BİLGİSAYAR BİLİMLERİ
VE
TEKNOLOJİLERİ DERGİSİ
JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGIES
http://dergipark.org.tr/tr/pub/bibted

Araştırma Makalesi

Comparative Analysis of Machine Learning Algorithms in Stock Price Prediction

Hakan Murat Karaca[1], Umut Dökmen[1]*

[1]Manisa Celal Bayar Üniversitesi, Bilgisayar Mühendisliği, Manisa, Türkiye

Keywords:
Machine Learning
Algorithms
Stock
Regression
Supervised Learning

ABSTRACT

Stock is part of a company's principal. A person who buys stock of a company shares the profit or loss of this company. Large volume transactions are made on stock exchanges where stocks are traded. Stock prices are difficult to predict because they are affected by many variables, but when they can be predicted, great benefits are provided. Prediction of stock prices is possible with today's computers using machine learning algorithms. Machine learning provides more successful results than fundamental and technical analysis in stock price prediction. In our study, daily closing price predictions were made by collecting approximately 5-years data of the top 5 stocks with the highest market value traded in BIST 100 between 2016 and 2020. Multiple linear regression, Bayesian regression, random forest, decision trees, support vector machines, artificial neural networks algorithms were applied to include maximum 22 features and the results were compared. The most successful result was obtained in the artificial neural networks algorithm. To achieve the highest success, data pre-processing, normalization, cross-validation, parameter optimization and feature selection were applied. It has been observed that using these methods together increases the success.

Hisse Senedi Fiyat Tahmininde Makine Öğrenimi Algoritmalarının Karşılaştırmalı Analizi

Anahtar Kelimeler:
Makine Öğrenmesi
Algoritmalar
Hisse Senedi
Regresyon
Gözetimli Öğrenme

ÖZ

Hisse senedi bir şirketin anaparasının bir parçasıdır. Bir şirketin hisselerini satın alan kişi, bu şirketin kar veya zararına ortak olur. Hisse senetlerinin işlem gördüğü borsalarda büyük hacimli işlemler yapılmaktadır. Hisse senedi fiyatları birçok değişkenden etkilendiğinden tahmin edilmesi zordur ancak tahmin edilebildiğinde büyük faydalar sağlanır. Hisse senedi fiyatlarının tahmini, makine öğrenmesi algoritmalarını kullanan günümüz bilgisayarları ile mümkün olmaktadır. Makine öğrenimi, hisse senedi fiyat tahmininde temel ve teknik analize göre daha başarılı sonuçlar sağlamaktadır. Çalışmamızda 2016-2020 yılları arasında BIST 100'de işlem gören en yüksek piyasa değerine sahip 5 hisse senedinin yaklaşık 5 yıllık verileri toplanarak günlük kapanış fiyatı tahminleri yapılmıştır. Çoklu doğrusal regresyon, bayesian regresyon, rastgele orman, karar ağaçları, destek vektör makineleri, yapay sinir ağları maksimum 22 özelliği dahil edecek şekilde uygulanmış ve sonuçlar karşılaştırılmıştır. En başarılı sonuç yapay sinir ağları algoritmasında elde edilmiştir. En yüksek başarıyı elde etmek için veri ön işleme, normalleştirme, çapraz doğrulama, parametre optimizasyonu ve özellik seçimi uygulanmıştır. Bu yöntemlerin bir arada kullanılmasının başarıyı arttırdığı gözlemlenmiştir.

***Sorumlu Yazar**

(hakanmkaraca@gmail.com) ORCID ID 0000-0003-2144-2994
*(umut.dokmen@cbu.edu.tr) ORCID ID 0000-0001-6919-4278

e-ISSN: 2717-8579

Geliş Tarihi: 19/12/2023; Kabul Tarihi: 16/12/2024

1. INTRODUCTION

A stock is a part of the principal of a company. People who buy a company's stock share in the profit and loss of that company. The stock signifies the special relationship between the company and the person who buys the stock (Summers, 2007). Companies open their stocks to investors through the stock market in order to increase their financial capacity and their capital. The expectation that the value of the stock will increase creates demand for that stock. This demand increases the value of the stock. On the contrary, the expectation that the value of the stock will decrease requires selling the stock and the price will decrease. Investors aim to make a profit by buying stocks that will rise in the future. For this reason, it is very important for investors to be able to predict the stock price.

Machine learning is widely used in the field of finance, as it is in many fields. Many companies use machine learning in stock trading. It is able to make very wise investment decisions and reduce financial risks for people. Many studies have shown that machine learning-based applications are more successful than traditional stock trading strategies. These results increase the applications of artificial intelligence and machine learning in the field of finance day by day (URL-1, URL-2).

Stock prices are volatile. There are many internal and external factors that affect the stock price. Internal factors, profit distribution policy, capital increase, financial structure, management, field of activity of the enterprise (Hürer, 1995). In this study, Opening Price, High Price, Low Price, Volume, Net Profit for the Period, Resource and Dividend Income Factors, which are internal variables affecting the stock price, are included in the calculations. External factors included in the calculation in this study: BIST Stars Tradded Value, BIST Stars Tradded Volume, BIST 100 Index, BIST 100 Volume (TL), BIST 100 Difference, Dollar-TL, Euro-TL, XAU-USD, Brent Oil, S&P 500 Index, Euro Stoxx 50 Index, Interest and Inflation. These factors affect the stock price differently. The effect of these factors on the stock price can be calculated by statistical methods. But there are other factors that affect the stock price. These are the political situation in the country and the world, financial expectations, sectoral expectations, unexpected events. The political situation in the country and the world or natural disasters cannot be predicted by numerical methods. However, since the policies of the country and the world will affect the external factors used in this study, the indirect effect on the stock can be calculated.

In the study of Ghana, Awaisa and Muzammula, they tried to predict the stock prices of Apple, Amazon and Google with time series forecasting algorithms and observed that the exponential smoothing results gave greater accuracy (Ghani, 2019; Muzammul 2019). Sarode, Tolani, Kak and Lifna used real-time data along with news analysis. With LSTM (Long Short-Term Memory), an artificial neural network architecture, they presented a system that decides whether to buy the stocks of different companies (Sarode, 2019; Tolani, 2019; Kak 2019; Lifna 2019). In their study, Usmani, Adil, Raza and Ali tried to predict the end-of-day closing performance of the Karachi Stock Exchange (KSE) using machine learning. In the study, it was found that the multi-layer perceptron showed the best performance and the feature that affected the index the most was the oil price (Usmani, 2019; Adil, 2019; Raza, 2019; Ali, 2019). Tipirisetty made stock price prediction using deep learning. In addition to quantitative analysis, his study also analyzed textual public news from online news sources and concluded that "accuracy increases when textual information is used in stock price prediction" (Tipirisetty, 2018). For stock price prediction, Sinngh collected 10 years of data from yahoo finance and used LSTM and linear regression for prediction. RMSE is used as the evaluation metric. RMSE was found 2.04 for linear regression and 0.43 for LSTM (Singh, 2021). In his study, Guo tried to predict the S&P 500 index. LSTM, arima and garch models are used and found that the model in which these 3 models were used together gave more successful results than the model in which LSTM was used alone (Guo, 2022).

In this study, the daily closing price of the stocks of the top 5 companies as seen table1 with the highest market value in BIST(Borsa Istanbul) 100 is predicted. For the training of machine learning models, approximately 5 years of historical data between 2016 and 2020 were collected and combined from borsaistanbul.com, investing.com and isyatırım.com. In the study, 80% of the data set was used for training and 20% for testing. In order for machine learning algorithms to give more successful results, data preprocessing has been implemented. The effect of the normalization methods used in the data set on the model success was investigated. The parameter changes in the models were examined and their effects on the results were investigated. Optimum parameters are selected to get the highest success. By using feature selection methods, the success of the model has been increased. The success of machine learning algorithms used in the study has been compared.

One of the aims of the study is to predict the stock price and give direction to the investors. Another purpose is to analyze the factors affecting stock prices with machine learning methods and to give an idea to financial analysts. Another purpose is to optimize prediction success by using internal and external factors that affect the price of stocks together as features.

In similar studies in literature, values close to the closing price such as opening, high, low or global features such as gold, dollar, euro, interest, inflation etc. were included in the model. Differently in this study, the performance of the machine learning algorithms aimed to increase by including the period net profit, dividend income, resource features obtained from the company's internal balance sheet.

2. MACHINE LEARNING

Machine learning is an artificial intelligence field that aims to give the machine the ability to learn without programming it directly. There are broadly three types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. In stock price prediction, the supervised learning technique, which covers all prediction problems, is used because the future price is predicted from the past, known data set. Since the prices which is the output, we get in the stock price prediction is numerical, the task is called prediction. To predict stock prices, the computer learns patterns from past stock prices. The difference between the predicted price and the actual price is called the loss function. The machine improves its performance a little more with each experience. In practice, experience means training data. Therefore, we cannot easily distinguish between machine learning and statistical approaches. The goal of supervised learning is minimizing the loss function. In the stock price prediction machine tries to minimize the difference between the actual stock price with predicted stock price. In supervised learning the machine learns a predictive model that maps the features of the data to an output. Machine aims to learn a model predicting parameters (Molnar, 2019; Goodman, 2019; Kaminsky, 2019; Lessler, 2019).

2.1 Machine Learning Algorithms Using in the Project

Multiple linear regression is a linear regression with multiple independent variables. The equation form is also similar to simple linear regression. Both types of regression are ultimately linear.

$$y_i = \beta_i + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + e_i \qquad (1)$$

The dependent variable y in this study is the “stock price” we aim to find. The independent variables, represented by x in formula(1) are the features in the model for “qnbfb model” such as Opening, High, Low, Difference, Star Market Transaction Volume, Star Market Transaction Amount, Bist 100 Index, Bist100 Volume, Bist100 Difference %, Dollar-TL, XAU-USD, Euro- TL, SP 500, Brent Oil, Euro Stoxx 50, Interest %, Inflation %, Net Profit for period, Resources, Dividend income.

Support vector regression (SVR) is a supervised machine learning method used in prediction problems. Regression analysis is performed to analyze the relationship between a dependent variable and one or more independent variables. SVR formulates an optimization problem by learning a regression function to map input prediction variables to observed output values. SVR is another version of support vector machines which is classification algorithm. However SVM produces a class label i.e. a binary output. SVR is the solution to the regression problem consisting of a real-valued function prediction. The aim in SVR is to find the optimal width hyperplane containing the most appropriate line, that is, the maximum data point as seen figure 1. SVR does not try to minimize the difference between the actual value and the predicted value as in other regression models. It tries to best fit the data within a certain threshold value. The distance between the boundary line and the hyperplane is called the threshold value (Zhang and Donnel 2020).

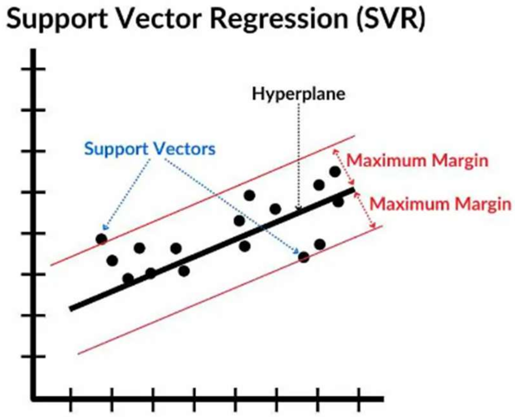


Figure 1. Support Vector Regression

Linear regression and logistic regression models cannot be successful when the relationship between the features and the dependent variable is not linear or when the features interact with each other. In such cases, tree-based models can be used. Tree models split data multiple times according to certain cutoff values in the features. Many subsets are created by splits from all data. The final, terminal subsets form leaves, and the other inner subsets form nodes. The average of the training data from this subdivided subset is used to estimate the outcome at each leaf node (Skinea. 2017).

The random forest algorithm is based on drawing more than one decision tree for the same dataset and using these decision trees together. Random forest algorithm can be used for classification and regression. While getting the regression result, the average of more than one separated decision tree is taken (Liu., Wang, Zhang 2012).

Bayesian regression is a type of linear regression based on Bayes’ theory. In the Bayesian approach, the uncertainty in the w vector is characterized by a probability distribution p(w). Bayes’ theorem applies this distribution through observations of data points and the probability function of the data.

$$P\ (\beta|y,X) = \frac{P(y|\beta,X)*P(\beta|X)}{P(y|X)} \quad (2)$$

P (β|y, X) is the posterior probability distribution of the parameters when the inputs and outputs are known in formula 2. This is found by multiplying the data probability P(y|β, X) by the prior probability of the parameters and dividing by a normalization constant (URL-10).

A neural network is an oriented structure that connects a simple input layer called a neuron to the output layer with weighted connections to larger structures. A neuron is connected with n input channels, each expressed in synaptic weight w_i. Each input from the neuron is multiplied by its weights and they are summed. An optional bias might be added to this sum. The summed result is then put into an activation(threshold) function. This function can be sigmoid, hyperbolic tangent, ReLU or any other function. The input produces an output after filtering it with the activation function (Bonaccorso, 2017).

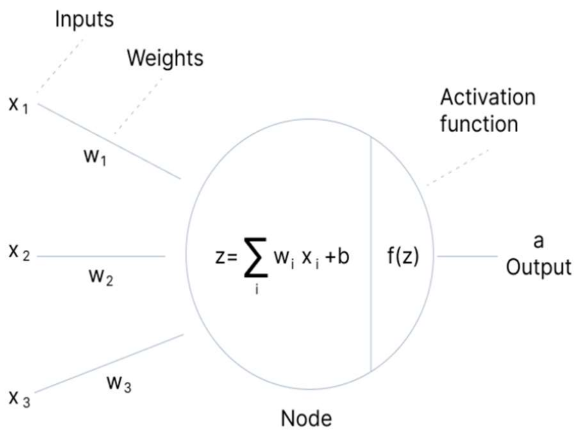


Figure 2 The structure of a simple neural network.

2.2. Data Set

Data has been collected from Borsa Istanbul formal website and investing.com. The daily index, opening, high, low, volume, difference, exchange rates, XAU-USD (golden ounce price), brent oil, S&P 500, data of each stock are taken from the 'investing.com' website. Market transaction volume, Bist 100 transaction amount, Bist100 index, Bist 100 volume, Bist 100 difference data are obtained from borsaistanbul.com. Net profit for the period, resource, dividend income are collected from www.isyatirim.com.tr [URL-3]. Data from different sources are combined in a single table. Since financial data is not open on holidays, only working days are added to the data set. While adding internationally valid features such as brent oil, S&P 500, lines corresponding to holidays in Turkey were not included in the data set.

Table 1. Names of stocks examined

Stock Code	Company Name
QNBFB	Qnb Finansbank
ENKAI	Enka İnşaat Ve Sanayi A.Ş
FROTO	Ford Otomotiv Sanayi A.Ş
EREGL	Ereğli Demir Ve Çelik Fabrikalari T.A.Ş
KCHOL	Koç Holding A.Ş

Table 2. Explanation of features

Feature	Explanation
Date	Indicates the date on which the relevant feature was obtained.
Close price	Indicates the closing value of the stock on the specified date(Aslan, 2020).
Opening price	Indicates the opening price of the relevant stock (Aslan 2020).
High	Refers to the highest value of the related feature (column) during the day (Aslan, 2020).
Low	Refers to the lowest value of the related attribute (column) during the day(Aslan, 2020).
Volume	Indicates the trading volume of the relevant stock during the day.
Difference %	Indicates the change in the day-to-day price of the relevant stock as a percentage (Aslan, 2020).
BIST Stars Traded Value (TL)	Number of transactions in the star market.
BIST Stars Traded Volume	It is the trading volume of the market in which the shares with a market value of 300 million TL and above of the portion offered to the public in the first listing to the stock exchange are traded (URL-4).
BIST 100 index	It consists of the 100 stocks traded in Borsa Istanbul with the highest market value and trading volume and is the main index of the Equity Market(URL-5).

BIST 100 volume (TL)	It is the total value of daily trading transactions in the BIST 100 (URL-6).
BIST 100 difference	It is the change of the BIST100 Index Value information announced by Borsa Istanbul at the end of the trading day according to the value of the next day (Karagöz, 2020).
Dollar-TL	TL equivalent of the dollar on the relevant date.
Euro-TL	TL equivalent of the euro on the relevant date.
XAU-USD	The dollar price of an ounce of gold on the relevant date.
Brent Oil	Dollar price of brent oil on the relevant date.
S&P 500	Stock market index of 500 major US stocks by Standard and Poor's(URL-7).
Euro Stoxx 50	Stock index of 50 stocks from 11 Eurozone countries designed by Stoxx (URL-8).
Interest	The policy interest rate used by the Turkey Central Bank is the interest rate applied in one-week repo transactions. Decisions on policy rates are taken by the Monetary Policy Committee (MPC) (URL-9).
Net profit for the period	It indicates the net profit of the company in that period.

2.3 Data Preprocessing

There are many factors that affect the success of machine learning algorithms. The most important of these is the representation and quality of the data set. Data must be preprocessed to improve quality. Machine learning suffers when there is too much irrelevant and redundant data. In machine learning studies, a significant amount of time is spent in data preprocessing. Data preprocessing is unavoidable as it is impossible to have a preprocessing algorithm that works on all datasets, providing reliable and effective performance. In data preprocessing, operations such as data cleaning, normalization, conversion, feature selection are performed (Kotsiantis, Kanellopoulos, Pintelas 2006; Alexandropoulos, Kotsiantis, Vrahatis 2016)

The normalization technique is used to transform the features to the same scale. In this way, it reduces the difference between the predicted value and the real value. Different feature normalization methods can be used when the actual distribution of features is not known beforehand.

Min max normalization is one of the most used methods to standardize data. For each component, it converts the element's base estimate to zero, the extreme value to 1, and the other values to a decimal between 0 and 1.

$$\hat{X}_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \qquad (3)$$

Where x_{min}: minimum value in X feature x_{max}: maximum value in X feature $\hat{X}_i$: scaled X

(Raju, Lakshmi, Jain, Kalidindi, Padma 2020)

The standard scaler(ss) is a method in which the distribution approaches normal by averaging each feature and scaling its variance to 1. In the formula, the mean is subtracted from the true value and divided by the variance.

$$\hat{X}_i = \frac{x_i - \bar{X}}{\sigma} \qquad (4)$$

Where $\hat{X}_i$: normalization version of x

σ: standard derivation

$\bar{X}$: mean

x_i: each observation from a sample

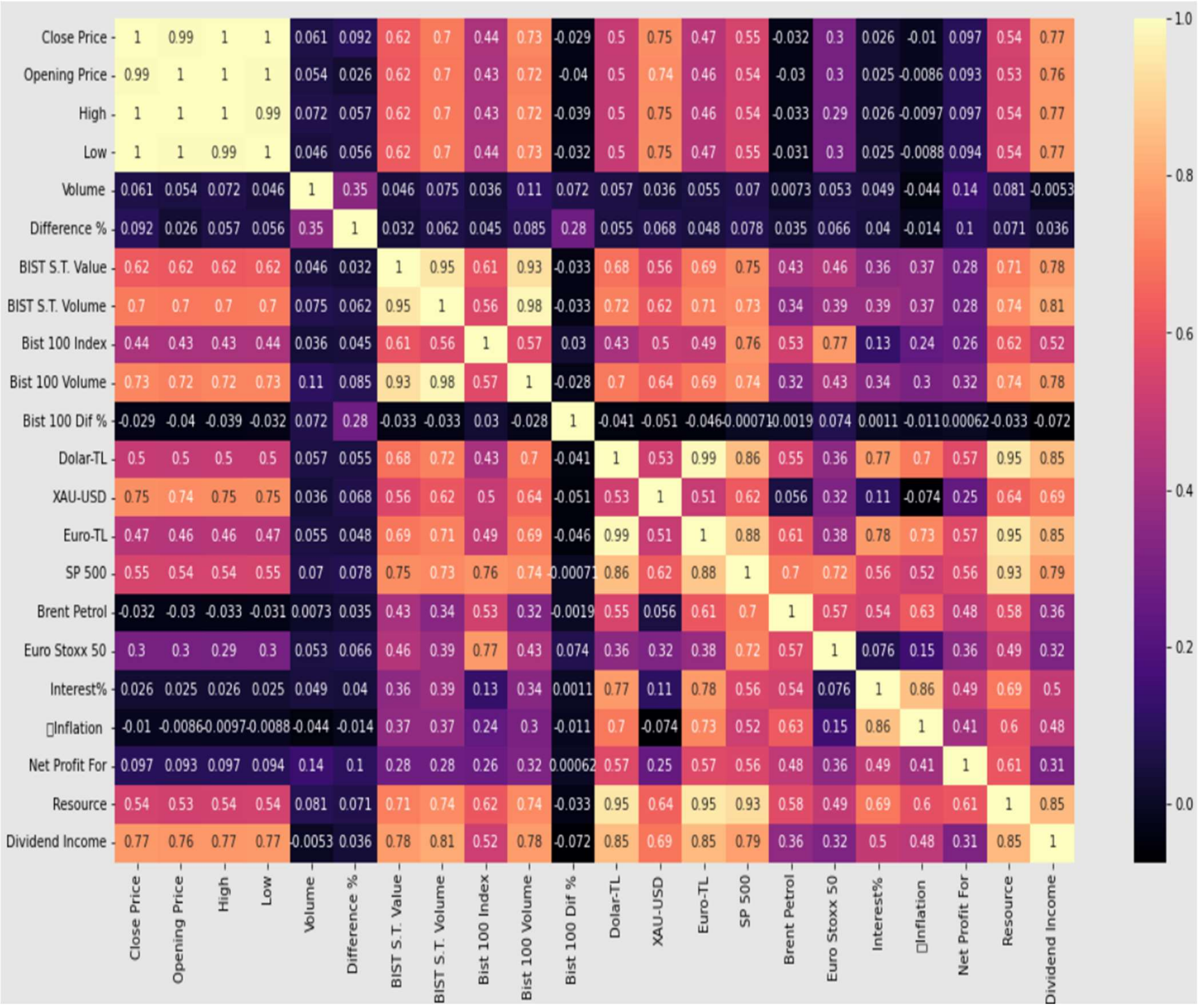
(Ferreira, Le 2019)

In order to the data to be of good quality, empty lines were removed during the data preprocessing stage, all the features were collected numerically, and dummy variables were not included in the model.

2.4 Analysis of Features

In this study, the target is the stock price as the dependent variable, since the desired value is the stock price. The remaining features are independent variables. It is important to analyze which features affect the stock price and how much.

Table 3. Heatmap of the data set



	Close Price	Opening Price	High	Low	Volume	Difference %	BIST S.T. Value	BIST S.T. Volume	Bist 100 Index	Bist 100 Volume	Bist 100 Dif %	Dolar-TL	XAU-USD	Euro-TL	SP 500	Brent Petrol	Euro Stoxx 50	Interest%	□Inflation	Net Profit For	Resource	Dividend Income
Close Price	1	0.99	1	1	0.061	0.092	0.62	0.7	0.44	0.73	-0.029	0.5	0.75	0.47	0.55	-0.032	0.3	0.026	-0.01	0.097	0.54	0.77
Opening Price	0.99	1	1	1	0.054	0.026	0.62	0.7	0.43	0.72	-0.04	0.5	0.74	0.46	0.54	-0.03	0.3	0.025	-0.0086	0.093	0.53	0.76
High	1	1	1	0.99	0.072	0.057	0.62	0.7	0.43	0.72	-0.039	0.5	0.75	0.46	0.54	-0.033	0.29	0.026	-0.0097	0.097	0.54	0.77
Low	1	1	0.99	1	0.046	0.056	0.62	0.7	0.44	0.73	-0.032	0.5	0.75	0.47	0.55	-0.031	0.3	0.025	-0.0088	0.094	0.54	0.77
Volume	0.061	0.054	0.072	0.046	1	0.35	0.046	0.075	0.036	0.11	0.072	0.057	0.036	0.055	0.07	0.0073	0.053	0.049	-0.044	0.14	0.081	-0.0053
Difference %	0.092	0.026	0.057	0.056	0.35	1	0.032	0.062	0.045	0.085	0.28	0.055	0.068	0.048	0.078	0.035	0.066	0.04	-0.014	0.1	0.071	0.036
BIST S.T Value	0.62	0.62	0.62	0.62	0.046	0.032	1	0.95	0.61	0.93	-0.033	0.68	0.56	0.69	0.75	0.43	0.46	0.36	0.37	0.28	0.71	0.78
BIST S.T. Volume	0.7	0.7	0.7	0.7	0.075	0.062	0.95	1	0.56	0.98	-0.033	0.72	0.62	0.71	0.73	0.34	0.39	0.39	0.37	0.28	0.74	0.81
Bist 100 Index	0.44	0.43	0.43	0.44	0.036	0.045	0.61	0.56	1	0.57	0.03	0.43	0.5	0.49	0.76	0.53	0.77	0.13	0.24	0.26	0.62	0.52
Bist 100 Volume	0.73	0.72	0.72	0.73	0.11	0.085	0.93	0.98	0.57	1	-0.028	0.7	0.64	0.69	0.74	0.32	0.43	0.34	0.3	0.32	0.74	0.78
Bist 100 Dif %	-0.029	-0.04	-0.039	-0.032	0.072	0.28	-0.033	-0.033	0.03	-0.028	1	-0.041	-0.051	-0.046	-0.00071	-0.0019	0.074	0.0011	-0.011	0.00062	-0.033	-0.072
Dolar-TL	0.5	0.5	0.5	0.5	0.057	0.055	0.68	0.72	0.43	0.7	-0.041	1	0.53	0.99	0.86	0.55	0.36	0.77	0.7	0.57	0.95	0.85
XAU-USD	0.75	0.74	0.75	0.75	0.036	0.068	0.56	0.62	0.5	0.64	-0.051	0.53	1	0.51	0.62	0.056	0.32	0.11	-0.074	0.25	0.64	0.69
Euro-TL	0.47	0.46	0.46	0.47	0.055	0.048	0.69	0.71	0.49	0.69	-0.046	0.99	0.51	1	0.88	0.61	0.38	0.78	0.73	0.57	0.95	0.85
SP 500	0.55	0.54	0.54	0.55	0.07	0.078	0.75	0.73	0.76	0.74	-0.00071	0.86	0.62	0.88	1	0.7	0.72	0.56	0.52	0.56	0.93	0.79
Brent Petrol	-0.032	-0.03	-0.033	-0.031	0.0073	0.035	0.43	0.34	0.53	0.32	-0.0019	0.55	0.056	0.61	0.7	1	0.57	0.54	0.63	0.48	0.58	0.36
Euro Stoxx 50	0.3	0.3	0.29	0.3	0.053	0.066	0.46	0.39	0.77	0.43	0.074	0.36	0.32	0.38	0.72	0.57	1	0.076	0.15	0.36	0.49	0.32
Interest%	0.026	0.025	0.026	0.025	0.049	0.04	0.36	0.39	0.13	0.34	0.0011	0.77	0.11	0.78	0.56	0.54	0.076	1	0.86	0.49	0.69	0.5
□Inflation	-0.01	-0.0086	-0.0097	-0.0088	-0.044	-0.014	0.37	0.37	0.24	0.3	-0.011	0.7	-0.074	0.73	0.52	0.63	0.15	0.86	1	0.41	0.6	0.48
Net Profit For	0.097	0.093	0.097	0.094	0.14	0.1	0.28	0.28	0.26	0.32	0.00062	0.57	0.25	0.57	0.56	0.48	0.36	0.49	0.41	1	0.61	0.31
Resource	0.54	0.53	0.54	0.54	0.081	0.071	0.71	0.74	0.62	0.74	-0.033	0.95	0.64	0.95	0.93	0.58	0.49	0.69	0.6	0.61	1	0.85
Dividend Income	0.77	0.76	0.77	0.77	-0.0053	0.036	0.78	0.81	0.52	0.78	-0.072	0.85	0.69	0.85	0.79	0.36	0.32	0.5	0.48	0.31	0.85	1

According to the heat map in table 3, the factors that affect the stock price the most are the features closest to 1 in the heat map. According to the analyzed data set, the current opening high and low values affect the stock price the most. The reason for this is that the stock price we are looking for is very close to the opening, high and low values of the same day. This was actually something we could see before the heatmap. Here there are other values close to 1. For example, with a value of 0.77 in dividend income, it is seen that it significantly affects the stock price. It is understood that the gold ounce price, which comes after that, with a value of 0.75, is also a feature that affects the stock price. Bist 100 volume and star market trading volume are among the features that significantly affect the dependent variable. These ratios can be used when selecting features to increase success.

2.5 Predicting Stock Price Using Machine Learning Algorithms

The stages applied in machine learning are:

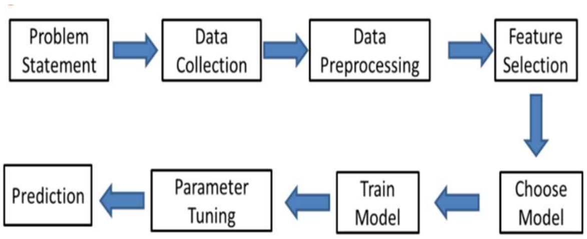


Figure 3. Machine learning steps

2.5.1 Creating machine learning model

In the machine learning model, the data set is divided into 80% training and 20% testing. The aim was to predict the closing price of the stock for current day. Therefore, the y dependent variable is the stock price, represented by x, the independent variables being the remaining features.

In the first stage, all the features were included in the system and the model was created. No feature selection and normalization has been done. The aim here is to determine which methods and which algorithms make the best predictions. Success found using no methods is compared to the success of predictions found using certain methods.

2.5.2 Cross validation

Cross validation is a resampling method to avoid memorization and generalize the model. In cross validation, the data set is divided into subsamples. Separate training and test samples are created for each sub-sample. The training and testing part of each sub-samples are different samples. The model learns from different parts of the data in each sub-sample. The model's estimate error is calculated for all sub-samples and their average is the model's error (King, 2021; Orhobor, 2021; Taylor, 2021). In this study, k-fold cross validation technique was used.

2.5.3. Feature selection

Through feature selection methods, the computation time of machine learning algorithms can be reduced, prediction success can be increased, and data can be better understood. There are many methods for feature selection in literature. These methods can be roughly classified as filter methods, wrapper methods, embedded and hybrid methods.

The purpose of feature selection is exclude unnecessary features that negatively affect the model that cause a decrease in success. Which feature combination will give the most successful result can be found by brute force method by trying one by one. However, this job is feasible only in models with very few features. It will be very expensive to calculate this in models with many features (Chandrashekar, 2014; Sahin, 2014; Jovic, 2015; Bogunovic, 2015).

3. RESULTS OF MACHINE LEARNING ALGORITHMS

This study and other studies in the literature give the result that there is a relationship between the stock market and macroeconomic variables. There are macroeconomic variables (such as interest rate, inflation, exchange rate, oil prices, gold prices) as well as intra-firm factors (such as firm performance, dividends, incomes, changes in the board of directors) that affect the stock price traded in the stock markets. In this study, internal factors and macroeconomic variables affecting stock prices were taken as features and it was investigated how much these features affect stock prices. Accordingly, machine learning models were created, and feature selections were made to increase success. Algorithm performances and used methods were compared.
The success of the test results in the studies created with the QNBFB data set was very high. The reason for this is that the target variable to be found is very close to 3 features. The QNB index values are very close to the “Opening, High, and Low” features. According to table 3, these 3 independent variables in the QNB index have a high correlation. Knowing the "Opening, High and Low" features of the model while training has greatly increased the success in the prediction. However, this situation is not effective and useful when predicting stock price in practice. Because in practice, the target is to predict the end-of-day closing index. The end-of-day features "low, high, open" may not be known at the beginning. Therefore, the model was created without using these 3 features for training while predicting the "Enkai" stock price in order to be more realistic in its application to daily life.

3.1. Effect of Machine Learning Algorithm to Result

In the study, when comparing algorithms, min-max normalization, which was the best normalization method before, was used, except for support vector regression and artificial neural network. In SVR and ANN, on the other hand, standard scaler(ss) normalization was used because it gave better results. Models were created with the parameters that gave the best results in parameter comparison before. No parameter optimization was done for MLR. For SVR, the highest success was achieved in the model where the kernel was determined linearly. The best results were obtained with default parameters for DTR and BRR. The model was created with various estimator numbers for RFR. The best result was obtained with the RFR created with 300 estimators. The effect of epoch number and learning rate on success for ANN was investigated and the optimum result was obtained with the parameters where learning rate was 0.0001 and epoch number was 1000.

Table 4 shows the parameters and normalization methods for which the algorithms were most successful, and the model was created with the same number of features and the performances of the algorithms were compared.

The success of the machine learning algorithms used in this study was mostly high. Reason for this may be enough data used in the models, using of optimum normalization, the appropriate parameter selection, the quality of the data set, and the appropriate selection of the features. In feature selection, the heatmap is basically used. Features that do not affect the closing price of the stock were removed from the model and more successful results were obtained with the remaining 18 features.

Table 4. Comparing the results of machine learning models.

Test Name	Algorithm	R Square	MAE	MSE	Feature Number	Methods	Parameters
Eregli_test1	MLR	0.99320	0.01010	0.00035	18	Min-max, cv	default
Eregli_test2	SVR	0.98759	0.07178	0.00865	18	ss, cv	kernel=linear
Eregli_test3	DTR	0.99538	0.00679	0.00022	18	Min-max, cv	default
Eregli_test4	RFR	0.99537	0.00762	0.00023	18	Min-Max, cv	n_estimators=300
Eregli_test5	BRR	0.99303	0.00993	0.00035	18	Min-Max, cv	default
Eregli_test6	ANN	0.99875	0.00477	0.00006	18	Ss, cv	epoch=1000, lr=0.0001

Figure 4. R square results for test data in machine learning model

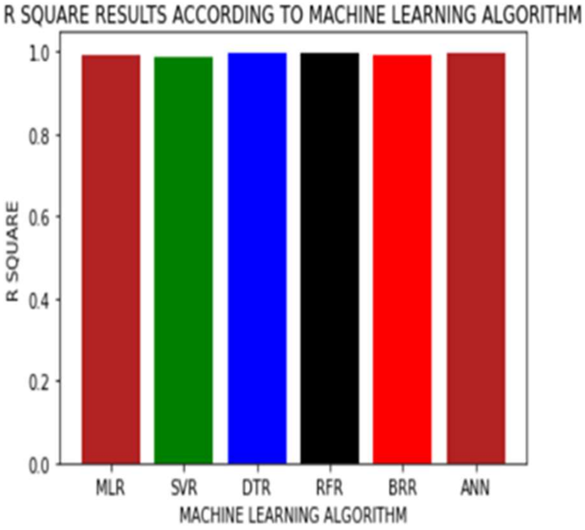


Figure 5. R square results for 104 unseen data in machine learning model

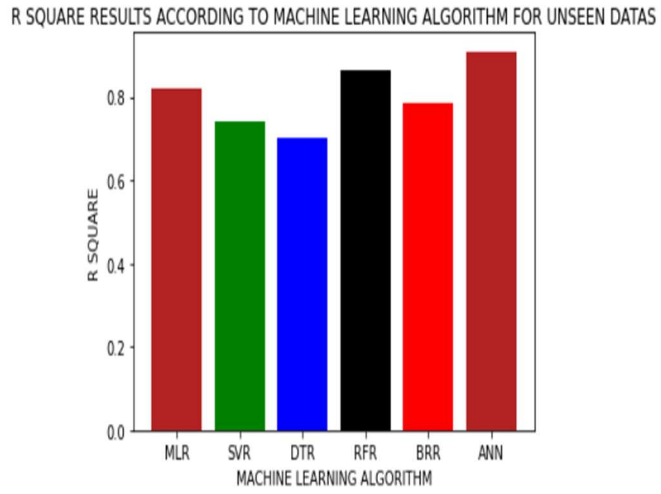
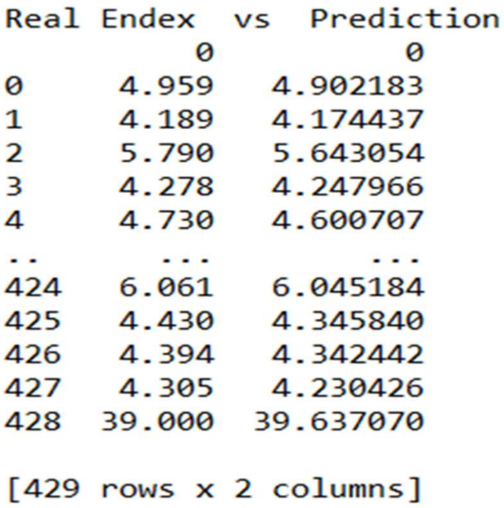


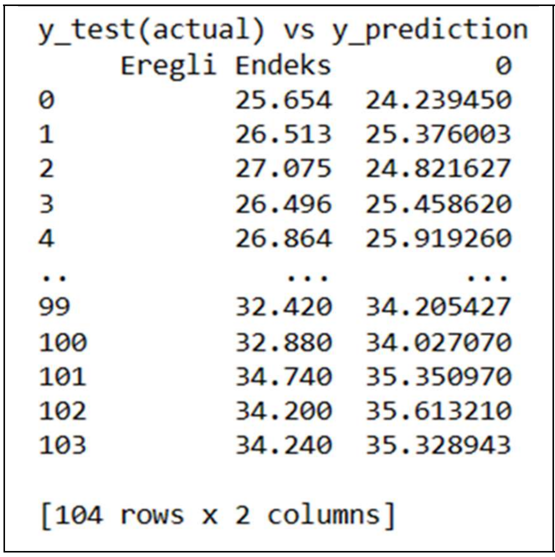
Table 5. Comparison of the actual values with the predicted values for QNBFB stock



```
Real  Endex   vs   Prediction
          0              0
0      4.959    4.902183
1      4.189    4.174437
2      5.790    5.643054
3      4.278    4.247966
4      4.730    4.600707
..       ...         ...
424    6.061    6.045184
425    4.430    4.345840
426    4.394    4.342442
427    4.305    4.230426
428   39.000   39.637070

[429 rows x 2 columns]
```

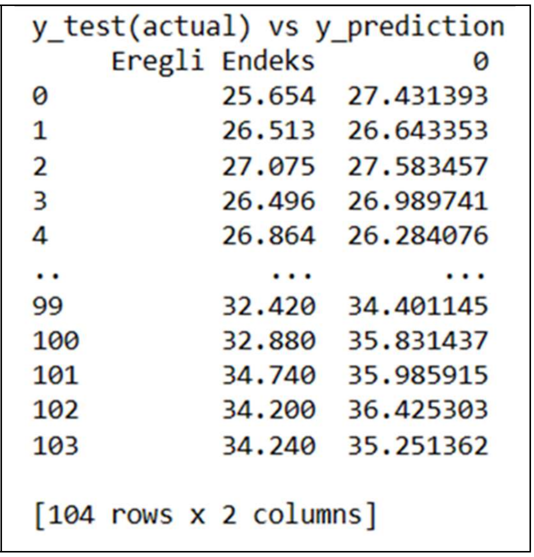
Table 6. Random Forest Regression Actual vs Predicted Values for Unseen Data



```
y_test(actual) vs y_prediction
     Eregli Endeks          0
0          25.654   24.239450
1          26.513   25.376003
2          27.075   24.821627
3          26.496   25.458620
4          26.864   25.919260
..            ...         ...
99         32.420   34.205427
100        32.880   34.027070
101        34.740   35.350970
102        34.200   35.613210
103        34.240   35.328943

[104 rows x 2 columns]
```

Table 7. Neural Network (MLP) Actual vs Predicted Values for Unseen Data



```
y_test(actual) vs y_prediction
     Eregli Endeks          0
0          25.654   27.431393
1          26.513   26.643353
2          27.075   27.583457
3          26.496   26.989741
4          26.864   26.284076
..            ...         ...
99         32.420   34.401145
100        32.880   35.831437
101        34.740   35.985915
102        34.200   36.425303
103        34.240   35.251362

[104 rows x 2 columns]
```

3.2 Effect of Feature Selection Methods to Result

In Table 8, the performances of the algorithms are compared before and after the feature selection method is applied. As it can be understood from Table 8, feature selection methods increased the success in all algorithms. The most significant increase in success has been in SVR, DTR and ANN algorithms. For all algorithms, backward elimination (bwe) and forward feature selection(ffs) methods have been tested. The success for MLR, SVR and DTR algorithms has increased with the forward feature selection method. The success for the RFR, BRR and ANN algorithms has increased with the backward elimination method.

Table 8. Comparing machine learning methods with and without feature selection method for unseen data.

Test Name	Algorithm	R Square	MAE	MSE	Feature Number	Methods	Parameters
Eregli_test7	MLR	0.8208	1.4129	2.7662	18	min-max, cv	default
Eregli_test8	MLR	0.8336	1.3587	2.5685	17	ffs,min-max, cv	default
Eregli_test9	SVR	0.7402	1.6551	4.0114	18	ss, cv	kernel=linear
Eregli_test10	SVR	0.8492	1.2793	2.3286	9	ffs, ss, cv	kernel=linear
Eregli_test11	DTR	0.7025	1.7568	4.5945	18	min-max, cv	default
Eregli_test12	DTR	0.7659	1.4455	3.6146	12	Ffs, min-max, cv	default
Eregli_test13	RFR	0.8629	1.2241	2.1161	18	min-max, cv	n_estimators=300
Eregli_test14	RFR	0.8656	1.2255	2.0748	17	bwe, min-max, cv	n_estimators=300
Eregli_test15	BRR	0.7867	1.5828	3.2934	18	min-max, cv	default
Eregli_test16	BRR	0.7868	1.5468	3.2913	16	bwe, min-max, cv	dafault
Eregli_test17	ANN(MLP)	0.9089	0.9544	1.4056	18	ss, cv	act=logistic, hidden layer=4 max_iter=1

Eregli_test18	ANN(MLP)	0.9424	0.7649	0.8885	7	bwe, ss, cv	act=logistic, hidden layer=4 max_iter=1000

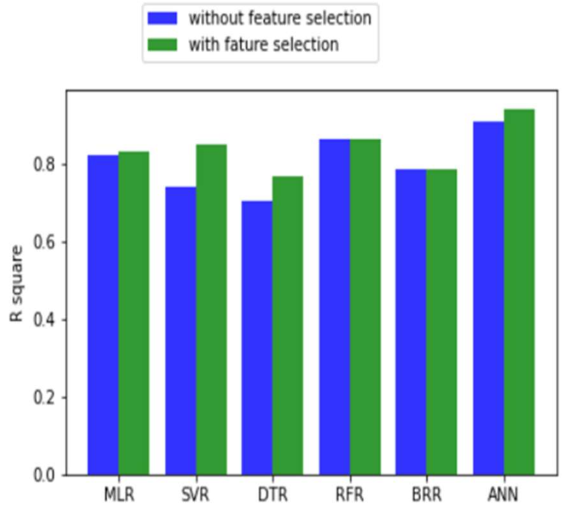


Figure 6. Effect of feature selection method in algorithm performances

4. CONCLUSION AND SUGGESTIONS

The success of the machine learning algorithms applied in this study was generally high. In order to the success to be high, 1070 data are used in machine learning models.

The effect of machine learning algorithm on success was investigated and the most successful results were obtained when ANN was applied according to table 4, figure 4 and 5. According to Table 5, the most successful algorithm, ANN, predicted the QNBFB stock dated 6 January 2022 as 39.63 with a real index value of 39.00. Since more successful results were obtained for unseen data, MLP, which is the ANN type, was applied. When the success was measured for the observed data, results were close to each other for all algorithms. While the SVR was given according to figure 4 for the data with the most unsuccessful result among the 6 algorithms applied, it gave DTR according to figure 5 for the data not seen.

The random forest algorithm has achieved better performance than the decision tree algorithm because the random forest consists of many decision trees trained from different subsets of the data. Taking averages over many decision trees reduces variance and over-fitting. Artificial neural networks can learn the non-linear relationship between dependent and independent variables. It does this by using a large number of neuron layers with non-linear activation functions. ANN can capture patterns that random forest algorithms and decision trees cannot capture (Robbach, 2018). Therefore, ANN have shown better performance than linear models and tree models.

In order to increase the success of machine learning algorithms in the study, forward selection and backward elimination methods, which are feature selection methods, were applied. According to Table 8, the feature selection methods increased the success of R square and decreased the error measures in the results obtained from testing the final stock closing price of 2021 for unseen data that is not in the training data set. For unseen data, the R square success of MLR algorithm increased from 82.0% to 82.3%, MAE and MSE decreased with forward feature selection method. Similarly, applying the forward feature selection method in the SVR algorithm increased the success of R square from 74.0% to 84.9% and greatly reduced the error measures. The forward feature selection method in DTR algorithm increased the success of R square from 70.2% to 76.5% and decreased the error measures. Since the forward feature selection method did not increase the success in the RFR algorithm, the backward elimination method was applied. Although the backward elimination method did not significantly increase the performance of the RFR algorithm, it increased the success of the R square from 86.2% to 86.5% and slightly decreased the error rates. The forward feature selection method for the BRR algorithm did not increase the success, but the success increased slightly with the backward elimination method. In the MLP algorithm, which is the ANN type, the backward elimination method increased the R square success of the model from 90.8% to 94.2% and was successful by reducing the error rates.

In the study, cross validation technique was applied to prevent overfitting. In order to show that machine learning models learn without overfitting, predictions are made for the 104-days index of 2021, which is not in the training dataset of the models. As seen in Table 6 and Table 7 prediction results close to the actual value were found in the prediction of these 104 unseen data. For this reason, it has been shown that the machine learning models applied in the study can be generalized.

In this study, machine learning models were created by taking 3 internal and 18 external features in the models with the highest number of features. The number of internal features can be increased in future work. The number of features can be increased by adding the features used in this study by performing sentiment analysis with daily data compiled from container data or other financial news sites. In this study, 5 years of data were collected. The success of machine learning algorithms can be increased by adding more rows to the training data set.

REFERENCES

Summers, D. (2007) Longman Business English Dictionary, Pearson Longman, London, 594 p.

URL-1: https://dataconomy.com/2023/01/11/stock-prediction-machine-learning.
[Access date: 20.04.2023]

URL-2: https://builtin.com/machine-learning/machine-learning-stock-prediction.
[Access date: 20.04.2023]

Hürer, E. (1995) Hisse Senedi Fiyatını Etkileyen Faktörler ve İMKB Üzerine Bir Uygulama, İstanbul University, İstanbul, 208 s.(Master Thesis)

Ghani, M., Awais, M., Muzammul (2019), Stock Market Prediction Using Machine Learning(ML) Algorithms, *Advances in Distributed Computing and Artificial Intelligence Journal*, 4, pp. 97-116.

Sarode, S., Tolani, H., Kak, P., Lifna, C. (2019) Stock Price Prediction Using Machine Learning Techniques, *International Conference on Intelligent Sustainable Systems* (ICISS), Palladam, India.

Usmani, M., Adil, S., Raza, K., Ali, S. (2016) Stock Price Prediction Using Machine Learning Techniques. *3rd International Conference On Computer And Information Sciences* (ICCOINS), Kuala Lumpur, Malaysia.

Tipirisetty, A. (2018) Stock Price Prediction using Deep Learning. San Jose State University, Department of Computer Science, California, 54s. (Master Thesis)

Singh, S. Stock Prediction using Machine Learning, California State University, Computer Science, California, 2021, 16s. (Master Thesis).

Guo, Y. Stock Price Prediction Using Machine Learning, Sodertorn University, School of Social ScienceMaster, Economics, Stockolm, 2022, 34. (Master Thesis).

Molnar C. (2019) Interpretable Machine Learning, Lulu.com, 314 p.

Bi, Q., Goodman, K. E., Kaminsky, J., Lessler, J.(2019) What is machine learning? *A primer for the epidemiologist, American journal of epidemiology*, 188(12), 2222-2239.

URL-3 //www.isyatirim.com.tr/tr-tr/analiz/hisse/Sayfalar/Temel-Degerler-Ve-Oranlar.
[Access date: 12.12.2022]

Aslan, B. (2020), Derin Öğrenme ile Borsa Verileri Üzerinde Tahminleme Yapılması, Ege Üniversitesi, İzmir, 61. (Master Thesis)

URL-4: https://borsaistanbul.com/tr/sayfa/506/pazarlar
[Access date: 18.12.2022]

URL-5: https://www.alnusyatirim.com/bist-100
[Access date: 18.12.2022]

Karagöz, S. (2020), Payların Kapanış Fiyatlarının Makine Öğrenmesi Yöntemleri ile Tahmin Edilmesi,, İstanbul , 118. (master Thesis).

URL-6: https://bigpara.hurriyet.com.tr/
[Access date: 21.11.2022]

URL-7: https://en.wikipedia.org/wiki/S%26P_500
[Access date: 22.12.2022]

URL-8: https://en.wikipedia.org/wiki/EURO_STOXX
[Access date: 22.12.2022]

URL-9 https://www.tcmb.gov.tr/
[Access date: 23.12.2022]

Kotsiantis, S.B., Kanellopoulos, D., Pintelas P.E.(2006), Data Preprocessing for Supervised Leaning. *International Journal of Computer Science Volume* 1, pp. 111-117

Alexandropoulos, S.N., Kotsiantis S.B., Vrahatis M.N. (2019), Data Preprocessing in Predictive Data Mining, Cambridge University Press 34 E1.

King, R., Orhobor, O., Taylor, C (2019) Cross-Validation is Safe to Use, Nature Machine Intelligence. 2021, 3, pp. 276-276.

Daniel, B. (2021) Cross-Validation, Data Science Laboratory, 2, pp. 542-545.

Chandrashekar, G., Sahin, F., A Survey on Feature Selection Methods. Computers & Electrical Engineering, 2014, 40, pp. 16-28.

Jović, A. and Brkić, K., Bogunović, N. A Review of Feature Selection Methods with Applications. *38th International Convention on Information and Communication Technology*, 2015, Croatia.

Zhang, F., O'Donnel, L. Support Vector Regression, Machine Learning Methods and Applications to Brain Disorders. 2020, 7, pp. 123-140

Raju, G., Lakshmi, K., Jain, V., Kalidindi, A., Padma V., Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. *Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, India.

Skinea, S., Data Science Design Manual, New York, USA, 2017, 453 s.

Liu, Y., Wang, Y., Zhang, J. New Machine Learning Algorithm: Random Forest. *International*

Conference on Information Computing and Applications, 2012, pp 246-252)

URL-10
https://towardsdatascience.com/introduction-to-bayesian-linear-regression
[Access date: 20.04.2023]

Bonaccorso, G. Machine Learning Algorithms., *Packt Publishing, Birmingham*, UK, 2017, 337s.

Ferreira, P., Le. D., Zincir-Heywood N., Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection. *15th International Conference on Network and Service Management (CNSM),* 21-25 October, 2019, Halifax, NS, Canada.

Robbach P. Neural Networks vs. Random Forests – Does it always have to be Deep Learning? *Computer Science*, 2018.