as mentioned in (Almeida and Vieira 2023), this line is considered the fastest and produces the most signals in this indicator. It provides short-term information on price changes or trends, as well as entry points.

$$T(t) = \frac{\max x_{\max}(t-i) + \min x_{\min}(t-i)}{2}, i = 0, \dots, 8$$

Kijun-sen (baseline): the average of the highest and lowest price from the last 26 period. This line also helps to identify trends but with a slow reaction (Almeida and Vieira 2023).

$$K(t) = \frac{\max x_{\max}(t-i) + \min x_{\min}(t-i)}{2}, i = 0, \dots, 25.$$

Senkou Span A (Leading Span A): the average of Tenkan-sen and the Kijun-sen, plotted for the following 26 period.

$$A(t) = \frac{T(t-25) + K(t-25)}{2}.$$

Senkou Span B (Leading Span B): the average of the highest and lowest price from the last 52 period, plotted for the following 26 period.

$$B(t) = \frac{\max x_{\max}(t-i) + \min x_{\min}(t-i)}{2}, i = 25, \dots, 76$$

Senkou Span A and Senkou Span B helps in the prediction of future market trends, Both together form the cloud, a thicker cloud may indicate greater volatility while a thinner cloud may indicate lower volatility. (Almeida and Vieira 2023)

Chikou Span (Lagging Span): closing prices of the last 26 period, plotted on the current price chart. This line allows us to observe the trend both in the past and in the present.

$$C(t) = x_c(t+25)$$

Standard deviation

This indicator quantifies the amount of variation or dispersion around the mean, a high standard deviation indicates that data is more spread out, while a low indicates that the points are close to the mean. For financial forecasting, normally it is useful to estimate the volatility of the return on investment. (Liu 2023).

Mathematical formula is shown below from (Liu 2023, 126):

$$\sigma_P = \sqrt{\frac{\sum_{i=1}^{N} (R_i - R_P)^2}{N}}$$

In summary, these are the indicators used and their advantages, disadvantages and how to handle the results:

Indicator	Pros	Cons
RSI (Relative Strength Index)	- Identifies if a stock is overbought or oversold.	- Give false signals during sudden price change.(Goold 2023) -Keep overbought/oversold signal for long periods (Goold 2023).

MACD (Moving Average Convergence Divergence)	- Indicates trend direction (Anghel 2015) - Helpful in combination with other indicators	- It is possible that signals may be delayed (Anghel 2015).
Moving Averages	- Good at identify trends.	- Delayed signals. - Same weight for all the days
Stochastic Oscillator	Indicates price turning points(Paik et al. 2024).	- Can give false signals (Paik et al. 2024)
EMA (Exponential Moving Average)	- It reacts faster to price fluctuations. - identify trends quickly.	- Possible false signals for reacting faster. - Sensitivity to the noise in the data
Bollinger Bands	- Indicator for volatility.	- It is possible that signals may be delayed (Ushman 2023).
Ichimoku Cloud	- Check market trends and momentum	- At the beginning can be difficult to interpret the graphs(Ojedokun 2022).
Standard Deviation	- Indicator for volatility	- Additional analysis is required to be able to decide

Table 5. Indicators advantages and disadvantages

Indicator	How to Handle Results
RSI (Relative Strength Index)	Above 70 will be deemed overbought, while a reading below 30 will be deemed oversold.
MACD (Moving Average Convergence Divergence)	if the MACD line crosses above the signal line it is a buy signal and if the MACD line crosses below it is a sell signal
Moving Averages (SMA)	SMA crosses the price from below and sells when it crosses the top value
Stochastic Oscillator	If K% value is bigger than D% value it is recommended to buy the stock, in the opposite case it is recommended to sell
EMA (Exponential Moving Average)	Buy when closing prices crosses above EMA; sell when it crosses below(Puchong Praekhaow 2010).
Bollinger Bands	- Buy when the prices touch the lower band, sell when it reaches the upper band(Prasetijo et al. 2017).
Ichimoku Cloud	-if the conversion line crosses over the base line is a buy signal and when the conversion line falls below is a sell signal(Byun 2021).
Standard Deviation	- Volatility Measurement

Table 6.How to Handle Results

Target Variable

Unlike other datasets in which the target variable is found in the data, for this research, it was necessary to create the target variable, which in other words is what the model seeks to predict.

The target variable was called daily change, and it is defined as the percentage of change in the opening price of the stock. If the percentage of change is greater than 1% positive, then it indicates the instruction to sell and is marked as 1. If the percentage of change is greater than 1% negative, then it indicates the instruction to buy and is marked as 2, and if it is equal or less than the absolute value of 1 then it is hold.

$$Daily\_Change_{\ t} = \left(\frac{\text{Open}_{\ t} - \text{Open}_{\ t-1}}{\text{Open}_{t-1}}\right) \times 100$$

Where:

- $\text{Daily\_Change}_{\ t}$ is the percentage change in the Open price on that day.t
- $\text{Open}_{\ t}$: is the Open price on that day.t
- $\text{Open}_{\ t-1}$: is the Open price on the previous day.

And for target variable creation:
sell_threshold=1.0
buy_threshold=-1.0

$$Target_t = \begin{cases} 1\ if\ Daily\_change_t > Sell\_threshold \\ 2\ if\ Daily\_change_t < Buy\_threshold \\ 0\ otherwise \end{cases}$$

After the creation of all the features already mentioned, the dataset for the modeling is formed, it contains the independent variables for the model consisting of the technical indicators and the target variable which is the one that will be predicted.

Indicator Name	Name in the dataset
RSI	RSI_14
MACD	MACD_12_26_9
MACD-Histogram	MACDh_12_26_9
Signal	MACDs_12_26_9
Moving Averages	SMA_50

Stochastic Oscillator %K	STOCHk_14_3_3
Stochastic Oscillator %D	STOCHd_14_3_3
Exponential Moving Average	EMA_9
Lower Band	BBL_20_2.0
Middle Band	BBM_20_2.0
Upper Band	BBU_20_2.0
Bandwidth	BBB_20_2.0
Percent B	BBP_20_2.0
Senkou Span A	ISA_9
Senkou Span B	ISB_26
Tenkan-sen	ITS_9
Kijun-sen	IKS_26
Chikou Span	ICS_26
Standard deviation	STDEV_30
Target	Target

Table 7. Features for the model

Data Description

Once the characteristics for the dataset have been created, the stocks that we are going to use in the models will be analyzed, specifically the behavior of the opening price, and then the distribution of the target variable will be reviewed.

ETF iShares MSCI World Index

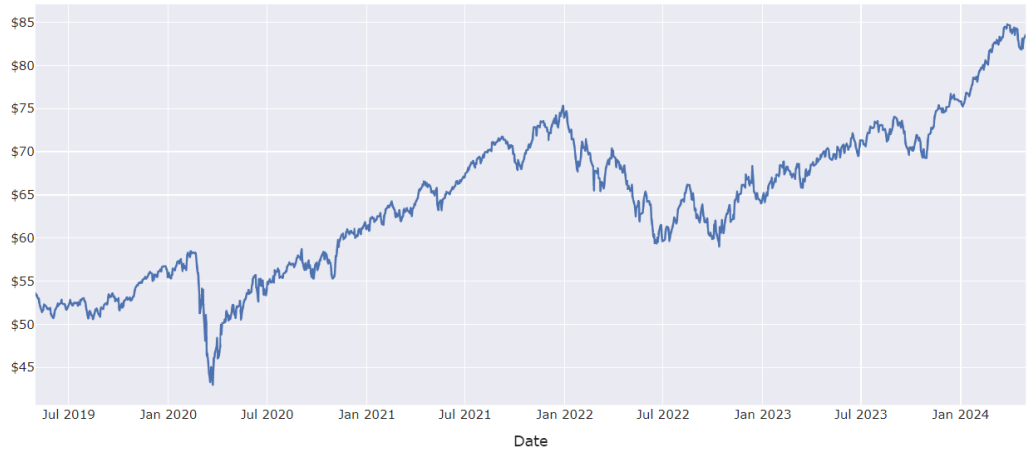

Figure 5. ETF iShares MSCI World Index

The previous graphic shows showing ETF iShares MSCI World Index between May 2019 to May 2024, The opening price in general is on the rise, however, there are 3 periods to analyze in this graph that are important to highlight, the first is before the pandemic, a constant growth in stock is seen between $53 CAD in July 2019 and $59 CAD in January 2020.

Then a drop was seen in March 2022 due to the pandemic with the price falling below $45 CAD. This drop reflects global uncertainty and massive stock sales. However, despite the fall, it had rapid growth from June 2022 until January 2022 in which it experiences a slight price drop but then recovers its upward trend as before the pandemic.

Figure 6. Meta

This graphic shows Meta's opening price between May 2019 to May 2024, at the beginning, the graph shows a constant behavior with values less than 200 USD for July 2019, then it shows a drop in March 2020 due to the pandemic less than USD 150, but it recovers quickly and has gradual growth with some fluctuations between July from 2020 to January 2022. then it presents an unprecedented new drop in October 2022 with less than 100 USD due to the group's decision to invest in large amounts in metaverse projects.

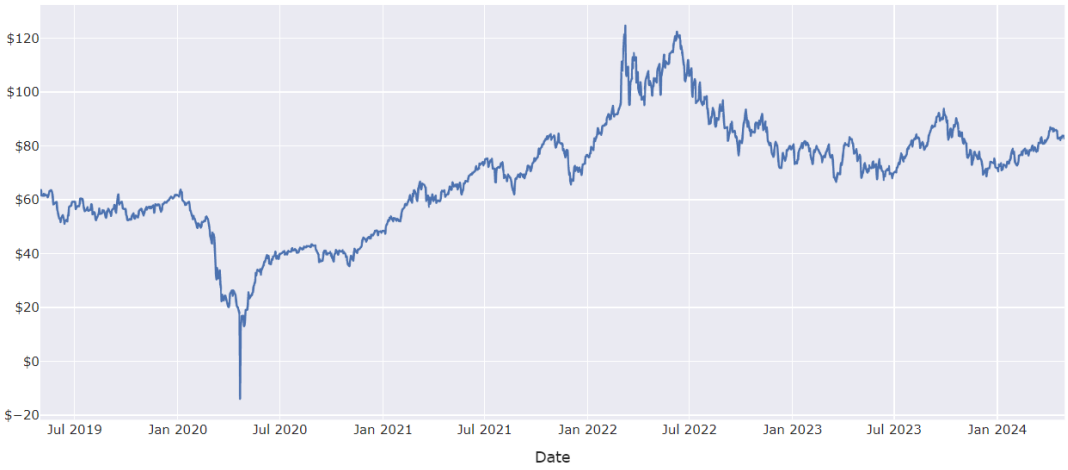

Figure 7. WTI Oil

This graphic shows the indicator WTI Oil for the same period as the previous one, The opening price, in general, is on the rise, however, presented a drop never seen before in which crude oil prices reached negative values of -20 USD, again due to the pandemic. After the sharp drop, recovery was observed in crude oil values with several fluctuations until reaching maximum points in the first quarter of 2022 with values greater than 120 USD.

Target Variable

The following graph shows the behavior of the target variable. In the case of Meta and WTI oil, a balanced target variable is observed; thus, the model will have enough of each class to learn and then predict, on the contrary, iShares is unbalanced with proportions of 79% class 0, 11% class 1 and 10% class 2 which can make difficult for the model to learn.

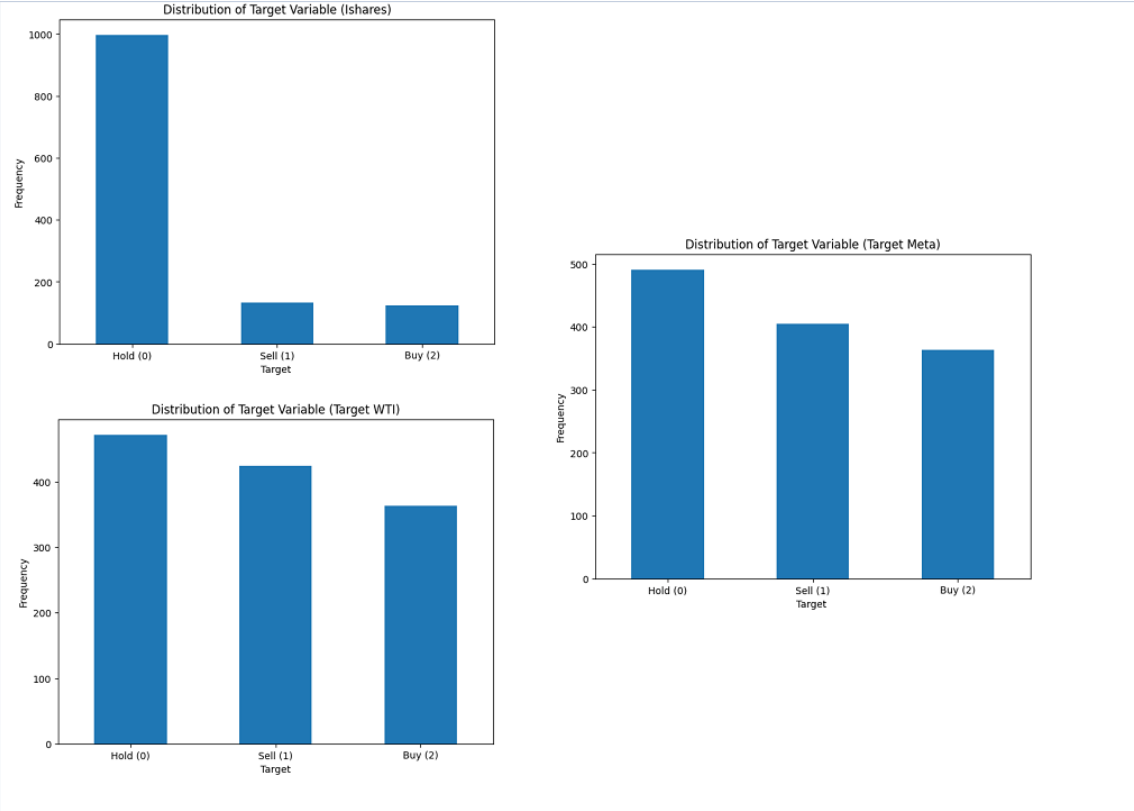

Figure 8. Distribution Target Variable

Data quality

Before proceeding with the preparation of the data, it is recommended to review the quality of the data within the Crisp DM process. This analysis consists of reviewing the data with descriptive statistics to understand the distribution of the data, the central tendency and volatility. specifically for this investigation, it was calculated the variance of the data to check the volatility of the data, a check was performed to review count, median, mean, 25th percentile, 50% and 75% for central tendency. Additionally, it can help identify possible outlier values.

Meta_indicators	count	mean	min	25%	50%	75%	max	std
RSI	1245	53.3	17.2	45.5	53.9	62.2	88.5	12.4
MACD	1234	1.7	-28.9	-2.2	1.9	6.3	29.6	7.7
MACD-Histogram	1226	0.0	-9.8	-1.0	0.1	1.1	7.8	2.0
Signal	1226	1.8	-27.1	-2.0	1.9	6.3	26.8	7.4
Moving Averages	1210	251.6	112.4	189.3	248.2	306.3	496.4	79.0
Stochastic Oscillator %K	1244	56.5	3.0	33.4	62.6	78.9	98.0	25.3
Stochastic Oscillator %D	1242	56.5	3.2	34.4	62.4	78.5	97.0	24.5
Exponential Moving Average	1251	254.5	99.2	188.4	240.8	313.0	514.4	85.9
Lower Band	1240	234.6	77.0	176.7	222.6	293.0	484.5	82.0

Middle Band	1240	253.8	103.6	188.3	242.4	311.7	507.3	84.4
Upper Band	1240	272.9	120.6	203.3	269.3	331.6	557.8	88.4
Bandwidth	1240	16.1	3.6	10.1	13.1	18.9	66.8	10.2
Percent B	1240	0.6	-0.4	0.3	0.6	0.8	1.4	0.3
Senkou Span A	1208	249.0	105.3	187.8	236.4	307.8	495.4	77.2
Senkou Span B	1182	246.4	111.0	187.4	255.2	300.3	439.5	71.8
Tenkan-sen	1251	254.4	95.3	188.5	241.4	312.3	514.4	85.9
Kijun-sen	1234	253.0	106.1	187.5	252.9	308.7	506.4	83.1
Chikou Span	1233	256.5	90.1	190.0	246.8	315.0	529.3	87.3
Standard deviation	1230	12.1	2.8	8.0	10.4	13.9	50.2	7.5

Table 8. Meta statistical analysis

WTI_indicators	**count**	**mean**	**min**	**25%**	**50%**	**75%**	**max**	**std**
RSI	1246	52.1	12.3	43.9	52.3	60.8	86.9	12.1
MACD	1235	0.1	-7.2	-0.9	0.3	1.4	8.5	2.0
MACD-Histogram	1227	0.0	-2.2	-0.3	0.1	0.3	2.9	0.6
Signal	1227	0.1	-6.6	-0.8	0.2	1.4	6.4	1.9
Moving Averages	1211	68.6	22.6	55.9	73.2	80.0	110.5	20.1
Stochastic Oscillator %K	1245	57.0	6.7	34.7	62.3	78.7	96.3	24.3
Stochastic Oscillator %D	1243	57.0	9.2	35.3	62.0	78.0	96.1	23.6
Exponential Moving Average	1252	68.7	14.1	55.4	71.6	81.3	119.4	20.4
Lower Band	1241	62.9	0.7	51.4	67.1	76.2	108.4	19.9
Middle Band	1241	68.6	17.3	55.6	72.1	81.3	116.4	20.4
Upper Band	1241	74.3	27.3	59.6	75.7	86.6	126.6	21.4
Bandwidth	1241	19.6	4.4	10.2	13.9	18.8	192.5	25.3
Percent B	1241	0.5	-0.5	0.3	0.6	0.8	1.4	0.3
Senkou Span A	1209	67.9	-8.0	55.2	71.0	80.3	115.2	21.7
Senkou Span B	1183	67.2	-2.8	55.9	73.0	80.6	111.7	23.8
Tenkan-sen	1252	68.3	-10.4	55.3	70.9	81.5	119.1	21.0
Kijun-sen	1235	67.9	-6.0	55.7	71.5	80.7	114.4	22.2
Chikou Span	1234	68.9	-14.0	55.7	71.8	81.7	124.7	20.7
Standard deviation	1231	3.6	0.7	2.1	2.9	4.4	12.0	2.1

Table 9. WTI statistical analysis

Ishares_indicators	**count**	**mean**	**min**	**25%**	**50%**	**75%**	**max**	**std**
RSI	1241	54.9	18.3	47.7	56.1	63.3	78.6	11.8
MACD	1230	0.2	-3.3	-0.1	0.3	0.6	1.2	0.6
MACD-Histogram	1222	0.0	-0.9	-0.1	0.0	0.1	0.8	0.2
Signal	1222	0.2	-2.9	0.0	0.3	0.5	1.1	0.6
Moving Averages	1206	64.0	49.1	56.7	64.7	70.3	83.0	7.9
Stochastic Oscillator %K	1240	63.3	3.8	38.3	72.6	88.4	99.8	27.9
Stochastic Oscillator %D	1238	63.3	7.1	38.8	72.1	87.4	98.2	27.0

Exponential Moving Average	1247	64.1	45.3	56.8	64.8	70.3	84.2	8.4
Lower Band	1236	62.3	40.3	55.2	62.9	68.7	83.2	8.5
Middle Band	1236	64.1	46.0	56.8	64.8	70.3	84.1	8.3
Upper Band	1236	65.9	49.6	58.5	66.8	72.0	85.5	8.3
Bandwidth	1236	5.8	1.6	3.6	4.8	6.4	35.0	4.0
Percent B	1236	0.6	-0.4	0.3	0.7	0.9	1.3	0.3
Senkou Span A	1204	63.6	46.8	56.6	64.5	69.9	83.0	7.9
Senkou Span B	1178	63.3	47.5	56.7	63.8	69.9	80.5	7.7
Tenkan-sen	1247	64.1	44.8	56.6	64.7	70.2	84.2	8.4
Kijun-sen	1230	63.9	46.4	56.7	64.5	70.1	83.5	8.3
Chikou Span	1229	64.4	43.0	57.0	65.2	70.4	84.8	8.4
Standard deviation	1226	1.1	0.3	0.7	1.0	1.4	5.5	0.7

Table 10. Ishares statistical analysis

In this chapter, business understanding, and data understanding were addressed together, as described in the introduction of this chapter. It begins describing the primary objective: to create a machine learning model capable of predicting the trends and classify 3 classes called "buy," "hold," or "sell", this objective contemplates 3 stocks; Meta, WTI and Ishares to represent diverse segments of the market.

In data understanding section, Yahoo finance is described, this data source is used to extract the opening prices for the technical indicator's calculation. This section covers the definition of the indicators and how the calculation of each of them is carried out. Additionally, prices' variation over 5 years period is described for each of the stocks, this analysis showed a general growth for Meta, WTI and iShares however all of them have an unexpected drop due to COVID-19.

This chapter ends with the distribution of the target variable and data quality analysis. For iShares, the target variable is unbalance meaning more hold signals than buy or sell, while Meta and WTI have a more balance target variable.

5 Data preparation and preprocessing

This chapter is crucial in the Crisp DM process because it prepares the data for the model, the preparation influences the accuracy of the model and provides reliable results.

The chapter begins with the separation of the data, the ways there are to separate the data and the one chosen for the investigation, and the process of cleaning the data will be described. This cleaning begins with the reduction of dimensionality, followed by the identification and treatment of outliers. To finish the cleaning, the identification of missing data and how to attribute it will be reviewed.

The chapter will end with feature selection or reduction to reduce the number of features of the model and balancing for the cases that apply.

5.1 Data split

An important aspect of machine learning models is the data splitting strategy, the historical data must be divided into training and test sets. A good separation of the data may prevent overfitting on the data, in other words, the model learns well with the training set but cannot accurately predict the test set (Korstanje 2021). There are different strategies for splitting, here are some of them:

Strategy	Definition
Random Split	This method consists of randomly split the data into a training set and test set, the most used split is 80% training set and 20% test set or 70% training set and 30% (Géron 2019).
Stratified sampling	This method contemplates the distribution of the classes in the target variable. It guarantees that each class for the target variable is represented in each set, in other words, it will take samples from each class and ensure that training and test set keep the proportion of each class (Géron 2019).
Temporal split.	In the context of trading, a common approach is to use splitting the data chronologically, that is, separating the training data and the test data while preserving the order of the dates, this helps

	capture seasonal trends and make better prediction(Botache et al. 2023). temporal split helps avoid data leaking in forecast trading, this data leaking can occur when data is divided randomly, the model can see future data on the training. Doing it in chronological order ensures that this does not happen.
K fold cross-validation	This method separates the data into k folds with equal size each one, the data is tested with one of these folds and trained with the rest of the folds, this process is repeated k times. Thus, at the end it will generate k evaluation metrics (Géron 2019). cross-validation will give a score of how well the model preforms, but also will give how reliable is that estimation. Normally if a model with random split is compared with k fold score, usually the performance of the model is lower with this last one (Géron 2019).
Walking forward or rolling window	This method addresses the problem that traditional approach normally has. it makes predictions with the new data that becomes available. This method performs the training and validation N times, each time increasing the training rate(Hasanov et al. 2022).

Table 11. Strategies for splitting

Traditional methods for data separation, such as random split and stratified split, can bring benefits for prediction because of the randomness. However, they can be very inconvenient for time series where chronological order is important, which is why temporal split and walking forward are more suitable for this research.

Temporal split and walking forward validation were applied; temporal split was slightly better for the model performance; thus, temporal split was selected. The separation was carried out in two periods, one for training and the other for testing. 70% of data were taken from the first years and 30% for the second respectively.

Dataset	X_train_shape	y_train_shape	X_test_shape	y_test_shape
meta	(881, 19)	(881,1)	(378, 19)	(378,1)
WTI	(882, 19)	(882,1)	(378, 19)	(378,1)
iShares	(878, 19)	(878,1)	(377, 19)	(377,1)

Table 12. Temporal split

5.2 Data cleaning

Cleaning the data is an important stage before making any prediction, as it ensures the quality of the data and improves the efficiency of the analysis. In this section, we will talk about 3 important steps before tackling modelling.

Dimensionality reduction

This technique consists of reducing the number of variables in a data set, eliminating variables that contribute little or nothing to data analysis or machine learning models. Having more features as input in the model, can add more complexity to it. Fewer features are better to predict accuracy of the model (Atsalakis and Valavanis 2009). One way to implement it is with low variance. The variance measures the dispersion of the data, a low variance indicates that the data is more concentrated in the mean, these values with low dispersion such as constants or columns with very few variations do not contribute or contribute very little to the learning of the model and may not be meaningful for the forecast (Géron 2019).For that, it is recommended to set a threshold for the low variance to remove this feature from the analysis (scikit-learn 2024).

For the investigation, the variance was calculated for each feature and a threshold of 0.1 was set. Features with a variance less than this threshold will be eliminated, and features greater than 0.1 will be retained.

Feature	META	ISHARES	WTI
RSI	154.6	138.7	145.6
MACD	59.0	0.4	4.1
MACD-Histogram	3.9	0.0	0.3
Signal	54.3	0.3	3.6
Moving Averages	6236.6	62.3	405.2
Stochastic Oscillator %K	639.2	776.6	592.4
Stochastic Oscillator %D	598.1	729.4	555.8
Exponential Moving Average	7374.3	70.8	415.9
Lower Band	6719.6	71.8	397.4
Middle Band	7130.3	68.9	415.0
Upper Band	7816.6	68.1	458.8
Bandwidth	104.0	16.2	638.2
Percent B	0.1	0.1	0.1
Senkou Span A	5962.9	62.4	469.4

Senkou Span B	5148.5	58.8	565.4
Tenkan-sen	7372.5	71.4	442.0
Kijun-sen	6897.7	68.5	491.9
Chikou Span	7616.3	70.9	429.1
Standard deviation	56.7	0.4	4.4

Table 13. Variance calculation

After calculating the variance, only 1 feature was removed for iShares, leaving it with 18 and meta and WTI with 19.

Outlier identification and treatment

Machine learning models require the identification of atypical values in the data, this anomaly can affect the model training and generate inaccurate predictions. The performance of the model can be increased by removing the Outlier from the data (Géron 2019). For that, it is important to identify the best option for the selected model.

In financial data, it is important to identify Outliers because strategies for purchasing assets are derived from the data. In trading, there are different types of outliers. it can be found atypical market events, such unusual events that only occur sporadically, for example, when the covid hit, the price of oil shares reached negative values. Additionally, Unexpected Falls can appear, prices can fall due to external events (example: geopolitical) that influence prices. (Botunac et al. 2024)

These short-term events, if integrated into the data without any treatment, may be inconvenient for learning the model. Additionally, atypical data may also be found due to registration or typographical errors.

For its identification, different strategies have been established, such as:

Strategy	Definition
Statistical Distribution Detection	distribution base methods that consider outlier, the points that exceed the standard distribution (Smiti 2020).
z-score	Is the calculation of the standard deviations that deviate from the mean, the points that exceed the chosen threshold are considered outliers, it is considered that for a normal distribution of +- 3 deviations, the data included in that interval is 99.73% (BHAT 2023).

Interquartile Range (IQR):	Is the difference between Q3 and Q1, if any observation moves 1.5 times away from the interquartile range, they are considered possible outliers (BHAT 2023).
Boxplots	This is a graphical tool that allows to visualize values that are very far from the majority of the data. "Boxplots are a simple way of representing the five-number summary which consists of five values, the extreme lower (min), the upper extreme (max), the first quartile(Q1), the median also known as the second quartile (Q2) and finally the third quartile(Q3)." (Smiti 2020, 3).

Table 14. Identification strategies

The outlier detection method chosen for this research is the z score, its simplicity allows outliers to be easily detected.

To treat the outliers previously identified, these are the recommended options that were considered in the research:

Strategy	Definition
Elimination	This is a common option, it will remove from the data these atypical values, it should be used carefully (BHAT 2023).
Winsorization	This technique consists of replacing the values identified as outliers with the closest non-outlier value (BHAT 2023).
Cap and floor	This technique consists of replacing the values identified as outliers with the thresholds established in the distribution (BHAT 2023).

Table 15. Outlier treatment

The best option must be determined depending on the data, if it is a small proportion, eliminating them may be an option, for larger percentages of outliers it is better to use a replacement tool. The method chosen for the treatment of outliers is cap and floor, this reduces the impact of outliers

and preserves the data, unlike elimination. Additionally, it works well with different types of data distribution.

Missing Values and imputation

In the prediction process of any machine learning method, the problem of missing values is often faced. In case of financial models, it may be that the data is not available due to errors in data loading or closing markets. These missing values can affect the performance of the model and for some models, cannot even be executed because they need a complete dataset for training (Jäger et al. 2021).

Features with many nulls contribute less to the accuracy of the model and are candidates for removal. This approach deletes all the columns with missing values, it is a simple process because any analysis with this data is required (Emmanuel et al. 2021). For missing values feature removal, it is required to establish a threshold to remove from the data. However, elimination is not a good choice when the proportion of missing values is lower, as the aim is to avoid the loss of information and preserve the sample size. For this research, a strategy was established for the treatment of missing values. Initially, the columns with more than 90% of missing values will be eliminated. In the remaining columns that require imputation, a strategy for filling the values will be applied. Below, a table is presented with the missing values imputation methods considered in the investigation:

Strategy	Definition
Imputation with an estimate	In this type of imputation, the missing values are replaced by mean, median or mode. For numerical values mean it is applied, in case of categorical value mode is applied (Jäger et al. 2021).
Simple Imputation	"Simple imputation treats the imputed values as true values." (Templ 2023, 161) Here each missing value is replaced with an estimate, this can be simple like the mean or something a little more complex like a regression model.
Multiple Imputation	For this type of imputation, different imputations are generated for each missing values, then each group of imputations created for each

	value is combined, a function is applied, and a result is given. This particular method minimizes the prediction error (Jäger et al. 2021).
K-nearest neighbors (KNN) imputation	This imputation uses the K closest neighbors of the missing value to perform the imputation, through a distance measure it identifies these points and the mean or median of the neighbors is calculated (Jäger et al. 2021).
Regression imputation	An independent variable is used (a variable for which the data is known) and another dependent variable (the one that contains the missing values that are to be imputed), then the relationship between these two variables is established through an equation with the following forms Y=mx + b, finally the model is used to calculate the missing values (Jäger et al. 2021).
Hot Deck and Cold Deck Imputation	Hot deck imputation uses random similar values from the dataset (donor) to impute missing values, hot deck uses recent data for the imputation, on the contrary cold deck uses previous period data (Jäger et al. 2021).
Imputation using machine learning	It is also possible to perform imputations with machine learning to predict missing values, which can be very useful when working with data with non-linear relationships As for Random Forests, XGBoost (Jäger et al. 2021).

Table 16. Imputation methods

After a review of the imputation methods, it was decided to proceed with KNN for missing values. This method is flexible since it does not assume data distributions, which makes it attractive for various types of data. Additionally, imputations are carried out considering the closeness of data, which suggests that the imputations maintain the structure of the data.

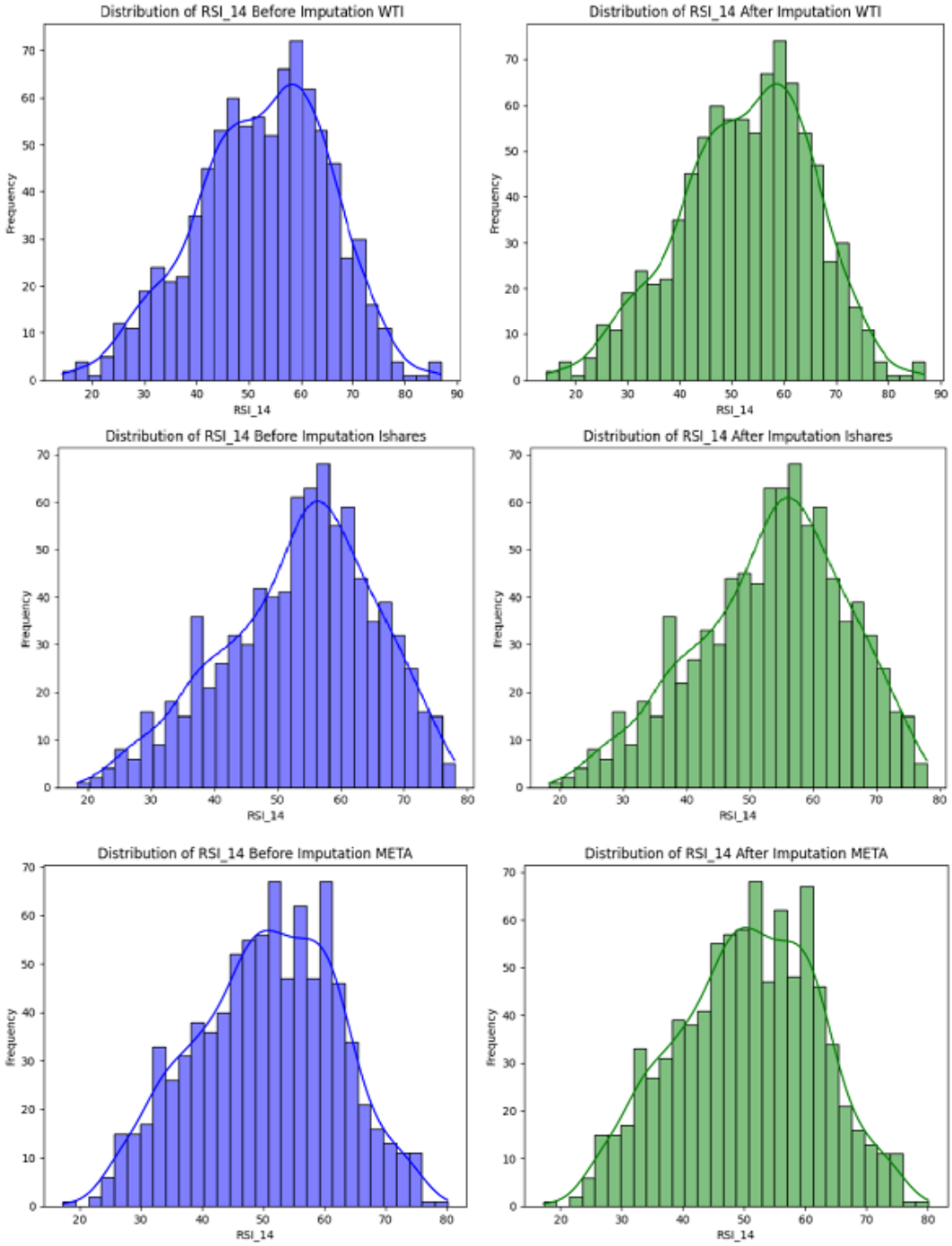

Figure 9. RSI indicator before and after imputation

The graph shows RSI, one of the technical indicators included in the 3 datasets (WTI, ISHARES and META) before and after imputation. In the graph before imputation, peaks between 50 and 60 are observed with a lower number of values in the tails. It can be observed that after imputation, the shape of the distributions is maintained.

5.3 Feature selection and Balancing

feature selection and reduction

The techniques of selection and reduction of characteristics are tools that allow us to reduce the number of features of a model. Many models have thousands of features as impute for the model. This big number of features can make the training slow and make it harder for the model to find a solution. However, it is possible to find a solution that reduces features and generates a good fit (Géron 2019).

The implementation of these techniques is recommended because it reduces the computational power required, since the number of features for training the model is being reduced. Additionally, by reducing these features, the noise of the data can decrease, and a better accuracy can be obtained for the model.

Models with many features have tendencies to overfit, it is recommended to opt for simpler models that are easier to understand and explain. The selection and reduction methods that were contemplated for the execution of the code will be described below:

selection of features:

This technique consists of choosing the most suitable characteristics for the model, instead of combining into new features. (Ramos-Pérez et al. 2024) the remaining data that is considered not necessary or redundant for the model is removed from it. Some methods will be mentioned here:

- Filtered methods: this technique consists of giving scores with statistical measures to the characteristics. These will be analyzed and those with the highest scores will be selected and the rest discarded (Ramos-Pérez et al. 2024).
- Wrapper Methods: This method uses any search algorithm to evaluate which combination of characteristics produces the best performance for a specific classifier (Ramos-Pérez et al. 2024).
- Embedded Methods: these methods are integrated into the modelling process. During the training of the model, weights are given to the features to produce the most optimal value for the accuracy. an example of them is lasso regression (Pudjihartono et al. 2022).

Feature Reduction

Also called dimensionality reduction, these techniques reduce the number of dimensions usually reduces the characteristics by transforming them into other variables, these new variables are a combination of the original variables. Among the best known are:

Principal Component Analysis (PCA): This technique helps to reduce the dimensionality of many inter-correlated variables while maintaining the highest possible variance. Principal components are new variables that are created from the transformations of the initial variables; these new variables are no longer correlated with each other. Among its most notable uses are the reduction of the number of variables in huge datasets, elimination of signal noise, and data compression (Mishra et al. 2017).

For the reduction of large amounts of data, PCA uses a transformation of the vector space. With a mathematical projection, it is possible to summarize the original data set in a few variables (Mishra et al. 2017).

For this step, a feature reduction method was decided. PCA reduces the number of features while maintaining the most significant amount of information, and it is computationally less demanding than other reduction methods. to ensure that it is appropriate for this case, it was applied The Kaiser-Meyer-Olkin Measure test, which helps measure whether the data can be used in PCA.

The following table contains values of reference for this test:

KMO Value Range	Description
0.00 to 0.49	Unacceptable
0.50 to 0.59	Miserable
0.60 to 0.69	Mediocre
0.70 to 0.79	Middling
0.80 to 0.89	Meritorious
0.90 to 1.00	Marvelous

Table 17. KMO Value Range

Dataset	KMO Model	Result
Meta	0.806	Meritorious
WTI	0.816	Meritorious
iShares	0.823	Meritorious

Table 18.KMO results

After the analysis was obtained KMO values around 0.8 for the 3 datasets, which indicates that the dataset is suitable for PCA. Additionally, the correlation matrix was created where correlated values are observed, which is also a sign that PCA may be appropriate.

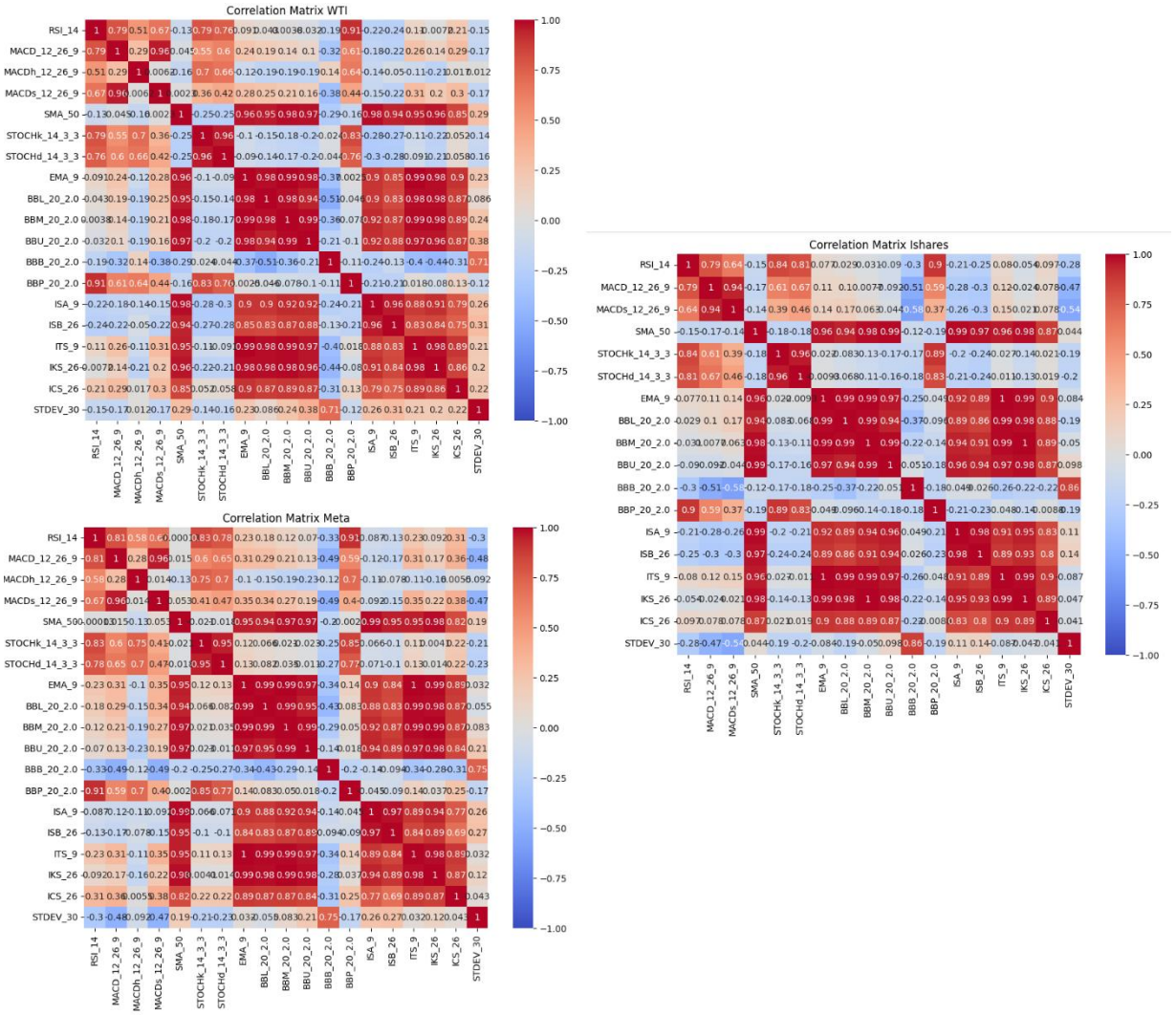

Figure 10.Correlation Matrix

To reduce dimensionality and lose the least amount of information, it was established that within the PCA execution the number of components that explain 95% of the variation will be determined. It was obtained that for iShares there were 4 components, while for Meta and WTI there were 5.

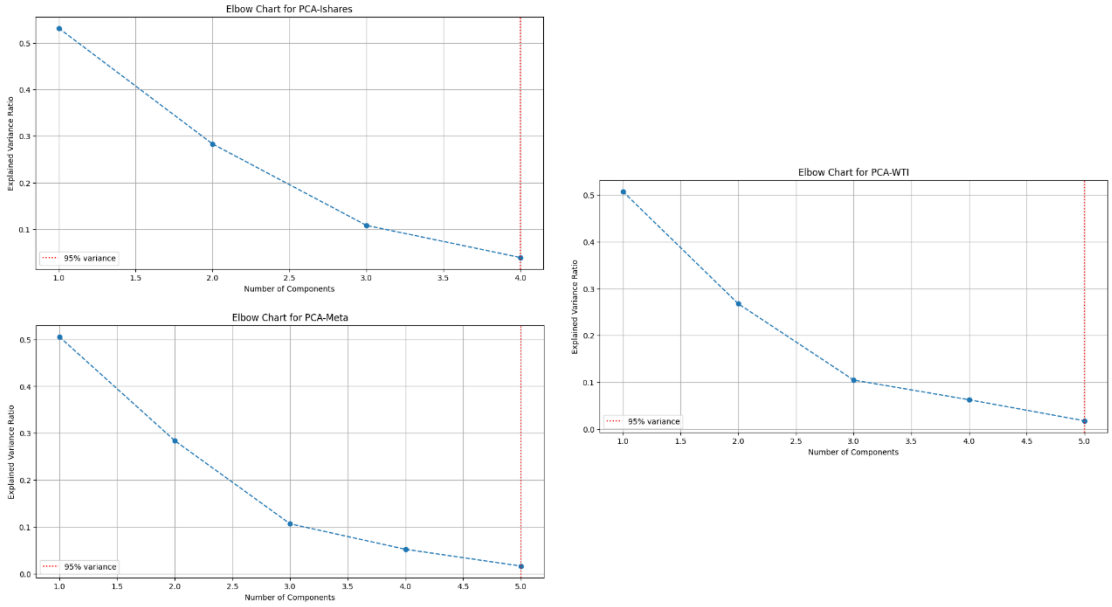

Figure 11. PCA chart

Balancing

Another challenges when using machine learning is unbalanced classes. This phenomenon occurs when there is a lot of difference between the proportions of the classes. This situation can cause the model to learn more from the majority class than from other classes, As a consequence the model tends to predict more the majority class and go unnoticed the possible signs of significant change towards other classes (Werner de Vargas et al. 2023).

Resampling Techniques

- Oversampling the Minority Class: this technique consists of increasing the records of the minority class in the data set. This method duplicated the minority class instance, which can possible lead to overfitting (Jadhav et al. 2022).
- Under sampling the Majority Class: this method reduces the records of the majority class until they are equal to the minority. The disadvantage with this method is that it is possible to remove important information for the model (Jadhav et al. 2022).

Synthetic Data Generation:

- SMOTE (Synthetic Minority Over-sampling Technique): as explained in (Jadhav et al. 2022) this technique creates a synthetic sample instead of creating copies as in the previous cases.

- ADASYN (Adaptive Synthetic Sampling): This method uses a weighted distribution to create the synthetic samples of the minority class and focuses on the minority classes that are more complicated to create.

For this investigation two out of three stocks did not require balancing since the classes of the target variable were balanced, however for the iShares stock Smote was applied giving the following results.

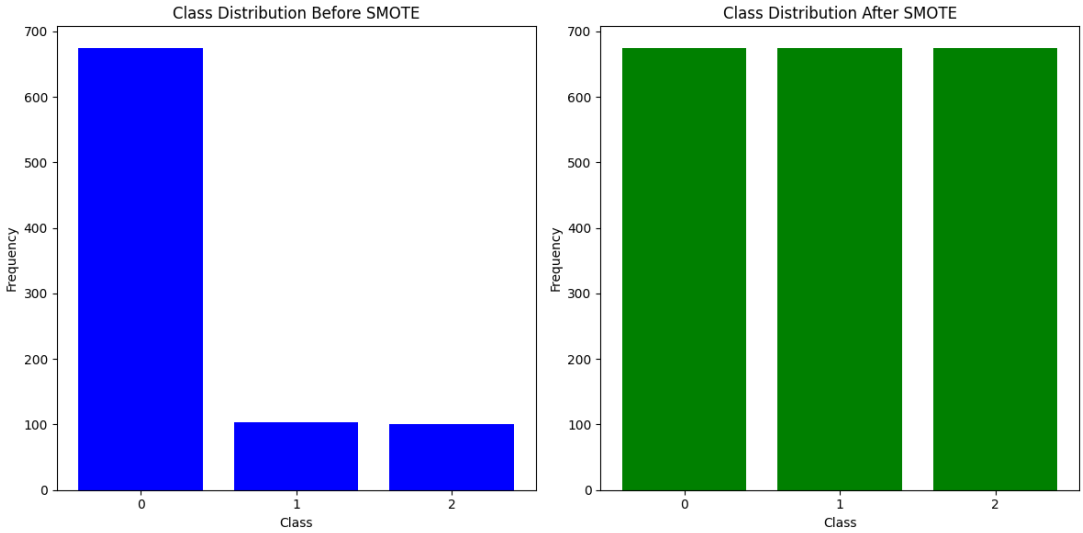

Figure 12.Smote before and after for iShares

This chapter describes the steps for data preparation and preprocessing and the methodologies that can be employed. It begins with the separation of the data, here a temporal split was chosen to preserve the chronological order.

Cleaning methodologies begin with dimensionality reduction, here the constant columns and those with variance less than 0.1 are eliminated as they don't add value to the model. Subsequently, the outliers are determined with the z score, and they are treated with the cap and floor technique.

For the missing data, KNN was used to perform the imputation and Pca was chosen for feature reduction. this technique seek to reduce the characteristics while conserving the variance. PCA suitability was reviewed by performing the Kaiser-Meyer-Olkin (KMO) test, obtaining more than 0.8 for each dataset, guaranteeing its use in the model. An elbow chart was also made reveling the number of characteristics that must be preserved, maintaining the variance of 95%.

Finally, to address the imbalance in iShares's target variable, smote is used to guarantee the learning of all classes in the modeling part.

Each step contributes to the construction of a robust model that can be used with different stocks, each technique was chosen considering the type of data, how they worked together, and which guaranteed best results for the model. Below is the table that summarizes the steps:

Steps	Methodologies selected
Split	temporal split
Dimensionality reduction	Remove constant values or variance less than 0.1.
Outlier identification and treatment	z score / Cap and floor
Imputation	KNN
Feature selection and reduction	PCA
Balancing	Smote

Table 19.Methodologies selected

6 Machine learning models

In this chapter, the machine learning models will be described, each one has different characteristics and applications. The purpose of this chapter is to describe each model, its strengths and weaknesses. We will begin by describing each model and end with a comparative table of advantages and disadvantages. The models used were Decision Trees, XGBoost, KNN (k-nearest neighbors), AdaBoost, and Ensemble Learning.

Machine learning models

Different machine learning models have been used to stock market predictions, providing patterns and trends in the data (Kumbure et al. 2022). Choosing which one is most suitable to use will depend on the type of data present, the training objective, and the sample size, among others. The models chosen and evaluated for this research are presented below.

Random forest

This is a model that combines various decisions trees, (Jiajian and Duan 2023) each tree will train a portion of the original data, also takes a random small set of features in our case, will be the technical indicators for each split for every tree. This promotes the diversity among the trees and extremely avoiding dependency on a particular feature. Also, reduce overfitting overall as it reduces bias.

Random forest will take a decision with the majority vote from the forest(Géron 2019), it is an iterative process that work the following way, the data is divided in multiple subsets, and a model is trained with each of this subset data, when all the trees are trained, it is exposed unseen features each tree will try to take a decision for the new data point. Thus, with this data, the model will choose the majority vote as mentioned before.

As random forests, utilize various trees for the model, it offers wisdom of a mixture of different decision trees that make this model an attractive for stock market prediction. As advantages, the model in general has high accuracy and it can easily handle many features. Additionally, it is good for classification and regression tasks.

Among its disadvantages is that it can be computationally expensive, finding the correct parameters requires a lot of validation, and it is not as easy to understand as decision trees, for example.

Decision trees

"This model can perform task for regression and classification" (Géron 2019, 195), is a model than generally does not need to preform transformation to use it, and it is so easy to understand that is considered a white box model (Géron 2019) basically is structure of a tree and each internal nodal represents a characteristic or feature, for this investigation will be the technical indicator such as RSI. And each branch represents a decision according to the values of the characteristics. For this, investigation will take a decision whether to buy, hold or sell.

One of the advantages of this model is that it is scalable, thus can handle large amounts of data efficiently (Kalcheva et al. 2020), which is good for stock markets predictions. It is suitable for data with missing values. As a disadvantage, it is prone to overfit, it also required large amount of data for the prediction (Kalcheva et al. 2020).

Xgboost

eXtreme Gradient Boosting, also known as XGBoost, is a very popular algorithm for its scalability and efficiency, particularly for classification and regression problems (Chen and Guestrin 2016). This model combines decision trees (weak learners) based on gradient boosting to establish a prediction (Bentéjac et al. 2021). An iterative process is generated in which the errors of the previous model are detected to improve the prediction iteratively.

Xgboost is designed to handle large amounts of data and is designed for large applications. includes L1 (Lasso Regression) and L2 (Ridge Regression) to avoid overfitting (Khan et al. 2024). It also handles missing values well (Bentéjac et al. 2021). XGBoost can be used to identify the most suitable features for the data type and the relationships between them (Khan et al. 2024).

On the other hand, This algorithm requires that its parameters be optimized for optimal performance (Khan et al. 2024), additionally it struggle with label-imbalanced classification tasks, which means that when the model encounters unbalanced data, it tends to have a bias towards the majority class (Wang et al. 2020).

KNN

This supervised algorithm is used for both classification and regression. Like the Decision Trees, it is a white box model. It consists of calculating the closeness or similarity of the data to establish a prediction, the process begins by establishing the distance that can be Euclidean or Manhattan and the k nearest neighbors (Hidayati and Hermawan 2021). For each point, KNN identifies the neighbors according to the established distance, the neighbors are the closer points to initial point. Finally, a prediction is made based on the majority vote (Guo et al. 2012).

One of the advantages of this method is its simplicity, since it is easy to understand both its operation and its results. It also does not assume distribution, which makes it very attractive for different data distributions (Guo et al. 2012).

As disadvantages, this method can be computationally expensive for large amounts of data, it also requires a lot of memory since it requires storing the data for prediction, finally it can be said that the accuracy of the model depends on the choice of the K neighbors (Guo et al. 2012).

ADABOOST

It is an algorithm designed for better accuracy that focuses on difficult cases to correct its predecessor and make a more accurate prediction. This uses a base classifier such as SVM or decision tree to make the predictions in the first iteration, subsequently iterating to rectify the errors of the previous iteration (Géron 2019).

Among its various characteristics, this algorithm is flexible to use any autonomous learning algorithm and is not restricted to a single option. It also focuses on error correction and improving the accuracy of the base classifiers (Géron 2019).

As a disadvantage, it is sensitive to noise in the data and outliers additionally it requires large amount of data for an accurate forecast (Kalcheva et al. 2020).

Ensemble Learning

Ensemble learning is a technique that is based on the principle that better results will be obtained with several models than with just one, it is possible to combine different models that each of them will have an area of application and different results. In general, this type of model usually has better results than just the original model (Géron 2019).

The assembly covers numerous techniques, here we will discuss those more frequently used in practice:

Bagging: smaller samples of the training data are created and trained with the same algorithm. The prediction is often made with the majority of the models votes. In this technique, each model is trained with a different initial sample, which adds different perspectives of the data (Géron 2019).

Boosting: this sequential technique consists of training each model one by one, assigning greater weights to the errors made. Sequentially, they will focus on these errors of the predecessor. improving performance in each iteration (Géron 2019).

Stacking: this technique combines the predictions of several models as input to bring together a better prediction as a group (Géron 2019).

The advantage of the assembly is that since it is a combination of several models, they can be less sensitive to data noise and become robust. They are also more precise models, since by combining multiple models they are less sensitive to overfitting (Hilpisch 2021).

As a disadvantage, it can be computationally expensive by having different models in its implementation (Mohammed and Kora 2023), this also means that it is not a suitable option for large data sets.

SUMMARY

In summary, we will work with the models mentioned above, and the following table contains the advantages and disadvantages of each of them:

Model	Advantages	Disadvantages
Random Forest	• high accuracy. • handle many features. • good for classification and regression tasks.	• computationally expensive. • is not simple to understand.
Decision Trees	• fast and scalable • easy to understand	• prone to overfitting. • less effective for small samples.
XGBoost	• can handle large amounts of data. • includes L1 (Lasso Regression) and L2 (Ridge Regression) to avoid overfitting.	• it struggles with label-imbalanced classification. • requires grid search for optimal performance.
KNN	• easy to understand. • does not assume distribution. • naturally handles multi-class cases.	• computationally expensive. • requires a lot of memory. • Accuracy depends on K neighbors

AdaBoost	• flexible to any machine learning algorithm. • focuses on error correction and improving the accuracy	• noise sensitive • less effective for small samples
Ensemble Learning	• less sensitive to data noise • robustness • less sensitive to overfitting	• computationally expensive. • not a good option for large data sets.

Table 20. Models' advantages and disadvantages

Each model offers advantages and disadvantages, understanding the concepts helps in the execution of the models for the type of data available and the objective of the project. For this thesis, the best model will be chosen as the one that has the best evaluation metrics that we will see in the next chapter. The training will be carried out with the training data, and it will be evaluated with the test data.

To run the models, the python libraries required for each model will be imported. It is assumed that the model will be feed with the opening price. It is also assumed that the opening value does not change much from the opening of the market until the decision is made. Here are the parameters select for the models:

Model	**Parameters**
XGBClassifier	objective='multi:softprob', n_estimators=100, learning_rate=0.1, max_depth=3, seed=42
RandomForestClassifier	n_estimators=100, random_state=42
KNeighborsClassifier	n_neighbors=5
DecisionTreeClassifier	max_depth=3
AdaBoostClassifier	base_estimator=DecisionTreeClassifier(max_depth=1), n_estimators=100, random_state=42

Ensemble Learning	estimators=[('rf', RandomForestClassifier(n_estimators=100, random_state=42)), ('xgb', XGBClassifier()), ('knn', KNeighborsClassifier(n_neighbors=5)), ('ada', AdaBoostClassifier(n_estimators=100))], voting='soft'

Table 21. Models Parameters

7 Empirical Analysis

This chapter will present a complete evaluation of the models for each of the stocks. The main objective is to evaluate the effectiveness of the models from the point of view of data science and a more practical application. For them, the available evaluation techniques are mentioned and those used to select the winning model are chosen.

Once the winning model is selected, backtesting is executed and the results are reviewed. To guarantee the effectiveness of the model, it is compared with a base strategy. Finally, the results are also analyzed.

7.1 Evaluation Metrics

Evaluation in machine learning is used to evaluate the performance of the model, this consists of comparing the model predictions with the real data. The evaluation not only measures the accuracy of the model, but also whether the model works with new data.

The evaluation allows us to select the best model and aims to validate that the model can predict beyond the trained scenarios. Evaluation metrics measure the performances of the models for classification or regression tasks, for this research, classification metrics will be described:

Accuracy: for classification metrics it is usual to measure the accuracy, it will be the number of correct predictions divided by the total number of predictions (Silhavy and Silhavy 2023).

$$\mathrm{A} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Precision: it is the prediction correctly classified (TP) divided to the all positive ones(true positive TP, and false positive FP) (Silhavy and Silhavy 2023).

$$\mathrm{P} = \frac{TP}{TP + FP}$$

Recall: also call sensitivity, are the prediction correctly classified (TP) divided to all the observation (True Positives TP + False Negatives FN) (Silhavy and Silhavy 2023).

$$\mathrm{S} = \frac{TP}{TP + FN}$$

F1 Score: this metric is the relationship between precision and recall (Silhavy and Silhavy 2023).

$$\mathrm{F\ 1\ Score}\ = \frac{(P * R)}{(P + R)}$$

ROC-AUC: it is a tool for visualization of the performance of each class. “Two variables are plotted along the ROC curve: TPR and FPR are plotted at different thresholds of classification. When the classification threshold is increased, more objects are classified as positive, which results in higher true positives and false positives” (Silhavy and Silhavy 2023, 21).

$$\text{True positive rate (TPR)} = \frac{TP}{TP + FN}$$

$$\text{False positive rate (FPR)} = \frac{FP}{FP + TN}$$

Confusion Matrix: this table is a visual representation where it is possible to validate model performance. the test set that was reserved will be checked with the model predictions. the diagonals are the values correctly predicted among the classes. As described in (Silhavy and Silhavy 2023).

		Actual values	
		Positive (1)	Negative(1)
Predicted values	Positive(1)	True positive (TP)	Fasle positive (FP)
	Nega-tive(1)	False negative (FM)	true negative (TN)

Table 22. Confusion matrix

Hypothesis Testing: after the data science metrics are calculated, it is recommended to observe if the results obtained were significant and if it is possible to take the results as truth. this can be done thorough statistical analysis.

Backtesting

This technique consists of simulating the chosen strategy, it uses historical data to recreate the scenario of what the situation would have been like if the strategy had been executed (Ni and Zhang 2005). As a result, the Profitability and the cash value can be determined during and after the evaluated period.

In the simulation, rules were also established for buying and selling. It was defined to buy as much as possible with the available cash, as well to sell all shares already bought.

7.2 Models Accuracy

Various metrics were used to choose the winning model, both traditional data science metrics and financial simulations to review its adaptation to the real world. For this research, two metrics were used especially for the evaluation of the winning model: the ROC curve and confusion matrix.

After choosing the winning model, we proceed with Backtesting for this model. After being executed, the winning model was compared with a base model. This will be only the RSI indicator for the decision to buy, sell or hold.

iShares Winning Model- 67% Accuracy

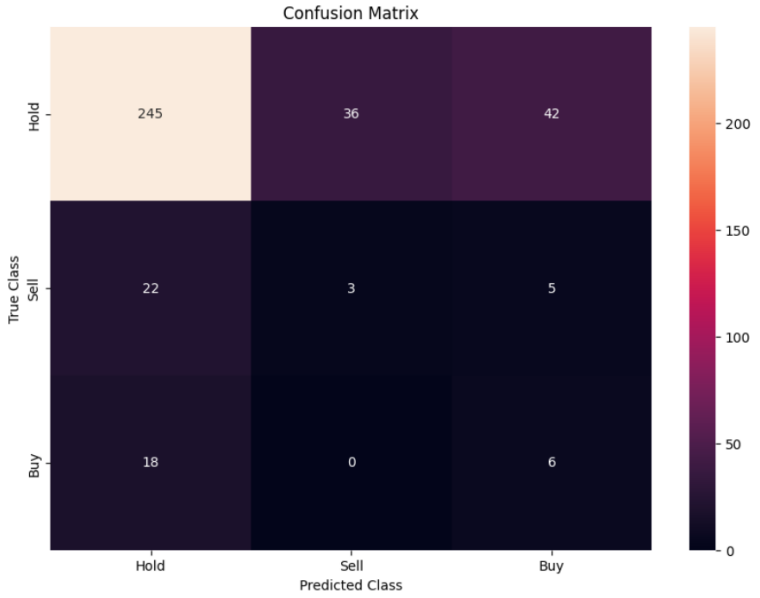

Figure 13.iShares confusion Matrix

Random forest was the winning model for iShares, This model is more precise for the Hold class than for the other classes with 245 true positives, 78 false negatives, these represent the times that the model did not correctly identify the hold category and classified it as another category.

The model has great difficulties in the sell and buy classes with values of 3 and 6 respectively. The values of false negatives and false positives are even higher than the true positives, which indicates that the model tends to predict the hold class when there are cases of buy and sell. The model has a conservative bias that could help in volatile markets but could cause losses in more stable ones.

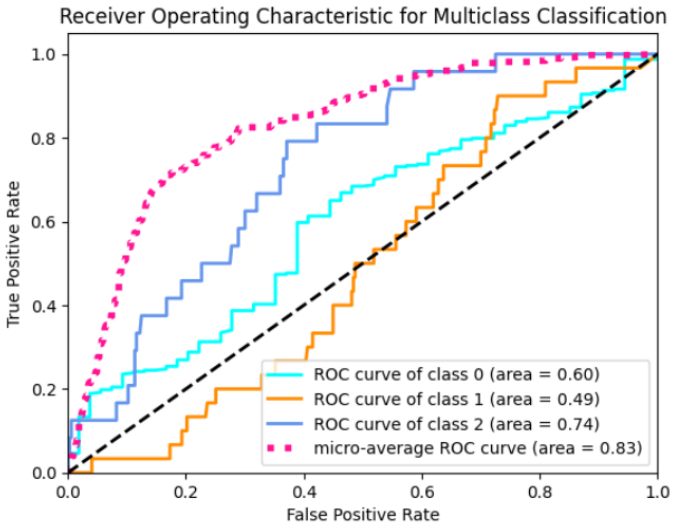

Figure 14. iShares Roc curve

The Roc curve as we saw previously helps to visualize the true positive rate and the false positive rate at different thresholds, for class 0 that represents Hold it shows 0.6 indicating that the model is a little better than a random decision. For class 1, which represents sell, it presents difficulties in differentiating the decision to sell. It is also observed that the model is equal to a random decision. Finally, for class 2, which is to buy, a value of 0.74 is observed, which suggests that the model can understand the decision to buy moderately well.

The Micro average for this model shows a value of 0.83 which indicates that the model in general can distinguish well between the classes; however, this is not uniform for all classes and is since some classes with better predictions dominate the overall score.

Meta Winning Model - 53% Accuracy

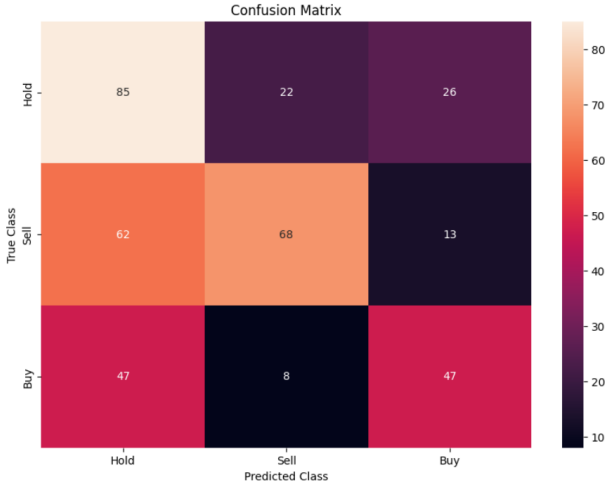

Figure 15.Meta confusion Matrix

The winning model was Decision tree for Meta. This model shows a more balanced picture than iShares. Its execution correctly predicted the hold class 85 times, which indicates that the model predicts well most of the time, however, it has many false predictions.

In the sell class, 68 true positives were obtained, the predictions for this class are better than the other classes; however, the error rate can be problematic in situations where the risk is high. In the buy class, on the other hand, we found 47 correct predictions, but the model is confused the same number of times with the hold class when it was actually buy.

The model in general has tendencies towards the decision to hold and sell, but it easily confuses the type of buy. This might be due to buy is a minority class.

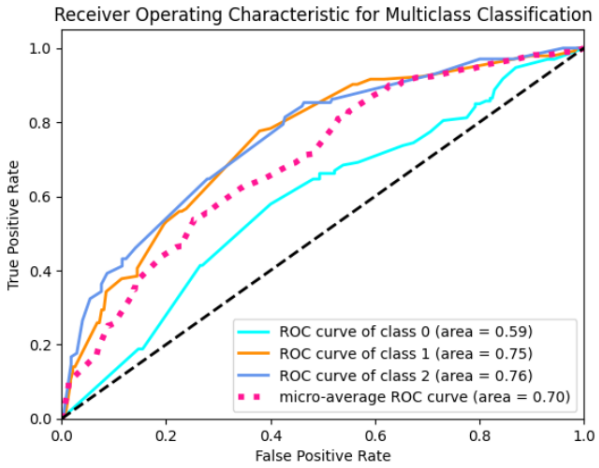

Figure 16. Meta Roc curve

Meta's Roc curve shows a more balanced performance than that observed for iShares. It is observed that no class is below the random classification. For class 0 hold the area under the curve is 0.59 which is a little better than a random classification.

For class sell 1, 0.75 was obtained, which indicates that the model is good at identifying class 1 sell with more efficiency. Likewise for class 2, the model can distinguish the class well, obtaining a 0.76 area under the curve. Finally, for the micro average Roc a 0.7 is observed, which is good but with potential for improvement.

WTI Winning Model- 60% Accuracy

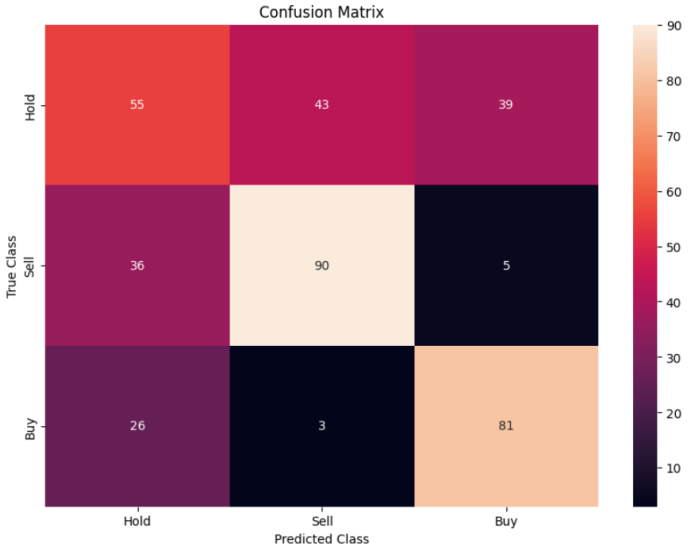

Figure 17. WTI confusion Matrix

Random forest was the winning model for WTI, this model is more precise for the buy and sell classes than for the hold class, with 90 true positive for sale and 81 for buy. However, both classes have slightly high false positive rate.

The model has difficulties in the hold class classification. The values of false negatives and false positives are even higher than the true positives, which indicates that the model tends to predict the buying and selling class when there are cases of hold. In general, the model is good with buy and sell; however, it doesn't predict accurately for all the classes.

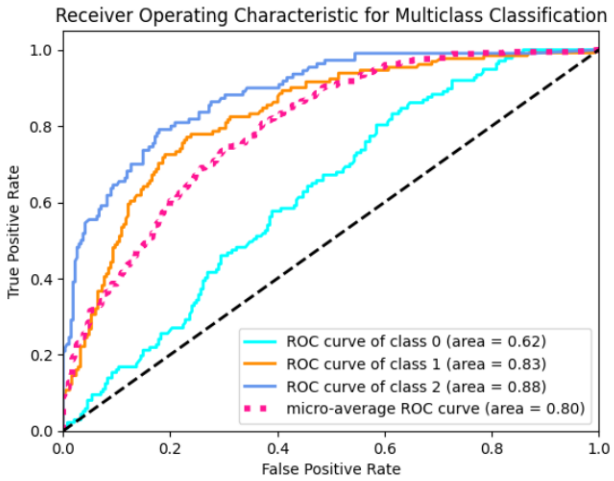

Figure 18. WTI Roc curve

The Roc curve for class 0 that represents Hold shows 0.62 indicating that the model is better than a random decision. For class 1, which represents sell, it shows strong performance for this class prediction. Finally, for class 2, which is to buy, a value of 0.88 is observed, which suggests that the model can understand the decision to sell really well.

The average Micro weights for each of the classes in general and for this model shows a value of 0.80 which indicates that the model in general can distinguish well between the classes.

7.3 Backtesting results

After the analysis from the perspective of data science, the simulation or backtesting was performed. The profitability results for the 3 stocks are presented below.

The profitability of iShares moves between 1% and 6% throughout the period, reaching maximum positive and negative points of 8% and -2% respectively. Periods of stability are observed without very abrupt changes, which can result in constant but not high profitability.

Meta's graph indicates that it moves between -10% and 20%, which reflects a more aggressive model compared to iShares, more pronounced peaks and more volatility are observed. This model seems to better capture market conditions, obtaining greater profitability; however, it is possible to have sudden falls with a higher risk.

The WTI graph shows an initial drop that does not improve over time, reaching values of -40%. It is observed that the model tries to recover in small proportions, but it is not enough to recover from the initial drop.

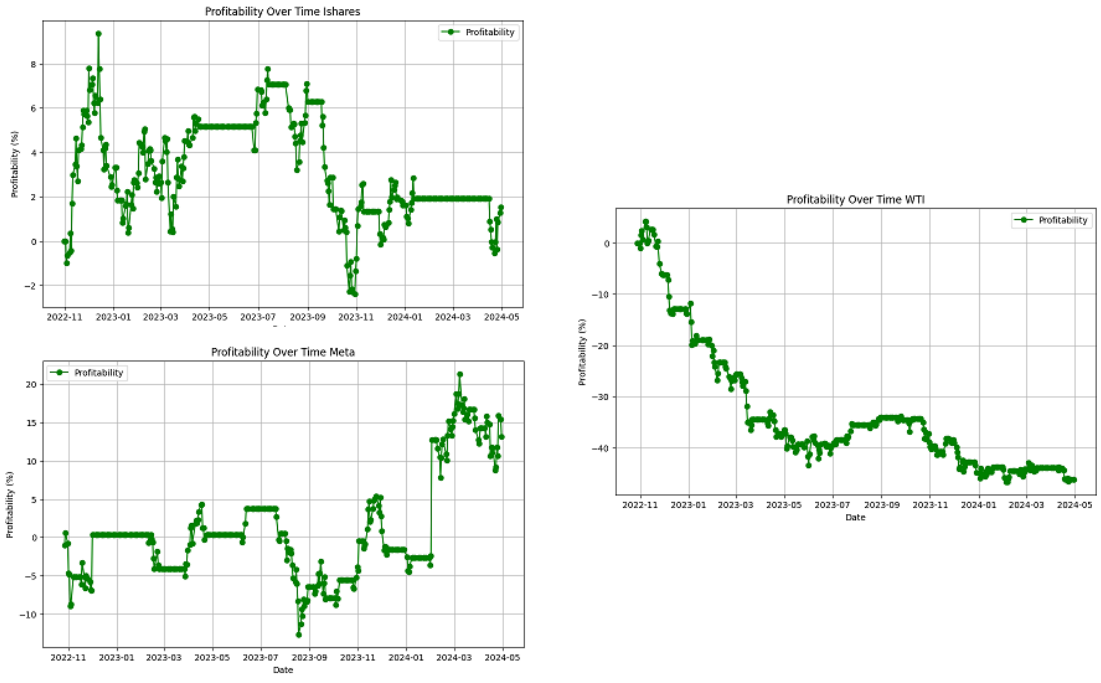

Figure 19. Model backtesting

Backtesting simulates every day and decides with several variables in the model. The following table shows 4 dates of the simulation for Meta as an example.

Date	Open	Target	Predicted Signal	Signal	Portfolio Value	Action	Profitability %	Holdings	Cash
4/22/2024	489.72	2	2	buy	108.80	hold	8.80	0.22	0.00
4/23/2024	491.25	0	2	buy	109.14	hold	9.14	0.22	0.00
4/24/2024	508.06	1	1	sell	111.75	sell	11.75	0.00	111.75
4/25/2024	421.4	2	2	buy	110.63	buy	10.63	0.26	0.00

Table 23. Backtesting example for Meta

Target: the real market signal.

Predicted Signal: the market signal predicted by the model.

Signal: the actual name representing that class, i.e. 0=Hold, 1=sell and 2=buy.

Portfolio Value: the total value of the portfolio that day (cash + holdings * price).

Action: the action taken that day according to the available balance or available stocks.

Profitability: the change in the portfolio due to the action taken.

Holdings: the number of shares held that day.

Cash: the money available that day.

The simulation starts with 100 monetary units and 0 shares. Throughout the simulation, cash is purchased, and shares are sold, according to the model's prediction. In the table above, dates were taken from the advanced stages of the simulation to explain the different situations presented.

On April 23, the market opens at a price of $491.25, with 0.22 shares held and 0 cash. The model predicted to buy that day. However, there is no cash available, so the decision in this case is to hold. The value of the portfolio grows slightly to $109.14, 0.22 of the purchased stock is held and profitability is 9.14%.

The next day, the market has an opening price of $508.06 and, as can be seen in the table, the model has a sell signal that matches the target variable. All shares are sold since there are shares to sell. The portfolio increases by $111.75, the cash goes from 0 to $111.75 and profitability of 11.75%.

Finally, on the 25th, the price dropped to $421.40 and the model predicts to buy again, also coinciding with the target variable. Therefore, the available cash is used to buy all possible stocks. The portfolio closes with $110.63, 0.26 shares held and profitability of 10.63%.

Base strategy

To compare the model and review its relevance in the market, a base strategy was chosen that consisted of using only RSI signals for the stock decision. In this strategy, no machine learning algorithm was used.

For the same period evaluated with iShares in the previous graph, it was found that with the base strategy the stock remains around 0% most of the time before a strong increase that at the end of the evaluated period was 4%. It is observed that the indicator opts for a more conservative strategy and apparently does not capture the small changes in the market; however, it did capture a significant movement that allowed it to extract profitability from negative values.

The profitability for Meta is observed to grow gradually and fluctuates from -10% to 40% at the end of the period. The base strategy in this case seems to capture the overbought and oversold market conditions well and adapts well to the high stock fluctuation.

Finally, WTI Oil shows a constant growth pattern with many fluctuations but with a good percentage of yield at the end of the period.

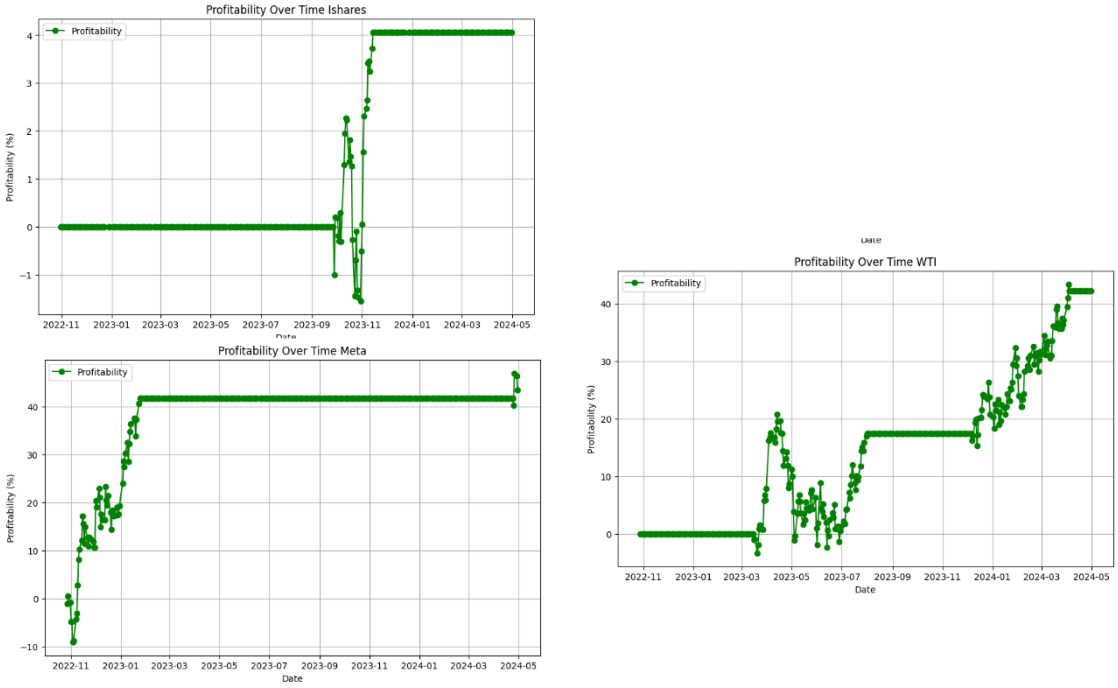

Figure 20. base strategy backtesting

This strategy in general seems to adapt better to different types of assets and to the abrupt changes in the market by capturing overbought and oversold market signals well; however, it is not as good with small changes.

This chapter described the evaluating methodologies for models. It begins by defining the metrics used as well as the calculation. Subsequently, the Backtesting financial strategy was explained, which allowed to simulate a real scenario.

To choose the best model for each stock, the metrics for all models were calculated and those of the winners were detailed in this chapter. Random forest was the winner for iShares and WTI, while decision tree was the winner for Meta. These models were chosen for having better performance in the confusion matrix and Roc curve over the others.

Winners model performance in general are good, but there is room for improvement. Other evaluation metrics such as backtesting are applied to these winning models and compared with the base model. Here it is observed that the base model with this backtesting strategy has a better profitability than the models evaluated in this thesis.

8 Conclusion and Recommendations

The purpose of this research was to create a machine learning model to predict the sell, buy and hold strategy for different sectors with the integration of technical indicators. The aim was also to contribute to the literature with a robust model that was replicable to other markets, instead of just improving the accuracy of the model. The combination of technical analysis with machine learning in this research resulted in the following conclusions.

Model Evaluation

PCA is effectively used for data dimensionality reduction for iShares with 4 components, while Meta and WTI Oil each with 5 components. The Kaiser-Meyer-Olkin (KMO) test was performed, showing that the data was suitable for PCA.

For iShares, the winning model was random forest among the 6 models executed, it is evident that the model can clearly differentiate the Hold class but is in problems with the sell and buy class. Similar results are seen in the average Micro ROC curve, in which the model in general is not bad but that the rating is influenced by the heavier classes.

The winning model for Meta was decision trees, which shows better performance than iShares. Higher rates are observed in the confusion matrix for true positives, however also numerous false predictions. In the ROC curve it is possible to see that in all classes they are better than a random classification.

The evaluation of the predictive model for WTI showed that the model is more accurate for the buy and sell classes; however, it has a slightly high false positive rate in these two classes. The Roc curve confirms a positive performance for buy and sell indicators at the different thresholds. In general, the model yields positive results, although it has difficulties in classifying the hold indicator.

The difference in performance (Roc and confusion matrix) for Meta and WTI models regarding iShares may be related to class balancing. Meta and WTI had their classes balanced, and no additional process was necessary, unlike iShares where the balancing was carried out. iShares, despite the synthetic creation of samples, could not correctly learn to differentiate the 3 classes.

Backtesting

The profitability analysis for the machine learning models showed a positive trend for iShares, with variations throughout the period. Meta offers the highest profitability margin with up to 20%

but with high risk. On the contrary, WTI showed constant falls throughout the period. This evaluation shows that in general the model can capture well the trends of certain sectors such as technology and the diversified market.

The results of the backtesting for the base strategy, which uses RSI to produce buy and sale decisions, position it as the strategy with the highest performance for each of the stocks studied, demonstrating its precision in the different sectors studied here.

iShares has a conservative growing tendency in backtesting due to its small and constant growth in the market, which allows it to have profits that are very stable but not high in comparison with the other stocks.

Roc curve and confusion matrix show consistent models, however, when the models are compared with the base strategy (Rsi indicator), the winning strategy in all 3 cases was the base strategy. From this can be interpreted that ML models needs more robust and balance technical indicators, as it can yield not optimal results due to the complexity of the input data.

Better performance of machine learning models is observed with stock with rapid changes and fluctuations, and not so good with stable models or with a constant upward trend.

Although the models do not have accuracies as high as shown in literature studies, their profitability range is higher than those shown in the literature.

Limitations

The model was trained with historical data, it does not have market information such as news, tweets and others that would allow to understand other factors that affect the market.

Three stocks were chosen to represent different sectors. For technology and diversified market, the model works better. however, it is possible that the results are not universal.

The study relied on traditional machine learning models for training the data, although the combination with the indicators worked well, other techniques such as deep learning can improve the accuracy.

The application of PCA worked well for the developed model and reduce the dimensionalities as intended, however during creation of the new components characteristics from the original features are lost that may be important.

Next steps

To improve the applicability of the models presented here, the following recommendations are proposed as following steps:

Expand the dataset by reviewing other indicators that can also work well together, and explore with new sectors.

Explore the data with deep learning to review model learning.

Perform a sentiment analysis to evaluate market sentiment.

Explore other functions or rules to create the target variable.