Scientific Annals of Economics and Business
69 (3), 2022, 459-475
DOI: 10.47743/saeb-2022-0024

Comparing Decision Trees and Association Rules for Stock Market Expectations in BIST100 and BIST30

Görkem Ataman*, Serpil Kahraman**

Abstract

With the increased financial fragility, methods have been needed to predict financial data effectively. In this study, two leading data mining technologies, classification analysis and association rule mining, are implemented for modeling potentially successful and risky stocks on the BIST 30 index and BIST 100 Index based on the key variables of index name, index value, and stock price. Classification and Regression Tree (CART) is used for classification, and Apriori is applied for association analysis. The study data set covered monthly closing values during 2013-2019. The Apriori algorithm also obtained almost all of the classification rules generated with the CART algorithm. Validated by two promising data mining techniques, proposed rules guide decision-makers in their investment decisions. By providing early warning signals of risky stocks, these rules can be used to minimize risk levels and protect decision-makers from making risky decisions.

Keywords: stock market; efficient market hypothesis; CART; Apriori; association.

JEL classification: G20; G14; C89.

1. INTRODUCTION

Stock markets play a crucial role in monetary stability has been a discussion issue for financial economists. As a part of financial markets, the primary function of the stock market is to channel funds flow efficiently. The Efficient Market Hypothesis (EMH), which was developed by E. Fama (1970), refers to stock prices fully reflecting all available information is known as "efficient." In other words, the knowledge of all economic agents possesses equally. However, almost all stock market transactions involve asymmetric which results from asymmetric information between investors because the controlling shareholder may know more accurate insider information that is not announced to the public yet but can provide a competitive advantage in stock transactions (Martins & Paulo, 2012).

* Department of Business Administration, Yasar University, İzmir, Turkey; e-mail: *gorkem.ataman@yasar.edu.tr*.
** Department of Economics, Yasar University, İzmir, Turkey; e-mail: *serpil.kahraman@yasar.edu.tr* (corresponding author).

According to economic theory, the asymmetric information problem is the primary source of instability and fragility that affect market efficiency (Akerlof, 1970). Moreover, various factors influence stock market efficiency, and determining these factors is crucial for portfolio management as a critical issue in investment decision-making. Portfolio management performs the essential economic function of channeling funds to the most profitable and less risky investment opportunities. Thus, the accurate forecasting of stock market movement is crucial to reducing information asymmetry. Although various techniques have been implemented for stock market prediction in literature, with the advances in data science technologies, a more current approach is data mining (Park & Chai, 2021; Xu *et al.*, 2022; Zhou *et al.*, 2022). Data mining methods help to determine key variables and eliminate irrelevant ones to monitor financial markets and generate efficient predictions on financial markets, given these key variables (Huang & Liu, 2020; Kumar *et al.*, 2021). Besides, in today's era of big data, data mining is a crucial tool to reduce information asymmetry by extracting knowledge from the data. Hence, it has been well adopted in stock market prediction.

Since the Stock market is considered the source of financial fragility and instability, financial institutions should decide on the optimum portfolio based on the potential risk levels of stocks. Potential risk levels are evaluated in the light of the main determinants for stocks, including index value, underlying index weight, closing value, etc. (Han *et al.*, 2011). Recent and promising data mining technologies also received increasing attention in evaluating stock markets' risk levels (Cao, 2021). In this context, classifying the stocks as risky and successful using classification techniques, which data mining technologies provide, helps researchers propose high-performing risk prediction models (Zhang & Zhou, 2004). These models can be appropriately adopted in financial institutions for decision-making to increase efficiency and liquidity and reduce economic instability.

As another promising data science technique, association rule mining can also be used for making decisions in financial institutions (Pawar *et al.*, 2019; Chen *et al.*, 2022; Kartal *et al.*, 2022). The association rule mining aims to discover a pattern of frequently used items appearing together in data sets. This technique reflects behaviors in which the same behavior often occurs. Since, with globalization, the potential of stock markets to exhibit common behaviors is increasing, it gains importance to discover these patterns or associations. In the context of stock market prediction, mining the associations between key variables affecting the risk levels of stocks is a novel point to be addressed.

Thus, in stock market prediction, current data science technologies, classification analysis, and association rule mining provide information to financial institutions on which group of stocks is most relevant for their customer segment (Bhasin, 2006). In other words, investment analysts can also determine a relatively more efficient portfolio (Zhang & Zhou, 2004).

Despite the geopolitical instability, attractive rate of returns, appreciation of Turkish liras (TL), and the policies of the Turkish Central Bank support the market capitalization of Borsa Istanbul (BIST). Besides, due to their importance in financial management, research on stock markets, especially in emerging countries, such as the Turkish stock market - BIST, receives considerable attention (Hashmi & Chang, 2021). In BIST, stock market expectations and portfolio decisions are strongly linked with BIST30 and BIST100 indexes, suggesting that the stock market is a critical part of financial intermediation. According to the theory of EMH, the nature of the stock market index movement places great emphasis on rational expectations. This study's theoretical premise is the Efficient Market Hypothesis. As well-known, market imperfection plays a crucial role in market efficiency and the formation of rational or adaptive

expectations. However, the financial economics literature provides relatively limited studies that perform different quantitative techniques rather than econometrics.

Finally, grounded in EMH theory, this study aims to model stock market risk levels for the BIST30 and BIST100 indexes of the Istanbul Stock Exchange and provide a data science technology implemented tool for understanding critical elements of stock risk levels using real data. This tool may provide information for investors, investment companies, and fund managers in their portfolio decisions when strong evidence of risk level is available. Although data science technologies have been well implemented in literature for stock market prediction, integrating and comparing two promising techniques of classification analysis and association rule mining is the main contribution of this paper in this context. In addition, since both the generated procedure for classifying the stocks as low risk and high risk and the discovered associations for labeling the stocks as low risk and high risk yielded very similar results, the validity of the proposed methodology is supported. Thus, these findings may provide valuable insights for decision-makers on which stocks are considered low risky or more efficient given the values of the critical variables. Besides using these findings in stock market risk prediction in practice, the methodological approach can guide researchers and decision-makers and be adopted for other issues in financial management.

The paper is organized as follows. After the introduction section, Section 2 introduces the theory of expectations and reviews existing studies in the context. Section 3 describes the data and research methodology. Section 4 presents empirical findings, followed by concluding remarks in Section 5.

2. PREDICTING STOCK MARKET RETURNS: RATIONAL OR ADAPTIVE EXPECTATIONS?

Stock markets sparked growing interest for economists. In economic theory, the prediction or expectation of the future value of an economic indicator is based on the two main expectation theories; adaptive and rational expectations. Under the theory of rational expectations, developed by J. Muth, future predictions and expectations are based on all the available information, and there is no information asymmetry. So that the future predictions are rational. However, in adaptive expectation theory, the expectations of stock market index movement are based on the average of past data and experiences. In a stock market, investors tend to believe the previous trend will continue until new information arrives.

A limited number of empirical studies consider financial data mining applications in the practical framework. A study by Hajizadeh *et al.* (2010) present an overview of the application of data mining techniques in stock markets. Similarly, Inidapo *et al.* (2017) surveyed over 100 studies examining soft computing techniques for stock market prediction.

Enke and Thawornwong (2005) examine the different neural network models for their ability to forecast stock market movements. They noted that the stock trading decisions based on classification models provide a higher return than the neural network and linear regression forecast models. Similarly, Ou and Wang (2009) compare ten data mining techniques to find the most practical forecasting model for stock price movements using the Hang Sang index from the Hong Kong Stock Market. They conclude that the SVM and LS-SVM models have relatively higher ability for forecasting, among other techniques. Moreover, Chen *et al.* (2007) investigate an efficient forecast model and methods for seven years NASDAQ100 index of Nasdaq Stock Market and 4-year NIFTY index of S&P stock data. The results indicate a

strong correlation between the behavior of both indexes and the local weighted polynomial regression (LWPR) model.

Some studies perform data mining techniques on BIST- Borsa Istanbul. Bastı *et al.* (2015) examined stock underpricing in the initial public offerings (IPOs) traded on Borsa Istanbul from 2005 to 2013 using decision tree algorithms and SVM. The authors explain the underpricing of IPOs of BIST stocks. A study conducted by Filiz and Öz (2017) employs classifier algorithms and logic regression analysis to examine factors affecting BIST.100 index values. The authors highlight that factors affecting the index values are crucial for the classification.

Moreover, Uzar (2014) examines data mining techniques' performance within the BIST. They selected 129 manufacturing and 109 financial sector firms to employ questionnaires. They noted that there is no meaningful causal relation between the knowledge of data mining techniques and financial information system performance. Albayrak and Koltan Yılmaz (2009) apply decision tree algorithms on 173 industry and service sector firms' data in Istanbul Stock Exchange (ISE)100 between 2004 and 2006. Their results show that capital to net sales ratio, stock turnover, and profit ratio are the main variables for the classification. Kara *et al.* (2011) apply and compare artificial neural networks (ANN) and support vector machines (SVM) to forecast the direction of ISE100 index movement by using daily data from 1997 to 2007. The results indicate that the average prediction performance of the ANN model (%75.74) is significantly higher than that of the SVM model (%71.52). In one of the most recent studies on the issue, Yiğit and Muzır (2019) examine the relations between corporate governance returns and stock market returns using the corporate governance index (CGI) and BIST30, BIST100, and BISTALL indexes. The authors employ Ordinary Least Square (OLS) and Multivariate Adaptive Regression Splines (MARS) techniques and found a positive correlation between two market returns. Another study conducted by Ekinci and Ersan (2018) investigated high-frequency trading (HFT) activity by comparing BIST30, BIST100, and BISTALL stock market indexes in Borsa Istanbul. Authors indicate that HFT is unequally distributed between June 2015 and November 2015.

In the Indian context, Paranjape-Voditel and Deshpante (2013) perform the ARM techniques and fuzzy classification to determine the stock market portfolio recommender system. The authors observe that BSE-30, S&P CNX-100, CNX-50 or NSE-50, and DOW-30 stock market indexes have surpassed the returns generated by top mutual funds. Another empirical study that considers the NSE of India is a study by Vaiz and Ramaswami (2016). They employ the decision tree classification method to investigate the effect of technical stock market indicators such as moving averages, trend, volume, and volatility indicators on stock price movements. The authors point out that tree-based classifier algorithms are more effective tools for predicting stock market behavior.

Bordalo *et al.* (2019) investigate the dynamics of fundamentals, expectations, and stock returns using different forecasting techniques in the theoretical framework. Unlike the adaptive expectations theory, the results indicate that the investors are forward-looking in forming their belief in the light of observed stock earning growth. A study performed by X. Wu *et al.* (2020) considers the increasing complexity in stock markets which leads to the inflexibility of trading by applying the Gated Recurrent Unit, which represents the adaptive trading strategies. They found that between the two adaptive trading strategies, the Gated Deterministic Policy Gradient trading strategy with an actor-critic framework is more stable than the Gated Deep Q-learning trading strategy with a critic-only framework.

One recent study by Giglio *et al.* (2020) focuses on the investor expectations of economic growth and stock returns during the Stock Market Crash of the COVID19 pandemic. They indicate that the survey results do not clarify whether the expectations have rational or behavioral elements due to the asymmetric disagreements among investors. Another study that also considers the COVID19 pandemic is a study by Onali (2020). The author examines the effect of the pandemic on the impact of the expectations on volatility and trading volume of the US Stock Markets (Dow Jones and the SP500 indices) by applying GARCH and VAR models. In this study, the results are mixed. While some countries impact the US Stock Markets, others do not (China, Italy, Spain, the UK, Iran, and France). Finally, Angeletos *et al*. (2020) explain the expectations for the stock market and broader concepts by referring to rational expectations theory as a bedrock of modern macroeconomics. They use different econometric techniques, including ARMA and VAR models. Authors indicate that the main lesson for the view is that shocks drive most of the reactions, and forecasts consider imperfect expectations.

In this context, the critical assumption of our analysis is the existence of adaptive expectations within the economic agents of potential stock market investors because of the asymmetric information and free-rider problem in stock markets. Freeriding indicates stock market transactions without having enough liquidity to cover the trading. It also refers to having the data and information without paying anything by creating asymmetric information (Bhide, 1993). Another reason is the probability of higher volatility and risk potential in a stock market. Investors are more pessimistic about stock price index movement than any other investment instrument. Importantly, optimistic expectations about future stock prices are related to positive past experiences, while a tendency to be uncertain and risky is strongly related to negative past experiences (Kezdi & Robert, 2009). A more plausible explanation is that decision-makers in the way they process their experiences and information are more critical input. Moreover, they may also believe that stock market index values may follow a random walk because of the speculations, investors' moods, political issues, etc. (Hurd *et al.*, 2010).

3. DATA AND THE RESEARCH METHODOLOGY

This study systematically analyzed the risk levels of Turkish companies issued and traded on Borsa Istanbul. Besides, the underlying factors affecting risks were determined. Borsa Istanbul (BIST), formerly called Istanbul Stock Exchange, started operations with 40 listed companies in 1986. BIST is a member of different international federations and associations such as the World Federation of Exchanges (2019), Federation of Euro-Asian Stock Exchanges, Federation of European Securities Exchanges, and International Capital Market Association (Borsa Istanbul). BIST Stock indexes have been developed to measure the price and the return performances of a group of stocks traded on Borsa İstanbul. BIST has been developing many aspects, such as the number of listed corporations, the daily trading volume, the total market capitalization of listed companies, and the number of markets.

The BIST100 index, as the leading stock index of the Borsa Istanbul Equity Market, consists of 100 stocks, which are selected among the stocks of companies traded on the BIST Stars, BIST Main markets, the stocks of real estate investment trusts, and venture capital investment trusts. In comparison, the BIST30 index consists of 30 stocks. BIST100 index covers BIST30 stocks. The composition procedure of the index is based on two values: the free float market value and the daily average traded value in the base period.

There were 421 companies listed on BIST in 2019. The total market capitalization of BIST companies was 795 billion TL by the end of 2018. Foreigners own 65% of BIST companies' free float shares. Its daily trading volume was TL 16.46 billion on August 10, 2018 (Borsa Istanbul, 2019a). The BIST is the world's second most liquid trading platform by the year 2018. Moreover, Borsa Istanbul ranks 4th among Emerging Markets regarding equity market traded value. The BIST Equity Market is ranked 21st in the world based on Market total trading value by the end of 2018.

3.1 Data

This study covered monthly closing values of two BIST indexes, BIST 30 and BIST 100, from 2013-2019. Since the primary goal of this paper is to predict and classify the companies based on their risk, the output variable is defined as the target risk level of companies. Three leading indicators used in modeling: index name (BIST 30 or BIST 100), index value, and stock price.

The BIST index calculations use the free float market capitalization ratio and the latest registered stock prices. The daily settlement price is calculated as the weighted average price of all the transactions at the end of each session. In more detail:

- The weighted average price of all trades executed in the last ten minutes of the regular session,
- The weighted average price of the last ten trades executed during the regular session if fewer than ten trades were realized in the previous ten minutes of the regular session,
- The weighted average price of all trades executed in the regular session if fewer than ten trades were realized in the regular session,
- If no trades were made, the daily settlement price would be based on the theoretical price level by considering the costs of the underlying asset and other contracts based on the same underlying asset.

Moreover, if the calculation mentioned above cannot calculate the daily price, Exchange may determine the daily settlement price level (Borsa Istanbul, 2022).

3.2 Data pre-processing

BIST index's target risk levels are divided into 10, 20, and 30. In this study, the middle-risk group (20) is excluded from the study to highlight the differences between the two extremes. The remaining two levels are defined by a dummy variable, where the low extreme (risk level 10) is coded as "low risk" to show the low risk is expected, and the high extreme (risk level 30) is coded as "high risk," which means high risk is expected. Thus, this study's output variable, risk level, has a nominal scale corresponding to two groups.

The index name is the first input variable of the data mining models of this study. This is a nominal variable and integrated into the models as they are. Thus, the index name variable levels are labeled "BIST 30" and "BIST 100".

The remaining two input variables, index value (IV) and stock price (SP), is continuous. However, to apply and interpret the classification analysis and association rule mining efficiently, these variables are transformed into categorical variables measured in ordinal scales. The main descriptive statistics on these two continuous variables for data transformation are calculated as represented in Table no. 1.

Table no. 1 – Descriptive statistics on index value and stock price

variables	Statistics						
	N	minimum	maximum	mean	median	1st quartile	3rd quartile
Index value (IV)	324	398.46	1,041.10	613.04	623.63	483.44	692.08
Stock price (SP)	324	88,434	228,127	147,762	143,101	120,803	167,555

The corresponding 1st and 3rd quartile statistics in Table no. 1 are obtained for transforming IV and SP to ordinally measured categorical variables. The monthly IV values, which are at most equal to the first quartile of IV, are labeled as "IV-Low," and monthly values between the first and third quartiles are marked as "IV-Medium," and the rest are labeled as "IV-High." A similar approach is used for transforming the stock price. Then, the stock price variable is marked as "SP-Low," "SP-Medium," and "SP-High." With these transformations, the structured data set of this study is designated. No outliers and redundant entries were identified in the data set.

Table no. 2 shows an instance (representing just four rows) taken from the raw data set. This table shows the data set's structured form and the variables' measurement scales.

Table no. 2 – A presentation of the raw and structured data set

	RAW DATA				STRUCTURED DATA			
variable	Index name	IV	SP	Risk level	Index name	IV	SP	Risk level
type	input	input	input	output	Input	input	input	Output
scale	nominal	ratio	ratio	nominal	nominal	ordinal	ordinal	Nominal
31.01.2013	BIST 30 RK %10	405.27	135,019.59	10	BIST 30	IV-Low	SP-Medium	Low Risk
31.01.2013	BIST 30 RK %30	625.70	135,019.59	30	BIST 30	IV-Medium	SP-Medium	High Risk
31.01.2013	BIST 100 RK %10	411.88	110,570.11	10	BIST 100	IV-Low	SP-Low	Low Risk
31.01.2013	BIST 100 RK %30	617.46	110,570.11	30	BIST 100	IV-Medium	SP-Low	High Risk

3.3 Applied data mining techniques

In this study, two main data mining functions, classification and association analysis, are comparatively used in modeling risk levels of BIST indexes. A decision tree algorithm, specifically Classification and Regression Trees (CART), is used for classification. Apriori algorithm used for mining association rules. Although different algorithms exist in data mining, since the output variable of interest is binary and the input variables are more practically used with these categorical definitions, these two algorithms are selected based on their superiority in performance and interpretability in the obtained findings (M. Wu *et al.*, 2018; Valente *et al.*, 2021).

3.3.1 Decision tree algorithms

Decision tree algorithms have gained significant attention in business analytics and data mining. Although many different decision tree algorithms exist in the literature, one of the most commonly used ones is CART. It produces easily understandable and interpretable trees and can process continuous and categorical input and output variables. CART was introduced

by Breiman *et al.* (1984) as a type of decision tree that utilizes some splitting rule based on the predictor variables by building a binary decision tree. CART algorithm works recursively where data is partitioned into two subsets to increase homogeneity within these subsets. These subsets are then partitioned again until either a node is reached at the level where splitting no more improves uniformity or other time-based stopping criteria are satisfied. The stopping node is a terminal node that determines prediction in the form of a class in classification problems or the average response of the response variable in regression problems (Denison *et al.*, 1998). Besides, a parent node is the starting point of analysis and contains the whole population or data. Thus, based on the splitting rules, CART generates an upside-down tree structure where the parent node locates at the up, and the stopping node is the downside of this tree. In this approach, members within each subgroup have the same properties affecting the probability of belonging to the related level of the output variable. The same input variable may be used repeatedly while the decision tree grows until the stopping node is reached.

3.3.2 Association rule mining

Association rule mining (ARM), which Agrawal *et al.* (1993), is used for mining past transactions to extract association rules those discover relationships and dependencies in the data set. The logic of ARM can be summarized as follows: Let D={T_1,T_2,…,T_n } be set of n transactions and let I be set of items, I={i_1,i_2,…,i_m }. Each transaction is a set of items, i.e. $T_i \subseteq I$. An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subset I$, and $X \cap Y = \emptyset$; X is called the antecedent, and Y is called the consequent of the rule. A set of items, such as X or Y is called an itemset. For an itemset $X \subseteq I$, $support(X)$ is defined as the fraction of transactions $T_i \in D$ such that $X \subseteq T_i$. The support of a rule $X \Rightarrow Y$ is defined as $support(X \Rightarrow Y) = support(X \cup Y)$. The rule has a measure of reliability, namely confidence which is calculated as $confidence(X \Rightarrow Y) = support(X \cup Y)/support(X)$. The other performance metric of ARM is lift measuring the predictive power of the rule $X \Rightarrow Y$, and statistically defined as $lift(X \Rightarrow Y) = confidence(X \Rightarrow Y)/support(Y)$. ARM aims to extract all rules embedded in the data set whose metrics are at least equal to specified values of minimum support and confidence. The user determines these specified values.

One of the literature's most widely used ARM algorithms is the Apriori algorithm, which was developed by Agrawal and Srikant (1994). To extract the association rules which demonstrate frequent item sets having support and confidence values higher than the user-specified levels, a dataset that includes many different transactions is given to Apriori as input. In the algorithmic process of Apriori, itemset I of length m is frequent if and only if every subset of I with length m-1 is also frequent. In this regard, the Apriori algorithm significantly reduces search space and allows rule discovery in computationally feasible time.

3.4 Data mining platform: WEKA

An open-source data mining platform: The Waikato Environment for Knowledge Analysis, WEKA (Hall *et al.*, 2009), developed by the University of Waikato in New Zealand, was used in this study. WEKA provides not only researchers but also practitioners easy access to cutting-edge techniques in data mining. Many different data mining functions such as regression, classification, clustering, association rule mining, and variable selection are included in the tool. Besides, WEKA can visualize data, trees, and graphs efficiently.

3.5 Outcome measures

The performance of the CART algorithm is evaluated based on sensitivity, specificity, and accuracy metrics. These metrics are measured with four possible outcomes of a classification algorithm:

- True positive (TP): While the actual value of the output variable is yes (high risk), the model correctly classifies it as yes (high risk).
- False positive (FP): While the actual value of the output variable is no (low risk), the model incorrectly classifies it as yes (high risk).
- True negative (FP): While the actual value of the output variable is no (low risk), the model correctly classifies it as no (low risk).
- False negative (FN): While the actual value of the output variable is yes (high risk), the model incorrectly classifies it as no (low risk).

With these outcomes, performance metrics of the CART algorithm are calculated as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

On the other hand, the accuracy of the Apriori algorithm is measured by using the confidence metric, which is used to order the extracted rules in Apriori (Agrawal *et al.*, 1993).

3.6 Proposed Model

To summarize the used methodological approach and the proposed model of this study, we present a flowchart in Figure no. 1. This figure summarizes the methodology of this paper step by step.

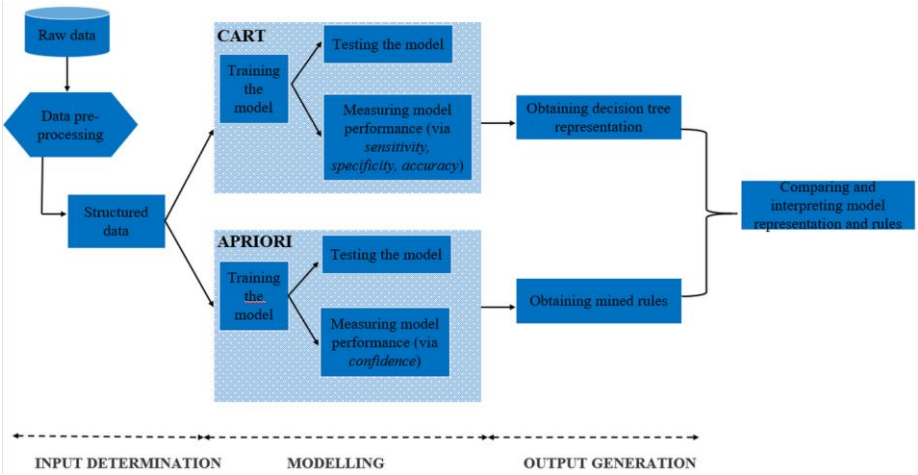

Figure no. 1 – Flow chart of the model

4. RESULTS

4.1 Frequencies and descriptive statistics

During the study period, 324 monthly values exist for BIST 30 and BIST 100 indexes. The data set is uniformly distributed based on the levels of the output variable. That is, while 50% of data set instances represent "low risk," the rest represent "high risk." Similar property exists based on the first input variable, index name. While half of the instances of the data set show the BIST 30 index, the remaining half relate to BIST 100 index.

As explained in data pre-processing (Section 4.2), first and third quartile statistics of these variables are used to transform them into nominal variables in the structured data set. With this transformation, levels of these variables and the frequencies of each level are shown in Table no. 3.

Table no. 3 – Transformed values of IV and SP in the structured data set

	Values	Levels	Frequency		Values	Levels	Frequency
	$IV \leq 483.44$	IV-LOW	81		$SP \leq 120{,}803$	SP-LOW	81
Index	$483.44 < IV$	IV-	162	Stock	$120{,}803 < SP$	SP-	162
value	< 692.08	MEDIUM		price	< 167555	MEDIUM	
(IV)	$692.08 \geq IV$	IV-HIGH	81	(SP)	$167{,}555 \geq SP$	SP-HIGH	81

4.2 Results of the CART algorithm

In classification analysis, 10-fold cross-validation is used in the training data set. The CART algorithm produced a decision tree where the number of leaves and size of the pruned tree are determined as 6 and 9, respectively. The generated tree is figured out in Figure no. 2.

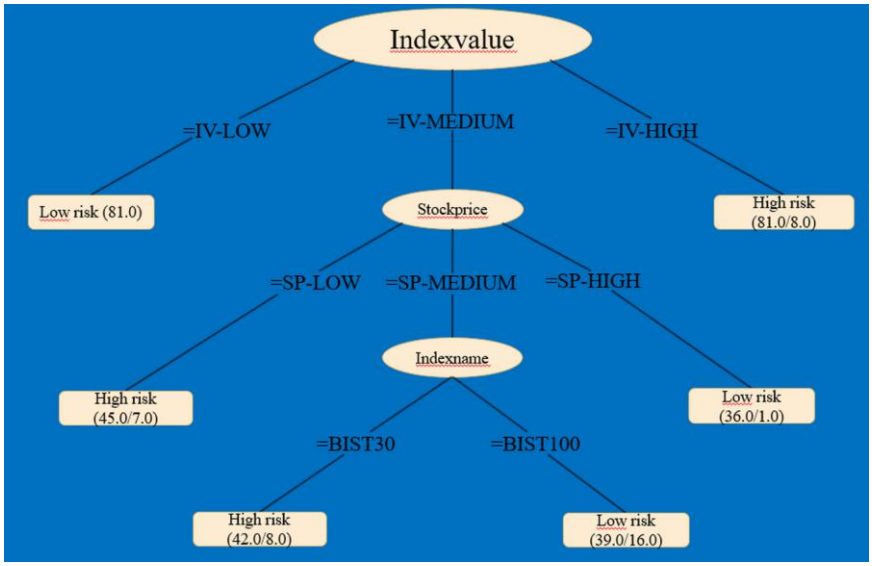

Figure no. 2 – Decision tree generated by CART algorithm

Figure no. 2 shows that starting node of CART is the index value meaning that the algorithm begins partitioning the data set based on this variable. In comparison, the instances with low index values are classified as "low risk," and those with high index values are classified as "high risk." The algorithm continues partitioning the data set for the instances labeled as medium-level index values. At this level, the stock price is used as the primary variable or criteria for classification. While the instances with medium index value and low stock price value are classified as "high risk," those with a medium index value and high stock price are labeled as "low risk." Similarly, the CART algorithm continues to partition the data set through the index name variable for the medium stock price levels. In this stopping level of classification, while BIST 30 index represents the "high risk" group, BIST 100 represents "low risk."

CART's outcome measures, sensitivity, specificity, and accuracy, are shown in Figure no. 3 to measure the performance of this algorithm.

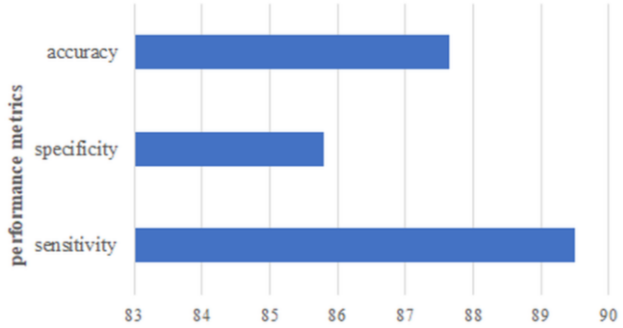

Figure no. 3 – Performance of CART in risk level classification (%)

Values of Figure no. 3 show sensitivity, specificity, and accuracy measures of the used classification algorithm are above 85%, meaning that the model performs efficiently in risk level classification of the stock market. Since sensitivity is a bit higher than specificity, it can be concluded that CART is more successful in classifying "high risk" values in the stock market.

4.3 Results of the Apriori algorithm

The Apriori algorithm is used to extract hidden rules of the stock market data set, and rules are generated for both "low risk" and "high risk" representations of the stock market. In the experimentation, the minimum support value is taken as 0.1, meaning that the algorithm shortlists the set of items observed at least in 32 instances. To discover rules that are pretty compared with CART results, the minimum confidence metric of the Apriori algorithm is taken as 0.85 (since the model specificity of CART is around 85%).

4.3.1 Extracted rules for "high risk" level representations

With the user-specified minimum confidence threshold of 0.85, Apriori discovered five rules for representing the "high risk" levels in the stock market. These rules are presented as follows:

Rule #1	IN=BIST 30 & IV=High →	High risk	(46,42; conf:0.91)
Rule #2	IV=High →	High risk	(81,73; conf:0.9)
Rule #3	IV=High & SP=High →	High risk	(46,40; conf:0.87)
Rule #4	IV=Medium & SP=Low →	High risk	(45,39; conf:0.86)
Rule #5	IN=BIST 100 &IV=Medium & SP=Low →	High risk	(41,35; conf:0.85)

Numbers in parenthesis respectively show the number of instances in which the left-hand side of the rule is satisfied in the data set, how many of these instances are identified as "high risk," and the corresponding confidence level of the generated rule. Thus, rule #1 represents that 46 instances of the data set are labeled as BIST 30 indexes having high index values, and 42 of those instances are identified as "high risk" (the remaining four instances have "low risk"), where therefore the confidence metric is calculated as 0.91.

4.3.2 Extracted rules for "low risk" level representations

Seven rules are extracted for demonstrating "low risk" levels in the stock market. These rules are as follows:

Rule #1	IV=Low →	Low Risk	(81,81; conf:1)
Rule #2	IV=Low & SP=Medium →	Low Risk	(47,47;conf:1)
Rule #3	IN=BIST 30 & IV=Low →	Low Risk	(43,43;conf:1)
Rule #4	IN=BIST 30 & IV=Low & SP=Medium →	Low Risk	(39,39;conf:1)
Rule #5	IN=BIST 100 & IV=Low →	Low Risk	(38,38;conf:1)
Rule #6	IV=Low & SP=Low →	Low Risk	(34;34;conf:1)
Rule #7	IV=Medium & SP=High →	Low Risk	(36;35;conf:0.97)

The confidence levels are higher in mined rules representing "low risk" in the stock market compared to "high risk" representations.

4.3 Discussion on algorithm performances

When the classification rules generated by the CART algorithm and mined rules by the use of the Apriori algorithm are comparatively analyzed, it can be seen that most of the results are consistent. While low-level index values are directly labeled as low-risk items, high-level index values are determined as high-risk ones in both algorithms. On the other hand, for the medium-level index values, the two algorithms continue partitioning by checking stock prices to categorize the risk levels of items. While the items with medium index values and high stock prices are categorized as low risk, those with medium level index values and low stock prices are classified as high risk. For items with medium levels of index values and stock prices, the index name becomes a significant predictor of the classification model. Between those items, while BIST 100 indexes are classified as low-risk items, BIST 30 are classified as high-risk items. However, the Apriori algorithm's index name is not identified as a significant predictor. Although in some of the mined rules (Rules #1 and #5 for high-risk representations and Rules #3, #4 and #5 for low-risk representations), index name levels are seen on the left-hand side of the association, these are indeed representing the associations which have already been mined in some other rules. For instance, when rules #1 and #2 in

high-risk representations are compared, it is understood that items with high index values are determined as high-risk items (Rule #2) regardless of the index name. For this mined rule, the index name contributes to the rule (Rule #1) by supporting it with a slight increase in the confidence metric. That is, while the high level of index values is identified as high-risk items with a 0.9 confidence level (i.e., in 73 of 81 items), the BIST 30 items having high index values are labeled as high-risk items with a confidence level of 0.91 (in 42 of 46 items). The consistency and similarity of the obtained findings of these two algorithms increased the validity of the proposed approach in stock market risk prediction. Besides, the investors can be more confident using these findings for risk evaluation in stock markets since two methods support these findings.

In addition to obtaining valid results using these two techniques, the algorithms are worth highlighting in terms of their performances. While the prediction accuracy of the CART algorithm is around 88%, the confidence metrics of the obtained rules are also varying between 85% to 97%. These performances are achieving or exceeding the research in stock market prediction (Waspada *et al.*, 2021; Parkash *et al.*, 2022).

5. CONCLUSION

In financial and economic theory, stock markets assume an efficient market hypothesis regarding resource allocation in a capital market mechanism. Efficiency indicates that the primary function of capital markets is an efficient fund transfer from lenders to borrowers, which plays a crucial role in stock markets. Asymmetric information is the main problem affecting market efficiency due to the lack of all the available information. Due to the complexity of stock market movements and other determinants such as macroeconomic conditions, speculations, asymmetric information, political factors, etc., predicting stock market index values is crucial for financial markets. Financial institutions mainly employ technical analysis to predict stock market index movements. However, prediction is difficult due to a large amount of data to be examined using traditional financial analysis. This study provides a pilot application for portfolio management decision-making tools using the advances in data science technologies.

This study mainly investigates the risk levels of BIST30 and BIST100 indexes based on the critical variables of index name, index value, and stock price. Methodologically, classification analysis and association rule mining techniques of promising data mining technologies are used. Empirical findings of this study mainly suggest a similar bias in stock market expectations. According to the results, it can be said that the BIST100 stock market index is a relatively more homogenous portfolio. Considering index value and stock prices, stock market efficiency is somewhat higher in the BIST100 index than in the BIST30 index. So, the rise in the trade volume and the foreign share in BIST100 as the leading index that determined Borsa Istanbul is remarkable. A plausible explanation of this empirical study may be that the BIST30 index is more critical for fund managers when new information arrives in the stock market. The obtained classification algorithm and rules in association rule mining provide crucial early warning signals based on the values of index value and stock price for both the BIST30 and BIST100 indexes. These warnings might be more brutal to detect for the portfolio analysis due to the financial dynamics. Besides, these findings may guide decision-makers in investing decisions on stocks. By using these techniques and the generated rules, the decision makers can choose the alternatives minimizing their risks and protecting

themselves from risky decisions. Since these two techniques generate similar findings on understanding the characteristics of the stocks having low or high risks, which increases the validity of the proposed method, the decision makers can be more confident in using these findings in decision making.

This paper's main limitation is analyzing only one stock market, Borsa Istanbul. The study is also limited to the selected three input variables: index name, index value, and stock price. However, there may be other key variables affecting stock risk levels. The study period is the further limitation of this study. The study period covers 2019 and excludes 2020 and 2021 data. Since stock markets are highly affected by COVID-19 in 2020 and 2021, and this external factor may significantly change the typical characteristics of stock markets, these periods are consciously excluded from the study period. In future research, the implementation of both classification analysis and association rule mining to understand the risk levels of stocks may be expanded to other stock exchanges of other emerging countries by using different and additional vital variables. Besides, to see the effect of the COVID-19 pandemic on stock markets, the proposed approach may be used by using pandemic period data, and obtained findings can be compared with the pre-pandemic period findings. Such a comparison of pre-and post- pandemic periods may provide insights into how the pandemic situation affects the risk levels of stock markets.

ORCID

Görkem Ataman https://orcid.org/0000-0002-8290-2248
Serpil Kahraman https://orcid.org/0000-0003-4570-1604

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *SIGMOD Record, 22*(2), 207-216. http://dx.doi.org/10.1145/170036.170072

Agrawal, R., & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules in Large Databases*. Paper presented at the Proceedings of the 20th International Conference on Very Large Data Bases.

Akerlof, G. A. (1970). The Market for "lemons": Quality, uncertainty and the market mechanism. *The Quarterly Journal of Economics, 84*(3), 488-500. http://dx.doi.org/10.2307/1879431

Albayrak, A. S., & Koltan Yılmaz, Ş. (2009). Veri madenciliği karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama. *Suleyman Demirel Universitesi Iktisadi ve Idari Bilimler Fakultesi Dergisi, 14*(1), 31-52.

Angeletos, G.-M., Huo, Z., & Sastry, K. A. (2020). Imperfect Macroeconomic Expectations: Evidence and Theory. *National Bureau of Economic Research Working Paper Series, 27308*. http://dx.doi.org/10.3386/w27308

Bastı, E., Kuzey, C., & Delen, D. (2015). Analyzing initial public offerings' short-term performance using decision trees and SVMs. *Decision Support Systems, 73*, 15-27. http://dx.doi.org/10.1016/j.dss.2015.02.011

Bhasin, M. L. (2006). Data Mining: A competitive tool in the banking and retail industries. *The Chartered Accountant*, 588-594.

Bhide, A. (1993). The hidden cost of stock market liquidity. *Journal of Financial Economics, 34*(1), 31-51. http://dx.doi.org/10.1016/0304-405X(93)90039-E

Bordalo, P., Gennaioili, P., La Porta, R., & Shleifer, A. (2019). Diagnostic expectations and stock returns. *The Journal of Finance, LXXIV*(6), 2839-2874. http://dx.doi.org/10.1111/jofi.12833

Borsa Istanbul. (2019a). BIST Stock Indices Ground Rules. Retrieved from https://www.borsaistanbul.com/en/sayfa/3621/equity-market-data

Borsa Istanbul. (2019b). Data. Retrieved from http://borsaistanbul.com/en/data/data/ipo-data

Borsa Istanbul. (2022). Daily settlement prices. Retrieved from https://www.borsaistanbul.com/en/sayfa/3066/daily-settlement-prices

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*: Chapman & Hall/CRC.

Cao, Y. (2021). *Application of machine learning algorithms in financial market risk prediction.* Paper presented at the International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy

Chen, Y., Mo, D., & Zhang, F. (2022). *Stock market prediction using weighted inter-transaction class association rule mining and evolutionary algorithm*: Economic Research-Ekonomska Istraživanja. http://dx.doi.org/10.1080/1331677X.2022.2043762

Chen, Y., Yang, B., & Abraham, A. (2007). Flexible neural trees ensemble for stock index modelling. *Neurocomputing, 70*, 697-703. http://dx.doi.org/10.1016/j.neucom.2006.10.005

Denison, D. G., Mallick, B. K., & Smith, A. F. (1998). A bayesian CART algorithm. *Biometrika, 85*(2), 363-377. http://dx.doi.org/10.1093/biomet/85.2.363

Ekinci, C., & Ersan, O. (2018). A new approach for detecting high-frequency trading from order and trade data. *Finance Research Letters, 24*, 313-320. http://dx.doi.org/10.1016/j.frl.2017.09.020

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications, 29*, 927-940. http://dx.doi.org/10.1016/j.eswa.2005.06.024

Fama, E. (1970). Efficient Capital Markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383-417. http://dx.doi.org/10.2307/2325486

Filiz, E., & Öz, E. (2017). Classification of BIST-100 index'changes via machine learning methods. *Marmara Universitesi Iktisadi ve Idari Bilimler Fakultesi Dergisi, 39*(1), 117-129. http://dx.doi.org/10.14780/muiibd.329913

Giglio, S., Maggiori, M., Stroebel, J., & Utkus, S. (2020). Inside the mind of a stock market crash. *NBER Working Papers, 27272*.

Hajizadeh, E., Ardakani, H. D., & Shahrabi, J. (2010). Application of data mining techniques in stock markets: A survey. *Journal of Economics and International Finance, 2*(7), 109-117.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl., 11*(1), 10–18. http://dx.doi.org/10.1145/1656274.1656278

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*: Elsevier.

Hashmi, S. M., & Chang, B. H. (2021). Asymmetric effect of macroeconomic variables on the emerging stock indices: A quantile ARDL approach. *International Journal of Finance & Economics*, ijfe.2461. http://dx.doi.org/10.1002/ijfe.2461

Huang, J. Y., & Liu, J. H. (2020). Using social media mining technology to improve stock price forecast accuracy. *Journal of Forecasting, 39*(1), 104-116. http://dx.doi.org/10.1002/for.2616

Hurd, M. D., Roojin, M., & Winter, J. (2010). Stock Market Expectations of Dutch households. *NBER Working Papers, 16464*.

Inidapo, I., Adebiyi, A., & Okesola, O. (2017). Soft computing techniques for stock market prediction: A literature survey. *Covenant Journal of Informatics & Communication Technology, 5*(2), 1-28.

Kara, Y., Acar Boyacıoglu, M., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications, 38*, 5311-5319. http://dx.doi.org/10.1016/j.eswa.2010.10.027

Kartal, B., Sert, M. F., & Kutlu, M. (2022). Determination of the world stock indices' co-movements by association rule mining. *Journal of Economics, Finance and Administrative Science*. http://dx.doi.org/10.1108/JEFAS-04-2020-0150

Kezdi, G., & Robert, J. W. (2009). *Stock Market Expectations and Portfolio Choice of American Households*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.506.1967&rep=rep1&type=pdf

Kumar, G., Jain, S., & Singh, U. P. (2021). Stock market forecasting using computational intelligence: A survey. *Archives of Computational Methods in Engineering*, *28*(3), 1069-1101. http://dx.doi.org/10.1007/s11831-020-09413-5

Martins, O. C., & Paulo, E. (2012). Information Asymmetry in Stock Trading, Economic and Financial Characteristics and Corporate Governance in the Brazilian Stock Market. *Revista Contabilidade & Finanças*, *25*(64), 33-45. http://dx.doi.org/10.1590/S1519-70772014000100004

Onali, E. (2020). Covid19 and stock market volatility. *SSRN, 3571453*, 2020.

Ou, P., & Wang, H. (2009). Prediction of stock market index movement by ten data mining techniques. *Modern Applied Science*, *3*(12), 28-42. http://dx.doi.org/10.5539/mas.v3n12p28

Paranjape-Voditel, P., & Deshpante, U. (2013). A Stock market portfolio recommender system based on association rule mining. *Applied Soft Computing*, *13*, 1055-1063. http://dx.doi.org/10.1016/j.asoc.2012.09.012

Park, M., & Chai, S. (2021). A machine learning-based model for the asymmetric prediction of accounting and financial information *Fintech with Artificial Intelligence, Big Data, and Blockchain* (pp. 181-190): Springer. http://dx.doi.org/10.1007/978-981-33-6137-9_7

Parkash, R., Ahmad, R., Qasim, S., & Nizam, K. (2022). Investor Sentiments and Stock Risk and Return: Evidence from Asian Stock Markets. *Competitive Social Science Research Journal*, *3*(1), 341-371.

Pawar, K., Jalem, R. S., & Tiwari, V. (2019). Stock market price prediction using LSTM RNN *Emerging trends in expert applications and security* (pp. 493-503): Springer. http://dx.doi.org/10.1007/978-981-13-2285-3_58

Uzar, C. (2014). The usage of data mining technology in financial information system: An application on Borsa Istanbul. *International Journal of Finance & Banking Studies*, *3*(1), 51-61. http://dx.doi.org/10.20525/ijfbs.v3i168

Vaiz, J. S., & Ramaswami, M. (2016). A study on technical indicators in stock price movement prediction using decision tree algorithms. *American Journal of Engineering Research*, *5*(12), 207-212.

Valente, F., Henriques, J., Paredes, S., Rocha, T., de Carvalho, P., & Morais, J. (2021). *Improving the compromise between accuracy, interpretability and personalization of rule-based machine learning in medical problems*. Paper presented at the 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society.

Waspada, I. P., Salim, D. F., & Fariska, P. (2021). An Application of the Smart Beta Portfolio Model: An Empirical Study in Indonesia Stock Exchange. *Journal of Asian Finance. Economics and Business*, *8*(9), 45-52.

World Federation of Exchanges, W. (2019). Retrieved from http://www.world-exchanges.org/statistics/monthly-query-tool

Wu, M., Hughes, M., Parbhoo, S., Zazzi, M., Roth, V., & Doshi-Velez, F. (2018). Beyond sparsity: Tree regularization of deep models for interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), •••. http://dx.doi.org/10.1609/aaai.v32i1.11501

Wu, X., Chen, H., Wang, J., Troiano, L., Loia, V., & Fujşta, H. (2020). Adaptive stock trading strategies with deep reinforcement learning methods. *Information Sciences*, *538*, 142-158. http://dx.doi.org/10.1016/j.ins.2020.05.066

Xu, H., Cao, D., & Li, S. (2022). A self-regulated generative adversarial network for stock price movement prediction based on the historical price and tweets. *Knowledge-Based Systems*, *247*, 108712. http://dx.doi.org/10.1016/j.knosys.2022.108712

Yiğit, F., & Muzır, E. (2019). Efficiency of the major Borsa Istanbul Indexes: An empirical investigation about the interaction between corporate governance and equity prices through a market model approach. *Ekonomi, İşletme ve Maliye Araştırmaları Dergisi, 1*(3), 237-245.

Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: data mining in financial application. *IEEE Trans. Syst. Man Cybern. Part C, 34*, 513-522.

Zhou, L., Chen, Q., & Zhu, T. (2022). An Improved Data Mining Model for Predicting the Impact of Economic Fluctuations. *Security and Communication Networks, 2022*, 1-11. http://dx.doi.org/10.1155/2022/2173402

To cite this article: Ataman, G., Kahraman, S. (2022). Comparing Decision Trees and Association Rules for Stock Market Expectations in BIST100 and BIST30. *Scientific Annals of Economics and Business*, 69(3), 459-475. https://doi.org/10.47743/saeb-2022-0024

Copyright