Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi • Cilt: 39 • Sayı: 1 • Haziran 2017, ISSN: 2149-1844, ss/pp. 117-129
DOI: 10.14780/muiibd.329913

CLASSIFICATION OF BIST-100 INDEX' CHANGES VIA MACHINE LEARNING METHODS

Enes FİLİZ*
Ersoy ÖZ**

Abstract

The changes in BIST-100 index are economically crucial. In this study, classifications will be made with the assumption that the changes in BIST-100 index are dependent on certain factors. The classifiers to be used are k-nearest neighbor algorithm, naive Bayes Classifier, logistic regression and C4.5 classifier from the machine learning methods. Factors affecting the change of BIST-100 index values are deemed as Euro/Dollar Parity, Gold value (ounce), Crude Oil Prices, Monthly Interest Rates, Inflation Data and DAX, FTSE, S&P 500 that are widely used in the literature. As a result of the transactions performed via Weka program, the most successful methods in order are C4.5 classifier algorithm (66.2%) and logistic regression analysis (65.9%).

Keywords: *BIST-100 Index, Machine Learning Methods, Classification.*

JEL Classification: C02, C19, C44,F31

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE BIST-100 ENDEKSİ DEĞİŞİMLERİNİN SINIFLANDIRILMASI

Özet

BİST-100 endeksinde meydana gelen değişimler ekonomik anlamda son derece önemlidir. Bu çalışmada, BİST-100 endeksinde görülen değişmelerin bazı faktörlere bağlı olduğu varsayımı ile sınıflandırmalar yapılacaktır. Kullanılacak sınıflandırıcılar makine öğrenmesi yöntemlerinden, k en yakın komşu algoritması, basit(naive) bayes sınıflandırıcısı, lojistik regresyon ve C4.5 sınıflandırma algoritmasıdır. BİST-100 endeksi değerinin değişimine etki eden faktörler ise literatürde en sık kullanılan Euro/Dolar Paritesi, Altın değeri(ons), Ham Petrol Fiyatı, Aylık Faiz Oranları, Enflasyon Verileri, DAX, FTSE, S&P 500 olarak alınmıştır. Weka programı kullanılarak yapılan işlemler sonucunda incelenen yöntemlerden en başarılı olanlar sırasıyla C4.5 sınıflandırma algoritması(%66,2) ve lojistik regresyon analizi(%65,9) olarak bulunmuştur.

Anahtar Kelimeler: *BİST-100 Endeksi, Makine Öğrenmesi Teknikleri, Sınıflandırma.*

JEL Sınıflaması: C02, C19, C44,F31

[*] Res. Assist. Enes FİLİZ, Yildiz Technical University, Department of Statistics, enesf@yildiz.edu.tr.

[**] Assoc. Prof. Dr. Ersoy ÖZ, Yildiz Technical University, Department of Statistics, ersoyoz@yildiz.edu.tr.

1. Introduction

It is evident that within the last few years, many developments are taking place in financial markets and therefore, in the stock markets. Due to globalization, it is observed that any changes or political developments affect all stock markets in the world. Therefore, Istanbul Stock Exchange (ISE) has also been affected from these developments. [1]

In 2013, Istanbul Stock Exchange merged with the gold exchange; future exchange and option exchange have merged and were named as Borsa Istanbul. Therefore, the ISE-100 index has changed as BIST-100 index. In this study, the original names used in the publications were used. In other words, Borsa Istanbul and BIST-100 terms have been avoided.

It is inevitable for the success of publicly-traded companies at ISE to be affected by the globalization problem, mainly the foreign capital inflows. [2]

There are many national and international factors that affect the stock market indexes. Some of these factors are Euro/Dollar Parity, daily foreign exchange rates, status of other stock markets, interest and inflation values, government debts, crude oil prices and gold prices. [3]

The estimation of yield of share certificates is important for analysts. Many analyses have been carried out with different assessments and methods based on these estimations in the past. With the technological advances in recent years, many studies using data mining in addition to conventional methods on ISE have been very successful; thus, there has been an increase in the interest in data mining. [4]

With the transfer of data to digital media, the amount of information in the world has been increasing incrementally. Thus, the number of databases also increases depending on such growth. Therefore, data storage has been simplified; data became cheaper and more accessible. Data mining is an area that connects many methods including database technology, artificial intelligence, machine learning, statistics, pattern recognition and data visualization. In addition to these areas, data mining is also applied in many other areas including economy, finance, marketing, insurance, health, and biology. [5]

Machine learning can be deemed as techniques based on learning from data or computers to learn how to solve problems on their own. Algorithms that will classifying and clustering of output are

[1] Bengü VURAN, **İMKB 100 endeksinin uluslararası hisse senedi endeksleri ile ilişkisinin eşbütünleşim analizi ile belirlenmesi**, Istanbul University Journal of the School of Business, 2010, p. 154-168.

[2] İlhan EGE- Ali BAYRAKDAROĞLU, **İMKB şirketlerinin hisse senedi getiri başarılarının lojistik regresyon tekniği ile analizi**, Zonguldak Karaelmas UniversityJournal of Social Sciences, 2009, p.139-158.

[3] Selçuk BALI- Mehmet Ozan CİNEL, **Altın fiyatlarının İMKB 100 endeksi'ne etkisi ve bu etkinin ölçümlenmesi**, Ataturk University Journal of Economics and Administrative Sciences, 2011, p. 45-63.

[4] Nezih TAYYAR- Selin TEKİN, **İMKB-100 endeksinin destek vektör makineleri ile günlük, haftalık ve aylık veriler kullanılarak tahmin edilmesi**, Abant İzzet Baysal University Journal of Social Sciences, 2013, p. 189-217.

[5] Serhat ÖZEKES, **Veri Madenciliği Modelleri Ve Uygulama Alanları**, İstanbul Ticaret Üniversitesi Dergisi, 2003, p. 65-82.

created. Thus, labor and costs decrease. Various algorithms are used in this method. If the output in the data set is known, supervised algorithms are used, if not, unsupervised algorithms are used. There are many studies in the literature that benefit from the method of machine learning through deterministic and randomized algorithms. [6]

The factors affecting machine learning respectively are the data set, a variable that may affect the results, determination of learning strategy, algorithm and parameters of the algorithm. In addition, this method is based on three fundamental studies. These are duty-oriented studies (development of the system to increase performance), cognitive simulations (transferring information belonging to humans to computers), and theoretical analysis (theoretical examination of algorithms and methods). [7]

Machine learning is used in almost all areas in our life. The algorithms suggested in the literature form the basis of machine learning. These algorithms have been instructive in classification and estimation process. The main algorithms used in machine learning are k-nearest neighbor algorithm, naive Bayes classifier, logistic regression analysis, decision trees, k-average algorithm and support vector machines.

In this study, classifications and required comparisons will be made through k-nearest neighbor algorithm, naive Bayes classifier, and logistic regression and C4.5 classifier from the machine learning methods for BIST-100 index changes.

In the second part of the study, literature review on ISE-100, machine learning and ISE-100 and machine learning are used together. Then, the data set and methods to be used will be explained and applied in the fourth section and finally in the last chapter, findings will be given in detailed.

2. Literature Review

2.1. Literature on ISE-100

Many analysis and inferences have been conducted on ISE-100 using various methods within the last few years. Vuran (2010) examined the relation between ISE-100 index with international share certificates. He used cointegration analysis in this study. As a result of the study, it was determined that ISE-100 index between January 2006 and January 2009 has a long-term relation with FTSE-100, Dax, Bovespa, Merval and IPC indexes.

In the studies that examine the relationship between share certificates and foreign exchange rates, mostly the foreign exchange rate and value of a foreign exchange in the national currency are used in calculations. However, some studies use the real foreign exchange rates. [8]

[6] Hatice NİZAM- Saliha Sıla AKIN, **Sosyal Medyada Makine Öğrenmesi İle Duygu Analizinde Dengeli Ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması**, XIX. Türkiye'de İnternet Konferansı, 2014, inet-tr.org.tr.

[7] Mehmet Erdal BALABAN- Elif KARTAL, **Veri Madenciliği Ve Makine Öğrenmesi**, İstanbul, Çağlayan Kitabevi, 2015, p. 29-30.

[8] İncilay SAVAŞ- İsmail CAN, **Euro-Dolar Paritesi ve Reel Döviz Kuru'nun İMKB 100 Endeksi'ne Etkisi**, Eskişehir

Balı and Cinel (2011)investigated the effect of gold prices on ISE-100 index in their study. As a result of the study spanning August 1995 and March 2011, it was determined that gold prices do not have a direct effect on ISE-100 index. However, they indicated that there is any other factor between parameters that explain the changes in ISE-100 index. Another important finding was that all changes in the balance of foreign trade are more determinant in terms of the effect than the foreign exchange rates.

Savaş and Can (2011) examined the effect of Euro-Dollar parity and real foreign exchange rates on the ISE-100 index. In this study, the relation was determined with multiple regression analysis and the direction of the relation with Granger causality test. The study used data collected between January 2000 and July 2009 and was used to revealthat Euro-Dollar parity and real foreign exchange rates explain the ISE-100 index by 77.5% and that its direction is positive.

Tayyar and Tekin (2013) used support vector machines to estimate the ISE-100 index. In their study, the classification success of support vector machines was compared with logistic regression. In total, 4226 datawas used between April 3, 1995 and March 19, 2012. It was determined that within 12 models of support vector machines, weekly model 1 (70%) estimated the direction of ISE-100 index the best.

2.2. Literature on Machine Learning

The history of machine learning dates back to recent times. Especially, after human brain was adapted to machines, many studies have been conducted following this method. There are applications of machine learning in many areas. The first studies on machine learning were conducted in 1990s.

Mitchell (1997) offered one of the best definitions in literature for machine learning. In addition, he revealed that machine learning and data mining are connected. [9]

The most important study on supervised learning was realized by Pang et al. (2002). By applying support vector machines, naive Bayes and maximum entropy classifiers to movie reviews, "Movie Review" data set. [10]

With the increasing interest in the last few years, many different studies have been conducted on machine learning. Kavzaoğlu and Çölkesen's (2010) study examined Kernel functions on the classification of satellite images by using support vector machines. The classification accuracy of four most widely used Kernel functions was determined in the analysis. Radial-based function and Pearson VII (<94%) had the highest performance. Normalized polynomial Kernel functions had the lowest classification accuracy (91.78%).

Osmangazi University Journal of Economics and Administrative Sciences, 2011, p. 323- 339.

[9] Mehmet Erdal BALABAN- Elif KARTAL, **Veri Madenciliği Ve Makine Öğrenmesi a.g.k**, p. 26-27.

[10] Aytuğ ONAN- Serdar KORUKOĞLU, **Makine Öğrenmesi yöntemlerinin görüş madenciliğinde kullanılması üzerine bir literatür araştırması**, Pamukkale University Journal of Engineering Sciences, 2016, p. 111-122.

Solmaz, Günay and Alkan (2013) revealed the effect of thyroid disorder diagnosis of Expert systems. In this study, they used a feed-forward artificial neural network and linear discriminant analysis with support vector machines. As a result of the study, it was determined that all methods had an acceptable success. Classification methods (93.48% - 96.57%) were determined to be more successful than clustering methods (88.83%-90.68%).

Meral and Diri (2014) conducted an emotion analysis on Twitter. In this study, they used naive Bayes, random tree and support vector algorithms used in machine learning. The findings of the study were provided as a comparison.

Erdal (2015) examined the benefits of using machine learning methods in construction sector based on compressive strength estimation. The study benefited from support vector machines and artificial neural networks. High estimation values were obtained at the end of the study. It was determined that support vector machines provide more successful estimations than artificial neural networks.

Bulut (2016) examined success evaluation of controlled students in unbalanced data sets. He used decision trees, k-nearest neighbor classifier, naive Bayes, support vector machines and logistic regression model. The most success classified was respectively determined to be logistic regression model, naive Bayes classifier and decision tree algorithms.

Onan and Korukoğlu (2016) conducted a literature review on using machine learning methods on opinion mining. They examined unsupervised, half-supervised and supervised methods. The strengths and weaknesses of machine learning on opinion mining was determined. It was revealed that supervised learning methods presented better results in larger data sets.

2.3. Literature on Machine Learning and ISE-100

ISE-100 data have been frequently used in machine learning methods. There are various related studies in literature. Koyuncugil and Özgülbaş (2008) determined the weaknesses and strengths of SMEs traded in ISE via CHAID decision tree algorithm. 697 SMEs traded in ISE between 2000 and 2005 were examined. SMEs traded in ISE were listed under 19 profiles. It was determined that equity productivity, asset productivity, financing of fixed assets, strategies for managing receivables and liquidity are factors on which SMEs strengths and weaknesses depend.

Albayrak and KoltanYılmaz (2009) used ISE-100 data and analyzed it via decision tree algorithms. In this study, annual financial indicators of 173 businesses in service and industry sectors between 2004 and 2006 according to ISE-100 index were examined. Most important variables that distinguish between firms' financial indicators were determined.

Ege and Bayrakdaroğlu's (2009) study analyzed the yield per share certificate of ISE companies via logistic regression. Twenty financial rates and nominal yields in 2004 of 30 companies traded in ISE were used. According to the findings of the study, Price/Earnings ratio, Cash ratio and Turnover Rate of Total Assets are independent variables effecting share certificate yields.

Özdemir, Tolun and Demirci (2011) benefitted from ISE-100 index while estimating the index yield via double classification method. They have reached the results by using logistic regression and support vector machine methods as well. Correct classification rates for both the modeling and estimation clusters were deemed as 75% and 86%.

3. Materials and Methods

3.1. Data Set

A 10-year period was used in this study. Daily data was collected between January 1, 2006 and December 1, 2016. In total, 2618 data was used. These data respectively are BIST-100 index, Euro/Dollar Parity, Gold values (ounce), Crude Oil Prices, DAX, FTSE, S&P 500, inflation (consumer price index) and interest rates. Instant stock market data, Euro/Dollar Parity, price of gold (ounce) and oil prices were gathered from a website called investing that examines market movements. [11]Inflation data determined on monthly basis were based on a website called bigpara. [12]Again current interest rates were obtained from the official website of the Turkish Central Bank. [13]As the BIST-100 index, Euro/Dollar Parity, Gold prices (ounce) Crude Oil Prices, DAX, FTSE and S&P 500 data change on daily basis, increases were shown as 1 and decreases as 0. As inflation data and interest rates are announced on monthly basis, they have been written equally for each day. BIST-100 index has been deemed as dependent variable, while Euro/Dollar Prity, Gold prices (ounce), Crude Oil Prices, DAX, FTSE, S&P 500, inflation (consumer price index) and interest rates have been deemed as independent variables. These variables were selected since they are the most widely used and encountered variables in the literature.

3.2. Methods

3.2.1. k-Nearest Neighbor Algorithm (k-NN)

k-NN is a widely used non-parametric classifier. The class of new sample is determined according to a k value determined for the sample by calculating its distance to other specimens in the sample. K value selected is crucial for the success of the algorithm. [14]

This method determines classification based on training set. The training set is consulted for each tested point. The majority the k-number of sample belong to determine the classification results. There are various studies for k value of k-NN algorithm in the literature. [15]

[11] Investing web page, *http://tr.investing.com*, Access Date: 13.12.2016.

[12] Bigpara web page, *http://www.bigpara.com*, Access Date: 02.12.2016.

[13] Turkish Central Bank web page, *http://www.tcmb.gov.tr*, Access Date: 29.11.2016.

[14] Mehmet Erdal BALABAN- Elif KARTAL, **Veri Madenciliği Ve Makine Öğrenmesi a.g.k.**,p.60-61.

[15] Faruk BULUT, **Dengesiz Veri Setlerinde Denetimli Öğrenicilerin Başarım Değerlendirmesi**, IEEE, 2016.

The basic formula is below;

$$Oy(x_i) = \begin{cases} \infty & , \quad if\ d(x_i, q) = 0 \\ \dfrac{1}{d(x_i, q)} & , \quad otherwise \end{cases}$$

In this formula x_i is a random sample received from space $x \in X$. [16]$d$shows distance. Neighbors get a higher vote. q is used for determining the opposite of the distance between the neighbor and the point of which the class is being researched. [17]

3.2.2. Naive Bayes Classifier

Naive Bayes Classifier assumes that all attributes of the samples are independent of given class. [18]This classifier is stands out as one of the most productive and efficient inductive learning algorithms within machine learning methods and in data mining. [19]Naive Bayes classifier examines conditional probability between random X situation and Y situation within a stochastic process. [20]This algorithm is usually used for classification of the available data with compound probability. [21]It especially leads to successful result in text classification. [22]

The general formula is below;

$$P(\vec{x}|c_j) = \prod_{i=1}^{n} P(x_{a_i}|c_j) \qquad i = 1, 2, \ldots, n\,; j = 1, 2, \ldots, k$$

In this formula; $X = \begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_m \end{bmatrix}$ is a sample space formed of m samples.

$\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \in \mathcal{R}^{m*n}$, n is the number of attributes and shows the observation matrix formed of m number of data. $c_1, c_2, \ldots\ c_k$are the class values in the sample space. [23]

[16] MİTCHELL, T., M., **Machine Learning (1st Edition)**, McGraw-Hill Science /Engineering / Math, 1997.

[17] Mehmet Erdal BALABAN- Elif KARTAL, **Veri Madenciliği Ve Makine Öğrenmesi a.g.k.**, p.61-62.

[18] McCALLUM, A., NİGAM, K., **A Comparison of Event Models For Naive Text Classification**, AAAI-98 Workshop On Learning For Text Categorization.

[19] ZHANG, H.,**The Optimality of Naive Bayes**, 2004.

[20] Faruk BULUT, **Dengesiz Veri Setlerinde Denetimli Öğrenicilerin Başarım Değerlendirmesi a.g.m.**, IEEE, 2016.

[21] M. Fatih AMASYALI ve diğerleri, **Farklı Özellik Vektörleri İle Türkçe Dökümanların Yazarlarının Belirlenmesi**, TAINN-2006.

[22] Hatice NİZAM- Saliha Sıla AKIN, **Sosyal Medyada Makine Öğrenmesi İle Duygu Analizinde Dengeli Ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması a.g.m.**, XIX. Türkiye'de İnternet Konferansı, 2014, inet-tr.org.tr.

[23] Mehmet Erdal BALABAN- Elif KARTAL, **Veri Madenciliği Ve Makine Öğrenmesi a.g.k.**, p.70.

3.2.3. Logistic Regression Analysis

The purpose of the logistic regression analysis is to obtain a model that can define the relation or relations between dependent and independent variables by using the least number of variables and that has the best conformity. [24]This method is chosen as it does not require the assumptions valid in linear regression and is more flexible. [25]The dependent variable to be categorical is the main difference between linear regression and logistic regression analysis. [26]Another difference is it has fewerconstraints than the least squares method when examined from the assumptions in logistic regression. [27]

Parameters are not obtained analytical in logistic regression. Therefore, maximum likelihood method, which is an iterative method, is used for making assumptions. [28]

Logistic function's formula can be written as below;

$$g(z) = \frac{1}{1+e^{-z}} \, , \mathcal{R} \rightarrow [0,1]$$

In this formula $g(z)$ value is between 0 and 1. If the value is less more than 0.5$0,5$ it approaches 1, if it is less, it approaches 0. [29]

3.2.4. C4.5 Classification Algorithm

When making classification via using decision trees method, data set is used to create a decision tree. The rules required for estimation are determined. These rules enable programmers to write programs in an easier manner. [30]It is one of the most important classification techniques easily understandable with its tree structure that sets rules and is used in information technologies. [31]

Algorithms such as ID3, C4.5, CART, Random Forest, or Raptree, etc. are used in decision trees. The main purpose of these algorithms is to minimize generalization error and to establish a decision tree from a given data set. [32]

[24] Ömay ÇOKLUK, **Lojistik regresyon Analizi: Kavram ve Uygulama**, Educational Sciences: Theory & Practice, 2010- Dinçer ATASOY, **Lojistik Regresyon Analizinin İncelenmesi ve Bir Uygulaması**, Graduate thesis, Cumhuriyet University, Institute of Social Sciences, Sivas, 2001.

[25] Kadir SÜMBÜLOĞLU, **Lojistik Regresyon Analizi**, 2015.

[26] Cengiz AKTAŞ, **Lojistik Regresyon Analizi: Öğrencilerin Sigara İçme Alışkanlığı Üzerine Bir Uygulama**, Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2009.

[27] A. Kerem ÖZDEMİR ve diğerleri, **Endeks Getirisi Yönünün İkili Sınıflandırma Yöntemiyle Tahmin Edilmesi: İMKB-100 Endeksi Örneği**, Niğde UniversityJournal of Economics and Administrative Sciences, 2011, p. 45-59.

[28] Ali Sait ALBAYRAK,**Uygulamalı Çok Değişkenli İstatistik Teknikler**, 2006.

[29] Mehmet Erdal BALABAN- Elif KARTAL, **Veri Madenciliği Ve Makine Öğrenmesi a.g.k.**, p.80.

[30] Mehmet Erdal BALABAN- Elif KARTAL, **Veri Madenciliği Ve Makine Öğrenmesi a.g.k.**, p.96.

[31] Serhat ÖZEKES, **Veri Madenciliği Modelleri Ve Uygulama Alanları a.g.m.**,İstanbul Ticaret Üniversitesi Dergisi, 2003, p. 65-82.

[32] ROKACH, L., MAİMON, O., **Decision Trees, Data Mining and Knowledge Discovery Handbook**, Springer, 2005.

C4.5 algorithm is a classical method used to represent information in machine learning algorithm. It also offers a strong solution for expressing data structures. It provides the highest accuracy for training data; however it may put extreme rules for some of the behaviors of certain data. [33]

3.3. Classification Criteria

Certain criteria related with classification should be examined in this study. The criteria to be examined will enable us to analyze the results of the classification in the best manner possible. The classification criteria are determined respectively as TP Rate, FP Rate, Precision, Recall, F-measure, MCC, ROC Area, and PRC Area. These classification criteria are calculated by using a confusion matrix. A confusion matrix is a matrix that indicates in comparison correct and incorrect classifications in a model. Table 4.1 shows the sections of the confusion matrix.

Tablo 4.1: Show that Confusion Matrix

		Real Value	
Estimate		**a**	**b**
	a	TN(True Negative Rate)	FP(False Positive Rate)
	b	FN(False Negative Rate)	TP(True Positive Rate)

The classification criteria calculated by using Table 4.1 are as below:

TP Rate(recall): determined by the ratio of positive samples correctly classified in the model to the total positive sample number.

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

FP Rate: determined by the ratio of normally negative but classified as positive samples to the number of total negative samples.

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

Precision: determined by the ratio of positive samples correctly classified in the model to the total positive estimated samples.

$$Precision = \frac{TP}{TP + FP}$$

F-measure: determined by the harmonic average of the precision value and TP Rate.

[33] Hatice NİZAM- Saliha Sıla AKIN, **Sosyal Medyada Makine Öğrenmesi İle Duygu Analizinde Dengeli Ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması a.g.m.**, XIX. Türkiye'de İnternet Konferansı, 2014, inet-tr.org.tr.

$$F - measure = \frac{2 * Precision * TP\ Rate}{Precision + TP\ Rate}$$

MCC(Matthews correlation coefficient): Obtained by using comparison matrix's elements and is between -1 and 1.

$$MCC = \frac{(TP\,Rate * TN\,Rate) - (FP\,Rate * FN\,Rate)}{\sqrt{(TP\,Rate + FP\,Rate) * (TP\,Rate * FN\,Rate) * (TN\,Rate * FP\,Rate) * (TN\,Rate * FN\,Rate)}}$$

ROC Area: Indicates the areas on X and Y axis between the TP rate and the FP rate. When one of them increases the other one decrease. The one above shows the TP rate area, the one below the FP area. [34]

PRC Area(Precision-Recall Curve): It is a two dimensional graphic similar to ROC curve. Precision value is shown on the Y axis and TP Rate on the X axis.

4. Application

In this study, classifications will be made with the assumption that the changes in BIST-100 index are dependent on certain factors. These factors were examined and included in the study. The success rates of methods have been evaluated. 10-year daily data between January 01, 2006 and December 01, 2016 was obtained and used. In total, 2618 data was used. Instant stock market data, Euro/Dollar Parity, price of gold (ounce) and oil prices were gathered from a website called investing that examines market movements. [35]Inflation data determined on monthly basis were based on a website called bigpara. [36]Again monthly instant interest rates were obtained from the official website of the Turkish Central Bank. [37]BIST-100 index has been deemed as dependent variable, while Euro/Dollar Parity, Gold prices (ounce), Crude Oil Prices, DAX, FTSE, S&P 500, inflation and interest rates were deemed as independent variables. Calculations were made by using Weka program. The output from this program is indicated in order in Table 4.2 and Table 4.3.

In machine learning methods, thedata set is divided into two groups as training and test. This operation can be determined by numbers such as 50%-50%, 70%-30%. In addition, through k-fold cross validation method, the data set can be divided into k equal parts and with k-1 training and the test with the one remaining can be carried out. This operation can be repeated for k times by using each part as a test cluster. Afterwards, the average of all the results is taken and the value for classification criteria is determined. In this study, k value was deemed as 10 and 10-fold cross validation was carried out.

[34] Mehmet Erdal BALABAN- Elif KARTAL, **Veri Madenciliği Ve Makine Öğrenmesi a.g.k.**, p.49-53.

[35] Investing web page, *http://tr.investing.com*, Access Date: 13.12.2016.

[36] Bigpara web page, *http://www.bigpara.com*, Access Date: 02.12.2016.

[37] Turkish Central Bank web page, *http://www.tcmb.gov.tr*, Access Date: 29.11.2016.

Tablo 4.2: Confusion Matrix for the Algorithms

	k-NN		Naive Bayes		C4.5		Logistic Regression	
classified as	**a**	**b**	**a**	**b**	**a**	**b**	**a**	**b**
a=0	726	537	785	478	720	543	729	471
b=1	606	749	430	925	342	1013	423	932

Tablo 4.3: Classification Results for the Algorithms

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
k-NN	0,575	0,447	0,545	0,575	0,560	0,128	0,583	0,551	0
	0,553	0,425	0,582	0,553	0,567	0,128	0,583	0,584	1
	0,563	0,436	0,564	0,563	0,564	0,128	0,583	0,568	**W.A.**
Naive Bayes	0,622	0,317	0,646	0,622	0,634	0,305	0,698	0,671	0
	0,683	0,378	0,659	0,683	0,671	0,305	0,698	0,681	1
	0,653	0,349	0,653	0,653	0,653	0,305	0,698	0,676	**W.A.**
C4.5	0,570	0,252	0,678	0,570	0,619	0,323	0,671	0,645	0
	0,748	0,430	0,651	0,748	0,696	0,323	0,671	0,629	1
	0,662	0,344	0,664	0,662	0,659	0,323	0,671	0,637	**W.A.**
Logistic Regression	0,627	0,312	0,652	0,627	0,639	0,316	0,699	0,672	0
	0,688	0,373	0,664	0,688	0,676	0,316	0,699	0,680	1
	0,659	0,344	0,658	0,659	0,658	0,316	0,699	0,677	**W.A.**

W.A.= Weighted Average

5. Conclusion

The main objective of this study is to carry out classifications by using machine learning methods with the assumption that changes observed in BIST-100 is dependent on certain factors. In line with this objective, methods were determined and factors affecting BIST-100 index were examined and the most widely used were selected.

In order to determine the success of classification, as a result of the analysis performed via Weka program, the algorithm values for the confusion matrix have been indicated in Table 4.2. "a" value shows "0" and "b" shows "1". These expressions have been used to indicate correct positive, false positive, correct negative and false negative values in classification. The success of classification can be determined from the weighted average section of TP Rates in Table 4.3. According to Table 4.3 the classification success of k-NN algorithm is 56.3%, the classification success of Naive Bayes Algorithm is 65.3%, classification success of C4.5 algorithm is 66.2% and classification success of logistic regression analysis is 65.9%. According to these results, the most successful classification was carried out by C4.5 algorithm. Logistic regression analysis is the second most successful algorithm for classification. The algorithm with the least success is found as k-NN.

The factors affecting BIST-100 act as important factors for achieving the classification success results. When values of factors affecting BIST-100 and included in this study are used; it can be asserted that the estimation success of decrease and increases in BIST-100 index of C4.5 algorithm is 66.2%. In other words, if the changes in the factors affecting the BIST-100 index are known, we have 66.2% success of estimating whether the index will increase or decrease when compared to the day before.

When the results are examined, the most import value is revealed as the correct classification rate. This value shows the success of the used machine learning method. Other results may be obtained by using different algorithms than the ones used in this study. By using these methods, different perspectives may be developed and used easily in daily life. In addition, it may be possible to increase the correct classification rate by using different factors affecting the BIST-100 index.

6. References

AKTAŞ, Cengiz, **Lojistik Regresyon Analizi: Öğrencilerin Sigara İçme Alışkanlığı Üzerine Bir Uygulama**,Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, (26), 2009, p. 107-121.

ALBAYRAK, Ali Sait, **Uygulamalı Çok Değişkenli İstatistik Teknikler**, Asil Yayın Dağıtım Ltd. Şti., Ankara, 2006.

ALBAYRAK, Ali Sait; Koltan Yılmaz, Şebnem, **Veri madenciliği: Karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama**, Süleyman Demirel University The Journal of Faculty of Economics Administrative Sciences, 14, 1, 2009, p. 31-52.

AMASYALI, M. Fatih; Diri, Banu; Türkoğlu Filiz, **Farklı Özellik Vektörleri İle Türkçe Dökümanların Yazarlarının Belirlenmesi**, TAINN-2006.

ATASOY, Dinçer, **Lojistik Regresyon Analizinin İncelenmesi ve Bir Uygulaması**, Graduate thesis, Cumhuriyet University, Institute of Social Sciences, Sivas, 2001.

BALABAN, Mehmet Erdal.; Kartal, Elif, **Veri Madenciliği Ve Makine Öğrenmesi**, Çağlayan Kitabevi, 2015.

BALI, Selçuk; Cinel, Mehmet Ozan, **Altın fiyatlarının İMKB 100 endeksi'ne etkisi ve bu etkinin ölçümlenmesi**, Ataturk University Journal of Economics and Administrative Sciences, 25, 3-4, 2011, p. 45-63.

BİGPARA Web Page, ***http://www.bigpara.com***, Access Date: 02.12.2016.

BULUT, Faruk, **Dengesiz Veri Setlerinde Denetimli Öğrenicilerin Başarım Değerlendirmesi**, IEEE, 978-1-4673-8654-8/15/$31.00, 2016.

ÇOKLUK, Ömay, **Lojistik regresyon Analizi: Kavram ve Uygulama**, Educational Sciences: Theory & Practice, 10(3), 2010, p. 1357-1407.

EGE, İlhan; Bayrakdaroğlu, Ali, **İMKB şirketlerinin hisse senedi getiri başarılarının lojistik regresyon tekniği ile analizi**, Zonguldak Karaelmas University Journal of Social Sciences, 5, 10, 2009, p.139-158.

ERDAL, Hamit; **Makine öğrenmesi yöntemlerinin inşaat sektörüne katkısı: basınç dayanımı tahminlemesi**, Pamukkale University Journal of Engineering Sciences, 21(3), 2015, p. 109-114.

INVESTİNG Web Page, ***http://tr.investing.com***, Access Date: 13.12.2016.

KAVZAOĞLU, Taşkın; Çölkesen, İsmail, **Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi**, Harita Dergisi, 144, 2010, p. 73-82.

KOYUNCUGİL, Ali Serhan; Özgülbaş, Nermin, **İMKB'de işlem gören kobi'lerin güçlü ve zayıf yönleri: CHAID karar ağacı uygulaması**, Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 23, 1, 2008, p. 1-21.

McCALLUM, A.; Nigam, K., **A Comparison of Event Models For Naive Text Classification**, AAAI-98 Workshop On Learning For Text Categorization, 752, 1998, p. 41-48.

MERAL, Meriç; Diri, Banu, **Twitter Üzerinde Duygu Analizi**, IEEE 22nd Signal Processing and Communicatios Applications Conference (SIU 2014), 978-1-4799-4874-1/14/$31.00, 2014, p. 690-693.

MİTCHELL, T.,M., **Machine Learning (1st Edition)**, McGraw-Hill Science /Engineering / Math, 1997.

NİZAM, Hatice; Akın, Saliha Sıla, **Sosyal Medyada Makine Öğrenmesi İle Duygu Analizinde Dengeli Ve Dengesiz Veri Setlerinin Performanslarının Karşılaştırılması**, XIX. Türkiye'de İnternet Konferansı, 2014, inet-tr.org.tr.

ONAN, Aytuğ; Korukoğlu, Serdar, **Makine Öğrenmesi yöntemlerinin görüş madenciliğinde kullanılması üzerine bir literatür araştırması**, Pamukkale University Journal of Engineering Sciences, 22(2), 2016, p. 111-122.

ÖZDEMİR, A. Kerem; Tolun, Seda; Demirci, Ebru, **Endeks Getirisi Yönünün İkili Sınıflandırma Yöntemiyle Tahmin Edilmesi: İMKB-100 Endeksi Örneği**, Niğde University Journal of Economics and Administrative Sciences, 4, 2, 2011, p. 45-59.

ÖZEKES, Serhat, **Veri Madenciliği Modelleri Ve Uygulama Alanları**, İstanbul Ticaret Üniversitesi Dergisi, 2003, p. 65-82.

ÖZKAN, Yalçın, **Veri Madenciliği Yöntemleri**, Papatya Yayıncılık, 2008.

ROKACH, L.; Maimon, O., **Decision Trees, Data Mining and Knowledge Discovery Handbook**, Springer, 2005, p. 165-192.

SAVAŞ, İncilay; Can, İsmail, **Euro-Dolar Paritesi ve Reel Döviz Kuru'nun İMKB 100 Endeksi'ne Etkisi**, Eskişehir Osmangazi University Journal of Economics and Administrative Sciences, 6(1), April 2011, p. 323- 339.

SOLMAZ, Ramazan; Günay, Mücahid; Alkan, Ahmet, **Uzman Sistemlerin Tiroit Teşhisinde Kullanılması**, XV. Akademik Bilişim Konferansı Bildirileri, 23-25 Ocak 2013.

SÜMBÜLOĞLU, Kadir, **Lojistik Regresyon Analizi**,http://78.189.53.61/-/bs/ess/k_sumbuloğlu.pdf.,2015.

TAYYAR, Nezih; Tekin, Selin, **İMKB-100 endeksinin destek vektör makineleri ile günlük, haftalık ve aylık veriler kullanılarak tahmin edilmesi**, Abant İzzet Baysal University Journal of Social Sciences, 13, 1, 2013, p. 189-217.

TURKISH CENTRAL BANKWeb Page, ***http://www.tcmb.gov.tr***, Access Date: 29.11.2016.

VURAN, Bengü, **İMKB 100 endeksinin uluslararası hisse senedi endeksleri ile ilişkisinin eşbütünleşim analizi ile belirlenmesi**, Istanbul University Journal of the School of Business, 39, 1, 2010, p. 154-168.

ZHANG, H.,**The Optimality of Naive Bayes**, A A, 1(2), 3, 2004.