A MACHINE LEARNING APPROACH FOR THE DETECTION OF TRADE-BASED MANIPULATIONS IN BORSA İSTANBUL

MAKİNE ÖĞRENMESİ YAKLAŞIMIYLA BORSA İSTANBUL’DA İŞLEM BAZLI MANİPÜLASYONLARIN TESPİTİ

NURULLAH CELAL USLU

DR. FUAT AKAL
Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of the Degree of DOCTOR OF PHILOSOPHY

in Computer Engineering

2022

ABSTRACT

A MACHINE LEARNING APPROACH FOR THE DETECTION OF TRADE-BASED MANIPULATIONS IN BORSA İSTANBUL

Nurullah Celal USLU

Doctor of Philosophy, Computer Engineering Department

Supervisor: Dr. Fuat AKAL

June 2022, 137 pages

Capital markets, one of the pillars of the financial system, play a vital role in transferring the excess funds of savers to investors who need funds in the medium and long term. Trust is essential for the safe and effective operation of capital markets and for the healthy transfer of resources from those who supply funds to those who request them. The capital market is constantly regulated and supervised in order to ensure it functions and develops in a reliable, transparent, efficient, stable, fair, and competitive environment and to protect the rights and interests of investors.

Manipulation, which prevents the capital market from operating regularly within trust and transparency norms, is an important issue that should be considered both for individual investors who offer funds to the securities markets and for companies that request funds by issuing stocks. This study examines the trade-based manipulations in Borsa Istanbul (BIST). Data on stocks that were manipulated between 2010 and 2015 were used in BIST, and a model consisting of supervised machine learning classification techniques and artificial neural networks was proposed to detect trade-based manipulation from the daily data of manipulated stocks. It has been shown that the proposed model is successful in detecting trade-based manipulations in the stock market based on accuracy, sensitivity, and f1 scores.

Experimental results show that an f1 score of 0.86, a sensitivity of 0.87, and an accuracy of 0.89 in market manipulation detection were achieved.

With this study, the manipulation in the stock market, the biggest obstacle for investors to make safe investments in the capital markets, will be minimized and the principles of transparency and trust, essential for the formation and development of the capital market, will be established. In addition, due to the success achieved in market manipulation detection, our study will benefit regulators especially in detecting stock market manipulation.

Keywords: Trade-based manipulation, machine learning, artificial neural networks, securities market, stock market

ÖZET

MAKİNE ÖĞRENMESİ YAKLAŞIMIYLA BORSA İSTANBUL'DA İŞLEM BAZLI MANİPÜLASYONLARIN TESPİTİ

Nurullah Celal USLU

Doktora, Bilgisayar Mühendisliği Bölümü

Danışman: Dr. Fuat AKAL

Haziran 2022, 137 sayfa

Finansal sistemin en temel unsurlarından olan sermaye piyasaları, tasarruf sahiplerinin ellerindeki fazla fonların orta ve uzun vadede fon ihtiyacı olan yatırımcılara aktarılması noktasında çok önemli bir görev yapmaktadır. Sermaye piyasalarının güven içinde ve etkin çalışabilmesinin ve kaynakların fon arz edenlerden fon talep edenlere sağlıklı bir şekilde aktarılabilmesinin en önemli temellerinden birini yatırımcıların piyasaya güveninin tesis edilmiş olması oluşturmaktadır. Sermaye piyasasının güvenilir, şeffaf, etkin, istikrarlı, adil ve rekabetçi bir ortamda işleyişinin ve gelişmesinin sağlanması, yatırımcıların hak ve menfaatlerinin korunması için sermaye piyasası devamlı düzenlenmekte ve denetlenmektedir. Bu düzenleme ve denetleme faaliyetlerinin temel amacı, piyasalarda şeffaflığı sağlamak ve yatırımcıları korumak olup, piyasalarda zaman zaman gözlemlenen kötü niyetli eylemleri önlemeye yöneliktir.

Sermaye piyasasının güven ve şeffaflık normları kapsamında düzenli bir şekilde faaliyet göstermesine engel olan faaliyetlerden biri olan manipülasyon, hem menkul kıymet piyasalarına fon sunan bireysel yatırımcılar, hem de hisse senedi ihraç ederek fon talep eden şirketler açısından dikkate alınması gereken önemli bir konudur. Bu çalışma, Borsa İstanbul'da (BİST) gerçekleştirilen işlem bazlı manipülasyonları incelemektedir. BIST'da 2010 - 2015 yılları arasında manipülasyona uğramış hisse senetlerine ilişkin veriler kullanılmış olup, manipüle edilen hisse senetlerinin günlük verilerinden işlem-bazlı manipülasyonu tespit etmek için denetimli makine öğrenmesi sınıflandırma teknikleri ile yapay sinir ağlarından oluşan bir model önerilmişitr. Çalışmanın sonucunda, önerilen modelin doğruluk, duyarlılık ve F1 skor ölçüm yöntemlerine dayalı olarak hisse senetleri piyasasında işleme dayalı manipülasyonları tespit etmede başarılı olduğu gösterilmiştir. Yapılan deneyler sonucunda manipülasyon tespitinde 0,86 F1 skoruna, 0,87 duyarlılığa ve 0,89 doğruluğa ulaşılmıştır.

Bu çalışma ile, yatırımcıların sermaye piyasalarında güvenli yatırım yapmalarının önündeki en büyük engel olan borsadaki manipülasyon en aza indirilerek, sermaye piyasasının oluşum ve gelişiminin temeli olan şeffaflık ve güven ilkeleri oluşturulacaktır. Ayrıca, piyasa manipülasyon tespitinin yüksek performansı nedeniyle çalışmamız özellikle borsadaki manipülasyonu tespit etmede düzenleyicilere fayda sağlayacaktır.

Anahtar kelimeler: İşlem bazlı manipülasyon, makine öğrenmesi, yapay sinir ağları, menkul kıymetler piyasası, sermaye piyasası

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Fuat AKAL, for his support, guidance, patience, motivation, and valuable comments throughout the all steps of this study. I am also thankful to my thesis monitoring committee, Prof. Dr. Erdoğan DOĞDU and Dr. Ahmet Ercan TOPCU, and thesis examining committee, Prof. Dr. Hayri SEVER and Assoc. Prof. Dr. Murat AYDOS for their valuable time and comments.

I want to express my sincere gratitude to my dear friend Mahmut OKUR for his valuable friendship and ongoing support throughout my entire doctoral journey.

Most importantly, none of this would have been possible without my family's love, patience, and endless support. This thesis is dedicated to my family, who have always been a source of love, support, and inspiration. My parents, Fadime USLU and Nuri USLU, deserve nothing less than my sincere appreciation.

I would like to express my heart-felt gratitude my sweet daughter Fatıma Zehra and my lovely cute son Nuri Selim. They have made my life much more joyful and peaceful since the day they opened their eyes to life and give me strength whenever I need it.

Last but not least, from the bottom of my heart I want to thank Handan USLU for such a wonderful wife. Without her support, love, and patience, I would not have been able to succeed. It would not be able to complete this study without her help and understanding.

CONTENTS

ABSTRACT i
ÖZET iii
ACKNOWLEDGEMENTS v
CONTENTS vi
FIGURES ix
TABLES x
ABBREVATIONS xi
1 INTRODUCTION 1
1.1 Motivation and Background 3
1.2 Contribution of the Proposed Research 5
1.3 Organization of the Proposed Research 6
2 BACKGROUND 8
2.1 Borsa Istanbul Equity Market 8
2.1.1 Markets 8
2.1.2 Electronic Trading System 9
2.1.3 Equity Market Trading Systems 9
2.1.3.1 Session 10
2.1.3.2 Trading Amount 10
2.1.3.3 Trading Volume 10
2.1.3.4 Trading Unit 11
2.1.3.5 Weighted Average Price 11
2.1.4 Price Determination Mechanisms in Equity Markets 11
2.1.4.1 Priority Rules 11
2.1.5 Order Types 13
2.1.6 Trading Methods 15
2.1.6.1 Continuous Auction – Multiple Price 15
2.1.6.2 Continuous Trading with Market Maker 18
2.1.6.3 Single Price 19
2.1.7 Circuit Breaker System in the Equity Market 20
2.2 Manipulations in the Stock Market 20
2.2.1 Manipulation Concept 22
2.2.2 Purpose of Manipulation 25
2.2.3 Various Concepts Confused with Manipulation 26
2.2.4 Types of Manipulation 27
2.2.4.1 Action-based Manipulation 28
2.2.4.2 Information-based Manipulation 29
2.2.4.3 Trade-based Manipulation 30
2.2.5 Trade-based Manipulation Types 32
2.2.5.1 Unreal Trades 33
2.2.5.1.1 Matched Orders 33
2.2.5.1.2 Washed Sales 34
2.2.5.2 Real Trades 35
2.2.5.2.1 High Frequency 35
2.2.5.2.2 Runs 35
2.2.5.2.3 Frontrunning 36
2.2.5.2.4 Short Selling 36
2.2.5.2.5 Corners 37
2.2.5.2.6 Capping and Pegging 37

2.2.5.2.7 Marking the Open and Marking the Close 38
2.2.5.2.8 Parking and Warehousing 39
2.2.5.2.9 Spoofing 39
2.2.5.2.10 Pumping and Dumping 40
2.2.5.3 Patterns of Manipulative Transactions 40
2.3 Machine Learning and Classification 42
2.3.1 Machine Learning 43
2.3.2 Machine Learning Methods 45
2.3.2.1 Supervised Learning 45
2.3.2.2 Unsupervised Learning 46
2.3.2.3 Semi-supervised Learning 46
2.3.3 Classification 46
2.3.3.1 Classification with Machine Learning 47
2.3.3.1.1 Naïve Bayes 47
2.3.3.1.2 Logistic Regression 48
2.3.3.1.3 Support Vector Machine 50
2.3.3.1.4 K-Nearest Neighborhood 52
2.3.3.1.5 Decision Tree Classifier 54
2.3.3.1.6 Random Forest 55
2.3.3.1.7 Artificial Neural Network 56
2.3.3.2 Classification Performance Measurement Criteria 62
2.3.3.2.1 Editing Dataset 62
2.3.3.2.2 Evaluation of Model Performance 62
2.3.3.2.3 Model Performance Evaluation Criteria 62
2.4 Data Science Methodology 65
2.4.1 Business Understanding 66
2.4.2 Data Understanding 66
2.4.3 Data Preparation 67
2.4.3.1 Data Selection 68
2.4.3.2 Data Cleaning 68
2.4.3.3 Data Integration 69
2.4.3.4 Data Reduction 69
2.4.3.5 Data Transformation 70
2.4.3.5.1 Min – Max Scaling 71
2.4.3.5.2 Standard Scaling 71
2.4.4 Modeling 71
2.4.5 Evaluation 72
2.4.6 Deployment 73
3 RELATED WORK 74
4 THE PROPOSED STOCK MARKET MANIPULATION DETECTION FRAMEWORK 83
4.1 Data Set 83
4.2 The Proposed Model 87
4.2.1 Algorithm 91
4.2.2 Preprocessing 92
4.2.2.1 Business Understanding 92
4.2.2.2 Data Understanding 92
4.2.2.3 Data Preparation 93
4.2.2.3.1 Data Selection 93
4.2.2.3.2 Data Cleaning 93

4.2.2.3.3 Data Integration ... 93
4.2.2.3.4 Data Transformation ... 93
4.2.3 Scaling ... 97
4.2.4 Training ... 97
4.2.5 Evaluation ... 99
4.2.6 Feature Engineering ... 99
4.2.7 Prediction ... 101
4.2.8 Program Output ... 103
5 EXPERIMENTAL ANALYSIS ... 106
6 CONCLUSIONS ... 118
REFERENCES ... 120
APPENDICES ... 132
Appendix 1 – Confusion matrix for proposed model ... 132
CURRICULUM VITAE ... 136

FIGURES

Figure 2.1 Machine learning working schema 44
Figure 2.2 Logistics function curve (Sigmoid) [136] 50
Figure 2.3 Hyperplane, margin, and support vectors in a two-class dataset 51
Figure 2.4 Selection of the three closest neighbors to the sample to be classified using the KNN algorithm 53
Figure 2.5 Decision tree example 55
Figure 2.6 Random forest method components 56
Figure 2.7 General structure of the neuron [158] 57
Figure 2.8 Artificial neuron 58
Figure 2.9 Multilayer feed-forward neural network 61
Figure 2.10 Phases of CRISP – DM diagram [163]. 65
Figure 4.1 Architectural diagram of the proposed model 90
Figure 4.2 Details of building models 98
Figure 4.3 Effects of attributes 100
Figure 4.4 Output sample 104
Figure 5.1 Trade-based manipulation detection performance of machine learning models 116

TABLES

Table 2.1 Sequence of orders sent to the trading System 12
Table 2.2 Continuous auction method order book recording 16
Table 2.3 Order book recording after first trading 17
Table 2.4 Order book recording after second trading... 18
Table 2.5 Single price order book and match price table 20
Table 2.6 Information-based manipulation definitions 30
Table 2.7 Real and unreal tradings 32
Table 2.8 Trade-based manipulative Patterns 41
Table 2.9 Equivalents of biological nervous system elements in artificial neural network 58
Table 2.10 Confusion Matrix... 63
Table 2.11 Steps followed within the CRISP-DM stages 73
Table 4.1 Information on transactions performed in 2010 - 2015....................................... 84
Table 4.2 Manipulative transactions performed in 2010 – 2015... 85
Table 4.3 Manipulated stocks.. 85
Table 4.4 Selected manipulative stocks.. 86
Table 4.5 Correspondence of CRISP-DM steps in our model 88
Table 4.6 Attributes and descriptions of the dataset generated on a daily basis 95
Table 4.7 F1 score list based on stocks 102
Table 5.1 Test results for the model created using the DTC algorithm 107
Table 5.2 Test results for the model created using the LR algorithm 108
Table 5.3 Test results for the model created using the KNN algorithm............................ 109
Table 5.4 Test results for the model created using the RF algorithm................................ 110
Table 5.5 Test results for the model created using the NB algorithm 111
Table 5.6 Test results for the model created using the SVM algorithm............................ 112
Table 5.7 Test results for the model created using the ANN algorithm............................ 113
Table 5.8 Test results for the purposed model 114
Table 5.9 Comparison of the performance of the models 116

ABBREVATIONS

Abbrevations	
ANN	Artificial Neural Networks
AUC	Area Under the Curve
BIST	Borsa İstanbul
BISTBD	Borsa İstanbul Board of Directors
BISTECH	Borsa İstanbul Technology
CESR	The Committee of European Securities Regulators
CHMM	coupled hidden Markov model
CMB	Capital Markets Board of Turkey
CML2499	Capital Market Law No. 2499
CML6362	Capital Market Law No. 6362
CRISP – DM	The cross-industry standard process model for data mining (CRISP - DM)
CSRC	China Securities Regulatory Commission
DTC	Decision Tree Classifier
FaK	Fill and Kill Orders
KNN	K-Nearest Neighbors
L3	Level 3
LR	Logistic Regression
MLP	Multilayer perceptron
NASDAQ	National Association of Securities Dealers Automated Quotations
NB	Naïve Bayes
RF	Random Forest
SEAQ	London Stock Exchange Automated Quotation
SVM	Support Vector Machine
WAP	Weighted Average Price

1 INTRODUCTION

Capital and money markets are basic elements of the financial system and play a vital role as indicators of countries' economic development. Capital and money markets are also named financial markets, and it is necessary to explain the market concept to understand the concept of financial markets. The market involves buyers and sellers directly, agents or communication tools, and is where exchange occurs [1].

The financial markets enable businesses to obtain short-, medium-, and long-term funding and savings holders to achieve high gain [2]. Moreover, financial markets are used to transfer funds between savers who spend less than their income (fund-supplying units such as households, firms, and government) and those who want to spend more than their income [3]. Therefore, financial markets play an important role in transferring the surplus funds of savers to investors needing funds [4]. In addition, in meeting this need, financial markets allow both the buyer and the seller to evaluate the information that influences the price in terms of supply and demand [5]. Money markets involve the supply and demand of funds over one year or less, while capital markets involve the supply and demand of funds over more than one year.

Money markets are where participants perform transactions in order to meet an urgent need for funds [6]. In these markets, which generally involve short-term loans, those with surplus savings and those in need of funds come together to transfer funds [7]. Capital market participants, on the other hand, are quite different from those in the money market. Capital markets involve long-term investments aimed at significantly increasing the investor's wealth [6]. Hence, financial assets with a maturity longer than one year are regarded as a part of the capital market [8]. Capital markets are divided into organized and unorganized markets. Organized capital markets refer to stock exchanges that have a specific place and working rules, where prices are determined under the most advanced competitive conditions [9]. Unorganized markets, on the other hand, are securities markets that are not under an institutional roof and do not need a central location with technological opportunities [10].

The main differences between capital markets and money markets can be summarized as follows [11]:

- The maturity of funds transferred or traded in capital markets is long.
- The financial instruments used in fund transfers in capital markets are bought and sold second hand. In other words, stocks and bonds offered to the public in capital markets change hands between individuals.
- Risk and expected return rates are generally higher in capital markets, because the financial instruments used in fund transfer are long term and their value and return vary over time.

Financial markets ensure that the needs of both parties are met, the level of welfare increases, and the economy progresses in this context by transferring funds from those that provide funds to those that request them. In financial markets, an effective system is needed when transferring these funds. As the efficiency of the financial system increases, the possibility of achieving fund supply and fund demand at affordable prices also increases. The manipulative transactions carried out to artificially change the price of capital market instruments, especially stocks, make manipulation an increasingly important phenomenon for investors trading. These malicious transactions, which are prohibited by laws and regulations, disrupt the functioning of financial markets and cause investors who invest in financial markets to turn to alternative markets [12]. Financial markets should be supervised and regulated in order for investors operating in financial markets to transfer resources mutually, for the financial system to fulfill its expected functions, and for the system to work effectively.

Regulations are created by the state in various sectors of the economy, and the most important of these concern the financial markets. The main reason for these regulations is to ensure macroeconomic and microeconomic stability. In particular, the bankruptcy of large banks, intermediary institutions, and insurance companies in financial markets may threaten macroeconomic stability. In order to prevent systemic risks that will cause financial institutions to suffer financial distress and affect the whole economy negatively, regulations are formulated for the financial sector. States do not consider the financial sector as a whole; they deal with institutions individually to ensure micro stability. Thus, the health of each institution ensures the health of the entire sector. Another purpose is to ensure transparency in markets and institutions and to protect investors, with regulations to prevent manipulation taking precedence.

Manipulation based on the buying and selling of securities aims to create an emerging market for the stock and to fix or stabilize the stock's market price. Thus, manipulation provides benefits to the relevant persons and institutions by maintaining the stock's stable position in the market and keeps its value above the market price [13]. The main aim of manipulation is to direct investors to buy and sell a security via external interventions in the functioning of the markets or to keep the price of the security at an artificial level. In other words, it is the deliberate inhibition of the process by which supply and demand freely determine prices in the securities markets. Investors who cannot properly evaluate their funds due to misdirection lose, while those who engage in manipulative transactions by creating an artificial price profit. Manipulation, which reduces market efficiency, harms public confidence by victimizing individual investors [14] and damaging investors' confidence in the market, causing them to turn to alternative markets. In other words, savers want to be sure that prices in the capital market in which they trade are generated under real supply and demand conditions [15]. On the other hand, as it affects the liquidity of the market, it also negatively affects the ability of companies to provide funds for commercial activities [16].

Not only are the financial markets adversely affected by this situation, but also the country's economy. For this reason, the proper functioning of the capital market system as well as its effectiveness and efficiency, and ensuring confidence among investors are very important for economic progress. In order for the capital market to function well and to provide confidence, the values of capital market instruments must be established under the market's own free supply and demand conditions, without misleading external interventions. Thus, the regulatory authorities implement measures to minimize or eliminate this loss of confidence resulting from manipulation. The Capital Market Law, regulating the functioning of capital markets in this country, aims to protect the rights and interests of investors by ensuring the market functions and develops reliably.

1.1 Motivation and Background

Stock buying and selling activities in the securities market now take place through the internet, with millions of transactions each day. Innovations in information technology enable the detection of manipulative market transactions and the emergence of new and inventive approaches. As a result of these developments, the quantity and types of transactions that distort the market are growing. Manipulations were divided by Allen and

Gale into three types: information-based, motion-based, and trade-based. Information-based manipulations provide the market with incorrect information, which can alter the value of capital market instruments and mislead investors, as well as spreading false rumors. Moreover, accurate information that needs to be declared is not shared with the public. Likewise, motion-based manipulation is generally done to influence and differentiate the expected real values in terms of stocks, and to change the market prices of the stocks and thus the company value [17]. Trade-based manipulation involves transactions carried out by buyers or sellers, those who manage accounts in order to produce a false or misleading impression of the prices, price fluctuations, and the supply of and demand for capital market instruments [18].

The main purpose of all manipulative attempts is to interfere with price mechanisms by inhibiting the free interaction of supply and demand, to encourage investors to trade in a security by deceiving them, and to keep the security price at an artificial level [5]. Manipulation, the most significant barrier to investors' safe participation in capital markets, hinders fair stock prices and harms investors by attempting to deceive them. As a result, stock market manipulation is rigorously monitored and manipulators face consequences.

Manipulation is a serious problem in the securities market and its detection is critical. As trade-based manipulation often appears to be a legitimate stock market transaction, it is difficult to detect [5]. Studies on a theoretical basis, such as researching information-based manipulation, examining specific techniques, and comparisons in country and legal dimensions, have received more attention in the literature; experimental studies, particularly on the detection of trade-based manipulation, remain rare. The main reason for this is that the transaction data needed to empirically evaluate the subject are kept confidential and academics have only limited access to these data [19]. In the studies on trade-based manipulations, those in which manipulation is detected are uncommon, and therein the manipulation is generally determined by measuring variables such as the return, amount, and volume of the stock during the manipulation period. The scarcity of studies on the detection of trade-based manipulations, in particular, encouraged us to examine this topic.

1.2 Contribution of the Proposed Research

We used data on stocks that were manipulated between 2010 and 2015 in Borsa Istanbul (BIST), and we employed supervised machine learning classification techniques and artificial neural networks to detect trade-based manipulations from the daily trading data. Our study varies from others in the manipulation literature as follows:

1. The vast majority of the work on detecting manipulation is theoretical, and overall manipulation is assessed by assessing factors such as volatility, return, amount, volume, and inventory turnover as a result of manipulation.
2. Studies on trade-based manipulation in this country mostly focus on closing price manipulation. However, we focused on estimating and detecting whether there was manipulation in the transactions of stocks that were manipulated between 2010 and 2015.
3. Unlike the majority of the research, we used real-time trading data of equities that were manipulated in BIST.
4. Our manipulation explanatory factors differ from those employed in earlier studies.
5. We detected anomalies from the daily data of manipulated stocks by using artificial neural networks with 6 supervised machine learning algorithms and two different scaling methods.

In addition to investors, manipulative activities also have significant effects on the stock market and the entire economy. Therefore, it is of great importance to identify manipulative transactions in order to protect investors who offer funds to the markets, to ensure the reliability of financial markets, and to eliminate such instability due to illegal activities. However, because of the strategies used, constantly changing technical breakthroughs, and legal gaps, it is extremely difficult to detect and prevent manipulative behavior in financial markets [20]. The current study will reduce stock market manipulation, which is the most significant barrier stopping investors from making safe investments in capital markets, and will establish transparency and trust, which are the foundations of capital market formation and development. Furthermore, because of the successful performance of market manipulation detection, our research will be useful to regulators, particularly in detecting stock market manipulation.

1.3 Organization of the Proposed Research

The remainder of the thesis is organized as follows: There are 6 chapters. In the first section of the second chapter the markets where transactions are made in the BIST equity market, the electronic trading platform (BISTECH) where transactions are made in the stock market, the types of stock market trading systems, the stock market price determination mechanisms, order types, trading methods, and circuit breakers are discussed.

The notion of manipulation is discussed in the second section of the second chapter, and information regarding the different types of manipulation, notably trade-based manipulation, is provided. Finally, patterns of manipulation related to trade-based manipulation types were extracted.

A full examination of machine learning and its methodologies, classification and machine learning algorithms used for it, and classification performance measurement criteria is given in the third section of the second chapter.

The fourth section of the second chapter provides thorough information on the stages of CRISP-DM methodology and data science methodology.

The third chapter reviews the literature and includes investigations on the detection of market manipulation on a national and international scale.

The first section of the fourth chapter discusses BIST stock data, which is fed into artificial neural networks and supervised machine learning classification algorithms. Moreover, statistical information on the quantity of the data utilized and the fields of the data are shown.

The second section of the fourth chapter, which is the primary section, describes step by step the proposed model, which is made up of supervised machine learning algorithms and artificial neural networks.

The fifth chapter summarizes the research findings and compares the performance of the trade-based manipulation detection models.

The findings obtained are analyzed in light of the theoretical knowledge explained in the preceding sections in the sixth chapter, the last portion, and the methodologies utilized during the experimental period are discussed.

2 BACKGROUND

The first section of this four-part chapter describes the operational principles and rules of the BIST equity market. The idea of manipulation is discussed in the second part, and a full examination of the types of manipulation in the capital market, specifically the types of trade-based manipulation that is the topic of this thesis, is provided. The third section investigates the machine learning algorithms and approaches utilized in classification in the literature. The chapter concludes with descriptions of data science methodology and a full investigation of CRISP-DM, a data science approach.

2.1 Borsa Istanbul Equity Market

Equity market is the general name for where the equities of companies in different sectors that are quoted or not quoted on BIST and the securities related to them (preemption rights coupon, exchange traded funds, warrants, certificates, etc.) are traded in various markets established by the Borsa Istanbul Board of Directors (BISTBD) [21].

2.1.1 Markets

Trading on the equity market is carried out in the following markets.

- **National Market:** Companies that meet BIST quotation requirements are traded in the National Market, which has been in operation since the establishment of BIST [21].
- **Collective Products Market:** The Collective Products Market began to operate in 2009 and some securities previously traded in the National Market were switched to this market. The following securities are traded here: securities investment trusts, real estate investment trusts and venture capital stocks, exchange traded funds participation certificates, and brokerage firm warrants and certificates [21].
- **Second National Market:** Small and medium-sized companies, which are the dynamo of the economy; companies that are temporarily or permanently removed from the National Market; and companies that cannot meet the quotation and trading conditions applicable to the National Market are traded in the Second National Market. The enterprises that meet the quotation conditions during the monitoring carried out at regular intervals in the National Market can be included, and the

enterprises that fail to meet the quotation conditions can be removed from the National Market [21].

- **Watchlist Companies Market:** Some companies traded on Borsa Istanbul may see extraordinary stock transactions from time to time, resulting in a failure to comply with current regulations and not sufficiently informing the public in a timely manner. The Watchlist Companies Market, established by a vote by the BISTBD on November 26, 1996, is where stocks are monitored, investors are regularly updated on time, and liquidity is offered to investors in the company. The aim of this monitoring and assessment is to protect investors while preventing potential negative risks [21].

2.1.2 Electronic Trading System

BIST, which is an important part of the vision of Istanbul International Financial Center, continues to regularly invest in technology. It renews its hardware, system, and application infrastructure to provide world-class performance and reliable service. Within the framework of the cooperation with National Association of Securities Dealers Automated Quotations (NASDAQ), all markets within Borsa Istanbul were moved to the new trading platform, called Borsa Istanbul Technology (BISTECH), to create a single trading platform. All BIST markets are run on the same platform from end to end.

With the launch of BISTECH, the order processing capacity of the exchange was maximized, with up to 100,000 orders processed per second and the round trip delay reduced to microseconds. Thanks to the globally known order entry and data dissemination protocols accompanying BISTECH, local and international investors accessing the stock market electronically have started to benefit from the high-performance and reliable service offered by this system without incurring development costs [22].

2.1.3 Equity Market Trading Systems

There are basically two types of trading systems: order-based and quotation-based. Those who trade in order-based systems send their orders to the trading system without the prices being determined by market makers or the central mechanism or auction. Investors' orders provide liquidity to the market and are met without any intermediaries. In order-based systems, transactions are usually carried out via an auction. This auction, on the other hand, is carried out by voice call method as in the New York Stock Exchange or electronically as in BIST. In quotation-based systems, market makers determine the buying and selling prices and transactions are carried out at this price. Examples of these markets are the NASDAQ and the London Stock Exchange Automated Quotation (SEAQ) system. The SEAQ system is a trading platform that emerged with the first steps taken in London in the mid-1980s for consolidation of equity markets in Europe [23].

2.1.3.1 Session

In general, trading in the markets takes place in sessions, defined as the period between the start and end of the transactions. Securities traded on the stock exchange can be traded between certain session hours every day. This is called a stock market session. The session of all securities traded in the same market with the same trading method starts and ends at the same time. In many stock exchanges in developed countries, transactions are carried out in a single session. With the launch of BISTECH, a single session lasts the whole day, and the new trading hours in BIST are 09:40-18:10 as of November 14, 2016 [24], [25].

2.1.3.2 Trading Amount

The number of securities bought and sold in a market in a certain period or session is defined as the trading amount. Each trading day in the stock market, the amount is published on a per security basis. When the amount of trading realized in different markets is summed numerically, the total transactions on BIST on that day is revealed [26]. In the BIST equity market, the amount of trading of equities is published for each session. In addition, the daily trade amount for all markets is announced at the end of the day. Trade amount information can be accessed instantly from data dissemination screens and the BIST daily bulletin.

2.1.3.3 Trading Volume

Trading volume refers to the amount found by multiplying the trade amount by the trade price. The sum of the trading volumes of all equities constitutes the total trading volume of the equity market [26]. The trading volume of all securities traded in the BIST Equity Market is announced instantly via data dissemination screens.

2.1.3.4 Trading Unit

Lot is a trading unit in the equity market. Trading unit refers to the minimum number or value of any capital market instrument that can be traded with itself or its multiples. 1 lot is equivalent to 1 equity or 1 TL nominal value share [26]. For example, in stocks where 1 lot equity is equal to 1 equity (1 TL nominal), selling 1 equity means selling 1 lot of equities. The BISTBD can change the value of the trading unit according to the characteristics of the equities [23].

2.1.3.5 Weighted Average Price

The basis for the calculation of the base price to be applied in the next session is the weighted average of the prices traded during the session. Before each session, the weighted average price (WAP) is calculated for each equity, taking into account the previous session's tradings. When calculating the WAPs of equities, the tradings carried out as a result of the normal lot orders are taken into account [26]. Information such as WAP trading unit and trading amount are announced by BIST via data dissemination screens and a daily bulletin.

2.1.4 Price Determination Mechanisms in Equity Markets

As in all markets in stock exchanges, prices in stock markets are set where the supply and demand of market participants match and become tradings. During the transformation of supply and demand orders into tradings, various price determination mechanisms are applied according to the characteristics of the markets. The priority rules, which are used in the stock markets when waiting for the order to turn into tradings, are among the most important rules for price determination mechanisms [23].

2.1.4.1 Priority Rules

During the matching of buy and sell orders in the equity market, some rules determine which buy order will be traded first in buy orders and which sell order will be traded first in sell orders and the ordering of subsequent orders in this way. These rules are called price and time priority rules. To execute orders, first the price priority rule is applied and then the time priority rule.

- **Price Priority Rules:** According to the price priority rule, the order of priority is from lowest price to highest price orders in sell orders, and from highest price to lowest price orders in buy orders. Accordingly, low-priced sell orders have the right to be traded earlier than high-priced ones, and high-priced buy orders have the right to be traded earlier than low-priced ones.

- **Time Priority Rules:** The time priority rule is applied only to orders with the same price, and if there is price equality in the orders sent to the system, the order sent earlier has priority for trading.
- **Application of Priority Rules:** Price and time priority rules are applied separately to buy orders and to sell orders. Orders sent to the trading system according to these rules are stored in electronic media called "order books". These orders are automatically sorted according to priority rules [27].

An example of the application of priority rules is given below:

The orders sent to the trading system for a sample equity are listed below. The order of the orders according to the price and time priority rules explained above is given in Table 2.1.

1. Quantity: 20 Lots, Price: 2.23, Sell Order (Check in Time: 10:00:00)
2. Quantity: 70 Lots, Price: 2.23, Sell Order (Check in Time: 10:00:03)
3. Quantity: 80 Lots, Price: 2.23, Sell Order (Check in Time: 10:00:04)
4. Quantity: 150 Lots, Price: 2.23, Sell Order (Check in Time: 10:00:04)
5. Quantity: 100 Lots, Price: 2.23, Buy Order (Check in Time: 10:00:00)
6. Quantity: 15 Lots, Price: 2.23, Buy Order (Check in Time: 10:00:01)
7. Quantity: 200 Lots, Price: 2.23, Buy Order (Check in Time: 10:00:02)
8. Quantity: 40 Lots, Price: 2.24, Buy Order (Check in Time: 10:00:03)
9. Quantity: 50 Lots, Price: 2.23, Buy Order (Check in Time: 10:00:04)

Table 2.1 Sequence of orders sent to the trading System

Buy – Lot	***Buy - Price***	***Sell - Price***	***Sell - Lot***
40	2.24	2.25	150
100	2.23	2.26	20
15	2.23	2.27	70
200	2.22	2.27	80
50	2.21		

As seen in Table 2.1, orders sent to the trading system for a sample equity are automatically sorted first according to price priority and then according to time priority.

Table 2.1 shows that on the buy side the highest priority order is the order of 40 lots at 2.24, which has the highest price among the buy orders. This order ranks 4th in terms of sending time among the orders on the buy side. However, since the order price is higher than the prices of other buy orders, it is the highest priority order. Out of the buy orders with the same price of 2.23, that with 100 lots has priority over that with 15 lots, since it was entered into the system before. These three orders are followed by orders sent at lower prices.

On the sell side, the order with 150 lots at a price of 2.25, which is the lowest price among the orders, has the highest priority on the sell side. While this order is followed by other orders with a higher price, the order with a quantity of 70 out of two orders sent at 2.27 is ahead of the order of 80 lots due to the time priority rule [26].

2.1.5 Order Types

There are 5 types of orders in the BIST stock market: normal orders, odd lot orders, special orders, quotation orders, and short selling orders.

- **Normal Orders:** These are orders given as a trading unit and consist of exactly one lot and its multiples. It can take the following forms [24], [28]:
    1. *Limit Orders:* In these orders, price and quantity must be specified. The unexecuted portion of the order is placed among the equity's pending (passive) orders based on price and time priority ranking until it is traded in the order book or until the expiration date [24], [28].
    2. *Fill and Kill Orders (FaK):* This is the type of order that is entered by specifying the price and quantity, and the part that is not executed after the order is entered is automatically canceled [24], [28].
    3. *Market Orders:* These are orders in which only the quantity and no price are required. Such orders are matched with the best bid or ask in the order book until the order quantity is met, at which point those leftover are removed (not kept among pending orders) if no order is left on the opposite side. Market orders placed during the call auction phase stay on the book until the matching state, at which point they are executed. If a market order does not

match, it is deleted and not transferred to the ongoing trading session. The maximum order value per equity applies to market orders. The ever-last price is used to compute the limit value. If there is no previous closing price, the previous trade price is verified, and in the absence of both any reference price established manually by BIST is checked; otherwise, the system does not allow the market order until a trade occurs [24], [28].

4. *Market to Limit Orders:* These are orders in which just the quantity must be stated, without the price, and can only match with the best bid or ask on the order book. The remainder of the execution is converted into limit orders and placed in the order book among pending (passive) orders with the latest trading price (if the type of order is not FaK). Market to limit orders are subject to the per-equity maximum order value limit, which is computed in the same way as market orders [24], [28].
5. *Conditional Orders:* Certain conditions can be established for the activation or execution of orders for various types of orders. There are four kinds of conditional orders: quantity conditional orders, price conditional orders, time conditional orders, and partial display conditional orders [24], [28].
   - Quantity Conditional Order: Such orders are not executed if the entire quantity contained in the order is not entirely matched at the given price level. They would, however, be carried out if the entire quantity contained in the request was matched [24], [28].
   - Price Conditional Order: Such orders are activated or executed in the order book only if the current price of a stock or the best bid and offer price in the order book meets the price level indicated in the order [24], [28].
   - Time Conditional Order: These orders are activated at a preset point in the session or entered to be valid during a specific point in the session (for example, orders are valid at the session's opening/closing) [24], [28].
   - Partial Display Conditional Order**:** Limit orders with the partial display condition can be entered. When the quantity shown is executed for such orders, the predefined quantity appears in the hidden area. The remainder will be added to the order book as a new order based on price and time priority. This continues until the full order has been matched, the time

limit has expired, or the order has been cancelled. On June 27, 2016, the equities market's order type went live. When entering or amending an order, the displayed quantity must be at least 20% of the total quantity of the order [24], [28].

- **Odd Lot Orders:** For quantities less than one unit per stock, odd lot trading will take place. Only the quantity is stated in these orders; the price is not specified and such orders are traded at the price of the most recent normal order transaction if there is an opposite order. Odd lot orders might be matched in whole or in part, a 0.3-unit trade on ABCDE, for example. The price of the most recent trade performed on ABCDE will be used for KE (odd lot) [24], [28].
- **Special Orders:** These are orders that contain a sum that exceeds the equities-based limits. They are traded on a different order book with different trading rules. The BISTBD determines the characteristics of such orders, as well as the price determination and trading rules applied to them [24], [28].
- **Quotation Orders:** This is an order type that comprises limit price buy and sell orders that can only be entered by market maker/liquidity provider members. This is the buy and sell order that generates the quotation. A quotation order can be canceled entirely or just on the buy or sell side [24], [28].
- **Short Selling Orders:** A short sell order is used to sell unowned capital market instruments. These are orders for short selling trades conducted in accordance with Turkey Capital Markets Board (CMB) and BIST laws. The CMB and BIST laws govern the pricing restrictions for certain types of order [24], [28].

2.1.6 Trading Methods

There are different methods for matching the buy and sell orders during the trading of securities on the stock exchange. In the BIST Equity Market, securities are traded via the electronic trading system, based on the price and time priority rule, according to "*continuous auction - multiple price*", "*continuous trading with market maker*", or "*single price*" [25].

2.1.6.1 Continuous Auction – Multiple Price

The continuous auction method is used in the major stock exchanges of the world. Trading orders that match each other during the session become tradings at that moment and price is constantly set throughout the session [29]. Continuous trading, defined as "*the method of* *matching the orders sent to the system for a capital market instrument at different price levels in accordance with the priority and trading rules, continuously during the trading periods determined by the exchange*", involves the transaction times determined by the exchange for the orders transmitted to the system for a capital market instrument. It matches different price levels continuously, in accordance with the priority and trading rules throughout. It is also called the "*Multi-Price Method*" [30].

In the event that a buy (sell) order sent to the trading system is met with a suitable sell (buy) order, it is called a "*continuous auction*", when it is matched and processed without delay. The most widely used method in order-based trading markets is the *continuous auction* method, and it is possible to match the buy and sell orders sent to the system where this method is used at different price levels during the session; thus, different trading prices can be established within the same session. As soon as the orders are entered into the system, those on the opposite side are checked by the system, and the transaction is executed when the price of the entered order matches those of the pending orders on the other side [23].

Table 2.2 Continuous auction method order book recording

Buy – Lot	***Buy - Price***	***Sell - Price***	***Sell - Lot***
40	2.24	2.25	150
100	2.23	2.26	20
15	2.23	2.27	70
200	2.22	2.27	80
50	2.21		

An example order book for the continuous trading method is seen in Table 2.2. On the left of the table, buy orders are ranked from best to worst price, that is, from highest to lowest price, and sell orders are ranked from lowest to highest price. The concepts of the best buying price and best selling price should be clarified. When trading in financial markets, investors want to trade at the best price. For transactions such as foreign exchange, commodities, and stocks, high profit is targeted by buying at the lowest price and selling at the highest. The buying and selling prices determined by the market for this profit target are called the best buying price (ask) and the best selling price (bid).

The numbers of buyers and sellers vary at each price level, and the trading will not take place unless a sell order equal to (2.24) or better than the pending bid price (below 2.24) is received. Likewise, the trade will not take place (above 2.25) unless there is a buy order equal to or better than the pending ask price (2.25). Since the buy and sell order prices pending in Table 2.2 do not match, the trading does not take place and the orders of both parties are pending.

How the records in the order book change when two transactions take place in the sample order book is shown in Table 2.2.

- When a sell order of 20 lots of price 2.24 is entered into the order book in Table 2, the sell order will be executed without entering the waiting list (passive), since there is a buy order that can match this order. Thus, 20 lots of orders are met with a pending order of 40 lots and 20 lots are traded.

Table 2.3 Order book recording after first trading

Buy - Lot	***Buy - Price***	***Sell - Price***	***Sell - Lot***
20	2.24	2.25	150
100	2.23	2.26	20
15	2.23	2.27	70
200	2.22	2.27	80
50	2.21		

- After trading, pending orders in the order book are listed in Table 2.3, which shows that 20 lots of the 40 lot pending order have been traded and the remaining 20 are pending.
- When a new buy order for 200 lots of price 2.26 is received, the transaction will start at that price level first, since there is a sell order at a price level (2.25) even lower than the price of the buy order entered. That is, first, 150 lots will be traded at 2.25 (the price of the buy order is 2.26, but there is a pending order at 2.25); then there will be a trade of 20 lots at 2.26. As a result, 150 lots will be traded at 2.25 and 20 lots at 2.26. Since the total amount traded is 170 lots, the remaining untraded 30 lots

from the 200 lot order will be entered in the order book at 2.26 on the buy side, resulting in the pending orders in the order book as shown in Table 2.4.

Table 2.4 Order book recording after second trading

Buy - Lot	***Buy - Price***	***Sell - Price***	***Sell – Lot***
30	2.26	2.27	70
20	2.24	2.27	80
100	2.23		
15	2.23		
200	2.22		
50	2.21		

Thus, the ask price level has changed and the 30-lot buy order is placed at the top of the priority list, as it is the best priced. On the sell side, since there are no orders left at 2.25 and 2.26, the order of 70 lots at 2.27, which has time priority among the remaining orders, becomes the highest priority order [23], [27].

2.1.6.2 Continuous Trading with Market Maker

In this method, equities are traded by continuous auction. Unlike continuous auctions, a market maker enters "*quotation*" orders into the trading system, determines the price limits to be traded, and sends a quantity of orders that can be traded other than quotation orders and market orders to the trading system. In the securities traded with the continuous auction method with a market maker, the price range in which the trading can take place for that security is determined by a continuous double-sided quotation (*price and quantity*) by a member appointed as the market maker [31]. In this way, the market maker contributes to the formation of the market and the realization of tradings by placing quotation or buy and sell orders on its own behalf [24].

Quotation orders are entered on both the buy and sell sides and determine the upper and lower price limits at which the shares can be traded. While market makers determine the price limits with the double-sided quotations they enter, they also offer certain amounts in order to be able to trade. The aim here is to determine the price range of the security in which the transaction can take place in a healthier way and increase liquidity via the transaction volume. Although there are various differences in the applications of continuous auctions with a market maker in the world stock markets, the aim is to eliminate the drawbacks caused by the transaction volume with the help of the market maker in the markets with less liquidity in general [23].

2.1.6.3 Single Price

In the single price method orders are accepted without matching in the BIST Equity Market trading system for a set period, and at the end of that period the price level that provides the greatest number of transactions is calculated and all transactions are carried out at this level [32].

In single price trading, orders are collected in a certain time interval; then they are converted into tradings by matching the price that will result in the greatest trading activity. The difference between this method and the continuous auction is that there is no matching in buy and sell orders unless the moment of action occurs. The main purpose of this trading method is to determine the price level at which the maximum trading activity can be achieved. This price is also called the trading price.

The single price method involves two parts. First, buyers and sellers send their orders to the market and orders accumulate without matching. A trading price is determined according to the accumulated orders, that is, according to supply and demand. Second, orders that are equal to or better than the trading price determined according to supply and demand become trades. Table 2.5 is an example order table showing the orders sent to the market in the single price method and their prices. In this example, 100 lot buy and 100 lot sell orders are sent to each price level between 41 and 45. According to single price trading, the orders must be matched at the price level at which the transaction will take place the most.

Table 2.5 Single price order book and match price table

Price	*Buy*	*Cumulative*	*Cumulative*	*Sell Orders*
45	100	100	500	100
44	100	200	400	100
43*	**100**	**300**	**300**	**100**
42	100	400	200	100
41	100	500	100	100

In the above example, 43 is the trading price. At this price, 300 lots are processed and, 300 are bought, and 300 are sold. At other prices in the table, the trades that can take place are under 300 lots and are rejected according to the single price method.

2.1.7 Circuit Breaker System in the Equity Market

The automatic session stop system was abolished as of 30/11/2015 with the launch of BISTECH, and the circuit breaker was introduced in its place. When trades in a capital market instrument are carried out using the multiple price method, if the price change calculated over a certain reference value reaches or exceeds the specified rates, the trades of the relevant capital market instrument are temporarily suspended. After this temporary suspension, the price can be determined by the single price method in the relevant capital market instrument or a certain period can be expected before the trade takes place [31], [33].

2.2 Manipulations in the Stock Market

Manipulation prevents the capital market from operating within the norms of trust and transparency. It is an important issue that should be taken into account both for individual investors offering funds to the securities markets and for companies seeking funds by issuing stocks. Major fluctuations and crises in the economies of countries where development and deepening cannot be achieved in the capital markets are inevitable. During such periods, capital markets are one of the most important resources for ensuring healthy and sustainable growth. These markets need to be developed and deepened in order to achieve the highest benefit for fund and fund demanders. This is only possible by attracting more investors into the market and increasing market volume. On the other hand, investors' lack of confidence in the market prevents more investors from participating. In this context, we encounter manipulation, one of the most important causes of distrust in the market [20].

Manipulation in capital markets comprises all actions attempting to shift the price above/below the demand and supply balance occurring in the normal market equilibrium and trading financial instruments at the determined price level. In other words, manipulation is the general name for attempts to artificially change prices outside the supply–demand balance and to ensure that the security is traded at this price. It includes all intentional actions aimed at deceiving or defrauding traders by influencing or controlling the price levels of financial instruments with market dynamics/unrealistic methods. For this reason, manipulation disrupts the functioning and order of the market, resulting in a violation of free competition conditions or defrauding investors [34]. Although manipulation has many different meanings, the term is frequently encountered in the securities market. It can be simply defined as artificial control of security prices or conducting artificial transactions to give the impression that a stock is actively traded [35], [36].

Manipulation in the capital markets can also be defined as the tradings conducted in order to give a misleading impression to investors or to create a misleading market regarding the stocks traded in those markets [37]. In other words, it aims to deceive and defraud investors by knowingly and willingly controlling capital market prices or artificially influencing prices [38]. In short, it is the artificial control of securities prices. In normal market conditions, all activities aimed at artificially raising, lowering, or maintaining the prices of capital market instruments determined according to supply and demand at a certain level are considered manipulation [39]. In other words, it is an illegal act in markets where supply and demand are often lacking in breadth and depth. It often occurs when most investors do not have enough information about the market structure and stocks and the market rules are unclear [20].

In capital markets and stock exchanges, the direction of capital market instruments is changed by buying transactions in violation of the public disclosure regulations and by false, misleading, or insufficient explanations that affect the decisions of investors [40]. Trading in bad faith in a way that harms the rights and interests of investors by giving the appearance of an active market leads to the formation of artificial prices and markets [41]. Those engaging in such manipulative trading causing stock price changes generate fake prices in the market and benefit while investors lose. Manipulation is defined as actions intended to interfere with price mechanisms by inhibiting the free interaction of supply and demand, trick individuals into trading a security, or keep the price of a security at artificial levels [42]. Manipulative activities damage the confidence of investors and have a devastating effect on market efficiency [43].

The primary goal of capital market manipulation is to direct investors to buy or sell securities or to keep the price of securities at an artificial level within the extent of external interventions [37]. In other words, manipulation is all kinds of deliberate actions that aim to deceive or mislead investors by artificially influencing and controlling the price of capital market instruments. From this point of view, manipulation can also be described as an important concept that includes more than one form of activity [44].

The Committee of European Securities Regulators (CESR) has listed three main actions in market manipulation [45]:

- To act in cooperation by directly or indirectly fixing the buying or selling price of the security or creating a dominant position regarding the supply or demand for a financial instrument,
- Buying or selling financial instruments to influence the market closing price,
- Making a profit on securities by taking a prior position on a financial instrument and expressing opinions through traditional or electronic media without considering the public interest.

2.2.1 Manipulation Concept

Manipulation, according to the dictionary definition, means a game, intrigue, cheating, deception, juggling, the activity of directing the market in line with one's own interests by spreading false and misleading news in the market [46]. According to the Turkish Language Association, manipulation is defined as *"changing or directing information by directing, selecting, adding, and removing"* [47]. Manipulation, which means deliberately tampering with the prices of securities for gain [48], is a French criminal concept and is used in the sense of manually designing or manipulating [5]. Originally, working a puppet with strings tied to the fingers for manual guidance was an act of manipulation, but now spreading inaccurate information using all kinds of written/visual/virtual channels and directing investors in a predetermined direction is also within the scope of manipulation [49]. If manipulation is defined as deliberate interference with the market balance, the issues of encouraging people to buy and selling, interfering with freely occurring supply and demand, and making efforts to increase/decrease the prices of financial instruments to levels incompatible with market dynamics are also included in the concept [42]. The word manipulate is defined as using skillfully, managing skillfully, directing in line with one's own purpose, and influencing [40].

Generally, in the markets of countries with deficient legal and institutional structures, a suitable foundation is created for manipulative activities, and manipulations can be seen in other market mechanisms in the economy as well as in securities markets [41]. Therefore, it can occur in every field from the industrial sector to the agricultural sector, from the most advanced financial investment instruments to futures [36]. Especially periods when there are no rules regulating the market and the majority of savers do not have the necessary knowledge and experience about financial markets and securities traded in the markets, manipulative practices are more common [50]. According to Ozbay, manipulation can also be defined as the transactions made with the intention of creating a false and misleading impression or creating a misleading market for the securities traded in the markets. Based on this definition, all activities aimed at artificially raising or lowering the market prices of securities or keeping them at a certain level are manipulation [50].

A manipulative attempt to deliberately interfere with freely occurring supply and demand, to encourage people to buy and sell, and to artificially bring the price of a security to a desired level causes artificiality in the supply–demand and prices of the assets in the market [42]. In this sense, although the real buying and selling transactions of a financially strong investor in the market are not artificial, the effect of the transaction on the related asset to increase or decrease the price is considered manipulative since it is artificial in nature [51].

Manipulation is not defined in either Capital Market Law No. 6362 (CML6362) or the repealed Capital Market Law No. 2499 (CML2499). However, tradings that can be considered manipulation are regulated as a type of crime in the relevant legislation. Namely, in clause A of article 47 of the repealed CML2499, while it is counted as "*crimes committed in the form of obtaining an unfair advantage or eliminating a loss in a way that disrupts the equality of opportunity among investors by using non-public information to gain benefit,* *which may affect the value of capital market instruments*"; in Article 107 of CML6362, in the section entitled capital market crimes, "*buying or selling, placing orders, canceling orders, changing orders or making account movements in order to create a false or misleading impression regarding the prices, price changes, supply and demand of capital market instruments*" is considered market fraud [52]. As stated in the relevant articles of the capital market legislation, legal regulations are in place to prevent manipulation in the capital markets, and various fines and prison sentences are foreseen for those who engage in manipulative practices. The main reason for the sanctions is the link between excessive fluctuations in the capital markets that do not match the economic realities in the markets and manipulative practices [53].

In order for an activity to be called manipulative, it must have some features, and some possible signals indicating a transaction is manipulation are as follows [53]:

- Interfering with price mechanisms by preventing the free interaction of supply and demand,
- Inducing people to trade in a security by deceiving them,
- Actions designed to keep the price of the security at an artificial level.

Investors in the market think that the supply–demand balance in the market reflects the real price level of the securities. Therefore, it can be inferred that investors act on the current situation in their evaluations and analyses regarding the pricing of securities. However, actions that are manipulative not only prevent investors' forming expectations, but also cause unfair losses. On the other hand, manipulators can gain excessive gains in transactions carried out in the context of these possible signals. In fact, the main purpose of the manipulator is to avoid a loss or make a profit by creating unusual price movements to the disadvantage of investors unaware of the manipulative actions [20]. Therefore, they attempt to manipulate the price of a security when they believe it will yield a profitable result [12].

The point to be considered in the determination of manipulative activities is not whether the applied transaction is artificial or real, but whether there is a conscious intervention in the supply–demand structure. For example, a market participant with financial strength for a stock with no depth can lower or raise prices by controlling real buy and sell orders. This method is not artificial. However, as the result of the buy and sell transactions is artificial, in other words, the price balance is created artificially, the transactions have manipulative characteristics. From this, it can be deduced that since prices are created and controlled artificially, manipulative features are added to trading transactions [20]. To sum up, the basic features of manipulation can be listed as follows [54]:

- There is an element of intent in manipulative activities involving a conscious and planned intervention in line with the purpose.
- In order to differentiate the equilibrium level of prices, the natural supply–demand balance is interfered with and artificial markets are created that do not reflect expectations.
- In this artificial market, it is aimed to deceive investors unaware of abnormal developments in prices and manipulative activities.

2.2.2 Purpose of Manipulation

The main purpose of market manipulation is to force investors to trade by artificially raising or lowering the prices of stocks or keeping them at a certain level [14] in order to gain high profits in the securities markets [55]. Since the main goal is to influence prices, manipulative activities are carried out in securities exchanges that allow every investor to trade, have an unlimited investor mass, are constantly traded, and have sufficient trading volume [5]. The aims of manipulative activities carried out in this context are as follows:

- For profitable sales, people or institutions with securities artificially increase or reduce the prices of securities,
- Blocking a general market collapse by slowing or stopping a rapidly falling market,
- To eliminate speculative raids concentrating on a single security and fix the price until it finds the normal market level of a newly issued security.

Basically, manipulative activities have three main purposes [41]. The first is to give the appearance of an active market by conducting buying and selling transactions to increase the stock prices and selling the portfolio at a more attractive price by attracting other investors into the market. Disposing of the existing portfolio by artificially raising stock prices is the simplest and most widely practiced manipulation method. The second purpose is to ensure the continuity and stability of stock prices. In such tradings, it is aimed to create a stable equilibrium in stock prices. This is generally applied in the mediation of public offerings. The third and final purpose is to bring prices down. Generally, this can be done by people who enter the market by short selling or take positions in futures markets [56].

2.2.3 Various Concepts Confused with Manipulation

Although manipulation in the capital markets is not a criminal offense, various elements have effects of the price formation processes on the market. These should not be confused with manipulation; speculation and price stability operations include revealing, repo, and reverse repo [57]. Manipulation is defined as artificially creating the prices in the markets using various techniques, while speculation is generally the trading of buying or selling the economic asset according to the predictions of the future price of the property, currency, or security. Thus, speculation is a spontaneous result in the market, while manipulation is an artificial and misleading action [58].

The concept of speculation is defined in the dictionary as "*the activity of making a profit from the price differences by buying the economic asset whose price is expected to rise based on personal predictions or by selling the one whose price is expected to decrease*" [59]. Speculation and manipulation are frequently confused in the finance literature [60], and the most basic point separating them is that speculation is not a crime in terms of the laws, regulations, and other practices [58]. However, manipulation that causes artificiality in security prices through deliberately interference with the supply and demand of any security is seen as a criminal action [12]. Speculation can be defined as the execution of buying and selling of any product with the hope that its current price may change in the future [61]. In other words, it is an effort to achieve an unguaranteed gain by assuming risk on the basis of expectations formed after the current market structure is evaluated [60]. The expectations of the speculators are to make a profit by buying if the price of the product decreases and by selling if it increases. Such tradings based on buying and selling take place mostly in stock exchanges and futures exchanges, which are more liquid than current markets and allow sufficient volume in short-term positions and are where speculators operate [44].

There are some exceptions in price stability operations. These operations can be defined as those that make price formation possible with a mechanism other than supply and demand in the market and prevent the negative effects that may occur in the market affecting the price [62]. However, these transactions are carried out with the aim of regulating the market when supply and demand do not stabilize after the public offering and there are uncertainties about price; therefore, since they aim to protect both the issuing institution and the investor, they fall outside the scope of criminal and legal liability arising from manipulation [57]. Repo is defined as the sale of securities with a commitment to re-buy, while reverse repo is defined as the buying of securities with a commitment to sell back. Like price stability transactions, repo and reverse repo are not manipulative transactions; they are considered in compliance with legal regulations. Short selling should not be confused with manipulation [20]. Short selling transactions take place within the scope of selling the unowned capital market instruments in line with the commitment to buy/replace later or placing an order for the sale [57].

2.2.4 Types of Manipulation

Stock trading transactions in the securities market are now carried out through the internet and millions take place per day. The differentiation of manipulative actions in markets is due to innovations in information technologies, as is the continuous emergence of new and creative methods, and the number of transactions that disrupt the market is increasing day by day with the development of information technologies.

Manipulative activities that disrupt the order and structure of competition in the markets are generally carried out in three ways [63]:

- Creating a suitable environment that will increase the prices by creating the belief that there is or will be a rising market, and selling the stocks at high prices,
- Buying stocks at low prices by making an effort to keep prices in a downward trend (*In this approach, it is aimed to replace stocks at low prices in short selling transactions*),
- Keeping the price at a certain level with transactions aimed at ensuring that the stock reaches its required value.

Two types of manipulation are mentioned in CML6362, which legally regulates manipulation. The first is trade-based market fraud in paragraph 1 of Article 107 and information-based market fraud in paragraph 2. In the academic literature, action-based manipulation is also a type of manipulation. The common point of all three types is that the functioning of the capital markets is negatively affected and the tradings must be conducted with an intention [64]. Manipulations classified under three headings [63], especially action and information-based manipulation attempts, are fraudulent, considered illegal under the Securities Exchange Act of 1934 [65].

2.2.4.1 Action-based Manipulation

Action-based manipulation is generally carried out to influence and differentiate the expected real values in terms of stocks, and it is aimed to change the market prices of the stocks and thus the company value [41]. The seizure of most of the publicly traded shares of a security, the use of company shares by company employees for their own interests, and the failure of intermediary institutions to fulfill customer orders in some cases are examples of action-based manipulation [66]. Similarly, the type of manipulation that ensures that the prices of the stocks traded in the market and the market values of the companies holding the shares are formed outside the equilibrium price can be defined as action-based manipulation. In the type of manipulation that aims to bring the stock prices and market value of the company to the desired levels, anyone aiming to manipulate must have a strong portfolio structure or cooperate with others having similar aims [20].

An example of action-based manipulation is a management that decides to close a profit-making factory in order to decrease the stock price, then buys large amounts of stocks that have fallen in price, and then reverses this decision [67]. In this example, short-term positions are closed profitably by buying the stocks at a low price, and then gains can be made in the following process by new negotiations or announcing the opening of a new factory [68]. Another example of action-based manipulation is the manipulative actions by the Harlem Railways at the beginning of 1863 [63]; at the beginning of 1863, the shares of Harlem Railways, which were between $8 and $9, increased to $50 in the following periods. In October 1863, the stocks rose from $50 to $75 when the New York Council decided to build a streetcar system along Broadway for the Harlem Railroad. Thereupon, the members of the council first sold their stocks short and then canceled this decision and put pressure on the price to go down [12]. One last example of this type of manipulation was in 1901; while the stocks of the American Cable Company were trading at $60, the company management took a decision that shocked the entire market and closed one of its factories. Subsequently, the value of the company's stock decreased to $40. The company executives collected their share certificates at $40, and then announced that the closed workshop had started production again. Upon this, the company's stocks returned to their previous level and the managers earned millions of dollars [63].

The basic behavior in action-based manipulation is to raise the takeover bid. Here, after the manipulator buys the stocks of a company, he bids to transfer these stocks. This causes the price of the company's stock to rise. Even after the prices rise, the manipulator makes a profit by selling his stocks at this high price. Naturally, the bid price decreases after that [69].

2.2.4.2 Information-based Manipulation

Information-based manipulation includes activities such as giving false and misleading information to investors or creating false rumors about the company in order to bring the supply–demand balance in the market to the desired point [70]. As a result of such manipulations, an artificial price and market are created depending on the changing of the direction of prices with the created fluctuations [71] and an unfair advantage is provided.

In the information-based manipulation type, investors who have the power to affect the prices of securities make statements that increase or decrease the price of the securities within the scope of the price level determined before the manipulation. In the process following the learning of the determined price, buying and selling transactions are carried out. In information-based manipulation processes, there are investors who have information on the gaining side and investors who do not have information or have incomplete information on the losing side [17]. Investors consisting of individuals or groups that have the power to affect the prices of securities with the information and data they have can be summarized as follows [41]:

- Economy writers and correspondents of newspapers,
- Those who have the power to affect prices the most, issue bulletins about the state of the market, and make purchases and sales transactions on their own and investment companies' accounts,
- Officials who can access important information about the company.

Information-based manipulation is penalized according to both the CML6362 (Article 107/2) and the CML2499 (Article 47/I.A-3).

Table 2.6 shows that there is no significant difference that could entail an intervention in market mechanisms in the regulation of information-based manipulation in CML6362 or CML2499.

Table 2.6 Information-based manipulation definitions

CML2499 Article 47/I.A-3	***CML6362 Article 107/2***
Any natural person, authorized person of legal entities, and those acting in concert that provides information, news, and comments that are misleading, false, or deceiving and that could influence the value of capital market instruments or fails to disclose information that he/she should disclose shall be imprisoned from two to five years and fined juridically from five thousand days up to ten thousand days.	Any person who provides or disseminates information, spreads rumors, gives news or prepares reports that are misleading, false, or deceiving, with an aim to influence the value or price of capital market instruments or investors' decisions shall be imprisoned from two to five years and fined juridically up to five thousand days.

The expression "*not disclosing the information they are obliged to disclose*" introduced in CML6362 has been removed from the definition and it is not considered an information-based manipulation crime [72].

The most typical example of information-based manipulation involves the applications known as “*transaction pools*” that emerged in the USA in the 1920s. Here, investors form a group and first buy stocks and then spread positive and misleading information about the company, creating a pool to sell their stocks at a certain profit level [69]. Especially with the development and spread of internet technology, examples of information-based manipulation are media figures acting together with traders presenting untrue news about the manipulated stock, leaving misleading messages on internet forums, and sending spam e-mails encouraging the buying of stocks [19].

2.2.4.3 Trade-based Manipulation

In the trade-based manipulation, the changes in stock prices are provided by buy and sell tradings. Unlike the other types of manipulation described above, in trade-based manipulative, activities aiming to differentiate the equilibrium price created in the market are carried out only by trading transactions, without trying to change the value of companies with stocks that are traded in the market or presenting false or incomplete information about these stocks [73]. Unlike action-based manipulation, there is no need for a very strong portfolio structure in trade-based manipulation, and markets can be manipulated with the help of small-volume but continuous transactions. With this type of manipulation, misleading market movements can be achieved by only buying and selling the stock, regardless of the developments, data, and information that may affect the value of the stock [20].

While trade-based manipulation is defined as behavior that consists of active transactions or transaction orders and that gives or is likely to give an incorrect, misleading impression to keep the price of one or more qualified investments at an artificial level [74] , it is also defined as the transactions carried out in order to provide risk-free positive real return by those who affect prices through large-scale trading [75]. Such manipulations are seen in markets where information flow is not very healthy, the investor profile does not have much information, the inspection mechanisms are insufficient, and there is asymmetric information distribution [53]. Artificial price movements desired to be created by trade-based manipulations can be achieved in three main ways [57]:

- Preparing an environment in which prices will rise with the impression that the market will rise,
- Collecting stocks at low prices in order to create the impression of low prices in the market,
- Keeping stock prices at a certain level to prevent them from rising or falling.

According to the definitions, the crime of trade-based manipulation had a specific goal in mind, which was to create a false or misleading impression about the prices, price fluctuations, and supply and demand of capital market instruments [76]. Trade-based manipulation is a criminal offense in this country, according to paragraph 47/A-2 of CML2499 and clause 107/1 of CML6362. It is a trade-based manipulation crime, according to paragraph 47/A-2 of CML2499, to intentionally alter the supply and demand of capital market instruments, to create the appearance of an active market, to hold their prices at the same level, or to increase or reduce them. In contrast, according to CML6362 paragraph 107/1, those who buy or sell, give orders, cancel orders, change orders, or conduct account movements in order to create a false or misleading impression about the prices, price changes, or supply and demand of capital market instruments commit the crime of transactional manipulation. Furthermore, in this statute, the crime of trade-based manipulation is called market fraud [12].

Tradings conducted in trade-based manipulations are formally legal. However, the actual intention of those performing the transaction is different. By hiding their true intentions, they try to influence the price and create an artificial market by buying or selling in the securities markets [53]. Unlike other types of manipulation, no incorrect or incomplete information is presented to the market in trade-based manipulations. Actions are taken to mislead the market only by buying and selling securities [77]. Economic or psychological factors also play a major role in the evaluation of the manipulative transactions. In particular, according to behavioral finance theory, investors act according to the "*herd*" psychology and perform transactions without analyzing financial information, being influenced by the decisions made by other investors [5]. It is necessary to evaluate the techniques used in trade-based manipulations according to each case and to investigate whether it is criminal. Moreover, although these techniques can be legal on their own, they can be manipulations with coordinated transactions [78].

2.2.5 Trade-based Manipulation Types

Trade-based manipulations can be used for many different techniques, as they are carried out through legal transactions. Therefore, it is possible to use technological developments in manipulative techniques. As a result, trade-based manipulations also occur in markets with sophisticated and stringent control measures [70]. Manipulative tradings commonly used in the stock market can be grouped under two main headings: unreal tradings and real tradings [5]. Examples of real and unreal tradings are given in Table 2.7.

Table 2.7 Real and unreal tradings

Trade-based Manipulations	***Manipulative Tradings***
Unreal Tradings	• Matched Orders • Washed Sales

Trade-based Manipulations	***Manipulative Tradings***
Real Tradings	• High Frequency • Runs • Frontrunning • Short Selling • Corners • Capping / Pegging • Marking the Open • Marking the Close • Parking and Warehousing • Spoofing • Pump and Dump

2.2.5.1 Unreal Trades

Unreal tradings do not create a real change in ownership. Such tradings are the most typical trade-based manipulation. In fact, unreal tradings cause the most damage to the market and form the starting point of manipulation response regulations. In US legal regulations prohibiting initial manipulation, unreal tradings were seen as the types of market manipulations that caused the most damage to the stock market [5].

2.2.5.1.1 Matched Orders

Matched orders, which occur when the buy and sell orders of the same security are entered into the system simultaneously [79], are defined as orders that aim to mislead other investors with misleading information about the demand or supply of a security [80]. Matched orders, in which similar or reciprocal orders for the buying or selling of stocks are sent to the system [81], consist of the activities of both parties. While the first party enters the buy order into the system, the second party enters the same or a very similar sell order [82]. The manipulator, who then places large volumes of purchase orders to increase the price of the stock, cancels these orders and aims to make a profit by placing the orders they want fulfilled [19]. Therefore, in this trading, an artificial trading volume is created, increasing the price of the stock, and it is aimed to draw the attention of investors to these stocks [83]. For this purpose, orders that are similar to each other in terms of time, price, and amount are sent to the system and an artificial activity and price movement is created in the market of the traded stock, spreading misinformation among investors [5].

The main purpose here is to direct other investors in the market to buy the relevant stock by creating an active impression about it. Accordingly, other investors who will buy this security will cause the price to rise, which can be considered a selling opportunity for the manipulator [20]. Tradings can take place between two different people or between two separate accounts [36]. When the tradings take place between different accounts belonging to the same person, "*tradings that do not create a change in ownership*" occur.

2.2.5.1.2 Washed Sales

Washed sales, the oldest and easiest manipulation technique in capital markets, is a type of unreal trading that does not result in a change in ownership. It is also called "artificial or fake sales". The manipulator(s) conducts trades with high intensity and narrow frequency ranges in order to create the impression that the trading volume of the stock has increased, to affect the price and therefore to mislead other investors [20]. This method, also known as wash trading, occurs as a result of the parties involved agreeing before executing the trading and entering the orders of the same price and the same amount into the system simultaneously [53]. The main purpose here is to increase the market trading volume by giving the stock an active appearance [84]. In this trade, since there is no change in the ownership structure of the relevant security, the trade consists of artificial sell orders that do not have a real economic result [79]. Therefore, without changing the ownership or market risk of the financial instrument, only the buying or selling of securities based on a confidential contract takes place between the parties [12].

Such trades are artificial or fake sales conducted with the specific purpose of reducing tax payments or manipulating balance sheet items. The basis of washing sales is to create a profit opportunity or avoid loss by providing artificial price formation in connection with tax [56], [85]. For example, if an investor fake sells his investment instruments to someone he trusts and then buys them back, a wash sale is the result [41]. Manipulators can also create artificial volatility in the stock by buying or selling the stocks themselves from time to time. For example, a speculator who takes a long position on a stock or works to create a market for a particular stock may attack other speculators with the price above the market value of a stock that he will sell and that his collaborator will buy. In this transaction, no change in ownership takes place and the buyer is not exposed to any financial obligations to the seller through the agreement. The aim is to bring about a movement in the stock with this sale transaction, which is completely fictional [86].

Washing sales are frequently used in emerging markets due to the inadequacy of legal regulations [58]. Regulators in the USA stated that investors who acted as both buyers and sellers in the same trade gave false information signals to the stock and futures markets in order to manipulate prices and they prohibited such behaviors [87]. In fact, the securities law numbered 1934 prohibited such transactions aimed at deceiving unaware investors who entered the market with the hope of making a quick profit by creating a misleading transaction image [88].

2.2.5.2 Real Trades

In real trades some investors are involved in manipulation, while others are not. Although the trades by investors not involved are essentially legal, transactions carried out by investors misguided by the creation of an artificial supply and demand market for stocks by the manipulators are considered manipulative [5]. In order to determine whether real trades are manipulation or not, factors such as the timing and amount of trades and account movements between traders should be taken into account [89].

2.2.5.2.1 High Frequency

High frequency orders are those placed in succession on the stock market, and the most basic feature that distinguishes them from other stock orders is that they are given at constantly rising or decreasing prices. By this method, it is aimed to create an artificial price with orders that are small in quantity and by decreasing or increasing the price by one step [53]. By ensuring that the stock price is above or below the last realized level, it is aimed to make the closing price appear high or low in the same way [90].

2.2.5.2.2 Runs

In stock markets, it is common to create an active trading image for the stock by conducting intense buying and selling transactions and to draw attention to the stock via its increasing trading volume [20]. Transactions carried out as runs in securities cover the intense buying and selling activities of an investor in order to bring vitality to the stock market. The purpose of the investor here is to attract the attention of other market participants to the relevant stocks and to increase the stock prices by encouraging them to buy [5], [53]. It is known informally as "*tipping the scales*" [62]. Next, the stocks in the portfolio are sold at the higher price levels and, accordingly, abnormal returns are obtained [20]. Therefore, these investors, who make manipulative buys by increasing the demand for stocks, gain by selling the stocks they bought cheaply at a higher price after the increase in price and demand [19].

Manipulators can also carry out intensive selling transactions by agreeing with each other regarding the stocks that they want to decrease the price of via the runs method. In this type of transaction, the market price of the stock with increasing supply decreases and the manipulators who buy again at the falling market price profit [12]. Runs transactions in securities are legal in appearance. However, when they are examined, it is seen that there are regular price increase or price decrease transactions. In order to determine whether such transactions are for manipulation, the investor or the intermediary institution performing the transaction should be examined and it should be investigated whether they are rational [91].

2.2.5.2.3 Frontrunning

This method aims to take advantage of the possible price effect by conducting a transaction before the purchase and sale order, which may affect the market price of a market actor, is transmitted to the stock market. These transactions can be carried out by investor institutions or persons or institutions authorized to act by proxy on investor accounts [5]. In fact, in CML, these transactions are not defined as manipulation crimes, but as market disruptors. They are not intended to create a perception among other investors, but to make small profits by using the order transaction information of large investors.

2.2.5.2.4 Short Selling

Short selling is based on the principle of “sell high first, then buy low”, without the investor owning the security at the time of the sale. That is, it refers to the sale of a borrowed stock by an investor. Then, in order to close this loan debt, the same stock must be bought [92]. In other words, it is when the seller sells a security that he does not own with the expectation that its price will decrease in the future, provided that it is replaced when the time comes

[93]. In a short sale transaction, the price of the stock in question is expected to decrease. Thus, those who want to sell short for manipulative purposes perform fictitious transactions to reduce the stock price [53]. Short selling is not a manipulative transaction per se, but at the time of the transaction, it can have manipulative effects due to the stock not being owned yet.

2.2.5.2.5 Corners

In this technique, people who try to take control of a company's stocks use their dominance, squeezing them into a corner and selling at a high price while fulfilling the obligations of short-selling people [62]. It is also called “pressing” or “cornering” [41]. Investors, who expect that the price of a security will decrease, need to close this debt by purchasing the stock that they have borrowed when the time comes when the security with which they have made a short sale transaction should be returned to the market. People who are aware of this situation, by purchasing large amounts of the said capital market instrument, oblige those who are required to fulfill their debts by restricting the supply (investors selling short) to purchase the debt securities at a high price [53].

The corners technique is generally applied in short selling. In these transactions, investors sell stocks that are not included in the portfolio, assuming that the stock prices exceed the required levels and that the prices will decrease again in a short time. The manipulator, on the other hand, takes advantage of this situation among investors and raises the prices artificially. Seeing the volatility in the market, the investors turn to short selling to make a profit, unaware of the existence of a manipulator in the market. However, the inflated prices do not fall, contrary to what is believed; they are further increased by the manipulators and the short-selling shares are bought by the manipulator. When manipulators take over a significant portion of the stock, short-selling investors cannot close their positions by buying low. Thus, the manipulators who control the price of the stock can corner short-selling investors [20]. However, corners are not a manipulative technique on its own; it paves the way for results from other manipulative transaction patterns.

2.2.5.2.6 Capping and Pegging

This is carried out by selling large amounts of securities in the stock market in order to keep the market price below a certain level (capping) or buying in order to keep it above a certain level (pegging) [57]. The most basic feature of this type of transaction is to keep the price of the stock in question at a certain level. Transactions conducted to keep the price of the security at a certain level, if not disclosed to the public, are considered manipulative to keep the price of such transactions at the level determined by the manipulator. These transactions are essentially legal and since there is no remarkable price increase or decrease, keeping prices constant at a certain level, they are very difficult to detect [5]. Although the actions aimed at fixing the market price of the stock artificially affect the stock price created according to the supply and demand in the market, the main goal is to fix the stock price at a certain point, not to increase or decrease it [55].

2.2.5.2.7 Marking the Open and Marking the Close

The price levels formed within the scope of the orders given at the opening and closing of the session have important effects on the buying and selling decisions of investors. In line with the orders given, it is possible to determine the price dynamics that will occur in the following processes of the session by controlling the trading volume and prices at the opening of the session. Such activities, which aim to artificially manipulate prices in line with session opening and closing, are manipulative [53]. The orders placed at the beginning of the session and the executed transactions are important as they reflect the judgments of the market participants regarding the events that took place between the sessions. The transactions at the closing of the session represent a summary of the day. Thus, participants who want to see the latest situation regarding the price level of the stock will look at the closing price [39].

Determining the closing of the session is a manipulative activity that consists of entering a buying or selling order or actively buying or selling in the last moments of the session to artificially affect the closing price of a stock [35]. This technique can be summarized as follows: most transactions conducted during the closing process are fictitious, usually small transactions are performed, and the price of the stock is closed at lower or higher levels than it would normally be [20]. In addition to the closing price, transactions that aim to determine the direction of the session by influencing the trading volume or prices at the opening of the session can also be used for manipulative purposes [53]. This strategy involves placing a buying order at slightly higher prices or selling at lower prices to raise the price of the security when the market opens [94].

2.2.5.2.8 Parking and Warehousing

According to most of the definitions of stock storage transactions (warehousing - parking) in the literature, this focuses on transactions to hide the real ownership of the security through fictitious tradings [95]. The first of the two prominent methods is "*warehousing*". It is expressed as "*the buying of a stock by an intermediary institution or investor in order to transfer it to another*" [5]. In other words, an intermediary institution or investor who wants to hide his shareholding undertakes to buy back securities in the market, in accordance with an agreement; someone else buys securities from the market, but then these securities are bought by the committing party [53]. According to Goldwasser, the process of buying securities by one party on behalf of another party is defined as warehousing [81].

In the second of the stock storage transactions, known as parking, intermediary persons or institutions such as brokers/dealers first transfer the stocks to another broker/dealer or the customer's account temporarily and then buy back these securities without loss [81]. In other words, the securities owned by a brokerage house are temporarily transferred to the accounts of another brokerage house or customers without any buy and sell transaction, and then transferred to the real owner's account [53]. Unlike warehousing, the operation performed in parking is not real. In this case, sell transactions are generally artificial; the seller's purpose is to buy back the securities or to re-charge them to the firm's accounts at a later date [96]. Stock storage transactions are not a manipulative transaction on their own, but pave the way for obtaining results from other manipulative transaction patterns.

2.2.5.2.9 Spoofing

Some investors may choose to affect the price of a stock by entering impracticable orders on their order books, creating the impression that there is an imbalance between supply and demand [80]. All orders entered with the aim of misleading are canceled by the software and bots just before they are executed. Manipulators aim to make profit by taking advantage of the changing price by placing the order they want to be realized, and such actions are called spoofing. These orders are a form of manipulation in which a trader places fake buy and sell orders that are not intended to be carried out by the market. This is often done through software and bots to manipulate the market and asset prices by creating a false perception of supply and demand. Market manipulation through false orders is illegal in Turkey, as in many other countries. Trade-based market fraud, which is defined as "purchasing and selling capital market instruments, placing orders, canceling orders, changing orders, or carrying out account movements" in Article 107/1 of CML6362, also includes misleading orders. According to the law, those who place misleading orders in the markets under the control of the CMB are punished with imprisonment from three to five years and a judicial fine from five thousand days to ten thousand days.

2.2.5.2.10 Pumping and Dumping

This involves the buying of a certain stock intensively in order to create a large demand for it and then selling it at a high price [97]. A person or a collaborator buys or sells to artificially raise or lower the price of a security. In order to raise the price of the security, large investments and earnings are promised and other people are encouraged to invest in this stock. When the price of the security reaches the determined target price, in order to artificially lower the price of the security manipulators sell the securities they bought at low prices in large quantities, causing investors outside the group to make a loss [98]. In other words, in this manipulative attempt, firstly, a large buying position is obtained on the asset and then market participants are persuaded and the long buying position is profitably closed by purchasing the asset at high prices [99]. Then the manipulator, who continues to buy the security at rising prices, creates a momentum in the relevant stock and can then make a profit by selling his shares at these high prices [100].

2.2.5.3 Patterns of Manipulative Transactions

Although the trade-based manipulation types were defined in the previous section, it is not possible to specify in a restrictive manner which transactions can be included in these definitions. However, some of the transaction patterns that are agreed on theoretically and that occur frequently in practice are given in Table 2.7. Not every transaction that is compatible with manipulation types is always manipulative. Things to consider for a transaction to be manipulative are as follows:

- To be used together with other types of transactions,
- Timing of transactions,
- Realization of the traded stock with a density proportional to the market depth,
- Creating groups of traders.

A manipulation may occur depending on the combination of one or more of these factors [53].

Transactions performed within the scope of trade-based manipulation generally have the appearance of a legal stock market transaction, but they can cause manipulation when they are carried out as part of the manipulation plan and when the necessary conditions are met. The diversity of these movements is increasing with the development of information technologies. Information technologies are also used to determine whether such acts are manipulative or not. Trade-based manipulations can be detected, especially by computational methods such as data mining and machine learning.

The manipulative transaction patterns according to national and international practices and the types addressed are presented in Table 2.8 in order to inform investors about what kinds of behaviors are involved in manipulation [5]. Short selling, corners, front running, and stock storage transactions (parking and warehousing), which are trade-based manipulation types, are not included in the table since they are not manipulative transactions on their own.

Table 2.8 Trade-based manipulative Patterns

Manipulative Transaction Type	***Transaction Patterns***
Wash sales	• Acting together of persons who will benefit from the rise or fall of the security price.
Matched orders	• Giving buy and sell orders at the same time or at close time intervals continuously during the session. • Matching of orders placed by himself (from buyer position to seller position or from seller position to buyer position). • Execution of transactions that do not cause any change in the ownership of the stock. • These orders and transactions occupy an important place in the total transactions of the person and in the total transaction volume.
Runs	• Transactions carried out by those who concentrate on securities cause significant changes in the price and value of securities.
Marking the open	• Performing transactions intensively at the beginning of the session.
Marking the close	• Performing transactions intensively at the end of the session.

Manipulative Transaction Type	***Transaction Patterns***
High frequency	• Orders placed and executed transactions coincide in a very short part of the session. • It causes an adverse change in the price of the stock.
Pump and dump	• Buying by one or more people acting together to artificially raise the price of a security. • Selling by one or more people acting together to artificially lower the price of a security.
Capping/pegging	• Conducting transactions to sell the stock in order to keep the market price below a certain level (capping) or to buy it in order to keep it above a certain level (pegging).
Spoofing	• Giving orders to change the pending best buy and best sell price in the stock's queue. • Giving buy and sell orders to change the composition of the stock's buy and sell orders. • Cancellation of these orders before they are executed.

2.3 Machine Learning and Classification

Today, with the rapid changes and advances in computer technology, millions of items of data are produced per second by real and legal persons. At the same time, with the development of technology, storing and accessing data have become both easier and less costly. Data have started to be considered as valuable as financing, and giant companies that provide services in the electronic environment do not charge their users for any service they provide. Behind this free provision lies the desire to accumulate a large amount of data. Therefore, these companies can make strategic decisions and forward-looking predictions if they expand their customer portfolios and gain more data. It is not possible to manually analyze the huge amount of data and make predictions for the future, and so machine learning, based on the philosophy of learning from data, comes to the fore in analyzing data. In addition, machine learning methods are used to classify big data and predict the classes of new data according to the classified data. Therefore, classification and machine learning approaches are closely related.

2.3.1 Machine Learning

The terms machine learning and artificial intelligence were first coined in 1950 by Turing in his article "*Computing Machinery and Intelligence*", in which he questioned whether machines can think. Many researchers and academics have developed areas that can bring different learning to machines for this problem [101]. Although the terms artificial intelligence and machine learning are often used interchangeably, artificial intelligence is actually a system that has the ability to make choices. Its main purpose is to ensure success, while machine learning, in contrast, finds the most useful result, in other words, reveals the intended value. Even if it is used to find the determined benefit to be maximized, machine learning methods do not take action on their own. Machine learning methods are thus regarded as part of artificial intelligence [102].

Machine learning allows computers to learn by creating mathematical models with different algorithms from data (numerical, text, visual, etc.). In other words, it is the use of statistical methods and the computing power of a computer to extract complex patterns from certain data and make rational decisions [103]. It is based on the development of computer algorithms transforming data into understandable actions [104]. The area is built on three basic pillars: accessible data, statistical methods, and developing fast computational power [105]. Patgiri et al. (2020) define machine learning as logical analysis that can work on large datasets [106]. Mitchell defines machine learning as follows: the success of performing a task increases as the amount of experience a computer program gains while performing it increases [107]. According to Ogucu, machine learning is an artificial intelligence field that allows the system to create a model by learning from its past experiences and make a prediction about the situations it will encounter in the future. It can be ensured that the computer can make decisions and produce solutions to related problems by using the models they have created with the help of a model developed using the data at hand and the new data they encounter in the future [108].

Figure 2.1 shows the working scheme of machine learning, where x is an input vector as data and y is an output that is appropriate according to the input. In this system, the y = f(x) function is a model whose definition is prepared according to certain situations. With the x inputs and y outputs defined for the system, machine learning creates a model to make the most accurate inferences. From the model created using the input and output values, z output values are obtained according to the inputs to be tested. These output values are compared with the actual output values [109].

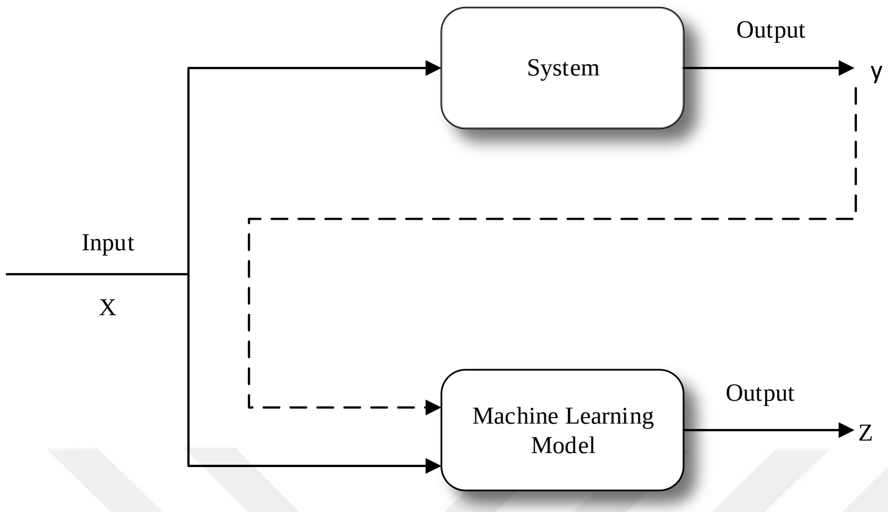


Figure 2.1 Machine learning working schema

The main difference between machine learning and other algorithms is that the machine can adapt to the new situation by using past experience or sample data in an environment where the programmer does not have to predict every situation and does not give all the results [110]. A machine learning algorithm finds the parameter values that will optimize the predetermined performance metrics to fulfill the determined purpose [111].

The fact that big data can be obtained and created more easily with today's systems makes the hypothesis generation and testing processes quicker than with traditional methods. To prove a mathematical theorem by traditional methods, the auxiliary propositions must be proven meticulously and this requires a deep mathematical background. Machine learning methods also enable this [112]. Machine learning has been applied in various areas in recent years, and there has been an increase in academic studies using it in bioinformatics, biochemistry, medicine, meteorology, economic sciences, robotics, aquaculture, food safety, and climatology.

2.3.2 Machine Learning Methods

The situation of the problem and the dataset are considered while choosing a machine learning approach. There are three types of machine learning: supervised, unsupervised, and semi-supervised.

2.3.2.1 Supervised Learning

Supervised learning includes the training process. There are training data with input and output information. It is aimed to get meaningful results by introducing these training data into the system by the person who applies it. During the learning process, known data with input and output information are introduced into the system as input. After the training process with these inputs and results given to the machine, predictions are made for data that have never been shown before using this learned information [113], [114]. In the evaluation of a problem, the data are divided into two: training and testing data. With supervised learning, a random selection is made from the available data for problem solving and, according to the selection, some training data are selected as the other part of the test data. A learning model is created by selecting and applying an algorithm that produces the most appropriate solutions from supervised learning methods for the part selected as the training data. This model is only adjusted for the training data. Next, the test data in the other part are used with the obtained model to measure the accuracy of the system's predictions. According to this accuracy, the system is also applied to the new test data [115].

In supervised learning, the problem is considered a classification problem and it is used for predictions and recognition involving the test set by using the model created by the trained system [116]. Supervised learning models are generally divided into regression and classification models depending on whether the dependent variable is continuous or categorical [117]. However, some resources supervised learning models, i.e., regression (prediction), classification (mathematical), hierarchical, layered etc., separate it into different models. Supervised learning is generally used in classification problems [118], [119]. Classification methods will be used in this thesis. The most frequently used supervised learning algorithms are decision tree classifier (DTC), logistic regression (LR), k-nearest neighborhood (KNN), random forest (RF), naïve Bayes (NB), support vector machine (SVM), and artificial neural networks (ANN).

2.3.2.2 Unsupervised Learning

In unsupervised learning, unlike in supervised learning, the system does not need a supervisor to perform the learning process. Only inputs are given to the system regardless of outputs. The algorithms are used to create a model that will enable the relationships between the samples entered into the system to be found [112]. The unsupervised system does not have any information about the output and, as the name suggests, it consists of observations with input variables without any controller. Only observations with input variables are given to the system [120]. The system learns by itself without using a controller and tries to find the relationship between the input variables. The aim is to collect samples with similar characteristics among the observations in the same cluster [121]. Unsupervised learning is used especially for clustering and dimensional reduction of relevant data in the big data structure [122]. After examining the product purchasing behavior of customers in a market, the assigning of those with similar shopping habits to the same group, updating of the stocks of their products, and arranging of the shelves in accordance with them are examples of unsupervised learning [111].

2.3.2.3 Semi-supervised Learning

This has partially been created to include the features of supervised learning and unsupervised learning. Most of the training data are unlabeled as in unsupervised learning. However, some are labeled. In order to make the best inference, the system tries to understand the structure by making predictions using all labeled and unlabeled data [113]. The amount of unlabeled data is usually greater than the amount of labeled data. Since it is difficult to label all unlabeled data, semi-supervised learning can be considered a practical solution [123].

2.3.3 Classification

Classification is the process of finding the classes of unlabeled data, that is, the unknown classes when the classes of the data we have are known [124]. While each element produced as output during classification is named a class, the algorithm used to solve the classification problem is called a classifier [125]. The purpose of classification is to ensure that the models created using data with certain classes have high accuracy on samples whose classes are not known and which have not been compared before by the system [126]. The algorithms used classify the data for the problem in accordance with their attributes. It is aimed to predict which class the new data will belong to when trained by separating them according to information such as type, condition, and class [127]. Classification methods are used in many problems such as credit approval, industrial quality controls, disease diagnosis, picture recognition, weather forecasting, separating necessary or unnecessary e-mails, and separating data according to languages [128]. The aim is to create a model to predict future customer behavior by classifying database records into predefined classes according to certain criteria [129]. Classification consists of two steps: the learning step, in which the model is created, and the classification step, in which the model is used to predict the unknown classes of the given data [130].

2.3.3.1 Classification with Machine Learning

In machine learning terminology, classification is a type of supervised learning [131] and is one of the best known. It is used to add a data group whose classes are not known to a data group with certain classes [132]. There are many algorithms used for classification among machine learning algorithms, the best known of which NB, LR, SVM, KNN, DTC, RF, and ANN. A classifier, i.e., a classification model, is a mapping from samples to predicted classes [125]. The classifying of data with machine learning consists of two steps. The first is to introduce a model suitable for the datasets. The model in question is implemented using the attributes of the records in the database. In order to establish this model, some of the data are randomly selected and used as training data. The rest are used as test data. Afterwards, a classification model is obtained by applying an algorithm to the training data. In the second step, classification rules are determined from the test data. These rules are tested this time by applying them to the test data. If the accuracy obtained by testing is accepted, this model is applied to other data [133].

2.3.3.1.1 Naïve Bayes

Classifiers based on Bayesian methods in machine learning calculate an observed probability for each class based on variable values, namely features, by using the training data and then use these observed probabilities in estimating the classes in the labeled test data. Thomas Bayes, an 18th century mathematician, came up with Bayes' theorem, which laid the foundation for Bayesian methods. The basis of this classification is Bayes' theorem, in which the data of the classified sample objects are used and the probability of an object of unknown class belonging to the determined classes is calculated. In other words, in this algorithm, in which the existence of an attribute in a particular class is assumed not to be related to the existence of any other attribute, the probability about which class new data belongs to is calculated by using the already classified data. This is a clear, understandable, and specific method based on natural tools for solving complex problems and is used especially when there are independent attributes and the database is important. This model is used in many different areas such as sentiment analysis, recommendation systems, spam filtering, text mining, and diagnosis of diseases.

NB is widely used because it is an easy, fast to implement, high classification performance algorithm. The NB algorithm can even outperform sophisticated classification methods in some cases. When the algorithm encounters data whose class is unknown, it calculates this value for each class and includes the data in the class with the highest probability value [134]. Bayes theorem;

$$P(A\backslash B) = \frac{P(B\backslash A)\,P(A)}{P(B)} \qquad (2.1)$$

According to Bayes' theorem, let A and B be events. P(A) is the probability of event A happening, and P(B) is the probability of event B happening.

- P(A), P(B) : probabilities of events A and B,
- P(A\B) : The probability of event A happening when event B occurs,
- P(B\A) : The probability of event B happening when event A occurs.

Bayesian classifiers have the advantages of speed, convenience, and high accuracy, as well as some disadvantages. For example, these classifiers do not consider relationships between features (fever, cough, etc.). In real life, some features are related to each other. Another disadvantage is that all features are considered equally important. Moreover, they do not work on datasets with a continuous class label [134].

2.3.3.1.2 Logistic Regression

LR is one of the most popular classification algorithms in supervised learning. It is frequently used in medical and social sciences research due to its possibilities and ease of use [135]. LR estimates the output of a categorical dependent variable. Therefore, the result must be a categorical value. “*Yes*” or “*No*” can be *0* or *1*, “*True*” or “*False*”, but instead of giving the exact value as *0* and *1*, it gives probability values between *0* and *1*. LR is very similar to linear regression except for the type of output (dependent variable) to be predicted. While linear regression is used to solve regression problems, LR is used to solve classification problems [136].

One of the most important differences between LR and linear regression is that the dependent variable in LR is categorical [137]. LR analysis gives successful results in classification studies [138]. There are three important differences between LR and linear regression analysis methods:

1. While the dependent variable to be estimated in linear regression takes continuous values, in LR the dependent variable takes discrete values.
2. While attempts are made to estimate the value of the dependent variable in linear regression, in LR the aim is to estimate the probability of realization of one of the values that the dependent variable can take.
3. In linear regression, the condition for the independent variable to show multiple normal distribution is sought, while in LR analysis there are no prerequisites for the distribution of the variables [139].

The value of the LR must be between *0* and *1* and cannot go beyond this limit, thus creating a curve shaped like an "*S*", called the sigmoid function or the logistic function. The sigmoid function is used to map predicted values to probabilities and in Figure 2.2 it is seen that the it maps any real value to another value in the range of *0* and *1*. In LR, the concept of threshold value is used, which defines the probability of *0* or *1*. Values above the threshold value are close to *1* and those below it are close to *0*.

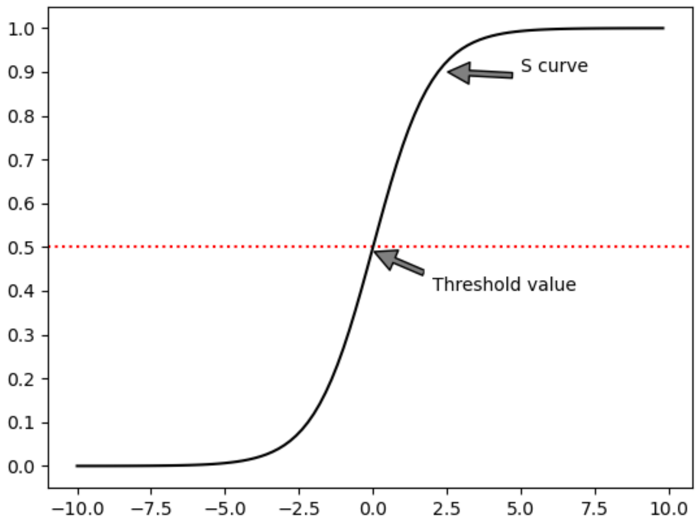


Figure 2.2 Logistics function curve (Sigmoid) [136]

2.3.3.1.3 Support Vector Machine

SVM is a powerful supervised learning algorithm devised by Cortes and Vapnik in 1995 for solving high-dimensional problems. It is mostly used in classification and regression processes of binary or multi-class data [140]. The algorithm finds the most appropriate function to classify examples belonging to two different classes through learning [115]. This method is based on statistical learning theory. With this algorithm, a hyperplane that is equidistant from the classes is created and the classes are separated thanks to this plane [141]. The learning curves closest to the plane used for classifying are called support vectors. The two closest border lines to this created hyperplane form and thus the hyperplane remains in the middle [134]. Figure 2.3 shows two class hyperplanes and their support vectors separated using a SVM.

When each input represented by x_i belonging to the D feature in the SVM is defined as belonging to one of the classes $y_i = -1$ or $y_i = +1$, all of the inputs are shown in equation (2.2), and this helps to find the linear hyperplane that will optimally separate the different classes from each other.

$$\{x_i, y_i\} | \ i = 1, \ldots n, y_i \in R^D \tag{2.2}$$

In equation (2.3), the weight vector is denoted by w and the constant value b.

$$wx + b \tag{2.3}$$

These boundary lines created are called margins. With the SVM, classes are separated by changing the position of these border lines and increasing or decreasing their widths. Although the SVM is mostly used to classify data with binary classes, it is sometimes used to classify data with more classes.

As seen in Figure 2.3, H separating the two classes is w for the hyperplane. $x + b = 0$, while the support vectors H_1 and H_2 are and $. x + b = -1$ and $. x + b = 1$.

The margin width, which represents the distance between the support vectors, is expressed as $2/w$. The smaller the w value, the more the margin width will increase; otherwise it will decrease. Since the purpose of the SVM algorithm is to obtain the largest margin value, it tries to minimize the w value [142].

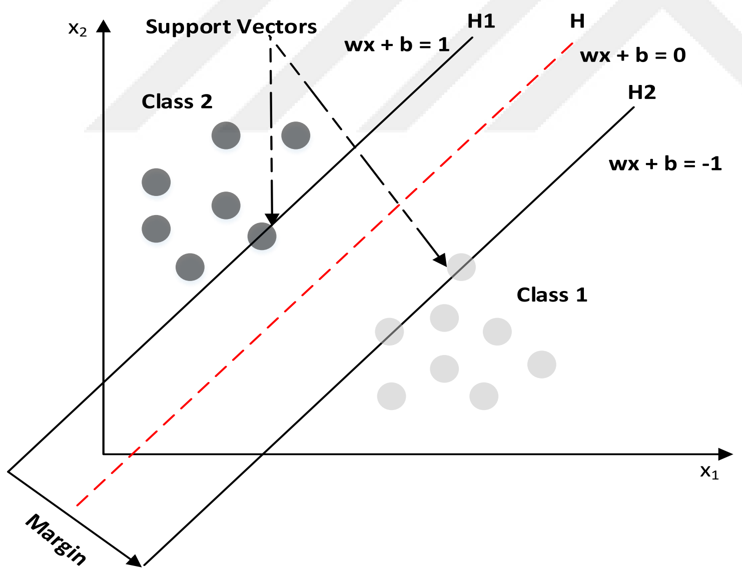


Figure 2.3 Hyperplane, margin, and support vectors in a two-class dataset

This method is generally used in cases in which the data can be linearly decomposed, but it can also be applied when there are data that cannot be separated linearly, since it is possible to make the data linearly separable by means of kernel functions such as on a sigmoid, polynomial, linear, and radial basis [143]. In other words, a larger size is applied for nonlinear datasets using kernel functions [144].

2.3.3.1.4 K-Nearest Neighborhood

The KNN algorithm was first introduced by Fix and Hodges in 1951. This algorithm has been frequently used to solve non-parametric classification and regression problems, and it is based on distance calculation [145]. It is based on the logic that the closest data to each other belong to the same class. The aim is to classify the new incoming data by making use of the previously classified data. It is a learning-based algorithm that applies the model learned from the training data. Data whose class is not known are called 'test', while previously classified data are called '*education*' [146].

In the algorithm, k is an integer expressing the number of neighbors, and the distance of the new sample whose class is to be determined from the features to be used in the classification to k of the samples in the database, whose class is already known, is checked. The K number is also a parameter that indicates how many classes the data will be divided into and it is a number that must be entered into the algorithm before classification. For example, when determining the class of an object whose class is unknown for $k = 3$, the 3 objects with known class closest to this object are taken, and the neighbor selection of the KNN algorithm is shown in Figure 2.4. In other words, in order to determine the class of the incoming data, its distance from other data is calculated. To calculate this distance, the Manhattan, Minkowski, Jaccard, or Euclidean distance is used. Since $k = 3$ is chosen in our example, the closest 3 of the measured distances are selected. If 2 of them are in the X class and one is in the Y class, the class of the new incoming data is X.

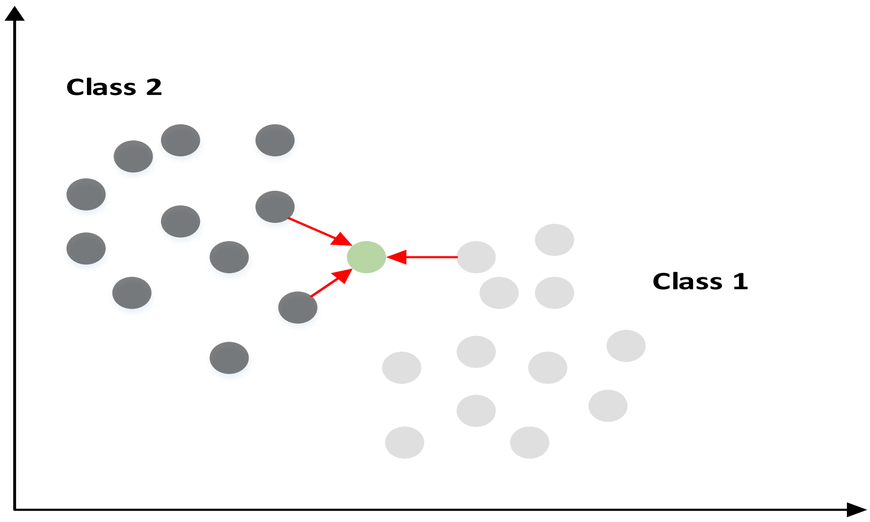


Figure 2.4 Selection of the three closest neighbors to the sample to be classified using the KNN algorithm

Different methods produce different results for distances. Moreover, although the most commonly used formula is Euclidean, it may be necessary to use different distance measurements for different types of data.

- **Euclidean distance:** It determines the straight line distance, that is, the distance between two points, in a plane with many points. It is the most widely used method.

$$d(p,q) = \sqrt{\sum_{j=1}^{n} (p_j - q_j)^2} \tag{2.4}$$

- **Manhattan distance:** It is the distance between a pair of points measured at 90 degree angles along the axes.

$$d(p,q) = \sum_{j=1}^{n} |p_j - q_j| \tag{2.5}$$

- **Minkowski distance:** It is used to calculate the distance depending on a certain number of variables. If 2 is written instead of x in the equation below, it is equal to Euclidean distance, and if 1 is written, it is equal to Manhattan distance.

$$d(p,q) = \left(\sum_{j=1}^{n} |p_j - q_j|^x\right)^{1/x} \tag{2.6}$$

2.3.3.1.5 Decision Tree Classifier

DTC is one of the most commonly used supervised machine learning algorithms in classification problems and it is also used in regression problems. It is quite easy to configure and understand compared to other classification algorithms. DTCs are frequently used because they are reliable and simple, have low computational complexity, and can work with high-dimensional datasets with both continuous and discrete variables [134]. In the literature, the different decision tree algorithms depending on the tree formation methods and data diversity [147] include AID, CHAID, CART, ID3, C4.5, C5.0, MARS, E-CHAID, SLIQ, SPRINT, and QUEST. They differ from each other by the paths followed in the selection of root, node, and branching criteria. The first decision tree-based algorithm is the AID algorithm developed by Morgan and Sonquist in the early 1970s. It is based on the best estimation and finding the independent variable with the strongest correlation. The CHAID algorithm, which is based on statistics, was developed by G. V. Kass in 1980 for classification and regression processes. A decision tree algorithm named ID3 was developed by J. Ross Quinlan in 1986. He released C4.5, an advanced version, and the C5.0 algorithm, which is faster, uses less memory, and creates more precise rules than C4.5, to eliminate the deficiencies of the ID3 algorithm in 1993 [148].

Decision trees have decision nodes and leaf nodes. While decision nodes are used to make decisions, classifications, or predictions in the dataset, leaf nodes store the decisions made [149]. The tree consists of roots, branches, and leaves. The root and internal nodes represent properties, the branches represent values that properties can have, and each leaf node is considered a class. Each branch is connected to the upper root. The number of nodes is important for classification performance, because large trees (with a large number of nodes) work slowly, while small trees are not successful in classification despite working fast [150]. Decision trees are among the most popular machine learning algorithms considering their intelligibility and simplicity, and since it is a type of supervised learning, it works in two stages: creation of a model from the data we have, which we call learning, and determination of the class by testing the test data on the model.

In the decision tree given in Figure 2.5, the decision about whether the football match will be played is based on the weather information. If the weather is cloudy, the match will be played; if it is rainy, the decision is made according to the wind; and if the weather is sunny, according to the air temperature.

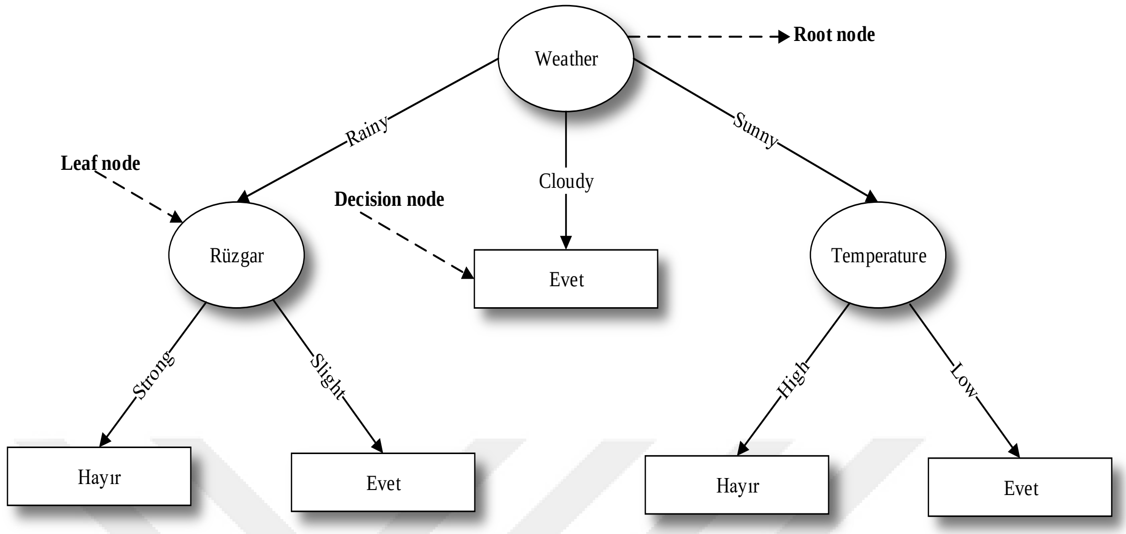


Figure 2.5 Decision tree example

2.3.3.1.6 Random Forest

RF, proposed by Breiman and Cutler, is a machine learning algorithm that aggregates predictions from many decision trees into different subsets of data. It is one of the most widely used ensemble classification algorithms and consists of multiple trees using randomly generated samples from existing situations. To classify a sample, an input vector is given to each tree in the forest and a result is obtained for each tree [142]. Ensemble classification methods try to classify new data by using the results produced by more than one classifier [151]. The RF algorithm chooses the class that gets the most votes from the given results [152].

This method is very successful in presenting correct generalizations and giving close results in estimations, because it includes optimized features of ensemble methods and random sampling. In Figure 2.6, the components of the RF component method consisting of n decision trees are shown. Accordingly, a classification result comes from each tree for the dataset to be classified, and each tree "*votes*" for the classification estimate in question. The class with the most votes out of all trees in the forest is the final result.

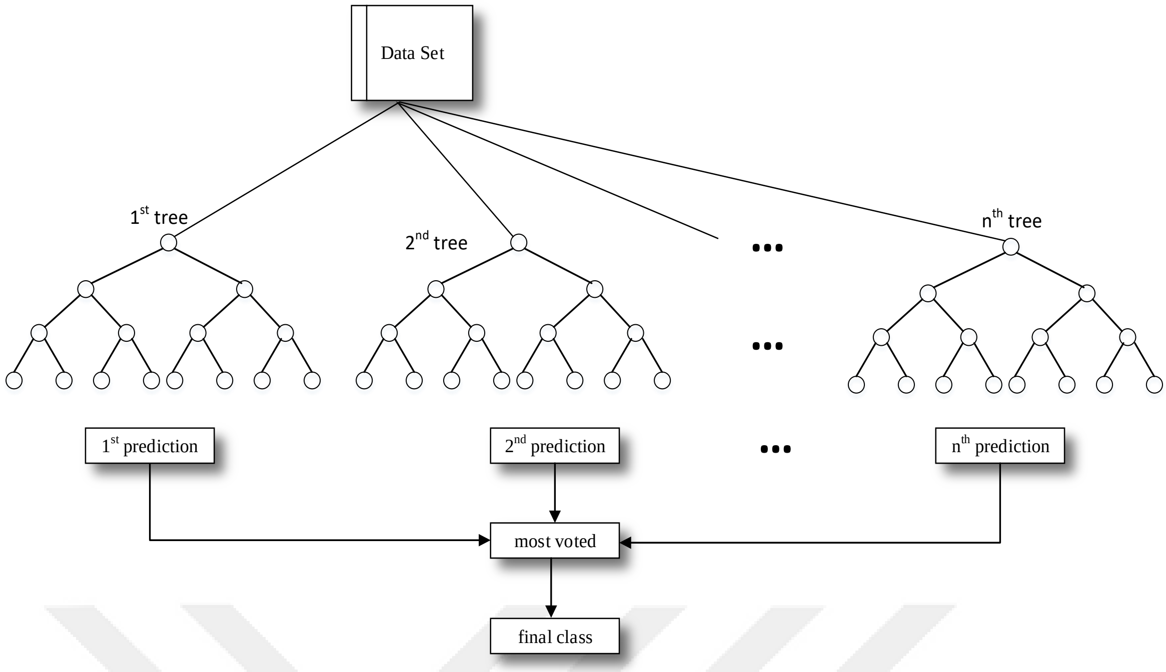


Figure 2.6 Random forest method components

This method can give better results in unbalanced datasets. Therefore, it can solve memorization situations. It works more effectively in large datasets and gives better results than other algorithms in datasets with missing observations. It makes the model more generalizable. In addition, as the number of trees increases, it becomes very difficult to examine RF trees [153].

2.3.3.1.7 Artificial Neural Network

The first information about ANNs was presented by William James in 1890. The first ANN model was developed by Warren McCulloch and Walter Pitts in 1943 and expresses the mathematical model of a nerve cell. After a pause in ANN research, it started to gain momentum again in the 1980s. In the same years, Hopfield showed that ANNs can be used to solve problems in many areas difficult to solve using computer programming. With the developments in computer technologies, ANNs have started to become more useful by being integrated into many systems used in daily life since the 1990s. The ANN obtains information by learning. It can produce new information from the information it has learned. Thanks to its ability to explore, it can respond to external influences similar to human behavior [153]. Generally, models created with ANN are used in time series analysis, optimization, classification, association or nonlinear system modeling [154].

The ANN is an improved information processing technique that imitates the behavior of the human brain and nervous system [155]. In other words, it is a computer system that takes the situations people encounter in real life as an example, creates situations that they do not encounter by multiplying these examples, and provides new capabilities with the situations it creates [156]. In addition, ANN is a machine learning method that can produce results by interpreting the situations that it has never encountered or the situations that it has learned incompletely [153]. The concept of learning for biological systems is through the connections between nerve cells. People update these links using outside information. The concept of learning in ANNs is achieved through the connection of events with results [157].

ANN is a concept produced from the features of the human brain; therefore, examining the structure and working logic of neurons provides a better understanding of ANNs. Biological neural networks are made up of millions of interconnected neurons. Neurons are the building blocks of the nervous system and they consist of dendrites, soma (cell body), and axons. Dendrites in neurons are structures that receive and carry signals and provide neural transmission. Axons are responsible for transmitting these signals to other neurons. The point where axons and dendrites meet is called the synapse. The general structure of a neuron is shown in Figure 2.7.

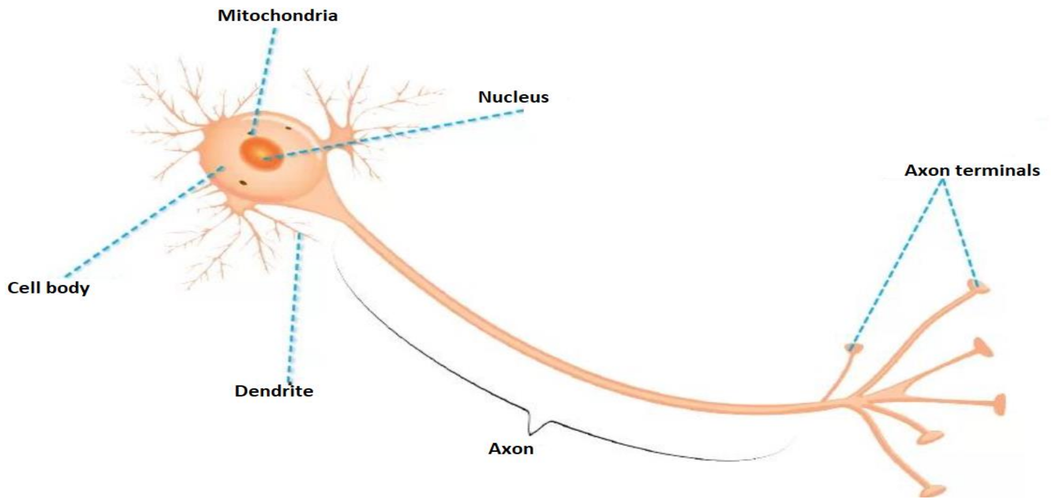


Figure 2.7 General structure of the neuron [158]

As in biological neural networks, the basic processing elements in ANNs are called neurons, artificial nerve cells, or nodes. ANNs consist of computational units called artificial neurons and the connections between these neurons. Each neuron in the network takes several input values to produce an output value. In these networks, each input of neurons, namely connections, has coefficients called weights [159].

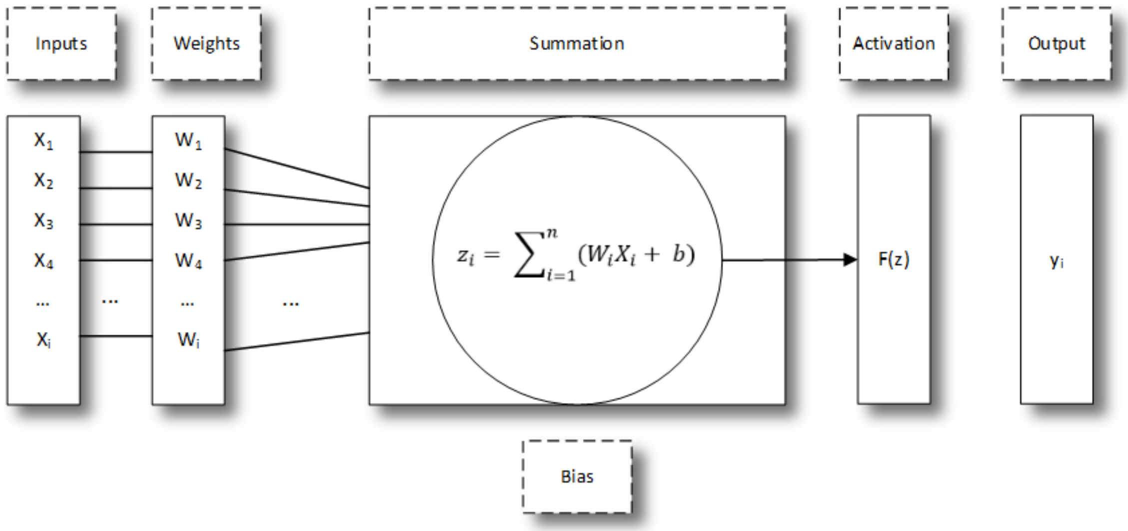


Figure 2.8 Artificial neuron

The dendrites in the biological nerve cell correspond to the activation functions in the ANN, the cell bodies to aggregation functions, the axon terminals to the output layer, and the synapses to the weights. As seen in Figure 2.8, the simplest artificial neuron consists of 5 main components: inputs, weights, summation function, activation function, and output.

The equivalents of the biological nervous system elements in the ANN are presented in Table 2.9. The biological nervous system is divided into parts and each element is given its equivalent in the ANN.

Table 2.9 Equivalents of biological nervous system elements in artificial neural network

Biological Nervous System	***Artificial Neural Networks***
Neuron	Processor element
Dendrite	Summation function
Cell body	Activation function
Axons	Artificial neuron output
Synapses	Weights

When we examine Figure 2.8 closely, the inputs are denoted by the symbol X_i and each of these inputs is multiplied by the weight (W_i) as shown in equation (2.7) and summed with the threshold value (b) of the activation function. The value generated as a consequence of the sum function is sent via a linear or nonlinear differentiable activation function to determine the processing element's output (y_i).

$$z_i = \sum_{i=1}^{n}(W_i X_i + b) \tag{2.7}$$

As seen in Figure 2.8, an artificial neuron consists of five main components: inputs, weights, summation (combination) function, activation function, and output [153].

- **Inputs:** Information from the outside world or the preceding layer to an artificial neuron is sent as input to the artificial neuron.
- **Weights:** The relevance of the information entering an artificial cell and its effect on the neuron is indicated by the weights. The weights (w_1, w_2, w_3, …, w_i) are coefficients that determine the effect of the inputs received by the artificial nerve on the nerve.
- **Summation Function:** This function calculates the net input by multiplying the input values by the weight values and adding them to the threshold value (bias). The sum function is shown in equation (6.9), and the output value y_i is calculated by passing the value obtained as a result of this function through a linear or nonlinear differentiable activation function. This is shown in equation (2.8).

$$y_i = \mathrm{f}(z_i) = f\left(\sum_{i=1}^{n}(W_i X_i + b)\right) \tag{2.8}$$

- **Activation Function:** The activation function creates net input values in the desired range and transmits them to the output. Depending on the activation function used, the output value is usually between [-1,1] or [0,1] and is usually a nonlinear function. The use of nonlinear activation functions enabled the application of ANNs to complex and very different problems. The most commonly used activation functions, sigmoid, are shown in equation (2.9) and hyperbolic tangent is shown in equation (2.10). The most widely used of these functions is the sigmoid function, which converts the inputs it receives into a real number in the range of 0 and 1. Therefore, it can be easily used in ANNs that should give positive results. Although the hyperbolic tangent function is similar to the sigmoid function, the outputs of this function are between -1 and 1.

$$f(z_i) = \frac{1}{1+ e^{-z_i}} \tag{2.9}$$

$$f(z_i) = \frac{e^{z_i} - e^{-z_i}}{e^{z_i} + e^{-z_i}} \tag{2.10}$$

- **Output:** It is the value obtained by applying the activation function.

ANNs are composed of many artificial nerve cells connected to each other. Today, many ANN models suitable for use in different fields and for specific purposes have been developed, and the most common use among these network structures is multi-layer feed-forward ANNs (multilayer perceptron - (MLP)). In MLP networks, neurons are organized in layers and there are basically three layers: input, output, and hidden. The input layer is the first and provides information about the problem to be solved to the neural network. The neurons of the hidden layer are not connected to the outside world. They simply receive and send signals to and from the input layer. The output layer is the final layer, and it is responsible for transmitting data to the outside world. The hidden layers are positioned between the input layer and the output layer.

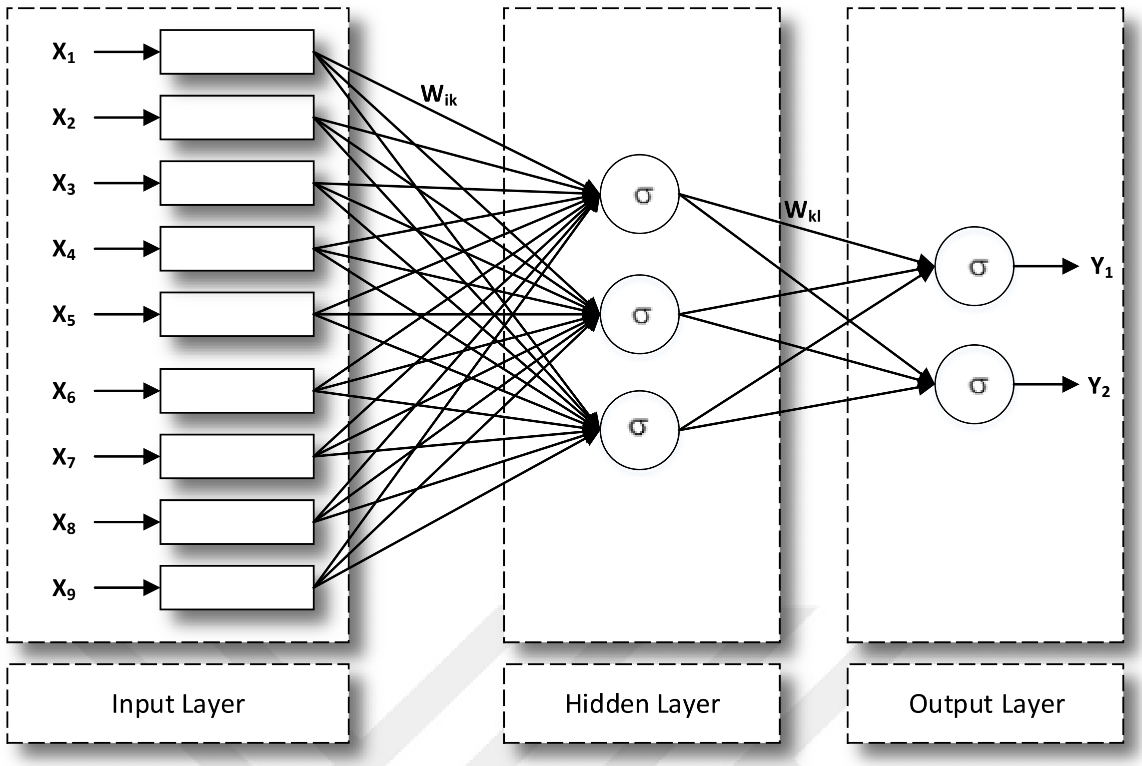


Figure 2.9 Multilayer feed-forward neural network

Figure 2.9 below shows a simple neural network model with input and output layers and one hidden layer. The neural network in this example consists of an input layer with 9 dependent variables, a hidden layer with 3 neurons, and an output layer with 2 categories. The σ sign indicates that the sigmoid function is used as the activation function. In addition, W_{ik} shows the weights between the input layer and the hidden layer, while W_{kl} shows the weights between the hidden layer and the output layer.

There is no restriction on how many hidden layers there will be in the neural network or how many neurons there will be in each layer. Depending on the designed architecture, it is possible to increase the number of hidden layers and neurons in the hidden layers. In this case, the training time of the model will increase since the complexity of ANNs will increase as the number of layers and neurons increases. The neural network model can be improved by finding the number of hidden layers and the number of neurons by trial and error.

2.3.3.2 Classification Performance Measurement Criteria

Different performance criteria are used to examine the success rate of models created using classification algorithms, that is, which model can achieve more accurate results. Although there are many performance metrics, most used for classification problems are based on the confusion matrix. Firstly, the confusion matrix is introduced and then information about the performance criteria in question is given.

2.3.3.2.1 Editing Dataset

When classifying with machine learning, the dataset is divided into three parts: training, validation, and testing. The training set is used to train the model; the test set to verify whether the model established during the training phase works efficiently with another labeled dataset, in other words, to measure model performance; and the validation set to improve the models to perform efficiently on data not encountered before. It is used to train and test under different conditions. The data in the test set are separated from those in the validation set with the feature of not being seen by the system before [160]. In scientific studies, for the training, validation, and test sets, the division rates are 70% - 15% - 15% and 80% - 10% - 10%.

2.3.3.2.2 Evaluation of Model Performance

Model performance measures are commonly regarded as the success of a model developed for classification, yet this figure is insufficient to determine the model's quality. A confusion matrix is frequently used to reveal the performance of the created model [161].

2.3.3.2.3 Model Performance Evaluation Criteria

In the error matrix, there are real situations and situations that reflect the predictions of classification algorithms. Therefore, the error matrix contains the correct and incorrectly predicted values, allowing the performance of classification algorithms to be evaluated, and the rows of the matrix contain the actual class values and the columns the predicted class values. The confusion matrix in question is shown in Table 2.10.

Table 2.10 Confusion Matrix

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

- **TP:** It represents the positive data that the algorithm predicts correctly. (*something that actually exists*)
- **TN:** It represents the negative data that the algorithm predicts correctly. (*something that does not actually exist*)
- **FP:** It indicates that an item of data that is actually negative is predicted as positive by the algorithm. That is, there is a misclassification of the negative sample. (*you have something that does not really exist*)
- **FN:** It indicates that an item of data that is actually positive is predicted as negative by the algorithm. That is, there is a misclassification of the positive sample. (*nothing that actually exists*)

In order for the classification performance to be high, it is desirable to increase the TP and TN regions of the confusion matrix and to decrease the FP and FN regions. Accuracy, precision, sensitivity, recall, and f1 score metrics created to measure the performance of classification algorithms using the error matrix are given in the following sections.

- **Accuracy:** The accuracy rate is obtained by dividing the correctly classified samples by the total number of samples. The accuracy rate formula is given in (2.11) and this rate is used to evaluate the performance of classification algorithms. However, it is not appropriate to evaluate the performance of classification algorithms based on accuracy only in cases in which the dataset is unevenly distributed. For example, in a model developed to predict whether a stock has been manipulated, if the number of days on which manipulation occurs in the sample dataset is only a small part of the dataset, such as 5%, even if a random guess is made it is possible to predict with 95% success that there is no manipulation in the stock in question. In such cases, high accuracy does not mean high performance.

$$accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{2.11}$$

- **Precision:** It is used to measure the algorithm's success in classifying negative samples. Therefore, the significance ratio is obtained by dividing the true negative value by the total number of negative samples, and the significance ratio formula is given in (2.12).

$$precision = \frac{\mathrm{TN}}{\mathrm{TN+\ FP}} \tag{2.12}$$

- **Sensitivity:** It measures the rate of correct prediction of the total positive predicted samples. It is obtained by dividing the correctly classified positive samples by the total number of positively predicted samples, and the sensitivity calculation formula is given in (2.13).

$$sensitivity = \frac{\mathrm{TP}}{\mathrm{TP+\ FP}} \tag{2.13}$$

- **Recall:** It measures the efficiency of the algorithm in predicting positive samples. The sensitivity is calculated by dividing the correctly predicted positive samples by the total positive samples, and the sensitivity calculation formula is given in (2.14).

$$recall = \frac{\mathrm{TP}}{\mathrm{TP+\ FN}} \tag{2.14}$$

- **F1 Score:** Generally, when the sensitivity rate is increased, the recall rate decreases, that is, the sensitivity rate is waived in order to increase the recall rate. Therefore, there is a balance between the two odds, and in order to find this balance, the f1 score, in which the sensitivity ratio and recall ratio are comprehensively addressed, is used [162]. In this context, the f1 score is calculated as the harmonic mean of sensitivity and recall metrics, and the f1 score calculation formula is given in (2.15).

$$f1 = 2 * \frac{sensitivity * \mathrm{recall}}{sensitivity + \mathrm{recall}} \tag{2.15}$$

2.4 Data Science Methodology

Machine learning is closely related to data mining. The cross-industry standard process model for data mining (CRISP-DM), which Shearer introduced to extract valuable information from large amounts of data with data mining, is among the most widely accepted [163]. The CRISP-DM process model is also used in different fields of study, and this process model is followed in problems solved using machine learning. In our study, the CRISP-DM process model was followed in the same way. The model is presented in Figure 2.10. It is very important to perform the steps mentioned here in order to successfully solve the problem within the targeted time.

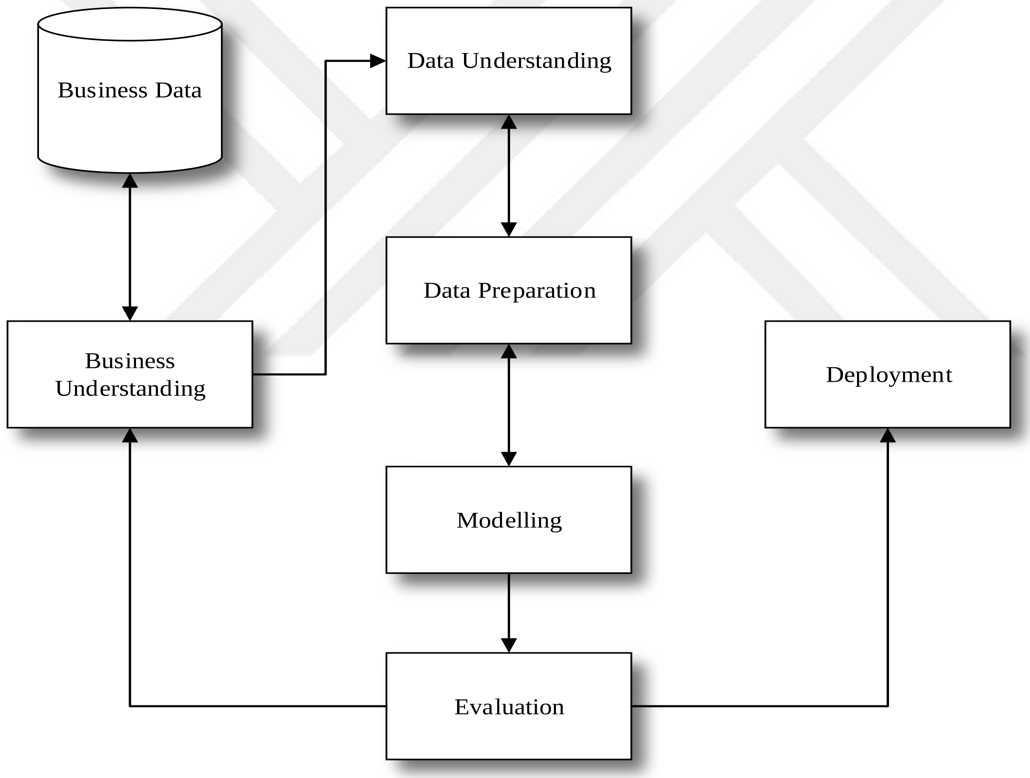


Figure 2.10 Phases of CRISP – DM diagram [163].

Leading data mining users such as DaimlerChrysler AG, SPSS, NCR, and OHRA need a standardized approach to help industries transform their problems into data mining tasks, recommend appropriate data transformations and data mining techniques, evaluate the effectiveness of results, and report experiences. It was developed by a consortium of providers [164]. The model was developed on the basis of knowledge discovery methods in databases followed in previous studies. Many widely used data mining tools such as Weka, RapidMiner, IBM SPSS Modeler, and StatSoft Statistica are based on this process [165].

The CRISP-DM process model attempts to reduce the cost, reliability, reproducibility, manageability, and to increase the speed of large data mining operations [164]. It consists of six stages: defining the problem, understanding the data, preparing the data, modeling, evaluating and selecting the model, and putting the model into practice [166].

2.4.1 Business Understanding

Understanding what the problem involves is the first and most important step in data mining. The focus is on understanding the goals and requirements of the project in accordance with the business's point of view, and then a pioneering plan is prepared in order to achieve the goals by defining the goals and requirements as a data mining problem [167]. These include a problem concerning agriculture, e.g., estimating the wheat harvest of Turkey in the coming year; a problem for a textile company, e.g., determining which customers have complaints in advance; a problem in the telecom field, e.g., predicting which customer will switch to a competitor; or a problem in the marketing field, e.g., which ad to show to which customer.

The first requirement for success in data mining studies is to clearly define the purpose of the application. The purpose should be problem-focused and expressed in a clear and concise language, and how to measure the success levels of the results obtained should be defined [168]. In addition, estimations regarding the costs incurred in the case of wrong estimations and the benefits to be gained from correct estimations should be included at this stage [169].

2.4.2 Data Understanding

The data understanding stage, which is the next step, involves collection of data suitable for the problem. In general, it is very important to define the problem correctly before this step because a common mistake is to process all the data unnecessarily. Data can be obtained in two ways: from existing databases or by survey, discussion, or measurement. By examining the data more closely, missing, noisy, and dirty items are detected. Afterwards, additions are made to the dataset according to need. Again, compliance with the criteria of accuracy, completeness, consistency, timeliness, credibility, added value, interpretability, and accessibility is checked in order to solve problems with the data and increase their quality

[149]. There is a very close relationship between understanding the business and understanding the data, and this stage includes activities to achieve various purposes such as recognizing problems related to data quality, getting first impressions about the data, identifying interesting sub-series, or developing hypotheses to extract the information hidden in datasets [167].

A good understanding of the available data is required before the model is set up. Before analysis of the dataset, some simple descriptive statistical calculations can give a preliminary idea about the data and graphs that can be obtained from them. All these processes provide preliminary information about which analyses should be performed during the data preprocessing [149].

2.4.3 Data Preparation

After obtaining and understanding the data, the next step is to make the data usable in mining. The data preparation phase covers all the operations from the initial raw data to the final dataset to be used in modeling. This is likely to be done more than once and not in a predetermined order. It consists of collecting, valuing, combining and cleaning, selecting, and transforming [167]. There is no pre-determined pattern; some preprocesses such as data filtering, transformation, and data reduction are required. The operation to be performed differs from data to data. It is possible to both create and define the dataset to be used for modeling.

The purpose of data preparation is to create a dataset that can be an input for the data mining algorithm. Problems during the establishment of the model may require returning to this stage and reviewing the data preparation. Hence, the steps must be carried out meticulously. The minimum subset of the selected data needs to be evaluated, considering the relevant attributes and the appropriate time period [170]. These processes, also considered preparation of the data before data mining, especially aim to remove corrupt values in the database and inconsistencies between the data [171].

Data preparation usually accounts for 70–80% of the data mining process. Decisions such as not including data with some missing parts in the system or completing the missing parts of the data, and what kind of method will be followed during this completion are made at this stage. A simple data conversion, such as converting birth dates to age, or operations such as extracting the person's information such as province and district in the address field are example operations performed at this stage [172].

2.4.3.1 Data Selection

Data selection is choosing the data thought necessary for the defined problem and the data sources from which these data will be obtained. Apart from the organization's own data sources, external sources can also be used [169]. Our goal is to select a subset of all available data to work with. It is more tempting to select all data, but this is not always appropriate.

2.4.3.2 Data Cleaning

Inconsistent and incorrect data in the database are called noisy data. In some applications, the data to be analyzed do not have the desired properties. For example, incomplete data and inconsistent data created by inappropriate data may be encountered [133]. In the data cleaning phase many techniques are used, such as correctly completing the missing data and reducing the noise by determining the extreme values. Missing data can have negative effects such as reducing the precision of the analysis result, and creating complexity in the calculation of gaps due to missing values [173]. In general, there are two operations to clean the data. First of all, since each dataset may contain data with inconsistent or incorrect information, these data are removed from the dataset. The other process is to discard the data known as missing values. Different values cannot always be measured. These data should be eliminated if the missing values have no effect on the outcome. If missing data add more significance to the dataset, the missing values can be given the feature's average value, the dataset's most repeating value, or the missing values can be incorporated into the dataset through regression [121].

To clean the noise in the data:

- Records with missing values can be discarded,
- Lost values can be replaced with a fixed value,
- The average of the other data can be calculated and this value can be written instead of the missing data,
- The average is calculated using all the data of the variable and this value can be used instead of the missing value,

- Instead of all the data of the variable, only the mean of the variable of the samples belonging to a class can be calculated and used instead of the missing value,
- The missing value can be estimated by estimating according to the data and can be used instead of the missing value [115].

2.4.3.3 Data Integration

During the creation of the data warehouses where the data are stored, integration of the data under the same roof is applied to not disturb the integrity between the data obtained from different sources. Consistency is ensured by matching all available data sources according to their characteristics and aggregation levels for data integration. Data integration is applied for keeping the same data in different formats at the stage of bringing together the data recorded in more than one information source owned by the institution [171].

The increase in data capacities increases the size of data warehouses and the access of stored data. Although it increases the process of accessing data provided by different users, it can cause some problems [174]. Combining data includes activities such as bringing together multiple tables containing information about the same object and collecting them into a single table, integrating more than one record by summarizing the data, and obtaining new records, especially in the case of multiple data sources. However, every data source will contain potential problems such as missing and inaccurate values. Combining data from multiple sources will cause new quality problems such as the same term having different meanings, using different terms for the same input, differentiating units and measurements, and using different descriptors [133]. For example, in a database used by two different users, the "*IDNumber*" attribute may be registered in two different ways as *ID_No* and *IDNo*. This inaccuracy can affect all data mining processes and causes some errors.

2.4.3.4 Data Reduction

Today, as a result of increasing memory capacities and much easier access to data, redundancy is a problem. Heaps of data whose storage is necessary/unnecessary at very large scales cause inefficient data analysis studies. If databases are included in studies without this data pollution being purified, the accuracy of the results is in doubt [175]. To overcome these problems, various data reduction methods have been developed [130]. They are applied to obtain an unreduced sample of the dataset with a smaller volume. Thus, more effective results can be obtained by applying data mining techniques to the reduced data set obtained [176]. For example, let's say we have an attribute that keeps students' grades. Data from source A can give students grades such as A, B+, B, C+, C, D+, D, and F. If the data are taken from source B, they include grades such as 66, 45, and 90 obtained by students. What should be done in this situation? Since they both belong to the same attribute, it would not make sense to keep these data in two separate attributes. It is imperative that this different format of data be properly reduced to a single format. By using the most widely used solution, it is necessary to find the degree equivalent of the actual grades received by the students, and to express the grades of all students in degrees and to consolidate the data from sources A and B.

2.4.3.5 Data Transformation

Data transformation involves transforming data into different values or scales in accordance with a defined function to obtain healthier results or to be compatible with the algorithms used [130]. It is the process of transforming the data into forms suitable for data mining by preserving its content according to the model to be used [177].

When transforming data into a format suitable for data mining, concatenation, normalization, and variable creation are often used. Concatenation means summarizing or combining data. For example, data can be combined to find the values of the daily total transactions regarding the transactions realized during the day in the stock market. In the variable creation method, it is aimed to create new variables by using the original variables. For example, in the present study, the trading volume of a stock is calculated by multiplying the transaction amount by the price at which the transaction takes place. Another example is that in time series models, variables change over time, and a new variable is produced by taking the differences of the values of a variable corresponding to a certain time interval and the next time interval. Such transformations are very important in terms of benefiting from knowledge discovery in data mining. In normalization, in some cases, there is a huge difference between the numerical values of the features in the datasets. In these cases, which occur in datasets, it may result that the features with high numerical value have more effect on the target variable than the features with smaller numerical values. In order to prevent this imbalance, scaling operations are applied.

2.4.3.5.1 Min – Max Scaling

Also called minimum–maximum normalization, this is the process of rescaling one or more features to a range of 0 to 1. That is, after the normalization process, the maximum value of the data contained in the relevant attribute will be 1 and the minimum value will be 0. This method is based on the principle of determining the largest and smallest numerical value in the data and transforming the others accordingly. The change relation in question is expressed as follows:

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2.16}$$

Here X* represents the transformed values, X is the observation values, X_{min} is the smallest observation value and X_{max} is the largest observation value [115].

2.4.3.5.2 Standard Scaling

This method, one of the most widely used for data transformation, involves the scaling of one or more features so that their mean values are 0 and their standard deviation values are 1. Thus, all observation units in the dataset receive values between -1 and 1. In the standardization process, it is assumed that the data have a Gaussian distribution. Although there is no obligation for the data to have this distribution, the standardization process is more effective for the data with it. In the transformation of these values, a relation is given as follows:

$$X^* = \frac{X - U}{S} \tag{2.17}$$

Here X* represents the transformed values, X represents the observation values, U represents the arithmetic mean of the observation values, and S represents the standard deviation of the observation values [115].

2.4.4 Modeling

At this stage, all the analyses such as choosing the modeling technique and adjusting the parameters of these models to optimal values are included. Finding the most suitable model for the defined problem is possible by establishing and testing as many models as possible [167]. Since there are multiple techniques that can be used in data mining, it is not possible to decide in advance which is the most appropriate. Therefore, the phase of model creation is an iterative process by applying different techniques until the model thought to be the best is established. In other words, the proper model and technique for analyzing the relationship between the variables in the dataset are determined at this step. More than one model should be constructed, more than one technique should be tried, and the best result should be tried in order to gain high efficiency from the dataset. When no model is considered adequate, the process is repeated after identifying any potential issues. The results are summarized and the superiorities of the developed models are provided at the end of this stage. If necessary, the model parameters are tweaked, and the process is repeated until the optimal model is found with high certainty.

2.4.5 Evaluation

Before the model is deployed, the model creation process should be carefully reviewed, evaluated in detail before proceeding to the implementation phase, and examined as to whether it achieves business objectives. At this stage, one or more models have been obtained and it is reviewed for the last time whether the model is sufficient for the current business objective. As a result of this step, actions that have been overlooked or that need to be repeated, if any, will be carried out and documented. For example, a system that recommends products to customers is expected to increase sales as a result of recommending the right product to the right customer, and how much the newly developed system provides in sales is measured. For this, experimental groups or different test techniques can be used.

In the model evaluation phase, it is tested to what extent the strategy created from the very beginning of the work achieves the determined targets. This is done using model performance criteria. Accuracy, sensitivity, specificity, precision, and f1 score are some of the criteria types [149]. After evaluating the results and reviewing the process, a decision will be made either to switch to productization to complete the data mining process or to iterate further or even go back to the very beginning of the data mining process.

2.4.6 Deployment

The models created are generally not the last stage of data mining projects. Although the purpose of modeling is to transform data into information, the information obtained should be organized and presented in a way that benefits the business. This involves turning the model into a form that the customer can use. The key steps are creation of the distribution plan, monitoring and maintenance of the plan, preparation of the final report, and finally the review of the project. The steps to be followed within these 6 stages are summarized in Table 2.11 [166].

Table 2.11 Steps followed within the CRISP-DM stages

Business Understanding	• Setting business goals • Assess the situation • Identifying data mining targets • Creating a plan
Data Understanding	• Data collection • Identification of data • Researching the data • Verification of data quality
Data Preparation	• Selection of data • Cleaning the data • Integration of data • Data reduction • Transformation of data
Modelling	• Selection of modeling technique • Creation of models • Evaluation of models
Evaluation	• Evaluation of results • Process review • Determining the next steps
Deployment	• Distribution plan • Plan monitoring and maintenance • Preparation of the final report • Review of the process

3 RELATED WORK

In the literature, there are numerous research on stock manipulation. Allen and Gale's research is the first to be published on this topic. Allen and Gale proved that a speculator who does not know the stock's future worth can earn from trading (buying and selling) by posing as a knowledgeable investor [63]. Jarrow developed an experimental model, claiming that it is impossible to argue about manipulation in the previous market if stock market players perform transactions in lockstep with one another [75]. In another theoretical study examining stock price manipulation, Allen and Gorton examined whether an uninformed investor could make a profit by buying a stock and then selling it at a higher price. They concluded that if the selling investor has less information than the buying investor, this causes an asymmetry in prices and creates a profitable manipulation opportunity [17]. In another theoretical study, Kumar and Seppi examined trade-based manipulation in futures markets. According to their study, manipulators who agreed with knowledgeable investors could make profits by taking a position in the futures contract and then trading in the spot market, but these profits decreased to zero with an increase in the number of manipulators [178].

Felixson and Pelli used a regression model to look at closing price manipulations in the Finnish stock market 15 minutes before and 15 minutes after the close of exchange-traded securities and found that it existed [179]. Stock price manipulations in the Chinese secondary market were investigated by Feng et al. They discovered that while a substantial change in stock prices isn't enough to show the presence of manipulation, a quick drop in the price and volume of the stock in issue invariably implies market manipulation [180]. Kucukkocaoglu examined whether the stock returns showed any manipulative activity towards closing, based on the intraday returns of 33 companies for the 2000-2002-time period. As a result of their regression analysis, they observed that the investors who traded through stock exchange representatives engaged in manipulative activities to determine the closing price. On the other hand, the main reason underlying the manipulative attempt of the agent who conducted the buy and sell transaction is showing the daily performance of the representative to the customer and increasing the potential profit of the customer [66]. Khwaja and Mian examined the characteristics of stock price manipulation in the Pakistan Karachi Stock Exchange for the period December 1998 to August 2001; they concluded that brokers trading on their behalf by pumping and dumping generate at least 50% - 90% higher abnormal cumulative returns than brokers trading on behalf of other investors [98].

Between January 1990 and October 2001, Aggarwal and Wu examined 142 examples of stock market manipulation in the United States. They discovered that illiquid equities are more likely to be manipulated, and that manipulation raises the liquidity, volatility, and return of the stock. They also discovered that a rise in stock price during the manipulation phase was followed by a drop afterward [181]. In BIST, Aktas and Doganay looked into trade-based manipulations. They stated that the manipulated stocks had increased transaction volume, return, and liquidity during the manipulation period, but that these values had dropped following the manipulation [182]. Koyuncugil proposed an early warning system (EWS) for securities markets that is based on fuzzy data mining and incorporates statistical learning perspectives. This EWS has three stages and is similar to the processes used in stock exchange markets. The system proposed by Koyuncugil includes the sequential determination of the shares according to their order, with a manipulative action by which the intermediaries may cause manipulation by the investors by performing the specified trading transactions [183].

Akyol and Michayluk investigated closing price manipulation using data from 2005 stocks on BIST. According to their study, the last trading returns in both the morning and afternoon trading sessions were high, but the returns in the second session were twice as large as those obtained in the first session, and this was caused by closing price manipulation [184]. Huang et al. looked into stock price manipulation on the Taiwan Stock Exchange. They look at the manipulated stocks' attributes as well as their effects on market quality. Their findings revealed that stock prices rose during the manipulation period, followed by a price reversal. Furthermore, during the manipulation phase, manipulated equities show greater return continuance and stock price volatility when trading volume is high. Stock manipulation can result in market inefficiencies, abnormally high trading volume and volatility, a reduction in market depth, and a negative influence on market quality [185].

Using data from 2005, Mongkolnavin and Tirapat discovered closing price manipulations in the Thai stock market. They claimed that association rules, one of the data mining methodologies, was utilized to detect tampering [186]. Many stock exchange manipulation instances, Palshikar and Apte discovered, featured coordination between groups of traders who trade heavily among themselves. They used graph clustering algorithms to detect collusion sets, and unsupervised learning algorithms, unlike other methods, were capable of separating legitimate stock market trading activities from fraudulent ones [187]. In terms of trade-based manipulation, Kamisli employed logistic regression and discriminant analysis to measure the use of financial ratios, which influence investors in their stock selection decisions, as an indicator. Financial ratios were developed utilizing data from trade-based manipulations at the Istanbul Stock Exchange between 1996 and 2005 for their study. The independent variables were these financial ratios, and the dependent variable was dichotomously coded as "0" or "1" depending on the trade-based manipulation's realization case [55]. Ogut et al. evaluated the performance of multivariate statistical techniques and data mining techniques in detecting stock price manipulations, finding that the latter were more suited [137].

Comerton-Forde and Putnins analyzed 184 closing price manipulation cases that were the subject of lawsuits in the USA and Canada between January 1997 and January 2009, by creating an index measuring profitability and the intensity of closing price manipulation by the authors. According to their study, in the event of manipulation strong evidence is obtained regarding large increases in end-of-day returns, yield reversal, trading activity, and price differences between trading, but it has been observed that manipulation has a negative effect on price accuracy [188]. Firm size, price/earnings ratio, information transparency, stock liquidity, and corporate partnership structure, according to Roodposhti et al., are useful variables in determining price manipulation of Tehran Stock Exchange companies [189]. When determining trade-based manipulations, Diaz et al. used data mining techniques to uncover stock price manipulations, taking into account both intraday trading prices and closing prices [74]. Over the course of a year, Sun et al. looked at the full transaction records of over 100 stocks. A stock trading network was developed to characterize the relationships between investors, and they discovered that the values of various attributes calculated for the nodes in the trading networks differed between manipulated and non-manipulated stocks [190].

Altinbas investigated the closing price manipulation processes in BIST using least squares regression and obtained statistically significant findings regarding the existence of closing price manipulation [90]. Using an unsupervised data mining method termed peer group analysis, Kim and Sohn suggested a method for detecting suspected stock price manipulations. By comparing the actions of study subjects to the behaviors of peers and measuring behavioral deviations, this form of analysis can discover anomalous behaviors in participants. This method has the advantage of detecting local outliers that would otherwise go undetected in the overall population [191]. Cao et al. studied coupled behavior analysis and demonstrated how to model and detect anomalous group-based trading behaviors using a coupled hidden markov model (CHMM). They suggested that buys, sells, and trades are inextricably linked and should therefore be analyzed collectively for anomaly detection. They suggested a CHMM-based model and analysis approach for aberrant linked trading behaviors [192]. Song et al. proposed a generic coupled behavior analysis approach for detecting group-based manipulation by capturing more complete couplings, which improved on the prior study. In terms of detecting aberrant collaborative manipulations, the experimental findings demonstrated that their suggested framework outperformed the CHMM-based technique [193]. To detect aberrant group-based trading activities, Song and Cao suggested a graph-based methodology. Without aggregating the behavioral data, the proposed framework represents the connected behaviors in a graph perspective. The proposed framework outperforms the CHMM-based technique in real-world stock market data, according to experimental results [194]. To classify whether a stock is manipulated according to specific key variables, Murugesan and Thoppan proposed a model based on discriminant analysis. In this model, first the linear classification function was used and it was seen that this function was not successful, and then the second order classification function that was more successful was used to categorize a stock as manipulated or non-manipulated [195].

Qui and Zhang examined the effect of insider trading on short selling restrictions, which is among the trade-based manipulation strategies. In their study, it was basically investigated how short selling restrictions affect stock price manipulation. According to the findings obtained, short selling restrictions increase manipulative activities in stock markets, and insiders can also make profits by taking short selling positions based on the information they have and change the direction of stock prices. In addition, it was concluded that short selling restrictions reduce the volatility of stock prices [196]. Imisiker and Tas analyzed the manipulation susceptibility levels of companies within the scope of trade-based manipulation cases in BIST for the period 1998-2006, using the panel probit regression method. Based on their results, they concluded that companies with lower market capitalization and free float ratio and higher leverage ratio are more prone to price manipulation of stocks [100]. Lee et al. found that 0.81% of the total orders were matched orders in the study they carried out using intraday orders and transaction data on personal accounts for the period of November 2001 to February 2002 in the Korean stock market. The returns and volatility of these stocks are high and they concluded that the market value is lower. In addition, it was observed that more matched orders were placed after the market opened and shortly before the market closed, and they found that investors following the matched order strategy achieved an extra 67 - 83 basis points more return in less than 45 minutes [80].

To identify suspicious transactions related to stock market manipulation, Golmohammadi et al. used supervised learning systems. With a f1 score of 53 %, sensitivity of 89 %, and specificity of 83 %, the naive bayes algorithm beat other learning approaches, according to their findings [197]. Gerace et al. looked at 40 manipulation cases in the Hong Kong Stock Market from 1996 to 2009, and used regression analysis to see if manipulators could influence the price and profitability of the stock exchange. According to their findings, an increase in market prices during manipulation times was discovered [14]. Imisiker and Tas investigated the profitability of wash sales, which is among the trade-based manipulation types, in BIST for the period 2003-2006. As a result of their research, it was determined that a significant amount of investors performed wash sales transactions and profit was obtained from these transactions, which constitute 30% of the total transactions [198]. For 15 equities, Kong and Wang looked at the consequences of manipulation using orders that were matched for price, volume, stock turnover rate, and liquidity on the China Stock Exchange. They discovered that the price, turnover, transaction volume, and volatility of the stocks were higher during the manipulation period, and that the price, turnover, transaction volume, and volatility of the stocks were lower following the manipulation period [199].

Kadioglu investigated the closing price manipulation with 102 stock data for the period of November 2006 to May 2012 regarding intraday return and volatility structures in BIST; he observed that the volatility was high at the opening of the session and showed significant decreases in the first 15-minute returns, and towards the close of the session both the returns and the volatility in the morning session decreased and he detected manipulative activities to increase the closing price [72]. Ozcomak and Gunduz investigated the relationship between the closing prices of 234 companies traded in the stock market on BIST in 2011 and the transaction volumes. They stated that stocks with high volatility in closing prices and trading volumes may be prone to speculation and manipulation [200]. Comerton-Forde and Putnins analyzed 184 closing price manipulations and the validity of closing price manipulations for the period January 1997 to January 2009 in the USA and Canada, and they concluded that 1% of the closing prices were manipulated. In addition, they found that stocks with high asymmetric information level and stocks with medium-low liquidity level were the most likely to be manipulated [201].

Golmohammadi and Zaiane focused on methods for detecting contextual outliers in time series, which can be used to spot securities fraud. They developed a prediction-based strategy for detecting contextual anomalies in complex time series that isn't well represented by deterministic models. When compared to kNN and random walk approaches, Golmohammadi and Zaiane's algorithm improves recall from 7 % to 33 % while preserving precision [202]. Huang and Cheng looked at stock price manipulation in Taiwanese stock markets, as well as manipulated stock patterns and market repercussions. A pump-and-dump trading technique and stabilization operations were used in the majority of incidents of manipulation. Pump-and-dump manipulations result in strong transient price impacts, heightened volatility, big trade volumes, short-term price continuation, and long-term price reversals during the manipulation period. As a result, their impact on market efficiency is enormous. In stabilizing circumstances, manipulation has no influence on market performance except that price drops and abnormal returns are considerably lower in the post-manipulation period than in the pre-manipulation period [203]. In their study analyzing the effects of bulk sales and purchases on stock prices for the 2004-2012 period on the Mumbai Stock Exchange of India and the National Stock Exchange of India, Chaturvedula et al. stated that approximately 80.5% of the investigations for the 2004-2008 period were related to market fraud [204]. Imişiker et al. used intraday transaction data from BIST from 2003 to 2006 to investigate pump-and-dump manipulation. They discovered that a significant percentage of the trades executed by the brokers were based on the pump-and-dump manipulation, and that the brokers, although rare, made large profits from this trade [205].

Leangaraun et al. looked into pump-and-dump and spoof trading tactics, developing mathematical models based on level 2 data, which included all of the information from level 1 data as well as buy/sell orders. They created feed-forward neural network models using level 1 data (excluding order cancellation data), which investors may access more easily as input. The algorithm was able to detect pump-and-dump actions with an accuracy rate of

88.28 percent, but it was unable to accurately reproduce spoof trading behaviors, according to their findings [206]. Martinez-Miranda et al. studied the core behaviors of fraudulent traders using a reinforcement learning framework within Markov decision processes, with the goal of discovering what motivated those behaviors [207]. Examining the effects of trade-based manipulation cases in BIST between 2005 and 2013 on returns, volatility, and trading volume, Ok found that stock returns, volatility, and trading volume increased during the manipulation period compared to before and after manipulation [62]. Zhang et al. used machine learning algorithms to detect stock price manipulation in China in order to increase market fairness and transparency. Using data provided by the China Securities Regulatory Commission, they calculated the difference in stocks between manipulated and ordinary time periods based on daily returns, trading volume, stock price volatility, and market value (CSRC). In their study, they used them as explanatory variables. They employed a single model, the support vector machine (SVM), and an ensemble model, the random forest, for detection (RF). The classification accuracy, sensitivity, and specificity tests of the SVM were compared to those of the RF. As a result, they discovered that both methods are highly accurate, with RF outperforming SVM. The influence of daily return and market value on detection is likewise stronger than those of other explanatory variables [208].

Using data from the CSRC, Li et al. applied supervised machine learning algorithms to detect market manipulations. Trained machine learning systems successfully recognized stock market manipulations from provided daily transaction data, but they performed badly when fed tick data, according to the results obtained for accuracy, specificity, sensitivity, and area under the curve (AUC). Out-of-sample evaluation was not used in their tests, which is why their accuracy (0.99) scores were so high [209]. Over the course of a year, Sun et al. analyzed transaction data from 8 manipulated stocks and 42 non-manipulated equities to uncover distinct trading tendencies. The manipulated stocks showed higher numbers of traders dealing in the same pairs on repeated days with high deviations from random networks, according to correlations between trading frequency and trading activity [210]. Gemici et al. investigated 273 trade-based manipulation cases in BIST in the period of 2001-2014. In the multiple logistic regression analysis performed by considering the period before the manipulation, the period of manipulation, and the period after the manipulation, the variables daily return, transaction volume, volatility, and stock turnover rate were used, and it was determined that daily return and volatility had a greater effect on manipulation [211]. Using data from the Indian stock market, Thoppan et al. investigated the accuracy of various classification techniques for detecting market manipulation. The data contained price, volume, and volatility statistics for a variety of equities. Techniques including discriminant analysis, a composite model based on artificial neural network – genetic algorithm, and SVM were employed to categorize equities into manipulated and non-manipulated categories. The SVM-based technique has the highest classification accuracy of the three techniques [212].

Pump and dump trading, quotation stuffing, and spoof trading are all examples of anomalous stock price manipulations that Abbas et al. proposed a model for identifying. They utilized an unsupervised learning technique, in which input data were empirically deconstructed and anomalies were detected using kernel density estimation-based clustering [213]. For detecting aberrant trading behaviors caused by stock price manipulations, Leangarun et al. used generative adversarial networks. The neural networks were not trained using manipulation cases. Instead, they trained them using real data, while simulated manipulation cases were only utilized for testing. Trading data from Thailand's Stock Exchange was used to test the detecting system. It detected pump-and-dump adjustments in unobserved market data with a 68.1 percent accuracy [214].

Based on transaction data, Shi et al. developed a generic technique to detect colluding traders. They looked into network cliques and discovered that while trading manipulated stocks, the quantity and weight of cliques is higher than when dealing non-manipulated equities. They proposed a strategy based on the weight of cliques to detect colluding traders [215]. To solve the irregular trade behavior identification problem in the stock market, Tran and Tran used three graph Laplacian-based semi-supervised ranking techniques. The un-normalized and symmetric normalized graphs, according to their findings, Random walk approaches were surpassed by Laplacian-based semi-supervised ranking methods. Semi-supervised ranking algorithm based on Laplacian [216]. By merging trade-based attributes acquired from trading records with typical properties of the list businesses, Quili et al. suggested an unique RNN-EL framework for identifying stock price manipulation operations. To perform empirical studies, they built a special dataset containing labeled samples with trading data and characteristic information based on prosecuted manipulation cases provided by the CSRC. The experimental results reveal that their proposed strategy outperforms state-of-the-art methodologies in identifying stock price manipulation by an average of 29.8% in terms of AUC value [217].

To identify and detect market manipulation, Sridhar et al. recommended utilizing an ensemble neural network. The Securities and Exchange Board of India provided affidavit information, which was used to create a daily trade dataset from the Bombay Stock Exchange website. The ensemble neural network model was tested with and without trainable sub-model layers using the daily trading dataset. The accuracy of the model with trainable submodel layers was 91%, while the model without trainable submodel layers was 96% [218]. Uslu and Akal devised a machine learning technique based on supervised machine learning classification models to detect trade-based manipulation from daily data of manipulated stocks. The data for the study was obtained from 22 instances of BIST manipulation between 2010 and 2015. In their study, supervised machine learning approaches were found to be successful at identifying trade-based manipulations in trading networks based on the measuring methods of accuracy, sensitivity, and f1 score. The proposed model has a f1 score of 91 percent, 95 percent sensitivity, and 93 percent accuracy when it comes to detecting market manipulation [37]. Youssef investigated the use of machine learning techniques to detect trade manipulations. To extract features from actual manipulation data, the continuous wavelet transform was utilized, while principal component analysis and factor analysis were used to reduce dimensionality. Following those steps, machine learning classifiers were trained and tested. It was discovered in this study that employing the continuous wavelet transform, it is possible to increase model accuracy while also drastically enhancing precision. Simultaneously, recall values fell little [219].

To construct stock market manipulation detection models, Liu et al. used machine-learning approaches. To compare the detecting abilities of SVM and a logistic model, they built a training set and a test set by integrating manually collected CSRC punishment cases from 2014 to 2016 with financial information from listed corporations. To boost the detection effectiveness of the algorithms, they integrated market sentiment indicators acquired from analyst rating reports, financial news, and Guba comments into their indicator collection. According to the data, the new indicators result in a significant marginal gain in model accuracy [220]. According to Leangarun et al. (2021), unsupervised learning was utilized to train deep neural networks for detecting stock price manipulation in order to discover unknown and previously unnoticed manipulation. It was excellent for detecting new or undiscovered manipulation kinds [221].

4 THE PROPOSED STOCK MARKET MANIPULATION DETECTION FRAMEWORK

4.1 Data Set

The dataset used in this study consists of stock trading transactions conducted in BIST between 2010 and 2015. It was obtained from the CMB Information Systems database, and the personal data of the investors were anonymized.

The kinds of orders that can be used on a stock exchange, as well as the sequence in which the matching process takes place, are the most important factors. In BIST, as in practically every other exchange, an electronic order book is used. Customers' buy and sell orders for market instruments are sent to the order book on a regular basis. Orders can take the form of a new order, a change to an existing order, or a cancellation.

Call auction and continuous auction procedures are used to complete the order matching process. Throughout the dataset, buy and sell orders are collected around 09:00 every day and forwarded to the order book for 30 minutes, after which a single price for each instrument is calculated using the call auction approach. In a call auction, the price is fixed at a level that allows the most volume to be exchanged. A continuous auction session runs from 09:45 to 17:00 or until 17:30, after the call auction session. Buy and sell orders are performed in a continuous auction based on price and time priority. The primary distinction between a call and a continuous auction is that in the latter investors can trade at any time the market is open. A new buy order is combined with the pending sell order with the lowest selling price and the longest duration in the continuous auction market. A new sell order is paired with a pending buy order that has the highest sale price and longest term. A trade is started and a buy–sell transaction is completed through the order matching process.

The order matching process initiates trades and a new trade takes place. In a Level 3 (L3) type trading book, stock trading transactions are ordered according to time and stock. Some of the attributes in the trading book are as follows:

- Stock name,
- Date and time of the transaction,

- ID of the transaction (a unique number for each transaction generated by an algorithm),
- Amount and volume of the transaction,
- Initiator of the transaction (whether it is an active attribute of the transaction),
- Information about whether the price of the transaction is higher than, lower than, or equal to the previous transaction price,
- Unique identification numbers that identify the buyer and seller.

Table 4.1 contains the summary of stock trading transactions in BIST from 2010 to 2015. The total number of transactions between 2010 and 2015 was 970,316,932, the average number of transactions was 161,719,489, and the total transaction data size was approximately 161 GB. The maximum number of transactions took place in 2011, the number of transactions was 199,560,506 and the number of stocks was 385.

Table 4.1 Information on transactions performed in 2010 - 2015

Year	*Number of Transaction Rows*	*Number of Traded Stocks*
2010	160.355.935	354
2011	199.560.506	385
2012	153.849.588	428
2013	154.139.091	443
2014	159.894.074	449
2015	142.517.738	441

The number of transactions shown in Table 4.1 is for all of the stocks traded in 2010–2015, and we focused only on the transactions related to the stocks manipulated during this period. In the bulletins published weekly by the CMB, there is information about the manipulation and non-manipulation periods of the manipulated stocks. These bulletins were examined for the specified period, and a new dataset was created by separating only the transactions of the manipulated stocks from the large dataset. The bulletins showed that the number of stocks subjected to manipulation was 75 days, the average manipulation period was 66 days, the longest manipulation period was 103 days, and the shortest manipulation period was 5 days.

Table 4.2 contains information about the transactions of stocks that were manipulated in BIST between 2010 and 2015. The highest number of manipulative transactions was carried out in 2011, with 17,860,780 transactions. The total number of manipulative transactions in this period was 51,916,018, and the total transaction data size was approximately 5.23 GB.

Table 4.2 Manipulative transactions performed in 2010 – 2015

Year	***Number of Transaction Rows***	***Number of Traded Stocks***
2010	10.438.540	17
2011	17.860.780	25
2012	6.821.272	11
2013	7.710.280	19
2014	6.875.908	18
2015	2.209.238	7

We picked 20 stocks with the highest manipulation period. We anonymously named them Stock1 to Stock20. They were exposed to 22 cases of manipulation in total during given period. Table 4.3 shows the number of days the stocks were manipulated for, the start and end dates of the manipulation, and the number of manipulation cases. According to Table 4.3, the longest manipulation was 280 days, while the shortest was 71 days. While Stock 1 and Stock 3 were manipulated twice during the manipulation period, there was 1 manipulation case each when the other 18 stocks were manipulated.

Table 4.3 Manipulated stocks

Stock	***Manipulation period (days)***	***Start date of manipulation***	***End date of manipulation***	***Number of manipulation cases***
Stock1	155	2011–10-03 2013–11-05	2011–11-24 2014–04-21	2
Stock2	82	2013-01-15	2013-05-10	1
Stock3	280	2012-03-28 2013-02-04	2012-10-23 2013-08-15	2
Stock4	140	2012-09-27	2013-04-16	1
Stock5	71	2013-09-16	2013-12-30	1

Stock	***Manipulation period (days)***	***Start date of manipulation***	***End date of manipulation***	***Number of manipulation cases***
Stock6	74	2013-02-22	2013-06-07	1
Stock7	115	2010-11-23	2011-05-02	1
Stock8	188	2010-06-01	2011-03-01	1
Stock9	101	2013-01-03	2013-05-27	1
Stock10	150	2012-04-18	2012-11-23	1
Stock11	172	2014-09-09	2015-05-15	1
Stock12	74	2010-09-06	2010-12-27	1
Stock13	111	2015-04-15	2015-11-24	1
Stock14	115	2012-12-20	2013-06-03	1
Stock15	134	2014-11-28	2015-06-11	1
Stock16	99	2010-12-29	2011-05-16	1
Stock17	143	2010-09-16	2011-04-11	1
Stock18	134	2011-06-24	2012-01-05	1
Stock19	209	2010-05-14	2011-03-16	1
Stock20	124	2012-10-30	2013-04-22	1

Table 4.4 contains information about the transactions of the stocks that were manipulated (20 selected stocks) in BIST between 2010 and 2015. The highest number of manipulative transactions was carried out in 2011, with 4,210,548 transactions, and the number of manipulated stocks was 6. The total number of manipulative transactions in this period was 16,870,752, and the total transaction data size was approximately 1.82 GB.

Table 4.4 Selected manipulative stocks

Year	***Number of Transaction Rows***	***Number of Traded Stocks***
2010	2.994.698	6
2011	4.210.548	6
2012	2.252.472	6
2013	4.145.316	8
2014	1.918.114	3
2015	1.349.604	3

4.2 The Proposed Model

Manipulative activities have significant effects on investors, financial markets, and, accordingly, the entire economy. The people who suffer because of manipulation most severely due to the techniques used and various malfunctions in the legal regulations are investors who play a major role in the financing of companies by supplying funds to the stock markets. While the manipulators, who create unusual price movements to the disadvantage of investors who are unaware of the existence of manipulative actions, gain excessive profits, investors incur losses. In general, although the markets of countries with legal and institutional deficiencies are subject to manipulation, manipulations take place in the securities markets during periods of bad economic performance. Trade-based manipulations, also the subject of this thesis, are different from other types. This type consist of legal transactions aiming to differentiate the equilibrium price in the market by performing only buying and selling transactions, without trying to change the value of the companies with the stocks traded in the market or presenting false or incomplete information to the market. As a result, manually detecting these modifications is becoming increasingly difficult. Although research has been conducted to detect trade-based stock market manipulations, with the advancement of technology, manipulation tools have become increasingly diverse, and effective detection methods remain a challenge. Data mining, machine learning, and deep learning technologies have been used successfully in recent years to detect trade-based manipulations. We used a machine learning technique and suggested a stock market manipulation detection model comprising supervised machine learning classification models for detecting trade-based manipulations in BIST between 2010 and 2015. DTC, LR, KNN, RF, NB, SVM, and ANN are among the supervised machine learning classification algorithms used in the model.

Experiments have shown that our model is suitable for detecting trade-based manipulations, and when our model is evaluated in terms of performance metrics, we achieved high accuracy, sensitivity, and f1 scores. We used a methodology similar to CRISP-DM. We made some changes to CRISP-DM, and these changes and the equivalents of CRISP-DM steps in our model are shown in Table 4.5.

Table 4.5 Correspondence of CRISP-DM steps in our model

CRISP-DM	***Proposed Model***
1. Business understanding	Preprocessing
2. Data understanding	
3. Data preparation	
3.1 Data selection	
3.2 Data cleaning	
3.3 Data integration	
3.5 Data transformation • Concatenation • Variable creation	
3.5 Data transformation • Scaling	Scaling
3.4 Data reduction	Feature Engineering
4. Modelling	Training
5. Evaluation	Evaluation
6. Deployment	Prediction

As can be seen in Table 4.5 and Figure 4.1, our model consists of 6 stages: preprocessing, scaling, training, evaluation, feature engineering, and estimation. According to Table 4.5, the steps related to our method can be summarized as follows:

- **Preprocessing:** By defining the problem, merging and variable creation processes were carried out from the steps of understanding, preparing, selecting, cleaning, integrating, and transforming the data.
- **Scaling:** Normalization (scaling) was conducted from the data transformation processes, and the data were converted into appropriate forms by preserving their content according to the model to be used.
- **Training:** This is the model creation phase, and our model was trained using DTC, LR, KNN, RF, NB, SVM, and ANN.
- **Feature engineering:** Data reduction was carried out and some unnecessary features in the datasets were removed.

- **Evaluation:** This is the evaluation phase of the model, and the building model process was reviewed and it was evaluated whether the model was successful in detecting manipulation.
- **Prediction:** The model was chosen and put into use, and reporting of the success rates (accuracy, clarity, precision, sensitivity, f1 score) in detecting trade-based manipulations was ensured.

These steps that make up our methodology were applied in detecting trade-based manipulations in the stock market, which is the subject of the study and they are explained in detail below.

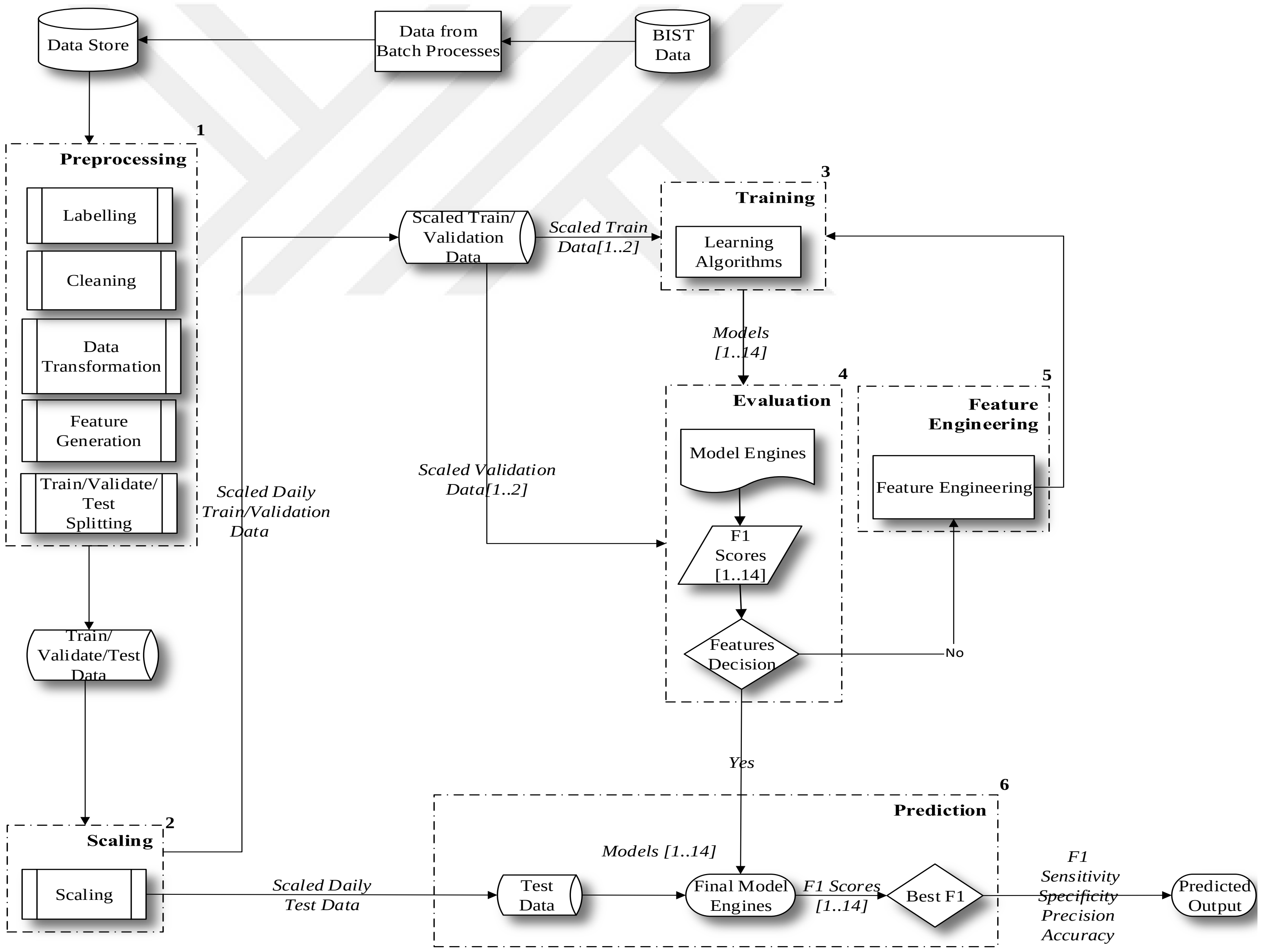


Figure 4.1 Architectural diagram of the proposed model

4.2.1 Algorithm

The pseudocode for the data transformation preprocessing, training, evaluation, feature engineering, and prediction procedures is provided below.

Pseudocode of the model

```
1.   Input D: original level 3 dataset
2.         N: number of the stock
3.         DSPV: defined success point value (dynamic)
4.   Output O: binary number indicating whether trade based manipulation exists and
     subsets of calculated performance metrics (accuracy, F1, sensitivity, specificity,
     and recall)
5.   preprocess D (preparing, selecting, cleaning, integrating, transforming D)
6.   scale D
7.   obtain train data, validation data and test data from step 6
8.   while ( N > 0 ) do
9.      N--
10.      implement training step \\ to obtain models by using train data (training)
11.      test the models with validation data \\ obtain f1 scores related to models
12.      do feature engineering \\ obtain better f1 scores related to models
13.      f1 = max([F1_scores])
14.      \\ evaluation
15.      if (f1 > DSPV) then \\ feature set is determined
16.         Decide final model engines
17.      \\ prediction
18.      test the models with test data \\ obtain f1 scores related to models
19.      f1 = max([F1_scores])
20.      select the model \\ the model of f1 is maximum
21.      calculate performance metrics \\ accuracy, f1, sensitivity and recall
22.      else
23.         \\ feature engineering
24.          implement feature engineering step \\ obtain optimum feature subset
25.          goto step 10
```

26. **end while**

27. **Output** subset of calculated performance metrics \\ for each stock

4.2.2 Preprocessing

The transactions carried out at this stage include important steps to be taken before our model is established. These steps consist of defining the problem, understanding the data, and preparing the data (selecting, cleaning, integrating, and transforming it).

4.2.2.1 Business Understanding

Manipulation, that is, market fraud, involves changing the prices and the market as desired by fraudulent methods, and to try to divert the market from its natural course and reality. In other words, manipulation is the process of deliberately trying to change the market prices of an asset, which is formed by real supply and demand, and to create an artificial, misleading, and deceptive market. Although there are different types of manipulation, trade-based manipulations consist only of activities aimed at misleading the market with trading transactions, artificially influencing the supply and demand of capital market instruments, creating the impression of an active market presence, and keeping prices at the same level. It causes changes in the market price via buying and selling transactions in order to increase or decrease it [41].

In this thesis, the problem of detecting trade-based manipulations in BIST with a machine learning approach and revealing the factors affecting the manipulation is handled with a holistic approach.

4.2.2.2 Data Understanding

The first thing when developing a model for solving a problem is to obtain data. Our L3 dataset on stock trading transactions in BIST between 2010 and 2015 includes all details of the transactions. It includes the stock name, the date and time of the transaction, the identity of the transaction, the amount and volume of the transaction, and unique identification numbers that identify the buyer and seller. With the acquisition of the dataset, the data are examined in general, and the attributes occurring when a trading transaction takes place are interpreted.

4.2.2.3 Data Preparation

This phase includes obtaining the data and making them suitable for the machine learning algorithms that we will use in modeling after the understanding phase. It covers all operations from the raw data we obtained at the beginning until we receive the final dataset to be used in modeling. The consecutive steps during the preparation of the data are selecting, cleaning, integrating, and transforming the data.

4.2.2.3.1 Data Selection

We collected data from the CMB Information Systems database. The L3 dataset we obtained includes stock transactions that took place in BIST between 2010 and 2015. We picked 20 of the manipulated stocks with the highest manipulated period and conducted our study with a subset of the large dataset.

4.2.2.3.2 Data Cleaning

In this phase, it is aimed to improve data quality by identifying and correcting inconsistencies and errors (invalid data, incorrectly entered data, duplicate data). Since our dataset consists of real transaction data that took place in BIST, data cleaning is not needed, but there are repeated records of some transactions. Each transaction record is kept with unique identity information and these repeated records are identified according to their identification numbers and removed.

4.2.2.3.3 Data Integration

Data obtained from different sources and having similar characteristics or related data are combined. Since our dataset was created from a single data source, this step did not need to be applied.

4.2.2.3.4 Data Transformation

This involves the conversion of the available data into a format suitable for modeling, and the types generally used are merging, scaling, and variable creation. We performed variable creation and merging from data transformation types in the preprocessing step, and scaling for our dataset in the next step (scaling).

- **Creating variables:** Our dataset is ordered according to the stock name and the date of the transaction, and contains many attributes of the transaction data. These attributes include the name of the stock, the date of the transaction, the unique identification numbers of the buyer and the seller, the amount and volume information of the transaction, and whether the buyer or the seller initiated the transaction. The stocks in our dataset were manipulated between 2010 and 2015, and the dates of the manipulations are included in the weekly bulletins published by the CMB. Since the dataset does not include an attribute on which dates the stocks were manipulated, a new feature called *manipulation* was defined. The date range in which each stock was manipulated was obtained from the weekly bulletins, and the *manipulation* attribute was filled with *1* in the period when the manipulation took place, and *0* in the period when the manipulation did not occur.
- **Merging:** The details of the merging operations we performed on our dataset are as follows:
    - In the L3 dataset obtained, the transactions regarding the stock are ordered according to the stock name and the date and time of the transaction, and the transactions take place in the order of microseconds. Whether there is manipulation in a transaction is decided not only according to a transaction that has taken place, but also by evaluating all transactions related to the stock in question via a holistic approach. While detecting manipulation, CMB experts use this approach and make their evaluations on a daily basis by converting transactions that take place in the order of microseconds into daily transactions. Based on this approach, we obtained a new dataset by transforming the dataset into daily operations. The attributes in the dataset obtained and their explanations are given in Table 4.6, and a few examples of the attributes are as follows:
        - trading volume (daily total trading volume),
        - amount of transactions (total amount of transactions per day),
        - number of buyers (total number of traders making daily purchases)
        - number of sellers (total number of traders who make sales transactions per day)

Table 4.6 Attributes and descriptions of the dataset generated on a daily basis

Feature	***Description***
Weighted average price	The weighted average price formed from the beginning of the day at the end of each transaction
Maximum price	Maximum price of the transaction during the period
Minimum price	Minimum price of the transaction during the period
Volume	The amount of stock in each transaction
Volatility	The natural logarithm of the maximum price and the minimum price during the period
Number of transactions	Number of transactions during the period
Number of buy transactions	Number of buying transactions during the period
Number of sell transactions	Number of selling transactions during the period
Number of price increasing transactions	Number of transactions for which the price is higher than the previous transaction price
Volume of price increasing transactions	Volume of the stock in transactions for which the price is higher than the previous transaction price
Number of price decreasing transactions	Number of transactions for which the price is lower than the previous transaction price
Volume of price decreasing transactions	Volume of the stock in transaction for which the price is lower than the previous transaction price
Number of transactions without effect on price	Number of the transactions for which the price is equal to the previous transaction price
Volume of transactons without effect on price	Volume of the stock in transaction for which the price is equal to the previous transaction price
Number of active buying transactions	Number of buying transactions for which match the selling transactions pending in passive buying
Volume of active buying transactions	Volume of the stock in transaction of in the form of active buying
Number of active selling transactions	Number of selling transactions for which match the buying transactions pending in passive selling
Volume of active selling transactions	Volume of the stock in transactions of in the form of active selling
Number of buyers	Total number of traders of buying transaction

Feature	Description
Number of sellers	Total number of traders of selling transaction
Manipulation (manip)	This indicates whether manipulation was detected by the administrative authority in the stock market in the period: manipulated period (= 1), non-manipulated period (= 0)

- Our new dataset is a time series ordered by stock name and date of transaction. We conducted an exploratory analysis on the dataset. The number of non-manipulative transactions is approximately 4 times the number of manipulative ones. In the literature, this is called the imbalanced data problem. Since we propose a learning-based model, this problem may affect the performance of our model. To avoid this, as seen in [137], we adjusted our dataset to be non-manipulated period, manipulated period, and non-manipulated period on the basis of stocks. Thus, for each capital market instrument, the number of manipulative transactions and non-manipulative transactions were adjusted to be 1:2.
- Before modeling with machine learning, we divided our dataset into three parts: training, validation, and test data. For this separation, we did not separate the dataset in chronological order. i) We divided our dataset, ordered by stock and time, into two datasets according to their manipulation values (0 = manipulation, 1 = no manipulation). ii) We separated both of these datasets separately as 70% training, 15% validation, and 15% test data. iii) We obtained 2 training datasets, 2 validation datasets, and 2 test datasets. One of the training, validation, and testing datasets includes manipulative operations (manipulation = 1), while the others have non-manipulative operations (manipulation = 0). All of the datasets we obtained are sorted by stock and date, and the manipulative training dataset and non-manipulative training dataset, the manipulative validation dataset and the non-manipulative validation dataset, and finally the manipulative test dataset and the non-manipulative test dataset will be ordered according to the stock and time. Thus, the 70% training, 15% validation, and 15% test data we obtained at the end became the output of the preprocessing phase.

4.2.3 Scaling

Scaling is performed in order to put the numerical features in the dataset into a certain order. The large imbalances among the features in the datasets may result in the large-valued features having a greater effect on the target variable. If there are very large numerical differences between the features in the datasets, the scaling operations performed contribute positively to the performance of the model. In other words, scaling operations are performed to bring all attributes to the same position in order to prevent numerical attributes in a dataset from getting ahead of other attributes during the training phase. Generally, scaling is applied in 2 ways, min-max and standard.

Min-max scaling, or minimum-maximum normalization, is the scaling of the features to a range of 0-1 and after this the maximum value of the data contained in the features is 1 and the minimum value is 0. Standard scaling is used more widely and is a statistical data transformation technique. By scaling the mean values of the features to be 0 and the standard deviation values to be 1, all values in the dataset will have a minimum value of -1 and a maximum value of 1.

The training, validation, and test datasets we received as input from the preprocessing phase were scaled by both methods at this stage. The outputs of this phase are 2 scaled training, 2 validation, and 2 test data, as shown in Figure 4.1. The test data are set aside to be used in testing the model at the prediction stage and will be input to the prediction stage.

4.2.4 Training

A large part of machine learning is training. In this stage, a model is formed for the solution of the machine learning problem by running a supervised machine learning algorithm on a dataset. The machine learning algorithm learns using the inputs and outputs of the training dataset; in other words, the model is formed. There are two main elements in the model formed by the completion of the learning phase: the weight and constant values of the attributes.

$$y = \mathrm{m} * \mathrm{x} + \mathrm{b} \tag{4.1}$$

Let the y line specified in equation (4.1) be our model to be created with the machine learning approach. Here, let x be the input value, m be the slope of the y line, b be the value that cuts the y line, and y be our output value. The values learned (trained) with the machine learning approach are m and b values. There is no other way to affect the position of the line, because the other variables are just our input x and our output y.

The dataset we used has many features and the effects of each on the manipulation variable are different. These effects are expressed as in equation (6.18) with the weights and constant values above for each feature. With the model created, an equation of (4.1) for each attribute will be obtained. In other words, we are trying to determine the weights and fixed values of the features with the model developed by the machine learning approach. The determined weights and fixed values build our model. Since we have more than one attribute, the weight and constant variables (w, b) are vectors. In addition, the success of the model is closely related to the machine learning algorithm used and the quality of the training dataset.

Fourteen different models were built after running DTC, LR, KNN, RF, NB, SVM, and ANN on 2 training datasets scaled with min-max and standard, which we received as input from the scaling during the training. Details of these models and the training phase are given in Figure 4.2.

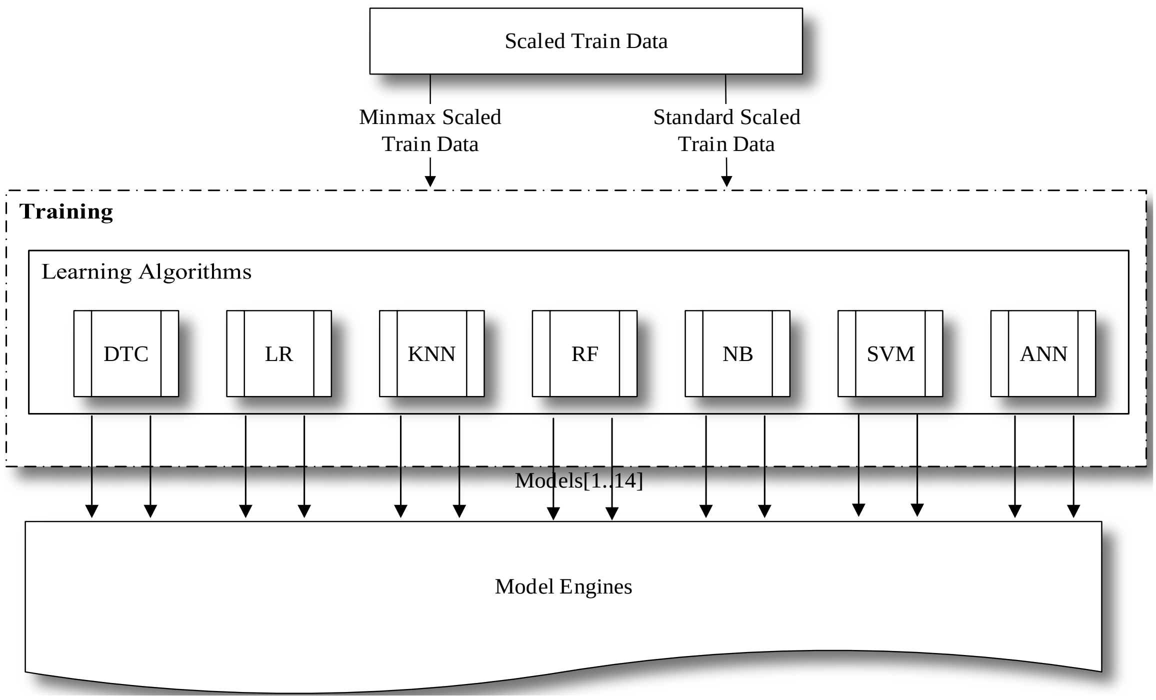


Figure 4.2 Details of building models

4.2.5 Evaluation

The performances of 14 machine learning models, which are the input of the evaluation stage, developed to detect trade-based manipulations taking place in BIST, are evaluated. Although the main purpose is to detect manipulations in the stock market, we used the f1 score, the most appropriate metric, since we also want an accurate classification for the non-manipulation period.

In machine learning studies, there are many features in the datasets and the effects of each feature on the output value differ. Some of the features in the datasets have a negative effect on the output value or no positive effect. Thus, removing the features that have effects can increase the model's performance. These processes are explained in the next section, they are expressed as feature engineering, and it is decided which features will be included in our models. As shown in Figure 4.1, the evaluation phase is completed by the feature engineering phase and then the training phase.

We obtained the f1 score values of the models by testing 14 machine learning models, the output of the training phase, which is also the input of this phase, and 2 scaled validation data (15% of our dataset), which is another input. We performed feature engineering to improve the f1 score values of our models. After the feature engineering, the training phase was performed again and we obtained 14 new machine learning models. We achieved new f1 scores by testing these models with revalidation data. We continued feature engineering and training until we got the best f1 scores for our models. After that, this stage was completed and the attributes of our models were determined. As the output of this stage and the input of the prediction stage, 14 machine learning models were built and their attributes were determined.

4.2.6 Feature Engineering

In the machine learning literature, feature engineering involves removing the features that have no effect or a negative effect on the output variable, instead of using all the features in the dataset. By applying feature engineering, we determined the features that have no or negative effect on the manipulation output variable and thus increased the performance of our models by removing these features. These operations are valid for supervised machine learning algorithms other than ANN, and feature selection is performed automatically in the ANN algorithm.

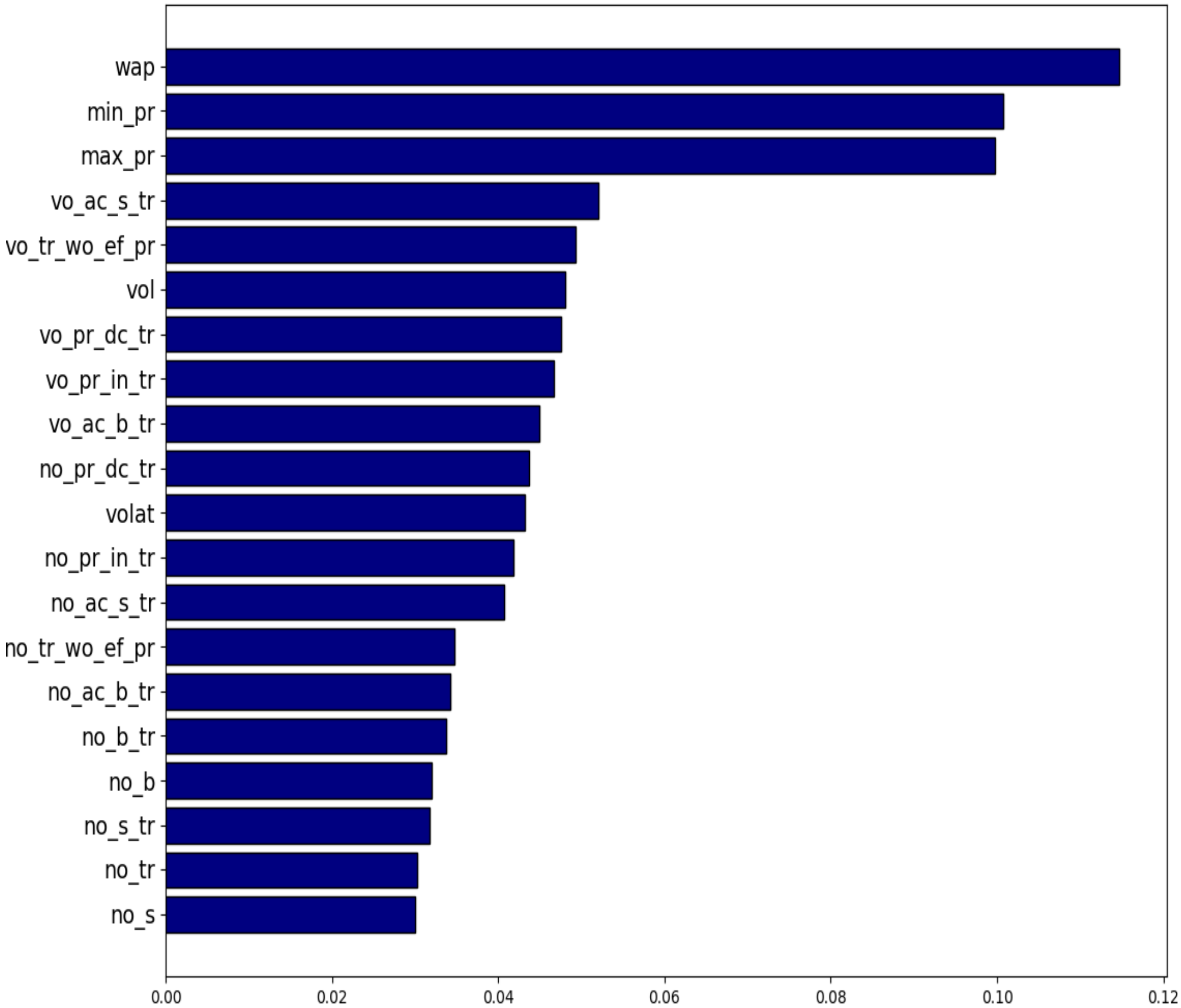


Figure 4.3 Effects of attributes

The effects of the attributes in the dataset on the manipulation output variable are shown in Figure 4.3. It is clear that the *weighted average price*, *minimum price*, and *maximum price* are the 3 most important attributes for trade-based manipulation detection.

It is important here is to determine which features other than these 3 should be selected for our models detecting manipulation. To select the other features, we repeated the evaluation, feature engineering, and training stages. The actions we perform at these stages are as follows:

1. The f1 scores obtained by testing the machine learning algorithms with the validation data were evaluated. Next, feature engineering was carried out by trial and error until the best f1 score was obtained.
2. After the evaluation, feature engineering was carried out, and the features that were not effective in detecting manipulation were determined by trial and error and removed.
3. After the feature engineering phase, some features were removed from the dataset, and the training phase was performed again to build new models.

After the evaluation, feature engineering and training stages were carried out repeatedly and the features to be found in our models were determined. Next, it was determined by trial and error that the first 12 features shown in Figure 4.3 were the most important features for the manipulation output variable. Accordingly, the attributes to be included in the models created for the detection of trade-based manipulations are as follows: *weighted average price, minimum price, maximum price, volume of active selling transactions, volume of without effect transactions in price, volume, volume of price decreasing transactions, volume of price increasing transactions, volume of active buying transactions, number of price decreasing transactions, volatility, and number of price increasing transactions.*

4.2.7 Prediction

We explained the working principle of the proposed model in order to detect the trade-based manipulations in BIST between 2010 and 2015. Our proposed model predicts whether the transaction it receives as input is manipulative. The processes performed consecutively at this stage are given below:

1. As a result of running 7 supervised machine learning classification algorithms on 2 different datasets scaled according to min-max and standard, we have 14 models. These models are the output of the evaluation phase and the input of the prediction phase and they have been tested with the scaled test data set aside during the scaling phase.
2. Our test data, like other training and validation data, are sorted by stock and date and testing is done on a per stock basis.
3. The f1 scores of each model were obtained in the test processes performed on the basis of stocks. In this case, there are 14 scores for each stock.

4. The f1 scores obtained for a stock are shown in Table 4.7. There are 2 different scores for each model. For example, the DTC_mm model is built by running the DTC on the dataset scaled according to min-max, and DTC_s model is built by running it on the dataset scaled according to standard. As a result of testing the DTC_mm and DTC_s models with test data, f1_1 and f1_2 score values were obtained.

Table 4.7 F1 score list based on stocks

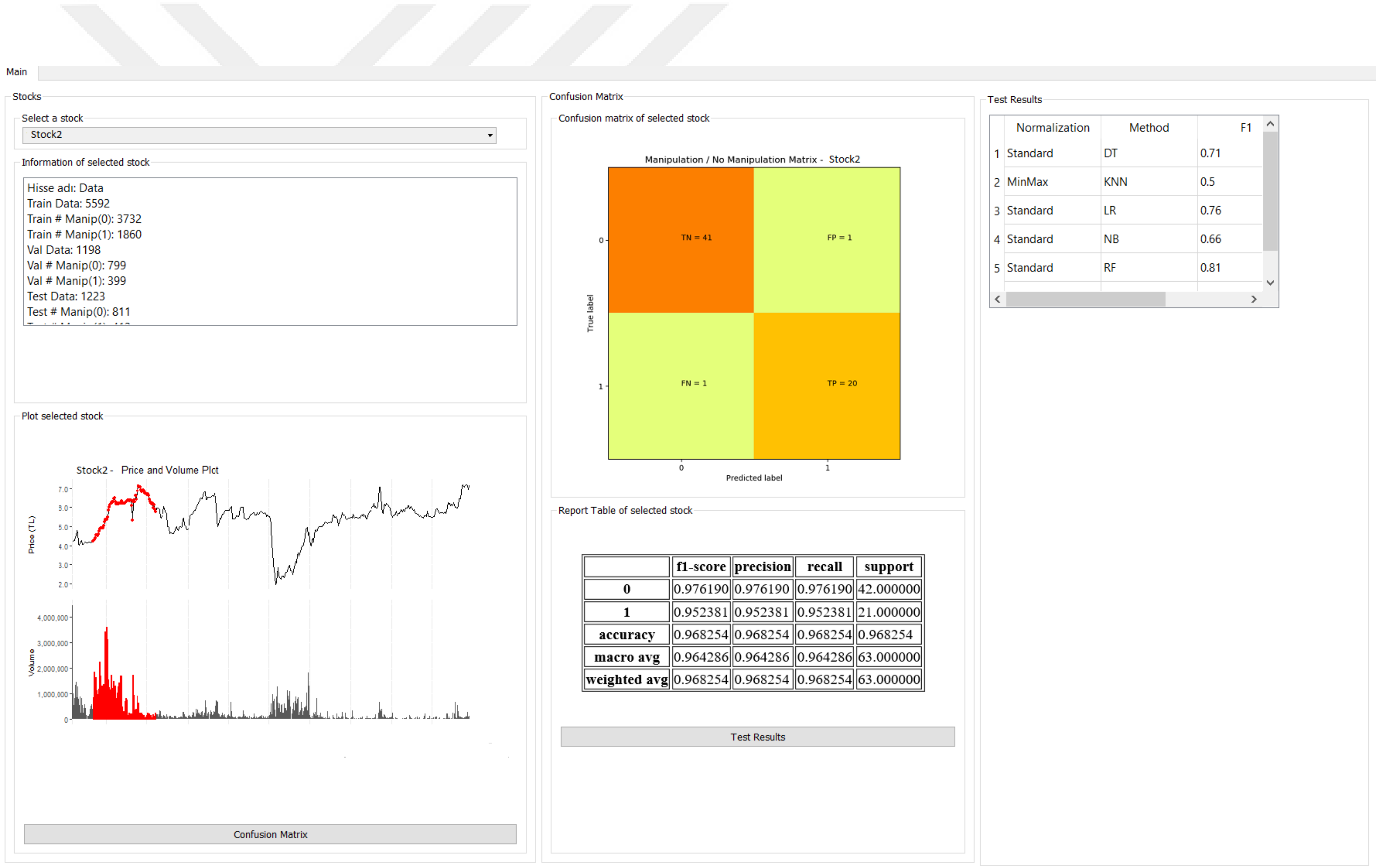
Model List	***Scaling Method***	***F1 Score List***
1. DTC_mm	Minmax	F1_1
2. DTC_s	Standard	F1_2
3. LR_mm	Minmax	F1_3
4. LR_s	Standard	F1_4
5. KNN_mm	Minmax	F1_5
6. KNN_s	Standard	F1_6
7. RF_mm	Minmax	F1_7
8. RF_s	Standard	F1_8
9. NB_mm	Minmax	F1_9
10. NB_s	Standard	F1_10
11. SVM_mm	Minmax	F1_11
12. SVM_s	Standard	F1_12
13. ANN_mm	Minmax	F1_13
14. ANN_s	Standard	F1_14

5. Our proposed model selects the model with the highest value from the f1 score list obtained on a per stock basis. Thus, the model with the best f1 value is selected for each stock. These selected models are listed in Table 4.7 and may consist of different supervised machine learning algorithms.
6. As the output of this phase, we get a value of 0 or 1 to show whether there is daily trade-based manipulation on a stock basis. An output value of 0 indicates that there is no manipulation in the stock in question, while 1 indicates that manipulation has taken place. In addition, the performance metric values (f1, sensitivity, specificity, precision, recall, and accuracy) of the selected model are presented with the output value.

4.2.8 Program Output

It is important to find solutions to the problems encountered while developing applications in machine learning. Thus, it is necessary to determine the environment and programming language in which the applications will be developed. The open source language Python is used in most studies on machine learning. An interface was developed using Python. Python is used on the Anaconda platform, an integrated platform prepared for those using Python or R for application development in areas such as data science and machine learning, and includes many ready-made packages. In addition to libraries that are frequently used in data science, machine learning, and artificial intelligence studies, this platform also includes development tools such as Spyder and Jupiter Notebook. Spyder was used herein.

The interface shows to what extent the proposed model can detect manipulations in the selected stock visually and numerically (Figure 4.4).



	f1-score	precision	recall	support
0	0.976190	0.976190	0.976190	42.000000
1	0.952381	0.952381	0.952381	21.000000
accuracy	0.968254	0.968254	0.968254	0.968254
macro avg	0.964286	0.964286	0.964286	63.000000
weighted avg	0.968254	0.968254	0.968254	63.000000

Figure 4.4 Output sample

The explanations of the program output shown in Figure 4.4 are as follows, consecutively:

1. Selection of the stock.
2. Information about the training, validation, and test datasets for the selected stock.
    a. Number of manipulative and non-manipulative transactions in the training dataset.
    b. Number of manipulative and non-manipulative transactions in the validation dataset.
    c. Number of manipulative and non-manipulative transactions in the test dataset.
3. Drawing the daily volume–price graph of the selected stock.
    a. The days of trade-based manipulations are shown with red dots.
    b. There is a sudden increase or decrease in the trading volume and price on the days when the manipulation takes place (dots with red dots).
4. It is a summary of the prediction phase, and the f1 values obtained as a result of testing the 14 models obtained after the evaluation phase with the test data are shown.
5. A confusion matrix example is given for the model chosen according to the model proposed in the prediction phase. The confusion matrices for the proposed model obtained are included in the appendix at the end of the thesis.
6. It shows the performance metrics of the selected model according to the model proposed in the prediction phase.

5 EXPERIMENTAL ANALYSIS

Our model consists of DTC, LR, KNN, RF, NB, SVM, and ANN. Transactions subject to trade-based manipulations in 2010-2015 were used as the dataset, which was obtained raw from the CMB Information Systems database. We used a methodology similar to CRISP-DM consisting of 6 steps. The preprocessing step includes the operations we perform on the dataset. The processes of defining the problem and understanding, preparing, selecting, cleaning, integrating, and transforming the data are performed in this step. In the scaling stage, the values of the numerical features in the dataset were scaled to a certain value range to prevent the features with higher numerical values in the dataset from gaining superiority over other features in the modeling stage. The training phase is the modeling phase and models are built for solving the defined problem. The evaluation and feature engineering stages are related. In the evaluation stage, the f1 scores are obtained by testing the model with the validation data and then these scores are evaluated. The prediction phase is where the approach of our proposed model is explained and it is the last step of our methodology. In this step, the final version of the models obtained in the evaluation phase is tested with the test data and a binary number (0 or 1) is output. If the output values are 0, there is no manipulation; if it is 1, manipulation has taken place. At this stage, performance metrics for our proposed model are presented by comparing the obtained 0 and 1 output numbers with the actual 0 and 1 output numbers in our dataset.

In this section, the performances of the models are presented in tables, both on the basis of stocks and the average values of the performance metrics for the whole dataset. As described in the previous section, we obtained two different f1 scores, that is, two different models, since the models obtained in the evaluation phase were tested with two different test datasets on a stock basis. Using the approach in our model, the tables in this section present the model with the best f1 score per stock. The performance metrics (*f1 score, precision, sensitivity, specificity, accuracy*) of the model are shown on a per stock basis. In addition, the performance metrics with the highest values on a column basis are shown in bold, while those with the lowest values are shown in italics and underlined. The last column contains the data scaling method used while building the model. In the last line, the average performance metric values are shown in red, considering all the stocks in the model's dataset. At the end of this section, f1 is compared between our model and the supervised machine learning models DTC, LR, KNN, RF, NB, SVM, and ANN.

Table 5.1 shows the performance metrics obtained by testing the trained model using the DTC. The values obtained by testing all stocks in the dataset with the model built using only the DTC are as follows:

• F1 score: 0.70

• Precision: 0.65

• Sensitivity: 0.79

• Specificity: 0.74

• Accuracy: 0.75

Table 5.1 Test results for the model created using the DTC algorithm

Stock Name	***F1 Score***	***Precision***	***Sensitivity***	***Specificity***	***Accuracy***	***Feature Scaling***
Stock1	**0.83**	**0.88**	0.78	0.79	0.78	Standard
Stock2	0.71	0.66	0.76	0.80	0.78	Standard
Stock3	0.76	0.76	0.76	0.88	0.84	Standard
Stock4	0.76	0.65	0.90	0.76	0.80	Standard
Stock5	*0.55*	*0.38*	**1.00**	*0.17*	*0.45*	MinMax
Stock6	0.78	0.75	0.81	0.86	0.84	Standard
Stock7	0.64	0.66	0.62	0.84	0.76	Standard
Stock8	0.78	0.76	0.81	0.87	0.85	Standard
Stock9	0.72	0.76	0.69	0.89	0.82	Standard
Stock10	0.68	0.62	0.76	0.77	0.76	Standard
Stock11	0.68	0.52	**1.00**	0.52	0.68	MinMax
Stock12	0.61	0.54	0.70	0.71	0.70	Standard
Stock13	0.61	0.61	*0.61*	0.80	0.73	Standard
Stock14	0.65	0.60	0.71	0.76	0.74	MinMax
Stock15	0.59	0.53	0.66	0.70	0.68	Standard
Stock16	0.70	0.63	0.80	0.77	0.77	Standard
Stock17	0.80	0.76	0.86	0.86	0.86	Standard
Stock18	0.75	0.71	0.78	0.84	0.82	MinMax
Stock19	0.82	**0.88**	0.76	**0.95**	**0.88**	Standard
Stock20	0.59	0.42	**1.00**	0.32	0.54	MinMax
Avg	0.70	0.65	0.79	0.74	0.75	

The data scaling column in Table 5.1 shows that *standard* scaling is chosen 15 times and *minmax* scaling 5 times, showing that *standard* scaling is more appropriate for this model. The highest f1 score value is 0.83, the highest accuracy value is 0.88, the lowest f1 score value is 0.55, and the lowest accuracy value is 0.45.

Table 5.2 shows the performance metrics obtained by testing the trained model using the LR. The values obtained by testing all stocks in the dataset with the model built using only the LR are as follows:

• F1 score: 0.82

• Precision: 0.85

• Sensitivity: 0.83

• Specificity: 0.89

• Accuracy: 0.87

Table 5.2 Test results for the model created using the LR algorithm

Stock Name	***F1 Score***	***Precision***	***Sensitivity***	***Specificity***	***Accuracy***	***Feature Scaling***
Stock1	0.91	0.91	0.91	0.96	0.94	Standard
Stock2	0.76	0.76	0.76	0.88	0.84	Standard
Stock3	0.80	0.90	0.71	0.96	0.88	Standard
Stock4	0.91	0.84	**1.00**	0.90	0.93	Minmax
Stock5	0.68	0.90	0.55	0.97	0.83	Standard
Stock6	0.95	**1.00**	0.90	**1.00**	**0.96**	Standard
Stock7	0.88	**1.00**	0.79	**1.00**	0.93	Standard
Stock8	0.94	0.88	**1.00**	0.94	0.95	Minmax
Stock9	0.80	0.94	0.69	0.98	0.88	Standard
Stock10	0.83	0.90	0.76	0.96	0.89	Standard
Stock11	0.52	*0.35*	**1.00**	*0.04*	*0.37*	Minmax
Stock12	0.77	0.85	0.70	0.94	0.86	Standard
Stock13	0.92	0.85	**1.00**	0.91	0.94	Standard
Stock14	**0.95**	0.91	**1.00**	0.95	**0.96**	Minmax
Stock15	0.82	0.70	**1.00**	0.78	0.85	Minmax
Stock16	0.57	**1.00**	*0.40*	**1.00**	0.80	Standard
Stock17	0.91	0.87	0.95	0.93	0.93	Standard
Stock18	0.90	0.82	**1.00**	0.89	0.92	Minmax
Stock19	0.87	0.90	0.85	0.95	0.91	Standard
Stock20	*0.66*	0.76	0.59	0.90	0.80	Standard
Avg	0.82	0.85	0.83	0.89	0.87	

The data scaling column in Table 5.2 shows that *standard* scaling is chosen 14 times and *minmax* scaling 6 times, showing that *standard* scaling is more appropriate for this model. The highest f1 score value is 0.86, the highest accuracy value is 0.90, the lowest f1 score value is 0.42, and the lowest accuracy value is 0.47.

Table 5.3 shows the performance metrics obtained by testing the trained model using the KNN. The values obtained by testing all stocks in the dataset with the model built using only the KNN are as follows:

• F1 score: 0.69

• Precision: 0.67

• Sensitivity: 0.76

• Specificity: 0.76

• Accuracy: 0.76

Table 5.3 Test results for the model created using the KNN algorithm

Stock Name	***F1 Score***	***Precision***	***Sensitivity***	***Specificity***	***Accuracy***	***Feature Scaling***
Stock1	0.78	0.94	0.66	0.98	0.87	Standard
Stock2	0.50	*0.37*	0.76	*0.32*	*0.47*	MinMax
Stock3	0.67	0.65	0.69	0.82	0.77	Standard
Stock4	0.68	0.70	0.66	0.86	0.79	MinMax
Stock5	0.66	0.61	0.72	0.77	0.75	Standard
Stock6	0.84	**1.00**	0.72	**1.00**	**0.90**	Standard
Stock7	0.65	0.73	0.58	0.89	0.79	Standard
Stock8	0.72	0.70	0.75	0.84	0.80	Standard
Stock9	0.83	0.80	0.86	0.89	0.88	MinMax
Stock10	0.70	0.77	0.65	0.90	0.82	Standard
Stock11	0.66	0.52	0.91	0.57	0.68	MinMax
Stock12	0.66	0.51	0.94	0.56	0.68	MinMax
Stock13	0.69	0.54	0.94	0.60	0.71	MinMax
Stock14	0.71	0.71	0.71	0.85	0.80	MinMax
Stock15	*0.42*	*0.37*	*0.50*	0.57	0.54	MinMax
Stock16	0.77	0.75	0.80	0.87	0.84	Standard
Stock17	**0.86**	0.83	0.90	0.91	**0.90**	Standard
Stock18	0.78	0.92	0.68	0.97	0.87	MinMax
Stock19	0.73	0.58	**1.00**	0.63	0.75	MinMax
Stock20	0.51	0.39	0.71	0.44	0.53	MinMax
Avg	0.69	0.67	0.76	0.76	0.76	

The data scaling column in Table 5.3 shows that *standard* scaling is chosen 9 times and *minmax* scaling 11 times, showing that *minmax* scaling is more appropriate for this model. The highest f1 score value is 0.86, the highest accuracy value is 0.90, the lowest f1 score value is 0.42, and the lowest accuracy value is 0.47.

Table 5.4 shows the performance metrics obtained by testing the trained model using the RF. The values obtained by testing all stocks in the dataset with the model built using only the RF are as follows:

• F1 score: 0.70

• Precision: 0.93

• Sensitivity: 0.58

• Specificity: 0.97

• Accuracy: 0.83

Table 5.4 Test results for the model created using the RF algorithm

Stock Name	***F1 Score***	***Precision***	***Sensitivity***	***Specificity***	***Accuracy***	***Feature Scaling***
Stock1	0.62	**1.00**	0.45	**1.00**	0.81	Standard
Stock2	0.81	**1.00**	0.69	**1.00**	0.89	Standard
Stock3	0.83	**1.00**	0.71	**1.00**	0.90	Standard
Stock4	0.80	**1.00**	0.66	**1.00**	0.88	Standard
Stock5	0.66	**1.00**	0.50	**1.00**	0.83	Standard
Stock6	0.90	**1.00**	**0.81**	**1.00**	**0.93**	Standard
Stock7	0.68	**1.00**	0.51	**1.00**	0.83	Standard
Stock8	**0.89**	**1.00**	0.81	**1.00**	**0.93**	Standard
Stock9	0.64	**1.00**	0.47	**1.00**	0.82	Minmax
Stock10	0.63	**1.00**	0.46	**1.00**	0.82	Standard
Stock11	*0.34*	*0.36*	*0.33*	*0.70*	*0.57*	Standard
Stock12	0.71	0.90	0.58	0.97	0.84	Standard
Stock13	0.84	0.93	0.77	0.97	0.90	Standard
Stock14	0.64	**1.00**	0.47	**1.00**	0.82	Standard
Stock15	0.66	**1.00**	0.50	**1.00**	0.82	Standard
Stock16	0.57	**1.00**	0.40	**1.00**	0.80	Standard
Stock17	0.84	**1.00**	0.72	**1.00**	0.90	Standard
Stock18	0.70	0.80	0.63	0.92	0.82	Minmax
Stock19	0.63	0.60	0.66	0.78	0.74	Minmax
Stock20	0.60	**1.00**	0.43	**1.00**	0.81	Standard
Avg	0.70	0.93	0.58	0.97	0.83	

The data scaling column in Table 5.4 shows that *standard* scaling is chosen 17 times and *minmax* scaling 3 times, showing that *standard* scaling is more appropriate for this model. The highest f1 score value is 0.89, the highest accuracy value is 0.93, the lowest f1 score value is 0.34, and the lowest accuracy value is 0.57.

Table 5.5 shows the performance metrics obtained by testing the trained model using the NB. The values obtained by testing all stocks in the dataset with the model built using only the NB are as follows:

- F1 score: 0.85
- Precision: 0.87
- Sensitivity: 0.84
- Specificity: 0.93
- Accuracy: 0.90

Table 5.5 Test results for the model created using the NB algorithm

Stock Name	***F1 Score***	***Precision***	***Sensitivity***	***Specificity***	***Accuracy***	***Feature Scaling***
Stock1	0.88	0.84	0.91	0.91	0.91	Standard
Stock2	*0.66*	*0.72*	*0.61*	0.88	*0.78*	Standard
Stock3	0.90	0.82	**1.00**	0.89	0.92	Standard
Stock4	0.89	**1.00**	0.80	**1.00**	0.93	Standard
Stock5	0.83	**1.00**	0.72	**1.00**	0.90	Standard
Stock6	0.90	**1.00**	0.81	**1.00**	0.93	Standard
Stock7	0.81	0.88	0.75	0.95	0.88	Standard
Stock8	0.89	**1.00**	0.81	**1.00**	0.93	Standard
Stock9	0.73	0.73	0.73	0.87	0.82	Standard
Stock10	0.90	0.88	0.92	0.94	0.93	Standard
Stock11	0.82	0.70	**1.00**	*0.78*	0.85	Standard
Stock12	**0.94**	0.94	0.94	0.97	**0.96**	Standard
Stock13	0.82	0.87	0.77	0.94	0.88	Standard
Stock14	0.90	0.86	0.95	0.93	0.93	Minmax
Stock15	0.92	0.85	**1.00**	0.91	0.94	Minmax
Stock16	0.83	0.81	0.86	0.90	0.88	Standard
Stock17	0.91	0.84	**1.00**	0.91	0.93	Standard
Stock18	0.81	**1.00**	0.68	**1.00**	0.89	Standard
Stock19	0.90	0.90	0.90	0.95	0.93	Standard
Stock20	0.72	0.75	0.68	0.89	0.82	Standard
Avg	0.85	0.87	0.84	0.93	0.90	

The data scaling column in Table 5.5 shows that *standard* scaling is chosen 18 times and *minmax* scaling twice, showing that *standard* scaling is more appropriate for this model. The highest f1 score value is 0.94, the highest accuracy value is 0.96, the lowest f1 score value is 0.66, and the lowest accuracy value is 0.78.

Table 5.6 shows the performance metrics obtained by testing the trained model using the SVM. The values obtained by testing all stocks in the dataset with the model built using only the SVM are as follows:

- F1 score: 0.70
- Precision: 0.93
- Sensitivity: 0.58
- Specificity: 0.97
- Accuracy: 0.83

Table 5.6 Test results for the model created using the SVM algorithm

Stock Name	***F1 Score***	***Precision***	***Sensitivity***	***Specificity***	***Accuracy***	***Feature Scaling***
Stock1	**0.96**	0.92	**1.00**	0.95	**0.97**	Minmax
Stock2	0.88	0.91	0.84	0.96	0.92	Standard
Stock3	0.88	0.79	**1.00**	0.87	0.91	Minmax
Stock4	0.95	0.95	0.95	0.98	0.96	Standard
Stock5	0.78	0.64	**1.00**	0.71	0.81	Minmax
Stock6	0.85	0.90	0.81	0.95	0.90	Standard
Stock7	0.93	0.90	0.96	0.95	0.95	Minmax
Stock8	**0.96**	**1.00**	0.93	**1.00**	**0.97**	Minmax
Stock9	0.81	0.85	0.78	0.93	0.88	Minmax
Stock10	**0.96**	0.92	**1.00**	0.96	**0.97**	Minmax
Stock11	0.77	0.63	**1.00**	0.70	0.80	Minmax
Stock12	0.75	*0.60*	**1.00**	*0.68*	0.78	Minmax
Stock13	*0.66*	0.61	0.72	0.77	*0.75*	Standard
Stock14	0.92	**1.00**	0.85	**1.00**	0.95	Minmax
Stock15	0.73	**1.00**	*0.58*	**1.00**	0.85	Minmax
Stock16	0.78	0.65	**1.00**	0.73	0.82	Minmax
Stock17	0.86	0.75	**1.00**	0.84	0.89	Minmax
Stock18	0.77	**1.00**	0.63	**1.00**	0.87	Minmax
Stock19	0.95	0.91	**1.00**	0.95	0.96	Minmax
Stock20	0.87	0.78	**1.00**	0.86	0.90	Minmax
Avg	0.85	0.84	0.90	0.89	0.89	

The data scaling column in Table 5.6 shows that *standard* scaling is chosen 4 times and *minmax* scaling 16 times, showing that *minmax* scaling is more appropriate for this model. The highest f1 score value is 0.96, the highest accuracy value is 0.97, the lowest f1 score value is 0.66, and the lowest accuracy value is 0.75.

Table 5.7 shows the performance metrics obtained by testing the trained model using the ANN. The values obtained by testing all stocks in the dataset with the model built using only the ANN are as follows:

• F1 score: 0.86

• Precision: 0.86

• Sensitivity: 0.87

• Specificity: 0.91

• Accuracy: 0.89

Table 5.7 Test results for the model created using the ANN algorithm

Stock Name	***F1 Score***	***Precision***	***Sensitivity***	***Specificity***	***Accuracy***	***Feature Scaling***
Stock1	0.96	0.96	0.96	0.98	0.97	Standard
Stock2	0.80	0.83	0.77	0.92	0.87	Standard
Stock3	0.87	0.88	0.86	0.94	0.91	Standard
Stock4	0.93	0.87	**1.00**	0.93	0.95	Minmax
Stock5	0.81	0.93	0.72	0.97	0.89	Standard
Stock6	0.90	**1.00**	0.82	**1.00**	0.94	Standard
Stock7	0.88	0.87	0.90	0.93	0.92	Standard
Stock8	**0.97**	**1.00**	0.94	**1.00**	**0.98**	Standard
Stock9	0.80	0.74	0.87	*0.84*	0.85	Minmax
Stock10	0.91	0.86	0.96	0.92	0.94	Standard
Stock11	*0.59*	0.41	**1.00**	0.26	*0.51*	Minmax
Stock12	0.80	0.92	0.71	0.97	0.88	Standard
Stock13	0.88	0.78	**1.00**	0.86	0.91	Standard
Stock14	0.98	0.95	**1.00**	0.98	0.98	Minmax
Stock15	0.86	**1.00**	0.75	**1.00**	0.91	Standard
Stock16	0.74	0.83	0.67	0.93	0.84	Standard
Stock17	0.96	0.92	**1.00**	0.95	0.97	Standard
Stock18	0.90	0.83	**1.00**	0.89	0.93	Minmax
Stock19	0.84	0.82	0.86	0.90	0.89	Standard
Stock20	0.74	0.84	*0.66*	0.94	0.84	Standard
Avg	0.86	0.86	0.87	0.91	0.89	

The data scaling column in Table 5.7 shows that *standard* scaling is chosen 4 times and *minmax* scaling 16 times, showing that *minmax* scaling is more appropriate for this model. The highest f1 score value is 0.97, the highest accuracy value is 0.98, the lowest f1 score value is 0.59, and the lowest accuracy value is 0.51.

When we evaluate the tables presented so far in this section together, the average f1 scores obtained by testing the various machine learning models with test data are as follows: 0.70 with DTC, 0.82 with LR, 0.69 with KNN, 0.70 with RF, 0.85 with NB, 0.85 with SVM, and 0.86 with ANN. Thus, the most successful result for these models can be obtained with ANN and the most unsuccessful result with KNN.

As explained in the previous section, our model consists of the supervised machine learning models DTC, LR, KNN, RF, NB, SVM, and ANN, and which of these models our model will include varies on a stock basis. In other words, while testing on a stock basis, each of these models is tested with test data and different f1 scores are obtained. Our model selects the model with the best f1 score. The performance metrics of our model are listed in Table 5.8.

Table 5.8 Test results for the purposed model

Stock Name	***Model Name***	***F1 Score***	***Precision***	***Sensitivity***	***Specificity***	***Accuracy***	***Feature Scaling***
Stock1	ANN	0.96	0.96	0.96	0.98	0.97	Standard
Stock2	SVM	0.88	0.91	0.84	0.96	0.92	Standard
Stock3	NB	0.90	0.82	**1.00**	0.89	0.92	Standard
Stock4	SVM	0.95	0.95	0.95	0.98	0.96	Standard
Stock5	NB	0.83	**1.00**	*0.72*	**1.00**	0.90	Standard
Stock6	LR	0.95	**1.00**	0.90	**1.00**	0.96	Standard
Stock7	SVM	0.93	0.90	0.96	0.95	0.95	Minmax
Stock8	ANN	0.97	**1.00**	0.94	**1.00**	**0.98**	Standard
Stock9	KNN	0.83	0.80	0.86	0.89	0.88	Minmax
Stock10	SVM	0.96	0.92	**1.00**	0.96	0.97	Minmax
Stock11	NB	*0.82*	*0.70*	**1.00**	*0.78*	*0.85*	Standard
Stock12	NB	0.94	0.94	0.94	0.97	0.96	Standard
Stock13	LR	0.92	0.85	**1.00**	0.91	0.94	Standard
Stock14	ANN	**0.98**	0.95	**1.00**	0.98	**0.98**	Minmax
Stock15	NB	0.92	0.85	**1.00**	0.91	0.94	Minmax
Stock16	NB	0.83	0.81	0.86	0.90	0.88	Standard
Stock17	ANN	0.96	0.92	**1.00**	0.95	0.97	Standard
Stock18	ANN	0.90	0.83	**1.00**	0.89	0.93	Minmax
Stock19	SVM	0.95	0.91	**1.00**	0.95	0.96	Minmax
Stock20	SVM	0.87	0.78	**1.00**	0.86	0.90	Minmax
	Avg	0.91	0.89	0.94	0.93	0.93	

The values obtained by testing all stocks in the dataset with the model built using the proposed model are as follows:

• F1 score: 0.91
• Precision: 0.89
• Sensitivity: 0.94
• Specificity: 0.93
• Accuracy: 0.93

The data scaling column in Table 5.8 reveals that *standard* scaling was chosen 12 times and *minmax* scaling 8 times among the data scaling methods used. The highest f1 score value is 0.98, the highest accuracy value is 0.98, the lowest f1 score value is 0.82, and the lowest accuracy value is 0.85. The proposed model selected the ANN model 5 times, the SVM model 6 times, the NB model 6 times, the LR model twice, and the KNN model once, and the RF and DTC models were never selected. Therefore, it is clear that the ANN, SVM, NB, and LR models are better than the others for detecting anomalies from daily trading data for the manipulated stocks.

Table 5.9 shows the performance metrics for manipulation detection of all models described so far. According to the f1 scores, the order of performance in detecting market manipulation is KNN, DTC, RF, LR, NB, SVM, ANN, and our model in ascending order. The f1 score of our model is 5% greater than that of the closest ANN and 21% greater than that of the farthest KNN. Likewise, the accuracy of our model is 3% greater than that of the NB closest to it and 18% larger than that of the farthest DTC. The ANN, SVM, NB, and LR models have very high performance in detecting trade-based manipulation. The f1 score of the ANN is 1% greater than that of the nearest SVM and 14% greater than that of the farthest DTC. The f1 score of SVM is 16% larger than that of the farthest KNN and has the same values as NB. The f1 score of NB is 16% greater than that of the farthest KNN. The f1 score of LR is 12% greater than that of the nearest DTC and RF and 13% greater than that of the farthest KNN. The f1 score values of DTC and RF are equal to each other and 1% greater than that of KNN.

Table 5.9 Comparison of the performance of the models

Model Name	***F1 Score***	***Precision***	***Sensitivity***	***Specificity***	***Accuracy***
DTC	0.70	0.65	0.79	0.74	0.75
LR	0.82	0.85	0.83	0.89	0.87
KNN	0.69	0.67	**0.76**	0.76	0.76
RF	0.70	0.93	0.58	0.97	0.83
NB	0.85	0.87	0.84	0.93	0.90
SVM	0.85	0.84	0.90	0.89	0.89
ANN	0.86	0.86	0.87	0.91	0.89
Our Model	**0.91**	0.89	0.95	0.93	**0.93**

Figure 5.1 compares the supervised machine learning models to detect the daily trading data of the manipulated stock used in trade-based manipulation detection, with f1 score values of 82% and above for ANN, LR, NB, and SVM. They perform exceptionally in terms of precision and accuracy of the performance indexes. Furthermore, Figure 5.1 clearly indicates that ANN, LR, NB, and SVM are better than DTC, RF, and KNN for market manipulation detection.

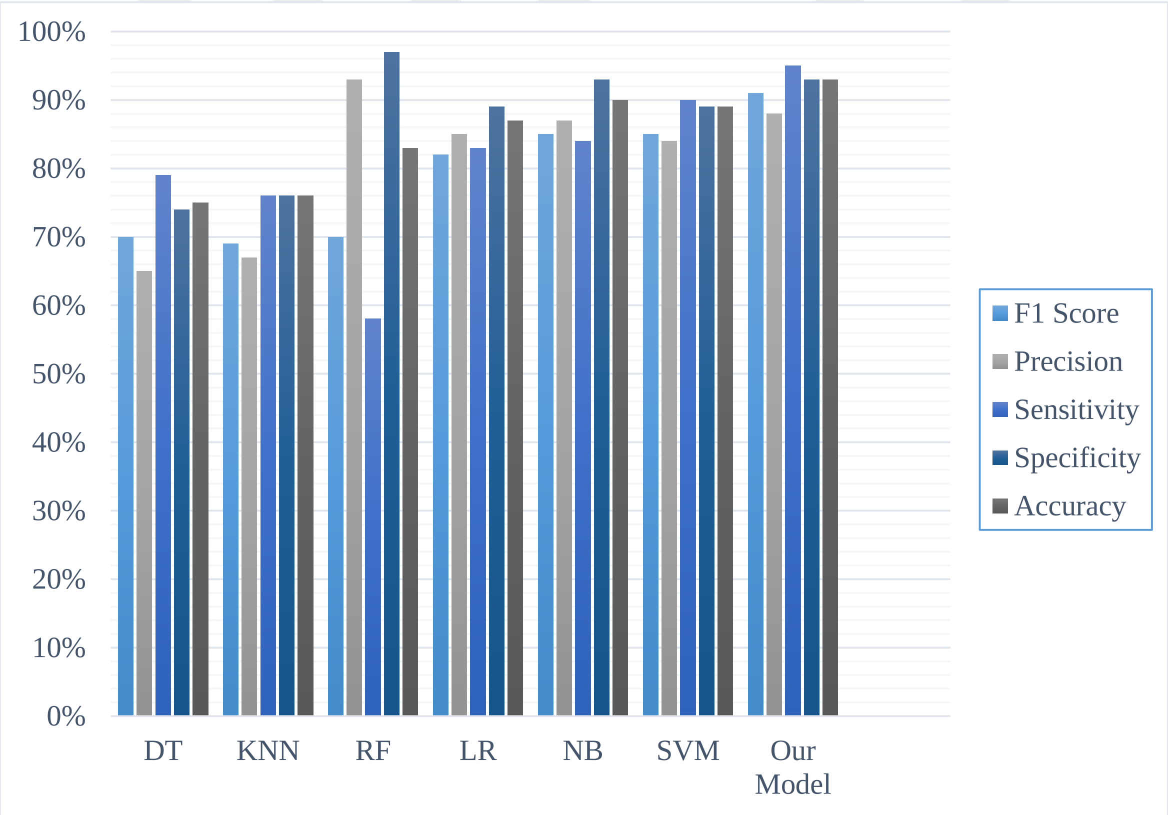


Figure 5.1 Trade-based manipulation detection performance of machine learning models

Moreover, our model has a higher f1 score than ANN, LR, NB, and SVM for detecting trade-based manipulation from the daily data. The experimental results showed that supervised machine learning models are suitable for daily trading data and have high market manipulation detection performance such as 86% f1 score, 90% sensitivity, and 90% accuracy for some supervised machine learning models. As a result of testing the daily trading data with these algorithms, the f1 scores obtained by the models were KNN 69%, DTC 70%, RF 70%, LR 82%, NB 85%, SVM 85%, ANN 86%, and our model 91%.

6 CONCLUSIONS

In this research, trade-based manipulations carried out in the capital market were examined. Our model consisted of supervised machine learning classification models to detect manipulations that took place in BIST between 2010 and 2015. We trained our model using daily trading data of manipulated stocks. The experimental results showed that it performed well in detecting anomalies from daily trading data.

Our research differs from the studies on manipulation in the literature in the following ways. First of all, studies on this subject in the literature are mostly theoretical and detect general manipulation by measuring variables such as volatility, return, quantity, volume, and stock turnover rate. These values of a manipulated capital market instrument are compared between the manipulative period and the non-manipulative period. The focus of our study is to predict whether manipulation has occurred or not. In order to determine manipulation, we calculated a wide variety of parameters for each stock on a daily basis. These parameters were not limited to return, amount, and volume data. After detailed analysis of the price formation rules in the stock markets, we searched for more suitable parameters to detect trade-based manipulations and included them in our model. For the reasons explained above, in addition to daily volume, amount, and return data, the effects of the transactions on the price, and whether the transactions are active or passive were also included. Thus, whether a transaction increases the price, decreases the price, or occurs at the previous price; number, volume, and quantities of transactions; and whether buying or selling is an active or passive transaction were the inputs of our model. Since the main focus of the manipulation in the stock market is the intraday price changes and the effects of the transactions on the price, it is important to use of the features of the price effect for the performance of the model. Secondly, while most of the studies on manipulation use synthetic data, we used transaction data from BIST between 2010 and 2015. Thirdly, while the features in the dataset used in other studies are quite limited, our dataset includes all the features related to the transaction data in BIST. Finally, our explanatory variables for stock market manipulation differ from those reported in other studies.

We proposed a new model with a machine learning approach for detecting anomalies from the daily data from manipulated stocks in BIST. The model consists of 7 different supervised machine learning classification algorithms. We increased the performance of our model by using two different scaling methods. We obtained two different training datasets, test and validation, through two different types of scaling, and we created the model by selecting the algorithm with the highest performance. Our model is suitable for detecting manipulations in daily trading data and has high detection performance with 91% f1 score, 95% sensitivity, and 93% accuracy.

Because of this strong performance, our work will aid regulators, particularly in detecting stock market manipulations. As a result of this research, stock market manipulations, the most significant barrier to investors making safe investments in the capital markets, can be reduced. Thus, transparency and trust, which are the foundations of capital market formation and development, will be established. Furthermore, it is critical to perform large studies in the future to study investor behavior patterns, since this will provide fresh insights into manipulation mechanisms from the perspective of trading behavior. Such insights would allow far faster detection of trade-based manipulations in the exchange than is currently attainable with the extant tools. This will make it easier to improve BIST's online surveillance systems.

REFERENCES

[1] M. Yasin, *Capital Markets Board of Turkey and Operations*. Seckin, 2002.

[2] N. Ergul, *Finance for Everyone*, 1. baskı. Istanbul: Literatur, 2004.

[3] F. S. Mishkin, *The Economics of Money, Banking, and Financial Markets*, 7. baskı. Pearson/Addison-Wesley, 2004.

[4] M. Gunal, *Money Bank and Financial System*, 4. baskı. Yeni Donem, 2012.

[5] C. Manavgat, *Trade-based Manipulation in the Capital Markets and Results in terms of Private Law*, Banking an. Ankara, 2008.

[6] S. G. Eakins, *Finance: Investments, Institutions, and Management*. AddisonWesley Educational Publishers Inc., 1999.

[7] P. S. Rose, *Money and Capital Markets: Financial Institutions and Instruments in a Global Marketplace*, 8. baskı. McGraw-Hill Higher Education, 2003.

[8] F. Fabozzi, F. J., Modigliani, *Capital Markets: Institutions and Instruments*. Prentice Hall, 1996.

[9] F. Seyidoglu, *International Finance*, 4. baskı. Guzem Can, 2003.

[10] K. Tunay, *Financial System: Structure, Functioning, Management and Economy*. Birsen, 2005.

[11] Sariaslan, H., & Erol, C., *Financial Management: Concepts, Institutions and Principles*. Ankara: Siyasal, 2008.

[12] E. Gemici, "The Analysis of Manipulations in Equity Market: An Applicatıon in Istanbul Stock Exchange", Osmaniye Korkut Ata University, 2016.

[13] A. Spitzer, J., & Sokolow, "Regulation of Stock Market Manipulation Comments", *Yale Law J.*, c. 56, ss. 509–533, 1946.

[14] P. Gerace, D., Chew, C., Whittaker, C., & Mazzola, "Stock Market Manipulation on the Hong Kong Stock Exchange", *Australas. Accounting, Bus. Financ. J.*, c. 8, sayı 4, ss. 105–140, 2014.

[15] B. Duman, "The Offense of Market Fraud Regulated in 6362 Numbered Capital Market Law", Kırıkkale University, 2020.

[16] A. Alkhamees, "Private action as a remedy against market manipulation in the USA", *J. Financ. Regul. Compliance*, c. 20, sayı 1, ss. 41–45, 2012.

[17] G. Allen, F., and Gorton, "Stock price manipulation, market microstructure and asymmetric information", *Eur. Econ. Rev.*, c. 36, sayı 2–3, ss. 624–630, 1992.

[18] CMB, "Capital Market Law", *http://cmb.gov.tr/Sayfa/Dosya/87*, 2012. http://cmb.gov.tr/Sayfa/Dosya/87 (erişim Kas. 21, 2021).

[19] R. Ozcan, "Manipulation Strategies in Stock Markets", *Istanbul Stock Exch. Mag.*, c. 49, ss. 19–41, 2013.

[20] Z. S. Karatepe, "Use of Extended Cox Models in the Analysis of Recurrent Events: A study on the Relationship between Trade-Based Manipulation and Financial Ratios", Yalova University, 2017.

[21] BIST, “Equity Market”, 2021. https://www.borsaistanbul.com/en/sayfa/2854/equity-market (erişim Kas. 21, 2021).

[22] BIST, “Electronic Trading System”, 2021. https://www.borsaistanbul.com/en/sayfa/2521/bistech-technology.

[23] M. Bal, “The closing session in the stock market and the analysis of the effects on the price formation in Borsa İstanbul”, Marmara University, 2014.

[24] BIST, “Equity Market Directive”, 2021. [Çevrimiçi]. Available at: https://www.borsaistanbul.com/files/borsa-istanbul-as-pay-piyasasi-yonergesi.pdf.

[25] B. Temir, “Istanbul Stock Exchange”, *Revue d’économie financière*, 1994. https://www.borsaistanbul.com/en/sayfa/2854/equity-market.

[26] BIST, *Capital Markets Exchanges Basic Information Guide*, 17. baskı. Borsa Istanbul Publications, 2002.

[27] Y. Percin, “The Applicability of Borsa İstanbul’s Working Based on Out of Session Working Principles and Examples from Abroad”, 2017.

[28] BIST, “Order Types”, 2021. https://www.borsaistanbul.com/en/sayfa/2857/market-functioning.

[29] B. Yalciner, “Over the Counter Stock Markets”, Ankara, 1996.

[30] BIST, “Equity Market Procedure”, 2021. .

[31] BIST, “Market Making”, 2021. https://www.borsaistanbul.com/en/sayfa/3585/market-making.

[32] BIST, “Market Making and Liquidity Providing”, 2021. https://www.borsaistanbul.com/en/sayfa/4443/market-making-and-liquidity-providing.

[33] BIST, “Circuit Breaker”, 2021. https://www.borsaistanbul.com/en/sayfa/2553/circuit-breaker.

[34] BIST, “Manipulation in Capital Markets (Specialization Training Material)”, Istanbul, 2011.

[35] Bromberg, A.D. & Lowenfels, L.D., *Securities Fraud and Commodities Fraud*, Shepard’s/. Colorado, 1994.

[36] I. Ozparca, “Manipulation and Speculation in Stock Markets”, Marmara University, 2000.

[37] Uslu, N.C. & Akal, F., “A Machine Learning Approach to Detection of Trade-Based Manipulations in Borsa Istanbul”, *Comput. Econ.*, 2021.

[38] E. Ugurlu, “The Effects of the Press on Stock Market”, Marmara University, 2002.

[39] Tarkun, S., “An Forecast for Detectıon of the Trading Based Manipulation: Istanbul Stock Exchange Application”, Mugla Sitki Kocman University, 2015.

[40] Evik, A.H., *Offenses of Making Deceptive Acts (Manipulation) Affecting the Value of Capital Market Instruments*. Ankara: Seckin, 2004.

[41] M. Tezcanli, *Insider Trading and Manipulations*. Istanbul: Ufuk Advertising and Printing Ltd., 1996.

[42] Fischel, D. R., & Ross, D. J., “Should the Law Prohibit ‘Manipulation’ in Financial

Markets?”, *Harv. Law Rev.*, c. 105, sayı 2, ss. 503–553, 1991.

[43] Benabou, R. & Laroque, G., “Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility”, *Q. J. Econ.*, c. 107, sayı 3, 1992.

[44] S. Bayindir, “Manipulation Crime in Turkish Capital Market Law”, Marmara University, 2010.

[45] CESR, “Market Abuse Directive Level 3 – preliminary CESR guidance and information on the common operation of the Directive”, 2005.

[46] E. Yilmaz, *Law Dictionary*, 8th baskı. Ankara: Yetkin, 2011.

[47] Turkish Language Association, “Manipulation”, 2021. https://sozluk.gov.tr/.

[48] A. S. Koyuncugil, “Data Mining and Application to Capital Markets”, Ankara, 2007.

[49] E. Bostanci, F. & Kadioglu, “Capital Markets Crimes (Manipulation and Insider Trading)”, *Stand. J. Econ. Tech.*, sayı 564, s. 99, 2009.

[50] R. Ozbay, “Rises and Falls in Stock Prices: Speculation and Manipulation in Stock Markets”, Ankara, 1990.

[51] N. Chambers, “Manipulation in Capital Markets and Its Examples in Borsa Istanbul”, *J. Account. Financ.*, ss. 62–72, 2004.

[52] K. Celik, “An Application on Financial Information Manipulation and Manipulatıon Detection in Borsa Istanbul”, Hitit University, 2016.

[53] CMB, “Manipulation in the Stock Market: Examples of Methods Used Examples of Manipulative Process Patterns Ways of Protection”, Ankara, 2003.

[54] Tuzcu, M.A., “Factors Affecting Stock Prices and Volatility in Borsa Istanbul”, Ankara University, 1999.

[55] Kamisli, M., “Determination of Trade Based Manipulation by Financial Ratios: An Application in the Istanbul Stock Exchange with Statistical Classification Analysis”, Eskisehir Osmangazi University, 2008.

[56] Jennings, R.W., Marsh, H., & Coffee, M., *The Securities Regulation: Cases and Materials*. Mineola: Newyork: Foundation Press, 1982.

[57] A. A. Yuce, “Manipulation in the Capital Market”, *Turkey Bar Assoc. J.*, ss. 363–388, 2012.

[58] E. Akanak, “XXI. Century Speculation and Manipulation Applications”, Atilim University, 2013.

[59] Turkish Language Association, “Speculation”, 2021. https://sozluk.gov.tr/.

[60] I. Kalayci, “Financial Infection and ‘Vaccination’ Theory: A Critique of General Crisis”, *Financ. J.*, sayı 160, ss. 25–56, 2011.

[61] G. Feiger, “What is Speculation?”, *Q. J. Econ.*, c. 90, sayı 4, ss. 677–687, 1976.

[62] H. Ok, “Analysis of Stock Market Manipulation Process: Istanbul Stock Exchange on an Empirical Application”, Eskisehir Osmangazi University, 2016.

[63] D. Allen, F., and Gale, “Stock-Price Manipulation”, *Rev. Financ. Stud.*, c. 5, sayı 3, ss. 503–529, 1992.

[64] K. Celik, “Manipulating the Financial Information and BIST (Istanbul Stock

Exchange) Practice towards the Detection of Financial Information Manipulation", Hitit University, 2016.

[65] Chatterjea, A., Cherian, J. A., & Jarrow, R. A., "Market Manipulation and Corporate Finance: A New Perspective", *Financ. Manag.*, c. 22, sayı 2, s. 200, 1993.

[66] G. Kucukkocaoglu, "Intra-day Stock Returns, Volatility and Close-end Price Manipulation in the Istanbul Stock Exchange", Baskent University, 2003.

[67] C. Korsmo, "High-Frequency Trading: A Regulatory Strategy", *Univ. Richmond Law Rev.*, c. 48, sayı 2, ss. 523–610, 2014.

[68] Stiglitz, J. E., & Ocampo, J. A., *Capital Market Liberalization and Development*. OUP Oxford, 2008.

[69] Mei, J., Wu, G., & Zhou, C., "Behavior Based Manipulation: Theory and Prosecution Evidence", 2004.

[70] D. Sensoy, "Manipulation; Market Fraud Crime, Measures to be Applied and Sanctions", *Ankara Bar Assoc. Mag.*, c. 3, ss. 371–400, 2013.

[71] V. Yanli, "Information-based Manipulation Crimes in the Framework of Capital Law", *J. Bank. Commer. Law*, c. 22, sayı 4, ss. 23–44, 2004.

[72] E. Kadioglu, "Effects of Changes in Micro-structure of Borsa İstanbul on the Intraday Returns and Volatility and Closing Prices", Baskent University, 2014.

[73] A. Z. Sayar, "Speculation and Manipulation in the Stock Markets; A Case Study of Istanbul Stock Exchange", Ankara University, 1995.

[74] P. Diaz, D., Theodoulidis, B., & Sampaio, "Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices.", *Expert Syst. Appl.*, c. 38, sayı 10, ss. 12757–12771, 2011.

[75] R. A. Jarrow, "Market Manipulation, Bubbles, Corners, and Short Squeezes", *J. Financ. Quant. Anal.*, c. 27, sayı 3, ss. 311–336, 1992.

[76] A. C. Yenidunya, "General Principles Regarding Crimes and Mades Regarded By Capital Markets Law", 2012.

[77] A. F. Tarkun, S., Ergur, H.O., & Aydin, "Investigation of Trade-based Manipulation Companies by Vector Autoregressive Analysis", *J. Acad. Approaches*, c. 5, sayı 1, 2014.

[78] G. Canbulut, "Financial Information Manipulation and a Case Study", Dokuz Eylul University, 2008.

[79] P. M. B. Hazen, T.L. & Johnson, *Derivatives Regulation*. Wolters Kluwer Law & Business., 2004.

[80] Lee, E. J., Eom, K. S., & Park, K. S., "Microstructure-based Manipulation: Strategic Behavior and Performance of Spoofing Traders.", *J. Financ. Mark.*, c. 16, sayı 2, ss. 227–252, 2013.

[81] V. R. Goldwasser, "Regulating Manipulation in Securities Markets: Historical Perspectives and Policy Rationales", *Aust. J. Leg. Hist.*, c. 5, ss. 149–200, 1999.

[82] Brodowski, D., de los Monteros de la Parr, M.E., Tiedemann, K., & Vogel, J., *Regulating Corporate Criminal Liability*. Springer International Publishing, 2014.

[83] Cassim, R., "An Analysis of Market Manipulation under the Securities Services Act

36 of 2004 (Part 1)”, *South African Merc. Law J. 33*, c. 20, sayı 1, 2008.

[84] Cao, Y., Li, Y., Coleman, S., Belatreche, A., & McGinnity, T. M., “Detecting price manipulation in the financial market”, *2014 IEEE Conf. Comput. Intell. Financ. Eng. Econ.*, ss. 77–84, 2014, doi: 10.1109/CIFEr.2014.6924057.

[85] Teweles, R. J., & Bradley, E. S., *The Stock Market*. Wiley, 1998.

[86] Eiteman, W.J., Dice, C.A., & Eiteman, D.K., Eiteman, W.J., Dice, C.A., & Eiteman, D.K., *The Stock Market*. McGraw-Hill, 1996.

[87] Kruse, T. A., & Todd, S. K., “Price Manipulation at the NYSE and the 1899 Battle for Brooklyn Rapid Transit Shares”, *Financ. Hist. Rev.*, c. 20, sayı 03, ss. 279–303, 2013, doi: 10.1017/s0968565013000218.

[88] N. S. Poser, “Stock Market Manipulation and Corporate Control Transactions”, *U. Miami L. Rev*, c. 40, ss. 671–735, 1986.

[89] Kaplan, M., & Beyoglu, C.U., *Market Fraud Crime in Turkish Capital Market Law*. Seckin, 2018.

[90] Altinbas, H., “An Examination of Closing Price Manipulation in Stock Market”, Dokuz Eylul University, 2012.

[91] Kutuk, H.I., “Manipulation in the Capital Market, Its Penal and Legal Outcomes”, Suleyman Demirel University, 2010.

[92] S. Karabacak, “Short Selling Operations”, Ankara, 2002.

[93] Egilmez, M., *Global Financial Crisis*. Istanbul: Remzi Bookstore, 2011.

[94] Cumming, D., Dannhauser, R., & Johan, S., “Financial Market Misconduct and Agency Conflicts: A Synthesis and Future Directions”, *J. Corp. Financ.*, c. 34, ss. 150–168, 2015.

[95] Cumming, D., & Johan, S., “Global Market Surveillance”, *Am. Law Econ. Rev.*, c. 10, sayı 2, ss. 454–506, 2008.

[96] Lowenfels, L. D., & Bromberg, A. R., “Securities Market Manipulations: An Examination and Analysis of Domination and Control, Frontrunning, and Parking”, *Albany Law Rev.*, c. 55, sayı 2, s. 293, 1991.

[97] Cross, F., & Miller, R., *The Legal Environment of Business: Text and Cases -- Ethical, Regulatory, Global, and E-Commerce Issues*. Cengage Learning, 2008.

[98] Khwaja, A. I., & Mian, A., “Unchecked Intermediaries: Price Manipulation in an Emerging Stock Market”, *J. financ. econ.*, c. 78, sayı 1, ss. 203–241, 2005.

[99] Kyle, A. S., & Viswanathan, S., “How to Define Illegal Price ManipulationNo Title”, *Am. Econ. Rev.*, 2008.

[100] Imisiker, S., & Tas, B. K. O., “Which Firms are More Prone to Stock Market Manipulation?”, *Emerg. Mark. Rev.*, c. 16, ss. 119–130, 2013.

[101] Turing, A.M., “Computing Machinery and Intelligence”, *Mind*, c. 49, ss. 433–460, 1950.

[102] Das, S., Dey, A., Pal, A. & Roy, N., “Applications of Artificial Intelligence in Machine Learning: Review and Prospect”, *Int. J. Comput. Appl.*, c. 115, sayı 9, ss. 31–41, 2015, doi: 10.5120/20182-2402.

[103] Kaya, C & Yildiz, O., “Intrusion Detection with Machine Learning Techniques:

Comparative Analysis”, *Marmara J. Sci.*, c. 3, ss. 89–104, 2014.

[104] Cui, G., Wong, M. L., & Lui, H. K., “Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming”, *Manage. Sci.*, c. 52, sayı 4, ss. 597–612, 2006.

[105] Lantz, B., “Equidistance of Likert-Type Scales and Validation of Inferential Methods Using Experiments and Simulations”, *Electron. J. Bus. Res. Methods*, c. 11, sayı 1, ss. 16–28, 2013.

[106] Patgiri, R., Hussain, S., & Nongmeikapam, A., “Empirical Study on Airline Delay Analysis and Prediction”, *arXiv Prepr. arXiv2002.10254*, 2020.

[107] Mitchell, T.M., “Machine Learning and Data Mining”, *Commun. ACM*, c. 42, sayı 11, ss. 30–36, 1999.

[108] Ogucu, M.N., “System Recognition with Artificial Neural Networks”, Istanbul Technical University, 2006.

[109] Forsyth, R., *Machine Learning: Principles and Techniques*, Chapman &. United Kingdom, 1989.

[110] Alpaydin, E., *Introduction to Machine Learning*. London: The MIT Press, 2014.

[111] Alpaydin, E., *Artificial Learning*. Istanbul: Bogazici University Publishing House, 2017.

[112] Nilsson, N. J., *Artificial Intelligence: A New Synthesis*. Burlington: Morgan Kaufmann Publishers, 1998.

[113] Huang, G., Zhu, Q., & Siew, C., “Extreme Learning Machine: Theory and Applications”, *Neurocomputing*, c. 70, sayı 1–3, ss. 489–501, 2006.

[114] Ayodele, T. O., “Types of Machine Learning Algorithms”, *New Adv. Mach. Learn.*, c. 3, ss. 19–48, 2010.

[115] Ozkan, Y., *Data Mining Methods*. Papatya Publisher, 2013.

[116] Chao, W. L ., “Machine learning Tutorial”, 2011.

[117] Marsland, S., *Machine Learning An Algorithmic Perspective*. New York: CRC Press Taylor & Francis Group, 2015.

[118] Hastie, T., Tibshirani, R., & Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. California: Springer Series in Statistics, 2001.

[119] Suthaharan, S., *Machine Learning Models and Algorithms for Big Data Classification*. New York: Springer International Publishing, 2016.

[120] Hinton, G., & Sejnowski, T. J., *Unsupervised Learning: Foundations of Neural Computation*. MIT Press, 1999.

[121] Efeoglu, E., “Explosive Detection Using Machine Learning Techniques From Microwave Measurement Data”, Trakya University, 2021.

[122] Celik, O., & Altunaydin, S. S., “A Research on Machine Learning Methods and Its Applications”, *J. Educ. Technol. Online Learn.*, c. 1, sayı 3, 2018.

[123] Huang, T. M., Kecman, V., & Kopriva, I., *Kernel Based Algorithms for Mining Huge Data Sets*. Springer, 2006.

[124] Amasyali, M.F., “New Machine Learning Methods and Their Applications to Drug Design”, Yildiz Technical University, 2008.

[125] Fawcett, T., “An introduction to ROC Analysis”, *Pattern Recognit. Lett.*, c. 27, sayı 8, ss. 861–874, 2006.

[126] Cakir, M., “Machine Learning Techniques in Identifying the Dynamics of Firm Failure: Empirical Applications and Comparative Analysis”, 2005.

[127] Ozekes, S., “Data Mining Models and Application Areas”, *Istanbul Commer. Univ. J. Sci.*, c. 2, sayı 3, ss. 65–82, 2003.

[128] Dunham, M.H., *Data Mining Introductory and Advanced Topics*. New Jersey: Prentice Hall/Pearson Education, 2003.

[129] Ngai, E. W., Xiu, L., & Chau, D. C., “Application of Data mining Techniques in Customer Relationship Management: A Literature Review and Classification”, *Expert Syst. Appl.*, c. 36, sayı 2, ss. 2592–2602, 2009.

[130] Akpinar, H., *Data Mining, Data Analysis*. Istanbul: Papatya Publisher, 2014.

[131] Alpaydin, E., *Introduction to Machine Learning*. MIT Press, 2020.

[132] Harrington, P., *Machine Learning in Action*. Manning Publications Shelter Island, 2012.

[133] Han, J., & Kamber, M., *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, 2006.

[134] Buyukkececi, M., “Evaluation of the Relationship Between the Stability of Feature Selection Techniques and Classification Performance in Data Mining”, Yasar University, 2019.

[135] Gursakal, N., *Machine Learning and Deep Learning*. Bursa: Dora Publishing and Distribution, 2017.

[136] JavaTpoint, “Logistic Regression in Machine Learning”, 2022. https://www.javatpoint.com/logistic-regression-in-machine-learning (erişim Şub. 04, 2022).

[137] Ogut, H., Doganay, M. M., & Aktas, R., “Detecting stock-price manipulation in an emerging market: The case of Turkey”, 2009.

[138] Girginer, N., & Cankus, B., “Measuring the Traveller Satisfaction of Tram Using Logistic Regression: A Case Study of Estram”, *Manag. Econ. J. Celal Bayar Univ. Fac. Econ. Adm. Sci.*, c. 15, sayı 1, ss. 181–193, 2008.

[139] Elhan, A.H., “Examination of Logistic Regression Analysis and an Application in Medicine”, Ankara University, 1997.

[140] Cortes, C., & Vapnik, V., “Support-Vector Networks”, *Mach. Learn.*, c. 20, sayı 3, ss. 273–297, 1995.

[141] Ayhan, S., & Erdoğmuş, S., “Kernel Function Selection for Solving Classification Problems with Support Vector Machines”, *Eskişehir Osmangazi Univ. J. Econ. Adm. Sci.*, c. 9, sayı 1, ss. 175–201, 2014.

[142] Bilgin, M., *Machine Learning Theory and Algorithms*. Papatya Publisher, 2018.

[143] Alpaydin, E., *Introduction to Machine Learning*. MIT Press, 2004.

[144] Filiz, E., “An Application on Machine Learning Methods and Training Data:

International Research on Mathematics and Science Trends 2015 The Case of Turkey”, 2019.

[145] Altman, N.S., “An Introduction to Kernel and Nearest-neighbor Nonparametric Regression”, *Am. Stat.*, c. 46, sayı 3, ss. 175–185, 1992.

[146] Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D., “Learning k for Knn Classification”, *ACM Trans. Intell. Syst. Technol.*, c. 8, sayı 3, ss. 1–19, 2017.

[147] Pala, M. A., Cimen, M. E., Boyraz, O. F., Yildiz, M. Z., & Boz, A. F., “Comparative Performance Analysis of Decision Tree and KNN Algorithms in the Diagnosis of Breast Cancer”, *Acad. Perspect. Procedia*, c. 2, sayı 3, ss. 544–552, 2019.

[148] Maimon, O., & Rokach, L. (Eds.), *Data Mining and Knowledge Discovery Handbook.* 2005.

[149] Balaban, M. E. & Kartal, E., *Data Mining and Machine Learning*. Caglayan Bookstore, 2015.

[150] Uysal, A.K., “New Approaches to Enhancing the Performance of Text Classification”, Anadolu University, 2013.

[151] Akar, O., & Gungor, G., “Classification of Multispectral Images Using Random Forest Algorithm.”, *J. Geod. Geoinf.*, ss. 139–146, 2012.

[152] Dogan, S., “N-gram Based Classification for Turkish Documents: Author, Genre and Gender”, Yildiz Technical University Institute of Science, 2006.

[153] E. Oztemel, *Artificail Neural Networks*. Papatya Publisher, 2012.

[154] Warner, B., & Misra, M., “Understanding Neural Networks as Statistical Tools”, *Am. Stat.*, c. 50, sayı 4, ss. 284–293, 1996.

[155] Chen, J., & Li, M., “Chained Predictions of Flight Delay Using Machine Learning”, içinde *In AIAA Scitech 2019 Forum*, 2019, s. 1661.

[156] Haykin, S., *Neural Networks: A Comprehensive Foundation*. New York: Prentice Hall PTR, 1994.

[157] Aydin, D.B., “Development of an Artificial Neural Network Based Decision Support System for Identifying Urinary Tract Infection in the Intensive Care Unit”, Bozok University, 2016.

[158] Fikir, “Structure, Function and Function of the Nervous System What is a Neuron?”, 2022. .

[159] LeCun, Y., Bengio, Y., & Hinton, G., “Deep Learning”, *Nature*, c. 521, sayı 7553, ss. 436–444, 2015.

[160] Uguz, S., *An Artificial Intelligence School with Machine Learning Theoretical Aspects and Python Applications*. Ankara: Nobel Publishing, 2019.

[161] Joshi, R., “Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures”, *Retrieved April*, c. 1, sayı 2018, 2016.

[162] Pillai, I., Fumera, G., & Roli, F., “Designing multi-label classifiers that maximize F measures: State of the art”, *Pattern Recognit. Lett.*, ss. 394–404, 2017.

[163] Shearer, C., “The CRISP-DM model: The New Blueprint for Data Mining”, *J. Data Warehous.*, c. 5, sayı 4, ss. 13–22, 2010.

[164] Wirth, R., & Hipp, J., “CRISP-DM: Towards a Standard Process Model for Data

Mining.", *Proc. 4th Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, ss. 29–39, 2000.

[165] Polat, A., "Examining Dropout and Graduation Status of Open High School Students Using Educational Data Mining", Sakarya University, 2021.

[166] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., *Step by Step Data Mining Guide*. SPSS inc., 2000.

[167] Tsiptsis, K. & Chorianopoulos, A., *Data Mining Techniques in CRM:Inside Customer Segmentation*. Wiley Publications ISBN: 978-0-470-74397-3, 2009.

[168] Doslu, A., "Market Basket Analysis and Determination of Association Rules in Data Mining", Yıldız Technical University, 2008.

[169] Ayik, Y.Z., Ozdemir, A., & Yavuz, U., "Analysis of the Relationship of High School Type and High School Graduation Success with the Faculty Earned By Data Mining Technique", *J. Ataturk Univ. Soc. Sci. Inst.*, c. 10, sayı 2, ss. 443–447, 2007.

[170] Raju, P.S., Bai, D.V.R., & Chaitanya, G.K., "Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries", *Int. J. Innov. Res. Comput. Commun. Eng.*, c. 2, sayı 1, ss. 2651–2653, 2014.

[171] Dondurmaci, G., & Cinar, A., "Data Mining Application in Finance Sector", *Acad. J. Soc. Stud.*, c. 2, sayı 1, ss. 258–271, 2014.

[172] Seker, S.E., & Esmekaya, E., "Completion of missing data (imputation)", *YBS*. ss. 10–17, 2017.

[173] Suthar, B., Patel, H., & Goswami, A., "A Survey: Classification of Imputation Methods in Data Mining", *Int. J. Emerg. Technol. Adv. Eng.*, c. 2, sayı 1, 2012.

[174] Namey, E., Guest, G., Thairu, L., & Johnson, L., *Handbook for Team-based Qualitative Research*. 2007.

[175] Kim, W., Choi, B. J., Hong, E. K., Kim, S. K., & Lee, D., "A Taxonomy of Dirty Data", *Data Min. Knowl. Discov.*, c. 7, sayı 1, ss. 81–99, 2003.

[176] Witten, I.H., Frank, E., & Hall, M.A., *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2005.

[177] Lomet, D.B., "Bulletin of the Technical Committee on Data Engineering", *Bull. Tech. Comm. Data Eng.*, c. 24, sayı 4, ss. 1–56, 2001.

[178] Kumar, P., & Seppi, D. J., "Futures Manipulation with 'Cash Settlement'", *J. Finance*, c. 47, sayı 4, ss. 1485–1502, 1992.

[179] Felixson, K., & Pelli, A., "Day end returns - Stock price manipulation.", *J. Multinatl. Financ. Manag.*, c. 9, sayı 2, ss. 95–127, 1999.

[180] Feng, Z., Rongqiu, C., and Xinping, X., "Fractal character of stock price-volume relation and regulation of stock price manipulation", *Fractals*, c. 11, sayı 2, ss. 173–181, 2003.

[181] Aggarwal, R.K., & Wu, G., "Stock Market Manipulations.", *J. Bus.*, c. 79, sayı 4, ss. 1915–1953, 2006.

[182] Aktas, R., and Doganay, M., "Stock-price manipulation in the Istanbul stock exchange", *Eurasian Rev. Econ. Financ.*, c. 2, ss. 21–28, 2006.

[183] A. S. Koyuncugil, "Fuzzy Data Mining and its application to capital markets",

Ankara University, 2006.

[184] Akyol, A., & Michayluk, D., "Is There Closing Price Manipulation on the Istanbul Stock Exchange", 2007.

[185] Huang, Y. C., Chen, R. C., & Cheng, Y. J., "Stock Manipulation and Its Impact on Market Quality", 2007.

[186] Mongkolnavin, J., & Tirapat, S., "Marking the close analysis in the Thai Bond Market Survaillance using association rules", *Expert Syst. Appl.*, c. 36, ss. 8523–8527, 2008.

[187] Palshikar, G.K., & Apte, M.M., "Collusion set detection using graph clustering", *Data Min. Knowl. Discov.*, c. 16, ss. 135–164, 2008.

[188] Comerton-Forde, C., & Putniņš, T. J., "Measuring Closing Price Manipulation", *J. Financ. Intermediation*, c. 20, ss. 135–158, 2011.

[189] Roodposhti, F. R., Shams, M. F., & Kordlouie, H., "Forecasting stock price manipulation in capital market.e", *World Acad. Sci. Eng. Technol.*, c. 80, ss. 151–161, 2011.

[190] Sun, X.-Q., Cheng, X.-Q., Shen, H.-W., & Wang, Z.-Y., "Distinguishing manipulated stocks via trading network analysis.", *Physica A*, c. 390, ss. 3427–3434, 2011.

[191] Kim, Y., & Sohn, Y.S., "Stock fraud detection using peer group analysis", *Expert Syst. Appl.*, c. 39, sayı 10, ss. 8986–8992, 2012.

[192] Cao, L., Ou, Y., & Yu, P.S., "Coupled Behavior Analysis with Applications", *IEEE Trans. Knowl. Data Eng.*, c. 24, sayı 8, ss. 1378–1392, 2012.

[193] Song, Y., Cao, L., Wu, X., Wei, G., Ye, W., & Ding, W., "Coupled Behavior Analysis for Capturing Coupling Relationships in Group-based Market Manipulations", içinde *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, ss. 976–984.

[194] Song, Y., & Cao, L., "Graph-based Coupled Behavior Analysis: A Case Study on Detecting Collaborative Manipulations in Stock Markets", içinde *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012, ss. 1–8.

[195] Murugesan, P. & Thoppan, J.J., "Detection of Stock Price Manipulation Using Discriminant Analysis", *Available SSRN 2700037*, 2012.

[196] Qiu, J., & Zhang, Y., "Effect of Short-Sale Constraints On Stock Price Manipulation", *Pacific Econ. Rev.*, c. 18, sayı 2, ss. 208–232, 2013.

[197] Golmohammadi, K., Zaiane, O.R., & Diaz, D., "Detecting Stock Market Manipulation using Supervised Learning Algorithms", içinde *The 2014 International Conference on Data Science and Advanced Analytics*, 2014, ss. 435–441.

[198] Imisiker, S., & Tas, B. K. O., "Wash Sales as a Stock Market Manipulation Tool", *Available SSRN 2476874*, 2014.

[199] Kong, D., and Wang, M., "The Manipulator's Poker: Order-Based Manipulation in the Chinese Stock Market.", *Emerg. Mark. Financ. Trade*, c. 50, sayı 5, ss. 73–98, 2014.

[200] Ozcomak, M. S., & Gunduz, M., "Analysis of the Relationship Between the Closing

Prices of the Companies Traded in Borsa Istanbul and the Transaction Amounts by Functional Canonical Correlation”, *Int. J. Econ. Adm. Stud.*, c. 7, sayı 13, ss. 233–252, 2014.

[201] Comerton-Forde, C., & Putniņš, T. J., “Stock Price Manipulation: Prevalence and Determinants”, *Rev. Financ.*, c. 18, sayı 1, ss. 23–66, 2014.

[202] Golmohammadi, K., & Zaiane, O.R., “Time Series Contextual Anomaly Detection for Detecting Market Manipulation in Stock Market”, *2015 IEEE Int. Conf. Data Sci. Adv. Anal.*, 2015.

[203] Huang, Y. C., & Cheng, Y. J., “Stock Manipulation and Its Effects: Pump and Dump versus Stabilization.”, *Rev. Quant. Financ. Account.*, c. 44, sayı 4, ss. 791–815, 2015.

[204] S. Chaturvedula, C., Bang, N. P., Rastogi, N., ve Kumar, “Price Manipulation, Front Running and Bulk Trades: Evidence from India”, *Emerg. Mark. Rev.*, c. 23, ss. 26–45, 2015.

[205] Imisiker, S., Ozcan, R., ve Tas, B. K. O., “Price Manipulation by Intermediaries”, *Emerg. Mark. Financ. Trade*, c. 51, sayı 4, ss. 788–797, 2015.

[206] L. T. T. P. and T. S., “Stock price manipulation detection using a computational neural network model.”, içinde *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*, 2016, ss. 337–341.

[207] Martinez-Miranda, E., McBurney, P., & Howard, M.J.W., “Learning Unfair Trading: a Market Manipulation Analysis From the Reinforcement Learning Perspective”, içinde *2016 IEEE Conference on Evolving and Adaptive Intelligent Systems*, 2016, ss. 103–109.

[208] M. Zhang, J., Wang, S., Xu, S., & Yu, “Stock Price Manipulation Detection Based on Machine Learning Technology: Evidence in China”, içinde *International Conference on Geo-Informatics in Resource Management and Sustainable Ecosystem. Springer, Singapore*, 2016, ss. 150–158.

[209] Li, A., Wua, J., & Liua, Z., “Market Manipulation Detection Based on Classification Methods”, *Elsevier Procedia Comput. Sci.*, c. 122, ss. 788–795, 2017.

[210] Sun, XQ., Shen, HW., Cheng, XQ., & Zhang, Y., “Detecting anomalous traders using multi-slice network analysis”, *Phys. A Stat. Mech. its Appl.*, c. 473, ss. 1–9, 2017.

[211] Gemici, E., Cihangir, M., & Yakut, E., “Trade-Based Manipulation: The Case of Turkey”, *Ege Acad. Rev.*, c. 17, sayı 3, ss. 369–380, 2017.

[212] Thoppan, J. J., Punniyamoorthy, M., Ganesh, K., “Competitive Models to Detect Stock Manipulation”, *Commun. IIMA*, c. 15, sayı 2, 2017.

[213] A. Abbas, B., Belatreche, A., & Bouridane, “Stock Price Manipulation Detection Using Empirical Mode Decomposition Based Kernel Density Estimation Clustering Method”, içinde *Proceedings of SAI Intelligent Systems Conference*, 2018, ss. 851–866.

[214] Leangarun, T., Tangamchit, P., & Thajchayapong, S., “Stock Price Manipulation Detection Using Generative Adversarial Networks”, içinde *2018 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE*, 2018, ss. 2104–2111.

[215] Shi, FB., Sun, XQ., Shen, HW., & Cheng, XQ., “Detect colluded stock manipulation

via clique in trading network”, *Phys. A Stat. Mech. its Appl.*, c. 513, ss. 565–571, 2019.

[216] Tran, L., & Tran, L., “To Detect Irregular Trade Behaviors In Stock Market By Using Graph Based Ranking Methods”, *arXiv Prepr. arXiv1909.08964.*, 2019.

[217] Wang, Q., Xu, W., Xinting, H., & Yang, K., “Enhancing Intraday Stock Price Manipulation Detection by Leveraging Recurrent Neural Networks with Ensemble Learning”, *Neurocomputing*, c. 347, ss. 46–58, 2019.

[218] Sridhar, S., Mootha, S., & Subramanian, S., “Detection of Market Manipulation using Ensemble Neural Networks”, içinde *2020 International Conference on Intelligent Systems and Computer Vision (ISCV). IEEE*, 2020, ss. 1–8.

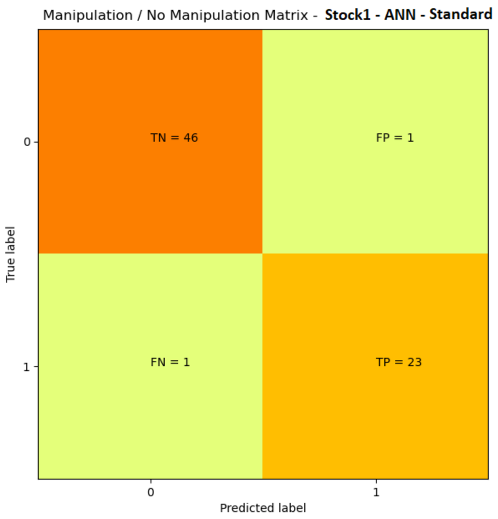
[219] S. M. Youssef, “Stock Market Manipulation Detection Using Continuous Wavelet Transform & Machine Learning Classification”, The American University in Cairo, 2021.

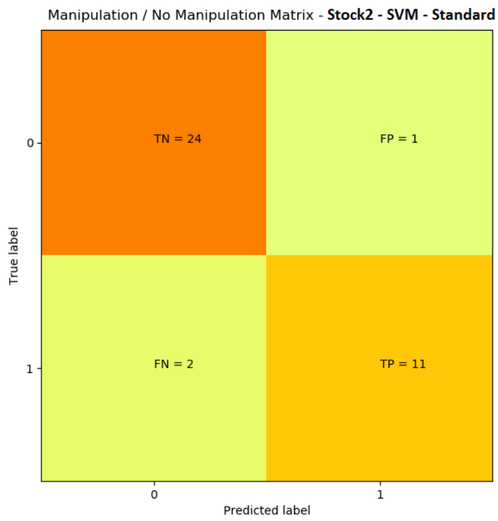
[220] Liu, Q., Wang, C., Zhang, P., & Zheng, K., “Detecting Stock Market Manipulation via Machine Learning: Evidence from China Securities Regulatory Commission Punishment Cases”, *Int. Rev. Financ. Anal.*, c. 78, sayı 101887, 2021.

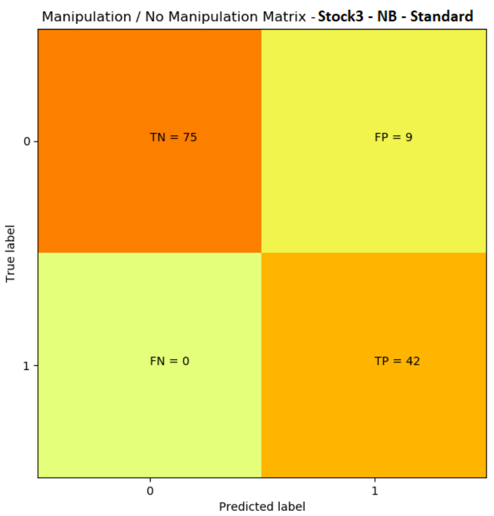
[221] S. Leangarun, T., Tangamchit, P., & Thajchayapong, “Stock Price Manipulation Detection Using Deep Unsupervised Learning: The Case of Thailand”, *IEEE Access*, c. 9, ss. 106824–106838, 2021, doi: 10.1109/ACCESS.2021.3100359.

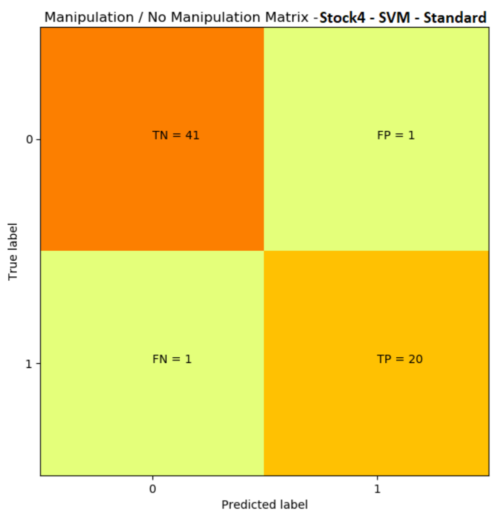
APPENDICES

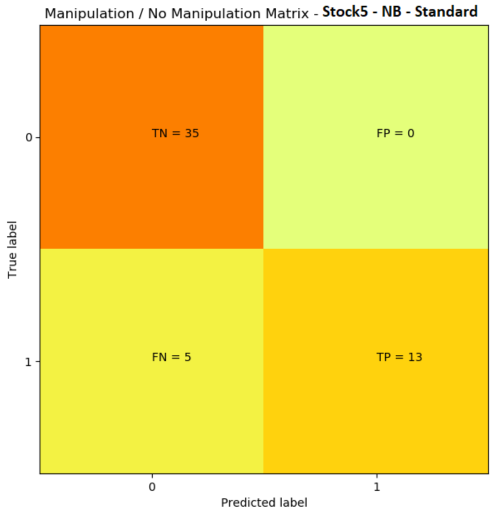
Appendix 1 – Confusion matrix for the proposed model

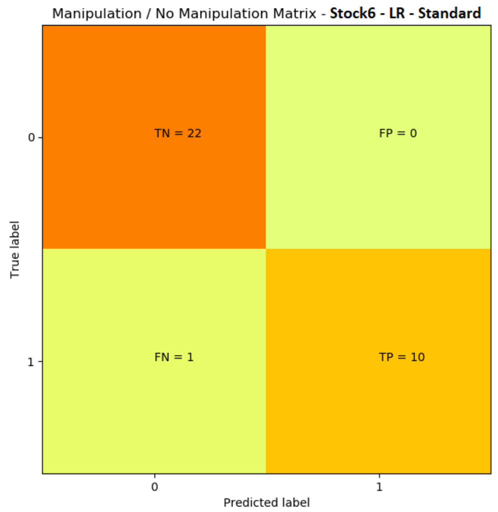


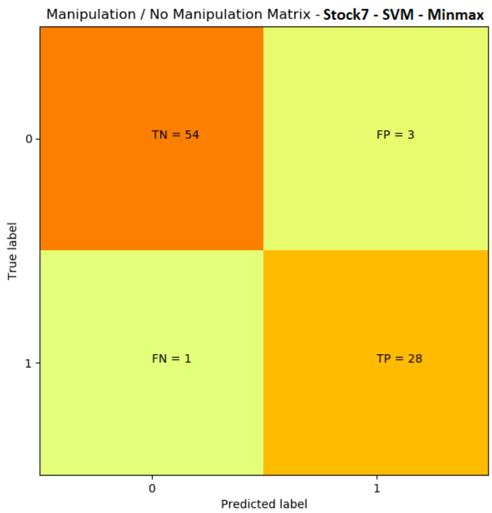


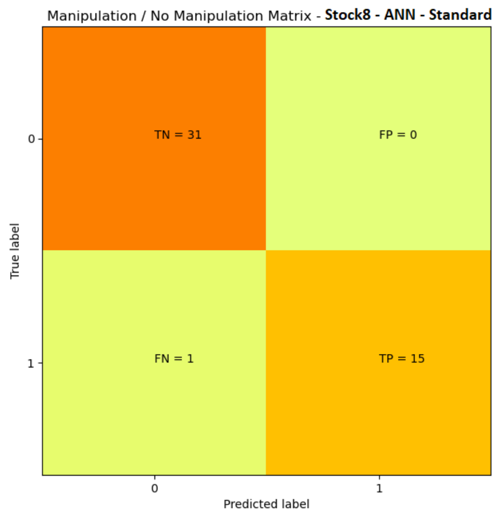


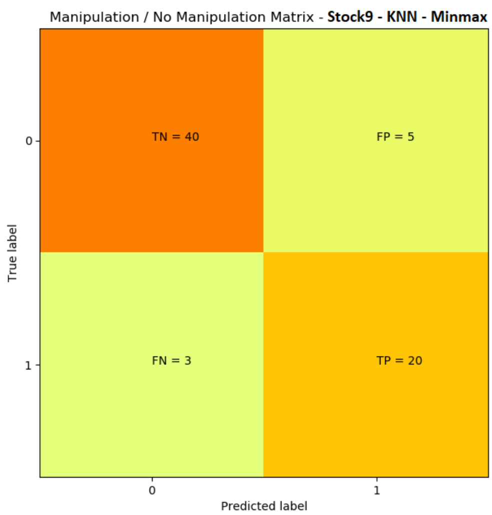


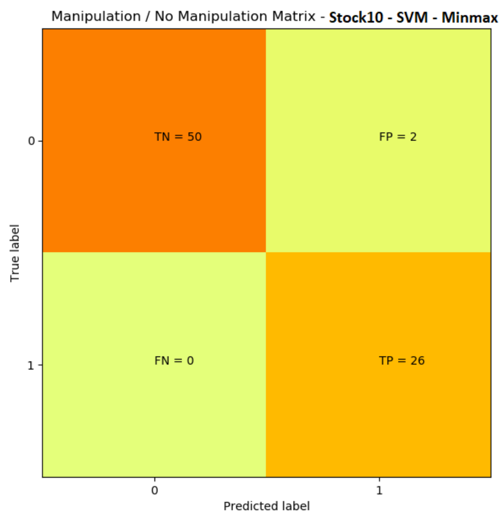

Manipulation / No Manipulation Matrix - Stock5 - NB - Standard
TN = 35
FP = 0
FN = 5
TP = 13
True label
Predicted label
0
1

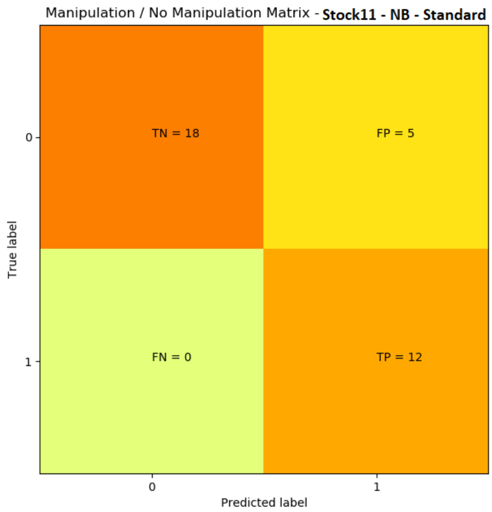

Manipulation / No Manipulation Matrix - Stock6 - LR - Standard
TN = 22
FP = 0
FN = 1
TP = 10
True label
Predicted label
0
1

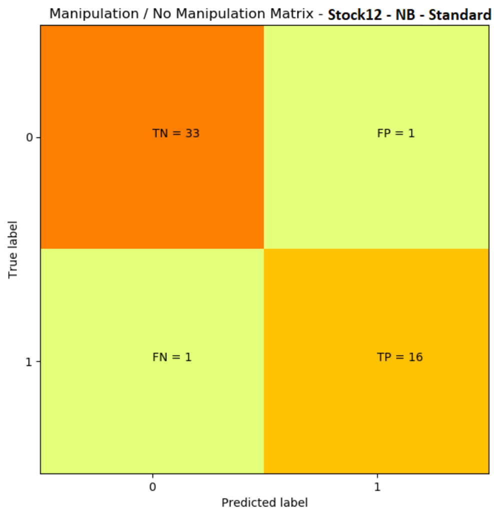

Manipulation / No Manipulation Matrix - Stock7 - SVM - Minmax
TN = 54
FP = 3
FN = 1
TP = 28
True label
Predicted label
0
1

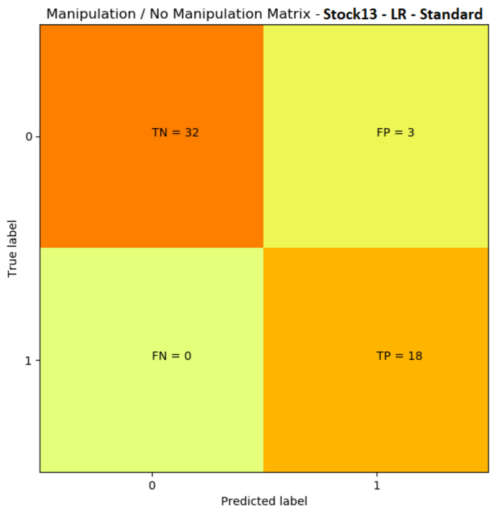

Manipulation / No Manipulation Matrix - Stock8 - ANN - Standard
TN = 31
FP = 0
FN = 1
TP = 15
True label
Predicted label
0
1

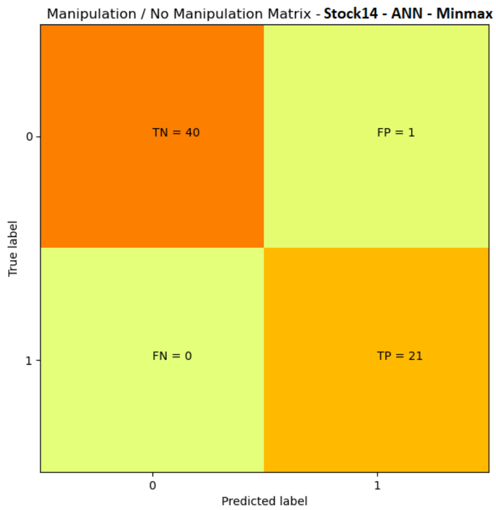

Manipulation / No Manipulation Matrix - Stock9 - KNN - Minmax
TN = 40
FP = 5
FN = 3
TP = 20
True label
Predicted label
0
1

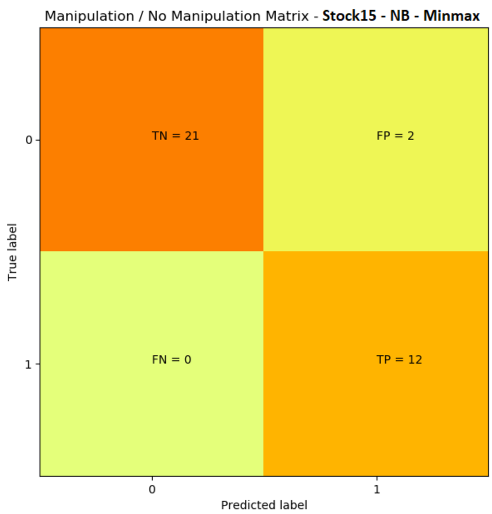

Manipulation / No Manipulation Matrix - Stock10 - SVM - Minmax
TN = 50
FP = 2
FN = 0
TP = 26
True label
Predicted label
0
1

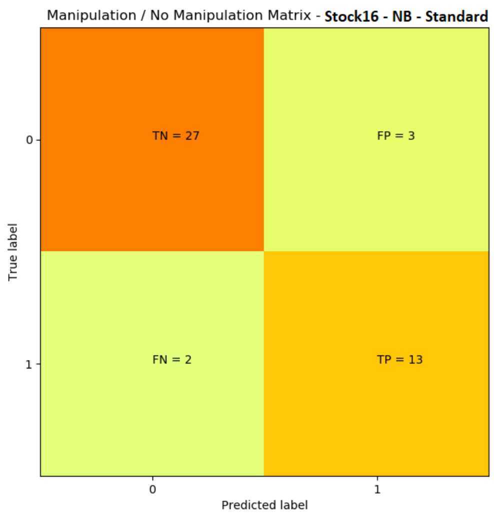

Manipulation / No Manipulation Matrix - Stock11 - NB - Standard
TN = 18
FP = 5
FN = 0
TP = 12
True label
Predicted label
0
1

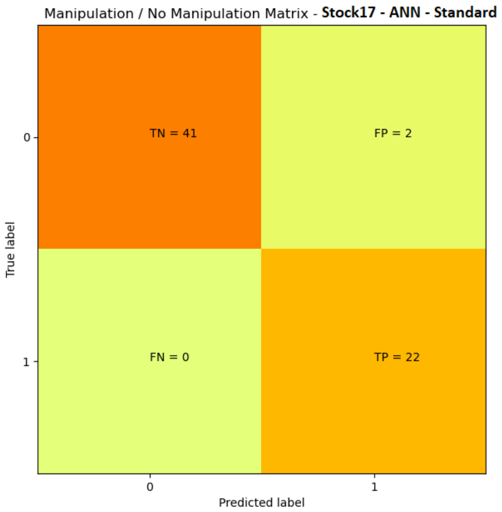

Manipulation / No Manipulation Matrix - Stock12 - NB - Standard
TN = 33
FP = 1
FN = 1
TP = 16
True label
Predicted label
0
1

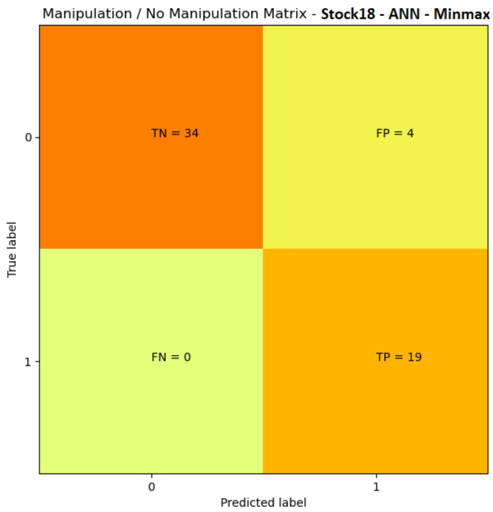

Manipulation / No Manipulation Matrix - Stock13 - LR - Standard
TN = 32
FP = 3
FN = 0
TP = 18
True label
Predicted label
0
1

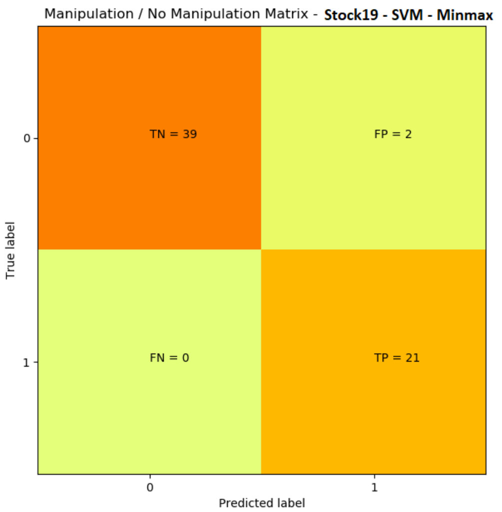

Manipulation / No Manipulation Matrix - Stock14 - ANN - Minmax
TN = 40
FP = 1
FN = 0
TP = 21
True label
Predicted label
0
1

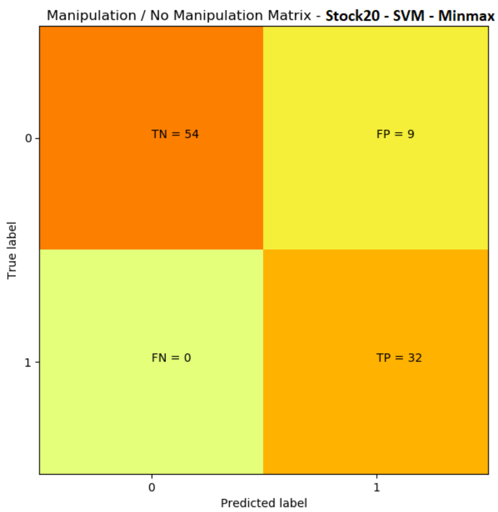

Manipulation / No Manipulation Matrix - Stock15 - NB - Minmax
TN = 21
FP = 2
FN = 0
TP = 12
True label
Predicted label
0
1


Manipulation / No Manipulation Matrix - Stock16 - NB - Standard
TN = 27
FP = 3
FN = 2
TP = 13
True label
Predicted label
0
1


Manipulation / No Manipulation Matrix - Stock17 - ANN - Standard
TN = 41
FP = 2
FN = 0
TP = 22
True label
Predicted label
0
1
0
1


Manipulation / No Manipulation Matrix - Stock18 - ANN - Minmax
TN = 34
FP = 4
FN = 0
TP = 19
True label
Predicted label
0
1
0
1


Manipulation / No Manipulation Matrix - Stock19 - SVM - Minmax
TN = 39
FP = 2
FN = 0
TP = 21
True label
Predicted label
0
1
0
1


Manipulation / No Manipulation Matrix - Stock20 - SVM - Minmax
TN = 54
FP = 9
FN = 0
TP = 32
True label
Predicted label
0
1
0
1