Akyol Özcan K. (2023). Borsa endeksi yönünün makine öğrenmesi yöntemleri ile tahmini: bist 100 örneği. Gümüşhane Üniversitesi Sosyal Bilimler Dergisi, 14(3), 1001-1018.

Borsa Endeksi Yönünün Makine Öğrenmesi Yöntemleri ile Tahmini: BIST 100 Örneği

Prediction of Stock Market Index Direction with Machine Learning Methods: Sample of BIST

Kübra Akyol Özcan[1]

Öz

Borsa endeksleri ve menkul kıymetler için yön (artış veya azalış) tahmininde bulunmak yatırımcıların ve araştırmacıların uzun zamandır dikkatini çekmektedir. Geçmiş verilerle gelecek veriler arasındaki bağlantının kurulması bu tahmini zorlaştırmaktadır. Söz konusu bağlantı ekonometrik modeller veya yapay zekâ modelleri yardımıyla kurulmaktadır. Yapay zekâ modelleri ekonometrik modeller gibi katı varsayımlar gerektirmez, nitel ve nicel verileri kullanabilir. Bu çalışmada Ocak 2002 - Eylül 2022 tarihleri arasında aylık ortalama BIST 100 endeks değerleri alınarak, bir önceki aya göre artış gerçekleşen durumlar için "1", azalış gerçekleşen durumlar için "0" şeklinde iki gruplu bir bağımlı değişken oluşturulmuştur. BIST 100, S&P 500, CAC40, FTSE10, NIKKEI225DAX, SHANGAICOMP, ONSUSD, USDTRY, VIX ve REPO değişkenlerinin 1. ve 2. gecikmeli değerleri bağımsız değişken olarak alınmıştır. Uygulamada BIST 100 endeksi için yön tahmininde Lojistik Regresyon Analizi (LR), Lineer Diskriminant Analizi (LDA), Naive Bayes Algoritması (NB), Rastgele orman Algoritması (RF), K-En Yakın Komşu Algoritması (KNN), Sınıflandırma ve Regresyon Ağacı Algoritması (CART), Yapay Sinir Ağları (NNET), Gauss Çekirdek Fonksiyonu ile Destek Vektör Makineleri (SVM-RBF), Polinomiyal Çekirdek Fonksiyonu ile Destek Vektör Makineleri (SVM-POLY) olmak üzere toplam dokuz farklı makine öğrenme metodu kullanılmıştır. Sonuç olarak lineer yöntemlerin daha başarılı tahmin sonuçları ürettiği görülmektedir.

Anahtar Kelimeler: *BIST 100, Makine Öğrenmesi, Borsa Endeksleri, Endeks tahmini,*

Abstract

Predicting the direction (increase or decrease) of stock market indexes and stock prices has long attracted the attention of investors and researchers. Establishing a connection between the past and future data makes this prediction difficult. The mentioned connection is established through econometric or neural network models. Neural network models do not require strict assumptions like econometric models and can utilize qualitative and quantitative data. In this study, monthly average BIST 100 index values were taken between January 2002 and September 2022, and a two-group dependent variable was formed as "1" for cases with an increase compared to the previous month and "0" for cases with a decrease. The 1st and 2nd lagged values of BIST 100, S&P 500, CAC40, FTSE10, NIKKEI225DAX, SHANGAICOMP, ONSUSD, USDTRY, VIX and REPO variables were taken as independent variables. A total of nine different machine learning methods which are Logistic Regression Analysis (LR), Linear Discriminant Analysis (LDA), Naive Bayes Algorithm (NB), Random Forests Algorithm (RF), K-Nearest Neighborhood Algorithm (KNN), Classification and Regression Trees Algorithm (CART), Artificial Neural Networks (NNET), Support Vector Machines with Radial Basis Function (SVM-RBF), Support Vector Machines with Polynomial Kernel Function (SVM-POLY) were used for direction prediction for BIST 100 index in practice. In conclusion, it is observed that linear methods produce more successful estimation results.

Keywords: *BIST 100, Machine Learning, Stock Market Indexes, Market Forecast*

Araştırma Makalesi [Research Paper]

JEL Codes: *C51, C52, C61*

Submitted: 22 / 06 / 2023
Accepted: 12 / 09 / 2023

[1] Dr. Arş. Gör, Bayburt Üniversitesi, Bayburt, Türkiye, kubraakyol@bayburt.edu.tr , Orcid No: https://orcid.org/0000-0002-1158-7017

Giriş

Menkul kıymetlerin değerlerinin yönünü tahmin etmek amacıyla çeşitli yaklaşımlar oluşturulmuştur. Bu bağlamda borsa tahmini hem ilginç hem de zorlu bir araştırma konusu olarak karşımıza çıkmaktadır. (Nasseri vd., 2015: 9192) Başarılı bir borsa tahmini için ana fikir minimum girdi ve en az karmaşık model seçimi ile en başarılı sonuçları elde etmektir. (Atsalakis ve Valavanis, 2009: 5932) Son zamanlarda yapılan araştırmalar diğer makroekonomik ve finansal değişkenlerden alınan verilerin yanı sıra önceki borsa hareketlerinden elde edilen verileri kullanarak gelecekte borsanın yönünü tahmin etmenin mümkün olduğuna dair kanıtlar içermektedir. Araştırmacılar borsanın öngörülebilir sonuçlar sağlaması nedeniyle öngörülebilirliğin nedenlerini incelemeye yönelmişlerdir. Hisse senetlerinin gelecekteki fiyatlarının tahmini siyasi olaylar, ekonomik koşullar, yatırımcı beklentileri ve hisse senedi fiyatlarını etkileyebilecek diğer çevresel faktörler gibi çok fazla faktörden etkilenmesi nedeniyle oldukça karmaşık ve zor bir meseledir. Buna ek olarak hisse senedi fiyat serileri doğası gereği genellikle oldukça dalgalı, değişken, doğrusal olmayan, parametrik olmayan, karmaşık ve kaotik bir veri yapısına sahiptir (Boyacioglu ve Avci, 2010: 7908).

Spekülatörlerin, yatırımcıların ve işletmelerin karşılaştığı en büyük zorluk finans ve emtia piyasalarındaki fiyat hareketlerini doğru bir şekilde tahmin etmektir. Bunlar, piyasaları tahmin etme arayışında gelecekteki olayların en azından kısmen mevcut veya geçmiş olaylara ve verilere dayandığını varsayarlar. Ancak finansal zaman serileri, tahmin edilmesi "en gürültülü" ve "en zor" seriler arasındadır. (Abu-Mostafa ve Atiya, 1996: 205) Hisse senedi fiyat endeksinin hareketini tahmin etmek, zaman serisi tahmininin en zor uygulamalarından biri olarak kabul edilmektedir. Etkili piyasa alım satım tekniklerini oluşturmak için hisse senedi fiyat endekslerinin hareketini doğru bir şekilde tahmin etmek oldukça önemlidir. (Kara vd., 2011: 5311) Son yıllarda bilgisayar yazılım ve donanım teknolojilerindeki ilerlemeler hesaba dayalı finansı mümkün kılmıştır. Hesaba dayalı araçların finansal alanda uygulanması, büyük kârlar ve para çekme potansiyeli nedeniyle ilgi görmektedir. Sonuç olarak hesaba dayalı finans önemli oranda ilgi kazanarak, çok disiplinli araştırma çabalarını teşvik etmiştir. Hesaba dayalı finans, akıllı araçların finansal görevlere uygulanmasının yanı sıra bilgisayar teknolojisi kullanılarak geleneksel tekniklerin geliştirilmesini de içeren matematik, ekonomi ve büyük ölçekli sayısal hesaplamaların bir entegrasyonunu oluşturmuştur. Bu finansal uygulamalar; alım-satım stratejileri, risk ve finansal yönetim, operasyon yönetimi, pazarlama, kredi riski, türev fiyatlandırması, tahmin, karar desteği, opsiyon fiyatlandırması, finansal piyasa modellemesi, arbitraj, portföy seçimi, kredi değerlendirmesi ve çok daha fazlasını içermektedir. (Tan vd., 2007: 236) Finansal zaman serileri gürültülü (noisy), durağan olmayan ve deterministik olarak kaotik olduğundan, finansal piyasa tahminini zorlaştırmaktadır. Gürültülü olma özelliği geçmişteki ve gelecekteki fiyatlar arasındaki bağlantıyı tam olarak açıklamada yeterli finansal piyasa davranışı verilerinin olmamasını ifade etmektedir. Dolayısıyla modelde yer almayan bilgiler gürültü olarak kabul edilmektedir. Finansal zaman serilerinin dağılımı durağan değildir. Ekonomik veya politik koşullar, tüccarın beklentileri, felaket veya çatışma gibi öngörülemeyen olaylar da dâhil olmak üzere birçok faktör borsa endeksi ve döviz kurları gibi finansal zaman serilerini değiştirebilir. Herhangi bir finansal zaman serisi ile diğer veri serileri arasındaki ilişki de zaman içinde değişebilir (Kumar ve Thenmozhi, 2006: 2).

Bir borsa endeksinin getirisini tahmin etmenin zor olan yönü geçmiş veriler ile gelecek arasında bağlantı kurmaktır. Bu bağlantı literatürde genellikle iki araştırma alanı tarafından ele alınmaktadır. Bunlardan birincisi ekonometrik modellerdir. Bu modeller doğrusal regresyon, AR, ARMA, ARCH ve GARCH gibi modelleri içermektedir. Bu modellerin uygulanmasındaki kilit nokta sonuçların kalitesini ve güvenilirliğini garanti etmek için bir finansal serinin yerine getirmesi gereken varsayımlara bağlıdır. İkincisi ise makine öğrenme modelleridir. Yapay zekâ tabanlı olan bu modeller yapay sinir ağları, genetik algoritmalar, bulanık mantık, destek vektör makineleri, rastgele orman yöntemi ve parçacık sürü optimizasyonu gibi modelleri içermektedir. Karmaşık, kesin olmayan ve çok büyük miktarda veriyi işleyebilirler ve bu da onları ilgi çekici hale getirmektedir. Ayrıca diğer modellere uygulandıklarında veri özelliklerinin anlamını belirsizleştirirler ve bilgi erişimini sınırlandırırlar. Yapay zekâ yöntemleri nitel ve nicel verileri kullanabilirler ve ekonometrik modeller gibi katı varsayımlar gerektirmezler (Paiva vd., 2019: 636; Sheta vd., 2015: 55).

Tahmin için kullanılan geleneksel istatistiksel teknikler verilerin doğrusal olmadığı uygulamalarda eksik kalmaktadır. Bir yapay sinir ağı, bağlantılarla birbirine bağlanan çok sayıda basit ve doğrusal olmayan bilgi işlem birimini veya düğümünü içeren bir bilgisayar sistemi türüdür. Bu, borsa analizinin kalıplarını tanımak ve optimizasyonu sağlamak için denenmiş ve doğrulanmış bir yöntemdir. Dolayısıyla borsaların nasıl çalıştığını doğru bir şekilde açıklayan doğrusal olmayan zaman serisi verilerinin kodunun, sinir ağları tarafından çözüldüğü kanıtlanmıştır (Zhu vd., 2008: 3043).

Bu çalışma BIST 100 endeksi yönünün 01:2002-09:2022 arası döneme ait aylık verilerle makine öğrenmesi metodları kullanılarak tahminini amaçlamaktadır. Bunun için öncelikle BIST 100 endeks değerleri alınarak, bir önceki aya göre artış gerçekleşen durumlar için "1", azalış gerçekleşen durumlar için "0" şeklinde iki gruplu bir bağımlı değişken oluşturulmuştur. Çalışmada Lojistik Regresyon Analizi (LR), Lineer Diskriminant Analizi (LDA), Naive Bayes Algoritması (NB), Rastgele orman Algoritması (RF), K-En Yakın Komşu Algoritması (KNN), Sınıflandırma ve Regresyon Ağacı Algoritması (CART), Yapay Sinir Ağları (NNET), Gauss Çekirdek Fonksiyonu ile Destek Vektör Makineleri (SVM-RBF), Polinomiyal Çekirdek

Fonksiyonu ile Destek Vektör Makineleri (SVM-POLY) olmak üzere toplam dokuz farklı makine öğrenme metodu kullanılmıştır. Çalışmanın birinci bölümünde hisse senedi piyasalarında fiyat tahminin önemi üzerinde durulmuştur. İkinci bölümde alan yazını sunulmuş, üçüncü bölümde veri seti ve metodoloji hakkında bilgilere yer verilmiş ve analiz sonuçları yorumlanmıştır. Son bölümde ise analiz sonuçları sunularak çalışma tamamlanmıştır. Çalışmada alan yazınındaki çalışmalardan daha fazla sayıda (dokuz farklı) makine öğrenme metodu kullanılmış ve nihai olarak hangi methodun BIST 100 endeksi tahmininde daha başarılı tahmin yöntemi olduğu ortaya konulmuştur. Bu yönüyle çalışmanın literatüre değer katacağı düşünülmektedir.

1. Literatür

Hill vd. (1994) çalışmalarında özellikle regresyona dayalı tahmin, zaman serisi tahmini ve karar verme alanlarında yapay sinir ağları ile istatistiksel modelleri karşılaştıran literatür taraması yapmışlardır. Yazarların nihai hedefi, tahmin ve karar verme modelleri için yapay sinir ağlarının potansiyelinin değerlendirmesini yapmaktır. Literatür araştırması sonucuna göre ampirik çalışmaların büyük çoğunluğu yapay sinir ağlarının istatistiksel modellerle karşılaştırılabilir olduğu yönündedir.

Khashei ve Bijari (2010) çalışmalarında otoregresif entegre hareketli ortalama (ARIMA) modellerini kullanarak yapay sinir ağlarının yeni bir hibrit modelini uygulamışlardır. Çalışmada üç farklı veri seti kullanılmıştır. Bunlar; Wolf'un güneş lekesi verileri, Kanada vaşağı verileri ve İngiliz sterlini/ABD doları döviz kuru verileridir. Çalışma sonucunda elde edilen ampirik sonuçlar, önerilen modelin yapay sinir ağları tarafından elde edilen tahmin doğruluğunu iyileştirmek için etkili bir yol olabileceğini göstermiştir. Bu nedenle özellikle daha yüksek tahmin doğruluğuna ihtiyaç duyulduğunda, tahmin görevi için uygun bir alternatif model olarak kullanılması mümkündür.

Guresen vd. (2011) çalışmalarında çok katmanlı algılayıcı (MLP), dinamik yapay sinir ağı (DAN2) ve yeni girdi değişkenleri çıkarmak için genelleştirilmiş otoregresif koşullu değişen varyans (GARCH) kullanan hibrit sinir ağları modellerini kullanmışlardır. Her model için karşılaştırma iki bakış açısıyla yapılmıştır. Bunlar; NASDAQ Menkul Kıymetler Borsası Endeksi'nin günlük reel kur değerleri kullanılarak Ortalama Kare Hata (MSE) ve Ortalama Mutlak Sapma (MAD)'dır. Çalışmada 7 Ekim 2008 ile 26 haziran 2009 arası döneme ait günlük verilerin 146 adedi eğitim için kullanılırken, 36 tanesi ise test amacıyla kullanılmıştır. Klasik YSA modeli MLP'nin küçük bir farkla DAN2 ve GARCH-MLP'den daha iyi performans gösterdiği sonucuna ulaşılmıştır. Hibrit modellerin (GARCH-ANN) tatmin edici sonuçlar vermediği de bir başka bulgu olarak sunulmuştur.

Kumar ve Murugan (2013) çalışmalarında Hindistan Ulusal Menkul Kıymet Borsası'na ve Bombay Menkul Kıymet Borsası'na ait 2007-2011 yılları arası dönemde yer alan 1237 gözlemin 866'sını eğitim ve 371'ini ise test verisi olarak kullanmışlardır. Çalışmada performans ölçümü olarak Ortalama Mutlak Hata (MAE), Ortalama Mutlak Yüzde Hatası (MAPE), Ortalama Mutlak Sapma Yüzdesi (PMAD), Hata Kareler Ortalaması (MSE), Hata Kareler Ortalamasının Karekökü (RMSE) kullanılmıştır. Çalışmada yapay sinir ağları kullanılarak, tahmin doğruluğu analiz edilmiş ve ölçülmüştür. Ayrıca çeşitli deneyler yapılarak tahmin ağı için doğru parametrelerin çağ sayısı, öğrenme oranı ve momentum sırasıyla 2960, 0.28 ve 0.5 olarak elde edilmiştir.

Akbilgic vd. (2014) çalışmalarında tahminci olarak Hibrit Radyal Temel Fonksiyonlu Sinir Ağları (HRBF-NN) adı verilen yeni bir tahmine dayalı istatistiksel modelleme tekniğini öne sürmüşlerdir. HRBF-NN; regresyon ağacını, ridge regresyonun ve radyal tabanlı fonksiyon (RBF) ile sinir ağları(NN)'nı bütünleştiren esnek bir tahmin tekniği olarak tanımlanmaktadır. En iyi tahmin edicilerin alt küme seçimini gerçekleştirmek için genetik algoritmayı (GA) ve uygunluk fonksiyonu olarak teorik bilgi ilkelerine dayalı model seçimini kullanarak yeni bir hesaplama prosedürü geliştirmişlerdir. Ayrıca BIST 100 ile diğer yedi uluslararası borsa endeksi arasındaki tahmin ilişkisini belirlemek için sayısal örnekler vermişlerdir. HRBF-NN modelinin doğrusal olmayan veri yapıları arasındaki ilişkileri yönetebilen oldukça esnek ve akıllı bir veri madenciliği tekniği olduğunu ve model tarafından yapılan tahminlerin yaklaşık %65 doğruluk oranı ile performans gösterdiğini belirtmişlerdir.

Kara ve Ecer (2018) çalışmalarında lojistik regresyon, destek vektör makineleri, yapay sinir ağları ve lineer diskriminant analiz modellerinin BIST Banka Endeksi hareketlerini belirlemede performanslarını sınıflandırmışlardır. Çalışmada 1995:03 ile 2018:03 dönemleri arası günlük veriler kullanılmıştır. Verilerin %90'lık kısmı eğitim ve %10'luk kısmı ise test verisi olarak ele alınmıştır. Çalışma sonucuna göre modeller performanslarına göre yapay sinir ağları (%81.74), lineer diskriminant analizi (76.87), lojistik regresyon (%76.70) ve destek vektör makineleri (%60,87) olarak sıralanmıştır.

Pabuçcu (2019) çalışmasında yapay sinir ağları, destek vektör makineleri ve naive-bayes sınıflandırma algoritması tekniklerini kullanarak bu modellerin sınıflandırma performanslarını, BIST 100 endeksini tahmin ederek analiz etmiştir. Çalışmada 2009 ile 2018 yılları arası teknik analiz göstergeleri kullanılmıştır. Sonuç olarak çalışmada kullanılan modellerin performansları sırasıyla yapay sinir ağları (0,998), destek vektör makineleri (0,991) ve naive- bayes (0,904) olarak sıralanmıştır.

Koç Ustalı vd. (2021) çalışmalarında BIST 30'da yer alan işletmelerin hisse senedi fiyatlarını tahmin etmek için yapay sinir ağları, rastgele orman ve XGBoost algoritmalarını kullanmışlardır. Hisse senedi fiyat tahmini için işletmelerin 2010-2019 yılları arası çeyrek dönemlik verileriyle likidite, faaliyet, kârlılık ve mali yapı oranları hesaplanmıştır. Elde edilen sonuçlara göre tahmin performansları sıralandığında birinci sırada XGBoost, ikinci sırada rastgele orman algoritması yer alırken, yapay sinir ağları en düşük tahmin performansını göstermiştir.

Demirel ve Hazar (2021) çalışmalarında BIST 100 endeksi fiyat hareketlerinin yönünü belirlemek amacıyla EURO STOXX50, FTSE 100, DAX ve DOW 30 endekslerine ait 2014-2019 yılları arası günlük verilere yapay sinir ağları yöntemini uygulamışlardır. Çalışma sonucunda BIST 100 endeksi tahmininde yapay sinir ağlarının %59.57 oranında başarı sağladığı görülmüştür.

Aksoy (2021) çalışmasında kurumsal yönetim ve BIST 30 endekslerinde yer alan beş imalat sanayi şirketinin 3 aylık ortalama hisse senedi fiyat yönü tahmini için 2010:3 ve 2020:3 dönemleri arası finansal tablo ve makroekonomik değişken verilerini kullanmıştır. Yapay sinir ağları, K-en yakın komşu algoritması, sınıflandırma ve regresyon ağacı yöntemlerinin kullanıldığı çalışma sonucunda hangi yöntemin daha başarılı tahmin yaptığı sıralanmıştır. Tahmin sıralaması en yüksekten düşüğe doğru yapay sinir ağları (%98,05), sınıflandırma ve regresyon ağacı (%96,10) ve K-en yakın komşu algoritması (%92,20) şeklinde oluşturulmuştur.

Filiz vd. (2021) çalışmalarında BIST 100 endeksinin değişim yönünü belirlemek amacıyla lojistik regresyon ve destek vektör makinesi PUK çekirdeği algoritması yöntemlerini kullanmışlardır. BIST 100 endeksi bağımlı değişken olarak alınırken, bağımsız değişken olarak Dünya'nın çeşitli ülkelerine ait borsa endekslerinin yanı sıra dolar kuru, euro kuru ve altın ons verileri 2006-2020 arası dönem için ele alınmıştır. Çalışma sonucunda destek vektör makinesi PUK çekirdeği algoritması %71.9 oranında başarı elde ederken, lojistik regresyon %70.6 oranında başarı elde etmiştir.

Jabeur vd. (2021) çalışmalarında çok sayıda açıklayıcı değişkenin ABD doları cinsinden verilen altın fiyatı üzerindeki etkisini araştırmışlardır. Veriler Ocak 1986'dan Aralık 2019'a kadar olan dönemi kapsayan 408 aylık gözlem içermektedir. Verilerin %80'lik kısmı eğitim ve %20'lik kısmı ise test amacıyla kullanılmıştır. Çalışmada kullanılan makine öğrenim modelleri; doğrusal regresyon, yapay sinir ağları, rastgele orman algoritması, Light gradyan artırma makinesi, CatBoost algoritması, XGBoost algoritmasıdır. Ayrıca bu çalışma, yorumlanabilir makine öğrenimi alanını birleştirmek için SHAP yöntemini sunmuştur. Altın fiyatını tahmin etmek için tasarlanmış XGBoost'un çıktılarını yorumlamak için Shapley katkı açıklamalarının nasıl kullanılabileceğini göstermişlerdir. Elde edilen analiz sonuçlarına göre XGBoost en iyi tahmin modeli iken onu sırasıyla CatBoost, Rf, lineer regresyon(doğrusal regresyon) ve yapay sinir ağları izlemektedir.

Sarı ve Saka Ilgın (2022) çalışmalarında 2008-2019 yılları arası aylık verilerle BRICS ülkeleri endeksleri yardımıyla BIST 100 endeks hareketlerini tahmin etmek için yapay sinir ağları modellerini kullanmışlardır. Verilerin %30'u modelin tahmin başarısını ölçmek için kullanılırken, %70'i eğitim verisi olarak kullanılmıştır. Elde edilen analiz sonuçları Rusya, Hindistan ve Güney Afrika ülkelerinin borsa endekslerinin BIST 100 endeksi tahmininde başarılı sonuçlar gösterdiği yönündedir.

2. Veri Seti ve Metodoloji

Çalışmanın bu kısmında kullanılan değişkenlere ve analiz yöntemlerine yer verilmiştir.

2.1. Değişkenler

Araştırma kapsamında makine öğrenme algoritmaları kullanılarak BIST endeksi için yön tahminleri gerçekleştirilmiştir. Veri setinde Ocak 2002 - Eylül 2022 tarihleri arasında aylık ortalama BIST 100 endeks değerleri alınarak, bir önceki aya göre artış gerçekleşen durumlar için "1", azalış gerçekleşen durumlar için "0" şeklinde iki gruplu bir bağımlı değişken oluşturulmuştur. Akbilgic ve diğ. (2014) hareketle, bağımsız değişken olarak BIST 100, S&P 500, CAC40, FTSE10, NIKKEI225, DAX, SHANGAICOMP, ONSUSD, USDTRY, VIX ve REPO değişkenlerinin 1. ve 2. gecikmeli değerleri alınmıştır.[2] Uygulamada kullanılan değişkenlere ait tanımlar tablo 1'de verilmiştir.

Tablo 1. Değişken Tanımları

Simge	Açıklama	Kaynak
BIST100	Borsa İstanbul 100 endeksi	investing.com

[2] Finansal parametreler genelde bir veya iki dönem önceki gecikmeli değerlerden etkilendiği için bu değerler seçilmiştir. Literatürde genelde 2 gecikmeye kadar değişkenler alındığı için çalışmada bu şekilde bir yol izlenmiştir.

S&P 500	Standard & Poor's 500 endeksi	investing.com
CAC40	Fransa Borsa Endeksi	investing.com
FTSE10	Londra Borsası Endeksi	investing.com
NIKKEI225	Tokyo Menkul Kıymetler Borsası	investing.com
DAX	Almanya Borsa Endeksi	investing.com
SHANGAICOMP	Şanghay Kompozit Endeksi	investing.com
ONSUSD	Ons Altın	investing.com
USDTRY	ABD Doları	investing.com
VIX	Vix Volatilite Endeksi	investing.com
REPO	Aylık Gecelik Repo Faizi	TCMB EVDS

Uygulamada BIST 100 endeksi için yön tahmininde toplam dokuz farklı makine öğrenme metodu kullanılmıştır:

- Lojistik regresyon analizi (LR)
- Lineer diskriminant analizi (LDA)
- Naive Bayes algoritması (NB)
- Rastgele orman algoritması (RF)
- K-en yakın komşu algoritması (KNN)
- Sınıflandırma ve regresyon ağacı algoritması (CART)
- Yapay sinir ağları (NNET)
- Gauss çekirdek fonksiyonu ile destek vektör makineleri (SVM-RBF)
- Polinomiyal çekirdek fonksiyonu ile destek vektör makineleri (SVM-POLY)

2.2. Lojistik Regresyon Analizi (LR):

Lojistik regresyon kategorik bağımlı değişkenin varlığı halinde bağımlı değişkenin bağımsız değişken(ler)le ilişkisinin incelenmesine yardımcı olan istatistik temelli bir algoritmadır. Lojistik regresyon,

$$y = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 x)}}$$

şeklinde formüle edilmektedir. Burada; $\propto_0$ elde edilecek eğrinin x ekseninde hareket etmesini sağlayan parametre iken, $\propto_1$ ise eğrinin dikliğini ayarlamasını sağlayan parametredir (Goy vd., 2019: 3).

Lojistik regresyon "regresyon" kelimesini içermesine rağmen aslında bir sınıflandırma modelidir ve birçok farklı özelliğe sahiptir. Veri dağılımı hakkında önceden herhangi bir varsayım gerektirmeden etiket olasılığını doğrudan modelleyip, uygun olmayan varsayımsal veri dağılımları gibi sorunları önlemek veya karar vermek amacıyla olasılığı kullanan görevler için gerekli olan ilişkili olasılıklarla birlikte etiketleri tahmin etmek bu özelliklerden birkaçıdır. Son olarak, lojistik regresyonun amaç fonksiyonu, birçok yararlı matematiksel özellik ile tüm mertebelerden türevlere sahip bir dışbükey fonksiyondur ve dışbükeylik, onu sayısal optimizasyon yöntemleriyle çözülebilir kılmaktadır. (Zhou, 2021: 63) Ayrıca lojistik regresyon yöntemi 0 ile 1 aralığında uzatılmış S şeklinde sınırlı lojistik eğriler üretebildiği için popülerliğini korumaktadır (Kleinbaum ve Klein, 2010: 6).

2.3. Lineer Dikriminant Analizi (LDA):

Lineer diskriminant analizi ilk olarak Fisher (1936) tarafından ikili sınıflandırma problemleri için önerildiğinden Fisher's Lineer Diskriminant (FLD) analizi olarak da bilinen klasik bir lineer yöntemdir. Basit bir temele sahip olan LDA, aynı sınıfa ait örnekleri birbirine yakın, farklı sınıflara ait örnekleri ise birbirinden uzak olacak şekilde sınıflandırmaktadır. (Zhou, 2021: 65) Ayrıca bu analiz, yeni ölçüm yapılarak elde edilen bir örneğin, birimlere ait m tane özelliği olan k adet sınıftan birine atanmasını sağlayan bir makine öğrenme yöntemidir. Birimler en az hata ile ait oldukları sınıfa diskriminant fonksiyonu

olarak adlandırılan eşitlikler yardımıyla atanarak sınıflandırılmaktadır. Lineer diskriminant analizinde kullanılan her değişkenin normal dağılıma sahip olması, LDA ile belirlenen ayırıcı fonksiyonların daha etkin sonuçlar vermesi için en temel varsayımdır. X, k boyutlu bir gözlem vektörünü ifade eden iki gruplu bir örnek için diskriminant fonksiyonu,

$$D_{xi} = d_{0i} + d_{1i}X_1 + \cdots + d_{ki}X_k$$

olarak gösterilmektedir. Eşitlikte D_{xi} diskriminant fonksiyonunun değerini, d_{ki} ise k. değişkenin katsayısını ifade etmektedir. (AltınYavuz ve Yavuz, 2021: 145)

2.4. Naive Bayes Algoritması (NB):

Bayes teoremini esas alarak oluşturulan Naive Bayes, sınıflandırma amacıyla kullanılan makine öğrenme algoritmalarından biridir. Söz konusu yöntem bir örneğin hedef özelliğinin sınıf değerine ait olma ihtimalini bulmak amacıyla kullanılmaktadır (KAYNAR vd., 2017: 7) Naive Bayes veya basit Bayesci sınıflandırıcı algoritmanın işleyişi aşağıdaki gibidir:

1. D'nin bir veri grubu olduğu varsayımı altında her bir grup, n boyutlu öznitelik vektörü ile temsil edilir ve $X = (x_1, x_2, \dots, x_n)$ sırasıyla n nitelikten, $A_1, A_2, \dots, A_n$ üzerinde yapılan n ölçümü ifade eder.
2. C_1'den C_m'e kadar m adet sınıf olduğu durumda ve bir X veri grubu verildiğinde sınıflandırıcı, X'in, X'e koşullanmış olarak en yüksek sonsal olasılığa sahip sınıfa ait olduğunu tahmin edecektir. Yani Naive Bayes sınıflandırıcısı X veri grubunun C_i sınıfına ait olduğunu tahmin eder. Böylece, $P(C_i|X)$'i maksimize edilir. $P(C_i|X)$'in maksimize edildiği C_i sınıfı, maksimum sonsal hipotez olarak adlandırılır. Bu durum,

$$P(C_i \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid C_i)P(C_i)}{P(\boldsymbol{X})}$$

şeklinde formüle edilir.

3. $P(X)$ tüm sınıflar için sabit olduğundan, yalnızca $P(\boldsymbol{X} \mid C_i)P(C_i)$'nin maksimize edilmesi gerekir. Sınıfın önceki olasılıkları bilinmiyorsa, o zaman genellikle sınıfların eşit derecede muhtemel olduğu varsayılır.
4. Birçok niteliğe sahip veri kümeleri verildiğinde, $P(\boldsymbol{X} \mid C_i)$'yi hesaplamak son derece pahalı olacaktır. $P(\boldsymbol{X} \mid C_i)$'deki hesaplamayı azaltmak için sınıf-koşullu bağımsızlık varsayımı yapılır. Bu, özniteliklerin değerlerinin, veri grubunun sınıf etiketi verildiğinde (yani, öznitelikler arasında hiçbir bağımlılık ilişkisi olmadığında) koşullu olarak birbirinden bağımsız olduğunu varsayar. Böylece;

$$P(\boldsymbol{X} \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$
$$= P(x_1 \mid C_i) \times P(x_2 \mid C_i) \times \cdots \times P(x_n \mid C_i)$$

olarak elde edilir. Veri gruplarından $P(x_1|C_i), P(x_2|C_i), \dots, P(x_n|C_i)$ olasılıkları kolayca tahmin edilir. Burada x_k'nın, X veri grubu için A_k özniteliğinin değerini ifade ettiği hatırlanmalıdır. Her öznitelik için özniteliğin kategorik mi yoksa sürekli değerli mi olduğuna bakılmalıdır.

5. X'in sınıf etiketini tahmin etmek için her bir C_i sınıfı için $P(\boldsymbol{X} \mid C_i)P(C_i)$ değerlendirilir. Sınıflandırıcı, yalnızca $P(\boldsymbol{X} \mid C_i)P(C_i)) >$ ise, veri kümesi X'in sınıf etiketinin C_i sınıfı olduğunu tahmin eder. Başka bir deyişle, tahmin edilen sınıf etiketi $P(\boldsymbol{X} \mid C_i)P(C_i)$'nin maksimum olduğu C_i sınıfıdır. (Han vd., 2012: 352-353)

2.5. Rastgele orman Algoritması (RF):

2001 yılında L. Breiman tarafından önerilen rastgele orman algoritması, genel amaçlı bir sınıflandırma ve regresyon yöntemidir. Rastgele birkaç karar ağacını birleştiren ve tahminlerini ortalama alarak toplayan yaklaşım, değişken sayısının gözlem sayısından fazla olması durumunda mükemmel sonuçlar vermektedir. Ayrıca, büyük ölçekli problemlere uygulanabilecek kadar çok yönlü bir yaklaşımdır. (Biau ve Scornet, 2016: 97) Sınıflandırma ve regresyon örnekleri, rastgele orman modellerinin genellikle lojistik regresyon ve lineer regresyon gibi parametrik modeller ile karşılaştırıldığında daha yüksek tahmin doğruluğuna sahip olduğu göstermektedir (Schonlau ve Zou, 2020: 24).

Rastgele orman, makine öğrenme algoritmasının bir birleşimidir. Bir dizi ağaç sınıflandırıcı ile birleştirilen her ağaç, en popüler sınıf için bir birim oy kullanır ve ardından bu sonuçları birleştirerek nihai sıralama sonucunu alır. RF, yüksek sınıflandırma doğruluğuna sahiptir, aykırı değerleri ve gürültüyü iyi tolere eder ve asla fazla öğrenme yapmaz. (Liu vd., 2012: 246) Çok sayıda karar ağacının torbalama yoluyla bir araya gelmesiyle elde edilen sınıflandırma yöntemi olan rastgele orman algoritmasında B torbalama sayısını, f_b sınıflandırma ağacını ve x' ise eğitim sonrasında görülmeyen örnekler için üretilmiş olan tahminleri göstermektedir.

$$f = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Orman yapısının oluşumunda N ağaç değerini ve m ise en iyi bölünmenin hesaplanmasında kullanılan değişken değerini ifade etmektedir. Eğitim ve test gibi iki gruba ayrılmış olan veri grubundaki tüm değişkenler arasından rastgele seçilen m değişkenleri ile en iyi dal belirlenmektedir. Gini indeksi seçilmiş olan değişkenlerin hangi değere göre ayrılacağını hesaplamaktadır. Ağaçta dal bitince bu işlem sona ermektedir. Gini indeksi,

$$\mathrm{Gini}(D) = 1 - \sum_{j=1}^{n} p_j^2$$

şeklinde hesaplanmaktadır. Formülde D, veri setinin tamamını, n seçilmiş olan veriyi, p_j veri grubunda yer alan verilerin her birinin kendisinden küçük veya büyük eleman sayılarına bölüm karesini temsil etmektedir. (Malkoçoğlu ve Malkoçoglu, 2020: 26-27)

2.6. K-En Yakın Komşu Algoritması (KNN):

K-en yakın komşu algoritmasında sınıflandırma işlemi eğitim setinde test nesnesine en yakın komşuların sayısı olan k değerine göre yapılmaktadır. Etiketlenmemiş bir nesneyi sınıflandırmak için bu nesnenin etiketlenmiş nesnelere olan mesafesi hesaplanır, k-en yakın komşuları tanımlanır ve bu en yakın komşuların sınıf etiketleri daha sonra nesnenin sınıf etiketini belirlemek için kullanılır. (Wu vd., 2008: 22) Bu algoritmada eğitim verilerini içeren kümedeki her bir örnekle tek tek işleme alınarak test edilen örneğin sınıfı, eğitim veri setindeki örneğe en yakın k adet örnek seçilerek belirlenmektedir. Bu örnekler arasındaki mesafe Öklid ile hesaplanmaktadır. (Küçük vd., 2013: 3) X'in $a_1(x), a_2(x), \ldots, a_n(x)$ vektörleriyle tanımlanan bir örneklem olduğunu varsayalım. Burada $a_r(x)$, x örneğinin r inci özelliğidir. $d(x_i, x_j)$) olmak üzere x_i ve x_j örnekleri arasındaki Öklid mesafesi,

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2}$$

olarak hesaplanmaktadır. (Şahan vd., 2007: 418)

2.7. Sınıflandırma ve Regresyon Ağacı Algoritması (CART):

En yaygın uygulanan veri madenciliği tekniklerinden biri olan CART işletme, tarım, tıp, endüstri ve mühendislik alanlarında yaygın olarak kullanılmaktadır. CART analizi tahmin problemlerinde etkilidir. Hedef değişkenin değeri kesikli olduğunda, bir sınıflandırma ağacı geliştirilir. Buna karşılık, sürekli hedef değişken için de bir regresyon ağacı geliştirilir (Chang ve Wang, 2006: 1021).

CART tarafından üretilen karar ağaçları kesinlikle ikili olup, her karar düğümü için tam olarak iki dal içermektedir. CART, eğitim veri kümesindeki kayıtları, hedef özellik için benzer değerlere sahip kayıtların alt kümelerine yinelemeli olarak bölümler. CART algoritması, her bir karar düğümü için mevcut tüm değişkenlerin ve tüm olası bölme değerlerinin kapsamlı bir aramasını gerçekleştirerek ve aşağıdaki kriterlere göre en uygun bölmeyi seçerek ağacı büyütür.

$$\Phi(s \mid t) = 2P_L P_R \sum_{j=1}^{\#\text{classes}} |P(j \mid t_L) - P(j \mid t_R)|$$

$\Phi(s \mid t)$, t. düğümdeki s. aday bölünmenin uygunluk ölçümü olarak hesaplanır. Burada,

$$t_L = t \text{ düğümünün sol taraftaki bölünmesi}$$
$$t_R = t \text{ düğümünün sağ taraftaki bölünmesi}$$
$$P_L = \frac{t_L \text{ deki kayıtların sayısı}}{\text{Eğitim setindeki kayıtların sayısı}}$$
$$P_R = \frac{t_R \text{ deki kayıtların sayısı}}{\text{Eğitim setindeki kayıtların sayısı}}$$
$$P(j \mid t_L) = \frac{t_L \text{ deki } j \text{ sınıflarının sayısı}}{t \text{ deki kayıtların sayısı}}$$
$$P(j \mid t_R) = \frac{t_R \text{ deki } j \text{ sınıflarının sayısı}}{t \text{ deki kayıtların sayısı}}$$

olarak hesaplanmaktadır. (Larose ve Larose, 2014: 168)

2.8. Yapay Sinir Ağları (NNET):

Sinir ağı, insan beynine benzer bir şekilde çalışan matematiksel bir model oluşturma girişimidir. (Gershenfeld, 1999: 150) Bir ağ, iletişim kanalları veya konektörlerle birbirine bağlanan birçok öğeden veya nörondan oluşmaktadır. Bu bağlayıcılar çeşitli yollarla ve katmanlar halinde düzenlenen sayısal verileri taşımaktadır. Sinir ağları, elemanlar arasındaki bağlantılara veya ağırlıklara belirli değerler atandığında belirli bir işlevi yerine getirebilir. Bir sistemi tanımlamak için modelin varsayılan bir yapısı yoktur, bunun yerine ağlar ayarlanır veya eğitilir, böylece belirli bir girdi belirli bir hedef çıktıya yol açar. (Minasny ve McBratney, 2002: 353) Yapay sinir ağı, girdi katmanı oluşturulabilecek 4 atomlu zincir sayısı ile değişebilecek şekilde geliştirilmiştir. Ek olarak potansiyel, girdi değişkenlerinin sürekli bir fonksiyonu ve girdilerin sıralanmasından bağımsız olacak şekilde yapılandırılmaktadır (Bholoa vd., 2007: 3).

Bir sinir ağının matematiksel modeli, ağırlıklarla birbirine bağlanan bir dizi basit fonksiyondan oluşmaktadır. Ağ, girişleri çıkışlara bağlayan bir dizi giriş birimi x, çıkış birimi y ve gizli birim z'den oluşmaktadır. Gizli birimler, girdilerden yararlı bilgiler çıkarır ve bunları çıktıları tahmin etmek için kullanır. Burada ele alınan NNET tipine çok katmanlı algılayıcı denir. $x_l(l = 1, \dots, N_i)$ öğelerin giriş vektörüne sahip bir ağ, gizli birim z_j $(j = 1, \dots, N_h)$'yi verecek şekilde ağırlıkla w_{jl} çarpılarak bir bağlantı aracılığı ile iletir.

$$z_j = \sum_{l=1}^{N_i} w_{jl} x_l + w_{jl}$$

Burada N_h, gizli birimlerin sayısı ve N_i, girdi birimlerinin sayısıdır. Gizli birimler, ağırlıklı girdi ve bir sapmadan (w_{j0}) oluşur. Bu sapma, ağırlığa eklenen bir sabit olarak işlev gören, sabit girdisi 1 olan bir ağırlıktır. (Minasny ve McBratney, 2002: 353)

2.9. Gauss Çekirdek Fonksiyonu ile Destek Vektör Makineleri (SVM-RBF) ve Polinomiyal Çekirdek Fonksiyonu ile Destek Vektör Makineleri (SVM-POLY):

Destek Vektör Makinesi (SVM) teorisi, 1990'larda Vapnik ve diğerleri tarafından yeni bir makine öğrenme yöntemi olarak geliştirilmiştir. Uyarlanabilir ve basit model yapısı, kolay optimizasyon özellikleri ile özellikle sınırlı örnekler için istatistik teorisine dayanmaktadır ve teorik olarak en iyi çözümü sunmaktadır. Yapısal risk minimizasyonu ilkesi altında iyi bir genelleme yeteneğine sahiptir. Ayrıca desen tanıma, görüntü işleme, yüz tanıma, yüz algılama ve tahmin gibi birçok alanda başarıyla uygulanmaktadır. Çekirdek işlevi, SVM için anahtar teknolojidir. Çekirdek işlevinin seçimi, makine öğreniminin öğrenme ve genelleme yeteneğini etkilemektedir. Farklı çekirdek farklı doğrusal olmayan dönüşüm ve özellik uzayını belirlemektedir. Bu nedenle SVM'yi eğitmek için farklı çekirdek seçmek farklı sınıflandırma anlamına gelmektedir. (Gao vd., 2008: 1) Çekirdeklerden bazıları; doğrusal çekirdek, polinom çekirdeği ve RBF çekirdeğidir. (Gopi vd., 2020: 5)

Genel olarak, RBF çekirdeği makul bir ilk tercihtir. Bu çekirdek, örnekleri daha yüksek boyutlu bir uzaya doğrusal olmayan bir şekilde eşlemektedir. Böylece doğrusal çekirdeğin aksine, sınıf etiketleri ve nitelikler arasındaki ilişkinin doğrusal olmadığı durumunu da işleyebilir. Polinom çekirdeği, RBF çekirdeğinden daha fazla model seçiminin karmaşıklığını etkileyen hiperparametreye ve daha az sayısal zorluğa sahiptir (Hsu vd., 2003: 4).

RBF çekirdek fonksiyonun matematiksel ifadesi,

$$\mathrm{K}(\mathrm{x}, \mathrm{x}') = \exp(-\gamma \|\mathrm{x} - \mathrm{x}'\|^2)$$

$$y = \frac{1}{2\sigma^2}$$

olarak gösterilmektedir. Denklemde yer alan x ve x' RBF çekirdeğinin iki örneğini ve σ ise serbest parametreyi simgelemektedir. (Çetin Taş ve Müngen, 2021: 384).

SVM-POLY durumu, polinom fonksiyon çekirdeğine dayanır. Formu,

$$K(x, y) = [x^T y + \theta]^p$$

şeklinde verilmektedir. Burada p, polinomun derecesidir ve θ, değerlerini genellikle tamsayı uzayından alan serbest bir parametredir. Ancak Hessian matrisinin sıfır olmasını engellediği için θ=1 tercih edilir. RBF'de olduğu gibi, serbest bir C parametresi tanımlanır. (Vafeiadis vd., 2018: 87) SVM-polinom çekirdeğinin formülü ise

$$K(x_i, x_j) = (\gamma x_i^T \cdot x_j + r)^d, \gamma > 0$$

şeklinde gösterilmektedir. (Mehdizadeh vd., 2017: 107).

BIST 100 endeksinin yön tahmininde makine öğrenme algoritmalarının performansını değerlendirmek üzere doğruluk oranı (ACC), F1-skoru (F1) ve Matthew korelasyon katsayısı (MCC) kullanılmıştır. Analiz bulguları çapraz geçerlilik yaklaşımı kullanılarak eğitim ve test verileri üzerinden değerlendirilmiştir. Araştırma verilerinde eğitim ve test kümeleri sırasıyla %70-%30, %80-%20 ve %90-%10 şeklinde üç dönem için ayrıştırılmıştır. Bu şekilde BIST 100 endeksinin yön tahmininde kısa, orta ve uzun dönem tahmin sonuçları değerlendirilmiştir.

Makine öğrenme algoritmalarının parametreleri, eğitim verileri üzerinden 10 katmanlı çapraz geçerlilik yaklaşımı üzerinden seçilmiştir. Eğitim ve test verileri ile analiz sonuçları 10 kez tekrarlanarak, seçilen parametrelere sahip modellerden hareketle ACC, F1 ve MCC değerleri elde edilmiştir.

Makine öğrenme uygulamalarında R programı kullanılmıştır (R Core Team, 2022). Analiz bulguları R programında yer alan psych Revelle (2022), caret Kuhn (2022) ve mlr Bischl ve diğ. (2016) paketleri üzerinden elde edilmiştir.

3. Bulgular

Araştırmada yer alan değişkenlere ait tanımlayıcı istatistikler tablo 2'de verilmiştir.

Tablo 2. Araştırma Verilerine Ait Tanımlayıcı İstatistikler

Değişken	Ort	SS	Min	Maks
BIST100	737.10	500.22	88.42	3179.99
S&P 500	1934.33	978.11	735.09	4766.18
CAC40	4541.00	1006.91	2618.46	7153.03
FTSE100	6034.05	1050.69	3567.40	7748.76
NIKKEI225	15980.06	5959.66	7568.42	29452.66
DAX	8519.93	3522.23	2423.87	15884.86
SHANGAICOMP	2646.88	887.11	1060.74	5954.77
ONSUSD	1141.54	466.51	317.80	1975.89
USDTRY	3.38	3.35	1.16	18.50
VIX	19.73	8.42	9.51	59.89
REPO	15.94	11.09	4.10	58.97

Not: Ort: Aritmetik ortalama, SS: Standart sapma, Min: Minimum değer, Maks: Maksimum değer

Tablo 2'de araştırmada bağımlı ve bağımsız değişken olarak kullanılan finansal verilerin tanımlayıcı istatistikleri verilmiştir. Tüm borsa verileri ve kur fiyatlamaları için ortalama, standart sapma, minimum ve maksimum değerleri hesaplanmıştır.

Tablo 3'te BIST 100 endeksinin yön tahmini için kullanılan dokuz farklı algoritma için seçilen düzenleme parametre değerleri verilmiştir.

Tablo 3. Makine Öğrenme Algoritmalarının Düzenleme Parametreleri

Algoritma	**Parametreler**
LR	-
LDA	-
NB	çekirdek=(var,yok), düzeltme parametresi=(0.1,0.2,0.3,0.4,0.5), fL=(0.1,0.2,0.3,0.4,0.5)
RF	mtry=(1,2,3,4,5,6,7,8,9,10)
KNN	k=(1,3,5,7,9)
CART	cp=(0.01,0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.09,1)
NNET	wdec=c(0.001,0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.010), nöron sayısı=c(2,3,4,5,6,7,8)
SVM-RBF	sigma=(0.01,0.02,.0.03,0.04,0.05), C=(0.25,0.50,0.75,1,1.25,1.50,1.75,2)
SVM-POLY	derece=(1,2), ölçek=(0.01,0.02,.0.03,0.04,0.05), C=(0.25,0.50,0.75,1,1.25,1.50,1.75,2)

Not:wdec: Ağırlık azaltımı, **mtry:** Deneme sayısı, **fL:** Laplace düzeltmesi, **cp:** Karmaşıklık parametresi, **Not:** NNET için tek katman kullanılmıştır.

Tabloya bakıldığında bu düzenleme parametreleri arasından 10 katlı çapraz geçerlilik tekniği ile en uygun değerler, doğruluk oranını (ACC) en iyi yapacak şekilde seçilmiştir.

Tablo 4. Eğitim Verilerine Ait Performans Ölçütlerinin İstatistiksel Bulguları

Algoritma	Uzun dönem (%70-%30)			Orta dönem %80-%20			Kısa dönem %90-%10		
	ACC	F1	MCC	ACC	F1	MCC	ACC	F1	MCC
LR	0.67	0.74	0.31	0.70	0.76	0.37	0.69	0.76	0.35
LDA	0.68	0.75	0.33	0.71	0.77	0.38	0.69	0.76	0.35
NB	0.65	0.72	0.27	0.70	0.75	0.38	0.66	0.71	0.29
RF	1	1	1	1	1	1	1	1	1
KNN	0.77	0.82	0.52	0.74	0.79	0.44	0.67	0.74	0.31
CART	0.69	0.77	0.36	0.71	0.79	0.39	0.60	0.75	0.04
NNET	0.94	0.95	0.87	0.92	0.94	0.84	0.87	0.90	0.73
SVM-RBF	0.62	0.76	0.12	0.64	0.77	0.17	0.61	0.75	0.08
SVM-POLY	0.63	0.47	0.17	0.62	0.59	0.16	0.60	0.82	0.09

Tablo 4'te makine öğrenme algoritmalarının eğitim verilerine yönelik performans ölçütü sonuçları gösterilmektedir. Eğitim verilerinde her üç dönem için de BIST 100 endeksinin yön tahmininde ACC, F1 ve MCC açısından en başarılı yöntem RF olarak görülmektedir.

Tablo 5. Test Verilerine Ait Performans Ölçütlerinin İstatistiksel Bulguları

Algoritma	Uzun dönem (%70-%30)				Orta dönem %80-%20				Kısa dönem %90-%10		
	ACC	F1	MCC		ACC	F1	MCC		ACC	F1	MCC
LR	0.66	0.77	0.22		0.65	0.77	0.16		0.63	0.69	0.24
LDA	0.66	0.77	0.22		0.65	0.77	0.16		0.63	0.69	0.24
NB	0.38	0.02	-0.01		0.44	0.25	0.11		0.30	0.03	-0.19
RF	0.47	0.31	0.12		0.39	0.11	-0.01		0.30	0.00	-0.24
KNN	0.62	0.71	0.17		0.45	0.28	0.00		0.52	0.54	0.14
CART	0.41	0.13	0.07		0.46	0.28	0.11		0.64	0.73	0.01
NNET	0.54	0.50	0.07		0.61	0.71	0.10		0.60	0.64	0.18
SVM-RBF	0.62	0.77	0.00		0.63	0.77	0.00		0.67	0.80	0.00
SVM-POLY	0.52	0.47	0.13		0.54	0.59	0.03		0.71	0.82	0.23

Tablo 5'te makine öğrenme algoritmalarının test verilerine yönelik performans ölçütü sonuçları yer almaktadır. Test verileri incelediğinde, uzun ve orta dönem BIST 100 endeksinin yön tahmininde ACC, F1 ve MCC açısından LR ve LDA algoritmalarının genel anlamda en başarılı sonuçlar ürettiği görülmektedir. Kısa dönemli BIST 100 endeksinin yön tahmininde ise SVM-POLY sonuçları, ACC ve F1 açısından diğer algoritmalara kıyasla daha başarılıdır. Ancak MCC açısından yine LR ve LDA teknikleri kısa dönemde de en başarılı performansa sahip yöntemler olarak göze çarpmaktadır.

Genel olarak test verileri incelendiğinde ele alınan değişkenler yardımıyla BIST 100 endeksinin orta ve uzun vadede yönünü tahmin etmede en başarılı olan algoritmaların lojistik regresyon analizi ve lineer diskriminant analizi olduğu görülmüştür. Bu analizlerin yanı sıra doğruluk oranının ve F1-skorunun da başarılı algoritmalar olduğu gözlenmiştir. Sonuç olarak kısa, orta ve uzun dönemde lojistik regresyon analizi ve lineer diskriminant analizi algortimalarının başarı performanslarının diğer algoritmalardan yüksek olduğu sonucuna ulaşılmıştır.

Bu bulgular, lojistik regresyon analizinin ve lineer diskriminant analizinin BIST 100 endeksinin orta ve uzun vadede yönünü tahmin etmede başarılı algoritmalar olarak tanımlayan önceki çalışmalarla da tutarlıdır (Lu, 2010; Aser ve Firuzan, 2022). Örneğin, Lu (2010) veri madenciliği, çıkarım ve tahmin de dâhil olmak üzere istatistiksel öğrenmenin unsurlarını tartışmakta ve lojistik regresyon analizi ile doğrusal diskriminant analizinin çeşitli uygulamalardaki etkinliğini vurgulamaktadır. Ayrıca, Bouwmeester ve diğerleri (2013) kümelenmiş veriler için tahmin modelleri üzerine karşılaştırmalı bir çalışma yürütmüş ve rastgele kesmeli lojistik regresyon gibi rastgele etkili regresyon modellerinin standart regresyon modellerinden daha başarılı sonuçlar verdiğini ortaya koymuştur. Bu durum regresyon tekniklerine dayanan LR ve LDA algoritmalarının BIST 100 endeks yönü için doğru tahminler sağlayabileceği fikrini desteklemektedir. Genel olarak, bu çalışmalardan elde edilen kanıtlar lojistik regresyon analizi ve lineer diskriminant analizinin kısa, orta ve uzun vadede BIST 100 endeksinin yönünü tahmin etmek için en başarılı algoritmalar olduğunu göstermektedir. Bu algoritmalar çeşitli çalışmalarda yüksek doğruluk, F1-skoru ve MCC göstererek BIST 100 endeksinin yönünü tahmin etmek için güvenilir seçenekler haline gelmiştir.

Sonuç ve Değerlendirme

Menkul kıymet borsaları mevcut finansal sistemin en önemli bileşenleri arasında yer almaktadır. Birçok yatırımcı takip ettikleri hisse senetlerinin açılış-kapanış-en yüksek-en düşük fiyatlarını ve işlem hacimlerini sürekli olarak tahmin etmeye çalışmaktadır. (Hu vd., 2018: 188) Borsa endeksinin yönü, fiyat endeksinin hareketini veya borsa endeksinin gelecekteki dalgalanma eğilimini ifade etmektedir. Yönü tahmin etmek bir yatırımcının bir finansal enstrümanı alma veya satma kararını büyük ölçüde etkilemektedir. Menkul kıymet borsası endeksi eğilimlerinin doğru tahmini, yatırımcıların borsada kâr elde etme fırsatlarına yardımcı olmaktadır. Bu nedenle, borsa endeks eğilimlerinin tahmin edilmesi yatırımcılar için son derece avantajlı olacaktır. (Qiu ve Song, 2016: 1) Bu çalışmanın amacı borsa endekslerinin artış ve azalışlarını tahmin eden makina öğrenmesi algoritmalarının karşılaştırılmasıdır.

Çalışmada Ocak 2002-Eylül 2002 tarihleri arası aylık BIST 100, S&P 500, CAC40, FTSE10, NIKKEI225DAX, SHANGAICOMP, ONSUSD, USDTRY, VIX ve REPO değişkenlerinin 1. ve 2. gecikmeli değerleri alınmıştır. BIST 100 endeksinin yön tahmini için dokuz adet makine öğrenme yöntemi kullanılmıştır. Bunlar; Lojistik Regresyon Analizi, Lineer Diskriminant Analizi, Naive Bayes Algoritması, Rasgele Orman Algoritması, K-En Yakın Komşu Algoritması, Sınıflandırma ve Regresyon Ağacı Algoritması, Yapay Sinir Ağları, Gauss Çekirdek Fonksiyonu ile Destek Vektör Makineleri ve Polinomiyal Çekirdek Fonksiyonu ile Destek Vektör Makineleri yöntemleridir. BIST 100 endeksinin yön tahmininde makine öğrenme algoritmalarının performansını değerlendirmek üzere doğruluk oranı (ACC), F1-skoru (F1) ve Matthew korelasyon katsayısı (MCC) kullanılmıştır.

Test verilerinin performans ölçümlerinin istatistiksel sonuçları incelendiğinde uzun dönemde LR ve LDA aynı başarı düzeyiyle birinci sırada yer almıştır. Orta vade de aynı algoritmik modellerin başarı düzeylerinin yüksek olduğu görülmektedir. Uzun ve orta vadede LR ve LDA modelleri en başarılı modellerken, kısa dönemde ise SVM-POLY modelinin performansının diğer modellerden fazla olduğu görülmektedir. Genel olarak bakıldığında LR ve LDA modellerinin BIST 100 endeksinin yön tahmininde öne çıktığı, bu algoritmaları SVM-POLY ve SVM-RBF modellerinin izlediği ve lineer yöntemlerin daha başarılı tahmin sonuçları ürettiği söylenebilir. Kısa dönemli yatırım yapmayı planlayanların SVM-POLY algoritmasını, orta ve uzun dönem için BIST 100 yatırımcılarının ise LR ve LDA algoritmalarını dikkate almaları başarı düzeyleri üzerinde etkili olacaktır. RF algoritması eğitim verisinde yüksek, test verisinde ise düşük değerli sonuçlar verdiği için aşırı uyum (overfitting) sorununa yol açmaktadır.

Genel olarak, makine öğrenme yöntemlerinin BIST 100 endeksi yönünün tahmininde kullanılabileceğini ifade etmek mümkündür. Bu sonuçlara göre yatırımcıların ve menkul kıymet borsasında faaliyet gösteren işletmelerin borsa endeksi yönünü belirlemek için makine öğrenmesi modellerini kullanmaları başarı düzeylerini artıracaktır. Gelecekte yapılması planlanan çalışmalar için daha fazla parametre kullanılması mümkündür. Hisse senedi fiyatlarının hükümet politikaları, şirket performansı, yatırımcıların ilgisi vb. birçok faktöre bağlı olduğu iyi bilinmektedir. Bu hususlardan herhangi biriyle ilgili çıkan bir haber, hisse senedi fiyatlarını doğrudan etkilemektedir. Bu haberler 'iyi', 'çok iyi', 'kötü' veya 'daha kötü' olarak kategorize edilerek analize dahil edilebilir.

Kaynakça

Abu-Mostafa, Y. S. ve Atiya, A. F. (1996). Introduction to Financial Forecasting, *Applied Intelligence*, *6*(3), 205-213. https://doi.org/10.1007/BF00126626

Akbilgic, O., Bozdogan, H. ve Balaban, M. E. (2014), A novel Hybrid RBF Neural Networks Model as a Forecaster, *Statistics and Computing*, *24*(3), 365-375. https://doi.org/10.1007/s11222-013-9375-7

Aksoy, B. (2021). Pay Senedi Fiyat Yönünün Makine Öğrenmesi Yöntemleri ile Tahmini: Borsa İstanbul Örneği, *Business and Economics Research Journal*, *12*(1), 89-110. http://dx.doi.org/10.20409/berj.2021.312

AltınYavuz, A. ve Yavuz, H. S. (2021). Denetimli Makine Öğrenme Yöntemleri ile Yüzey Su Kalitesinin Sınıflandırılması. *Biyoloji Bilimleri Araştırma Dergisi*, *14*(2), 142-155.

Aser, D. and Firuzan, E. (2022). Improving forecast accuracy using combined forecasts with regard to structural breaks and arch innovations. *Ekoist Journal of Econometrics and Statistics*, 37, 1-25. https://doi.org/10.26650/ekoist.2022.37.1183809

Atsalakis, G. S., & Valavanis, K. P. (2009). Surveying Stock Market Forecasting Techniques – Part II: Soft Computing Methods. *Expert Systems with Applications*, *36*(3, Part 2), 5932-5941. https://doi.org/https://doi.org/10.1016/j.eswa.2008.07.006

Bholoa, A., Kenny, S. D. ve Smith, R. (2007). A New Approach to Potential Fitting Using Neural Networks. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, *255*(1), 1-7. https://doi.org/https://doi.org/10.1016/j.nimb.2006.11.040

Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. *TEST*, *25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Bouwmeester, W., Twisk, J., Kappen, T., Klei, W., Moons, K., & Vergouwe, Y. (2013). Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Medical Research Methodology, 13*(1). https://doi.org/10.1186/1471-2288-13-19

Boyacioglu, M. A., ve Avci, D. (2010). An Adaptive Network-Based Fuzzy Inference System (ANFIS) For The Prediction of Stock Market Return: The case of the Istanbul Stock Exchange. *Expert Systems with Applications*, *37*(12), 7908-7912. https://doi.org/https://doi.org/10.1016/j.eswa.2010.04.045

Chang, L.-Y., ve Wang, H.-W. (2006). Analysis of Traffic Injury Severity: An Application of Non-Parametric Classification Tree Techniques. *Accident Analysis & Prevention*, *38*(5), 1019-1027. https://doi.org/https://doi.org/10.1016/j.aap.2006.04.009

Çetin Taş, İ., ve Müngen, A. A. (2021). Yapay Sinir Ağları ve Destek Vektör Makineleri Yöntemleri İle Bölgesel Trafik Yoğunluk Tahmini. *Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi*, *8*(15), 378-390. https://doi.org/10.54365/adyumbd.971461

Demirel, A. C., ve Hazar, A. (2021). Borsa Endekslerinin Birbirleriyle Etkileşimi ve Endeks Yönünün Tahmini: BİST100 Üzerine Bir Uygulama. *Ekonomi ve Finansal Araştırmalar Dergisi*, *3*(1), 1-8.

Filiz, E., Akoğul, S., ve Karaboğa, H. A. (2021). Büyük Dünya Endeksleri Kullanılarak BIST-100 Endeksi Değişim Yönünün Makine Öğrenmesi Algoritmaları ile Sınıflandırılması. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *10*(2), 432-441. https://doi.org/10.17798/bitlisfen.889007

Gao, H. S., Guo, A. L., Yu, X. D., ve Li, C. C. (2008, 12-14 Oct. 2008). RBF-SVM and its Application on Network Security Risk Evaluation. 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing, https://doi.org/10.1109/WiCom.2008.1110

Gershenfeld, N. A. (1999). *The Nature of Mathematical Modeling*. Cambridge University Press.

Gopi, A. P., Jyothi, R. N. S., Narayana, V. L., ve Sandeep, K. S. (2020). Classification of Tweets Data Based on Polarity Using Improved RBF Kernel of SVM. *International Journal of Information Technology*. https://doi.org/10.1007/s41870-019-00409-4

Goy, G., Gezer, C., ve Gungor, V. C. (2019, 11-15 Sept. 2019). Credit Card Fraud Detection with Machine Learning Methods. 2019 4th International Conference on Computer Science and Engineering (UBMK), https://doi.org/10.1109/UBMK.2019.8906995

Guresen, E., Kayakutlu, G., ve Daim, T. U. (2011). Using artificial Neural Network Models in Stock Market Index Prediction. *Expert Systems with Applications*, *38*(8), 10389-10397. https://doi.org/10.1016/j.eswa.2011.02.068

Han, J., Kamber, M., ve Pei, J. (2012). 8 - Classification: Basic Concepts. In J. Han, M. Kamber, ve J. Pei (Eds.), *Data Mining (Third Edition)* (pp. 327-391). Morgan Kaufmann.

Hill, T., Marquez, L., O'Connor, M., ve Remus, W. (1994). Artificial Neural Network Models for Forecasting And Decision Making. *International Journal of Forecasting*, *10*(1), 5-15. https://doi.org/https://doi.org/10.1016/0169-2070(94)90045-0

Hsu, C.-W., Chang, C.-C., ve Lin, C.-J. (2003). A Practical Guide to Support Vector Classification. In: Taipei, Taiwan.

Hu, H., Tang, L., Zhang, S., ve Wang, H. (2018). Predicting the Direction of Stock Markets Using Optimized Neural Networks With Google Trends. *Neurocomputing*, *285*, 188-195. https://doi.org/https://doi.org/10.1016/j.neucom.2018.01.038

Jabeur, S. B., Mefteh-Wali, S., ve Viviani, J.-L. (2021). Forecasting Gold Price With The XGBoost Algorithm and SHAP Interaction Values. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04187-w

Kara, İ., ve Ecer, F. (2018). BIST Endeks Hareket Yönünün Tahmininde Sınıflandırma Yöntemlerinin Performanslarının Karşılaştırılması. *ASOS Journal*, *6*(83), 514-524. http://dx.doi.org/10.16992/ASOS.14460

Kara, Y., Acar Boyacioglu, M., ve 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*(5), 5311-5319. https://doi.org/https://doi.org/10.1016/j.eswa.2010.10.027

Kaynar, O., Tuna, M. F., Görmez, Y., Ve Deveci, M. A. (2017). Makine Öğrenmesi Yöntemleriyle Müşteri Kaybı Analizi. *Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi*, *18*(1), 1-14.

Khashei, M., ve Bijari, M. (2010). An Artificial Neural Network (p,d,q) Model for Timeseries Forecasting. *Expert Systems with Applications*, *37*(1), 479-489. https://doi.org/10.1016/j.eswa.2009.05.044

Kleinbaum, D. G., ve Klein, M. (2010). Introduction to Logistic Regression. In D. G. Kleinbaum ve M. Klein (Eds.), *Logistic Regression: A Self-Learning Text* (pp. 1-39). Springer New York. https://doi.org/10.1007/978-1-4419-1742-3_1

Koç Ustalı, N., Tosun, N., ve Tosun, Ö. (2021). Makine Öğrenmesi Teknikleri ile Hisse Senedi Fiyat Tahmini. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, *16*(1), 1-16. https://doi.org/10.17153/oguiibf.636017

Kumar, D. A., ve Murugan, S. (2013, 21-22 Feb. 2013). Performance Analysis of Indian Stock Market Index Using Neural Network Time Series Model. 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, https://doi.org/10.1109/ICPRIME.2013.6496450

Kumar, M., ve Thenmozhi, M. (2006). Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest. Indian institute of capital markets 9th capital markets conference paper,

Küçük, H., Tepe, C., ve İ, E. (2013, 24-26 April 2013). Classification of EMG Signals by k-Nearest Neighbor Algorithm and Support Vector Machine Methods. 2013 21st Signal Processing and Communications Applications Conference (SIU), https://doi.org/10.1109/SIU.2013.6531240

Larose, D. T., ve Larose, C. D. (2014). Decision Trees. In *Discovering Knowledge in Data* (pp. 165-186). https://doi.org/https://doi.org/10.1002/9781118874059.ch8

Liu, Y., Wang, Y., ve Zhang, J. (2012). New Machine Learning Algorithm: Random Forest. Information Computing and Applications, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34062-8_32

Lu, Z. (2010). The elements of statistical learning: data mining, inference, and prediction. *Journal of the Royal Statistical Society Series a (Statistics in Society)*, *173*(3), 693-694. https://doi.org/10.1111/j.1467-985x.2010.00646_6.x

Malkoçoğlu, A. B. V., ve Malkoçoglu, Ş. U. (2020). Comparative Performance Analysis of Random Forest and Logistic Regression Algorithms. 2020 5th International Conference on Computer Science and Engineering (UBMK), https://doi.org/10.1109/UBMK50275.2020.9219478

Mehdizadeh, S., Behmanesh, J., ve Khalili, K. (2017). Using MARS, SVM, GEP and Empirical Equations for Estimation of Monthly Mean Reference Evapotranspiration. *Computers and Electronics in Agriculture*, *139*, 103-114. https://doi.org/10.1016/j.compag.2017.05.002

Minasny, B., ve McBratney, A. B. (2002). The Neuro-m Method for Fitting Neural Network Parametric Pedotransfer Functions. *Soil Science Society of America Journal*, *66*(2), 352-361. https://doi.org/10.2136/sssaj2002.3520

Nasseri, A. A., Tucker, A., ve de Cesare, S. (2015). Quantifying StockTwits Semantic Terms' Trading Behavior in Financial Markets: An Effective Application of Decision Tree Algorithms. *Expert Systems with Applications*, *42*(23), 9192-9210. https://doi.org/10.1016/j.eswa.2015.08.008

Pabuçcu, H. (2019). Borsa Endeksi Hareketlerinin Tahmini: Trend Belirleyici Veri. *Selçuk Üniversitesi Sosyal Bilimler Meslek Yüksekokulu Dergisi*, *22*(1), 246-256. https://doi.org/10.29249/selcuksbmyd.487862

Paiva, F. D., Cardoso, R. T. N., Hanaoka, G. P., ve Duarte, W. M. (2019). Decision-making for Financial Trading: A Fusion Approach of Machine Learning and Portfolio Selection. *Expert Systems with Applications*, *115*, 635-655. https://doi.org/10.1016/j.eswa.2018.08.003

Sarı, S. S., ve Saka Ilgın, K. (2022). BIST-100 Endeks Hareketlerinin BRICS Endeksleri Aracılığıyla Tahmin Edilmesi: Yapay Sinir Ağları Uygulaması. *Abant Sosyal Bilimler Dergisi*, *22*(1), 350-366. https://doi.org/10.11616/asbi.1096346

Schonlau, M., ve Zou, R. Y. (2020). The Random Forest Algorithm for Statistical Learning. *The Stata Journal*, *20*(1), 3-29. https://doi.org/10.1177/1536867x20909688

Sheta, A. F., Ahmed, S. E. M., ve Faris, H. (2015). A Comparison Between Regression, Artificial Neural Networks and Support Vector Machines for Predicting Stock Market Index. *International Journal of Advanced Research in Artificial Intelligence,*, *4*(7), 55-63. https://dx.doi.org/10.14569/IJARAI.2015.040710

Şahan, S., Polat, K., Kodaz, H., ve Güneş, S. (2007). A New Hybrid Method Based on Fuzzy-Artificial Immune System and K-Nn Algorithm for Breast Cancer Diagnosis. *Computers in Biology and Medicine*, *37*(3), 415-423. https://doi.org/10.1016/j.compbiomed.2006.05.003

Tan, T. Z., Quek, C., ve Ng, G. S. (2007). Biological Brain-Inspired Genetic Complementary Learning for Stock Market and Bank Failure Prediction. *Computational Intelligence*, *23*(2), 236-261. https://doi.org/https://doi.org/10.1111/j.1467-8640.2007.00303.x

Vafeiadis, T., Dimitriou, N., Ioannidis, D., Wotherspoon, T., Tinker, G., ve Tzovaras, D. (2018). A Framework for Inspection of Dies Attachment on PCB Utilizing Machine Learning Techniques. *Journal of Management Analytics*, *5*(2), 81-94. https://doi.org/10.1080/23270012.2018.1434425

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., ve Steinberg, D. (2008). Top 10 algorithms in Data Mining. *Knowledge and Information Systems*, *14*(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

Zhou, Z.-H. (2021). Linear Models. In Z.-H. Zhou (Ed.), *Machine Learning* (pp. 57-77). Springer Singapore. https://doi.org/10.1007/978-981-15-1967-3_3

Zhu, X., Wang, H., Xu, L., ve Li, H. (2008). Predicting Stock Index Increments By Neural Networks: The Role of Trading Volume Under Different Horizons. *Expert Systems with Applications*, *34*(4), 3043-3054. https://doi.org/10.1016/j.eswa.2007.06.023

Extended Abstract

Aim and Scope

Various approaches have been established to accurately predict the direction of the values of securities. In this context, stock market forecasting is both an interesting and challenging research topic. The biggest challenge facing speculators, investors and businesses is to accurately predict price movements in financial and commodity markets. The difficult aspect of forecasting the return of a stock market index is linking past data with the future. This linkage is generally addressed by two research areas in the literature. The first one is an econometric model. These models include models such as linear regression, AR, ARMA, ARCH and GARCH. The key point in applying these models depends on the assumptions that a financial series must fulfill in order to guarantee the quality and reliability of the results. The second is machine learning. These artificial intelligence-based models include artificial neural networks, genetic algorithms, fuzzy logic, support vector machines, the random forest method, and particle swarm optimization. This study aims to predict the direction of the BIST 100 index using machine learning methods with monthly data for the period between 01:2002-09:2022. For this purpose, firstly, BIST 100 index values are taken, and a two-group dependent variable is created as “1” for the cases where there is an increase compared to the previous month and “0” for the cases where there is a decrease. Following Akbilgic et al. (2014), the 1st and 2nd lagged values of BIST 100, S&P 500, CAC40, FTSE10, NIKKEI225, DAX, SHANGAICOMP, ONSUSD, USDTRY, VIX and REPO are taken as independent variables.

Methods

A total of nine different machine learning methods which are Logistic Regression Analysis (LR), Linear Discriminant Analysis (LDA), Naive Bayes Algorithm (NB), Random Forests Algorithm (RF), K-Nearest Neighborhood Algorithm (KNN), Classification and Regression Trees Algorithm (CART), Artificial Neural Networks (ANNET), Support Vector Machines with Radial Basis Function (SVM-RBF), and Support Vector Machines with Polynomial Kernel Function (SVM-POLY) were used. In this study, more machine learning methods (nine different methods) were used than in previous studies in the literature and finally, it was revealed which method is the more successful forecasting method for BIST 100 index.

Findings

It is seen that LR and LDA algorithms generally produce the most successful results in terms of ACC, F1 and MCC in the long- and medium-term BIST 100 index direction prediction. In the short-term BIST 100 index direction prediction, SVM-POLY results are more successful than other algorithms in terms of ACC and F1. However, in terms of MCC, LR and LDA techniques stand out as the methods with the best performance in the short term. In general, when the test data are analyzed, it is seen that logistic regression analysis and linear discriminant analysis are the most successful algorithms in predicting the direction of the BIST 100 index in the medium and long term with the help of the variables considered. In addition to these analyses, the accuracy rate and F1-score were also found to be successful algorithms. As a result, it is concluded that the performance of logistic regression analysis and linear discriminant analysis algorithms is higher than that of other algorithms in the short, medium and long term.

Conclusion

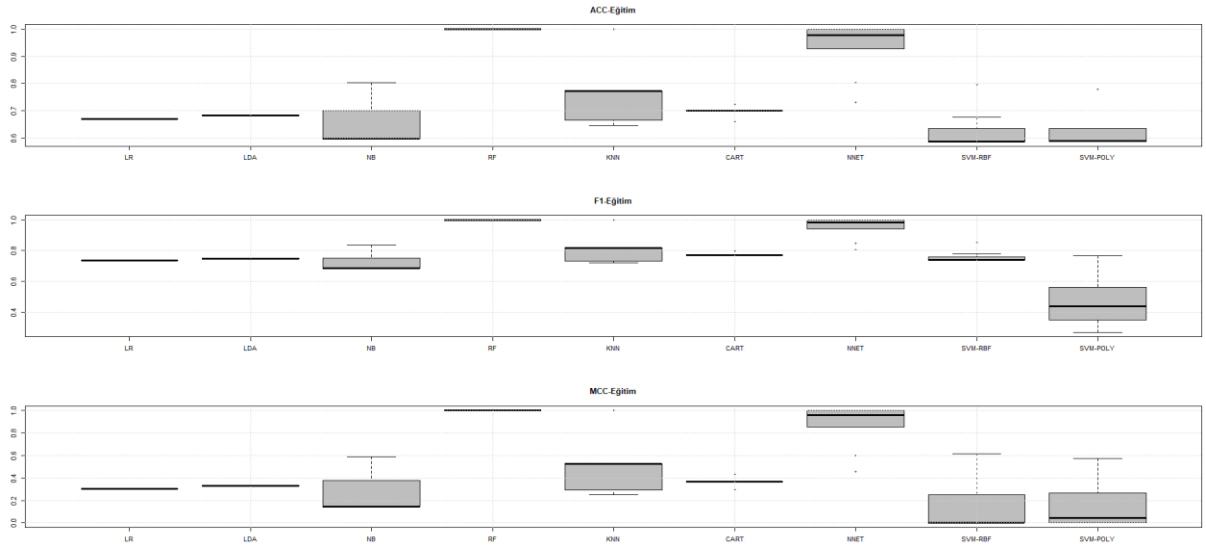
When the statistical results of the performance measurements of the test data are analyzed, LR and LDA rank first in the long run with the same success level. In the medium term, it is seen that the same algorithmic models have high success levels. While LR and LDA models are the most successful models in the long and medium term, the SVM-POLY model outperforms the other models in the short term. In general, it can be said that LR and LDA models stand out in the direction prediction of the BIST 100 index, followed by SVM-POLY and SVM-RBF models, and linear methods produce more successful prediction results. Short-term investors should consider the SVM-POLY algorithm, while medium and long-term BIST 100 investors should consider the LR and LDA algorithms. Since the RF algorithm gives good results on the training data and bad results on the test data, it leads to the problem of overfitting. In general, it is possible to state that machine learning methods can be used to predict the direction of the BIST 100 index.

According to these results, the use of machine learning models to determine the direction of the stock market index by investors, policymakers, and businesses operating on the stock exchange will increase their level of success. For future studies, it is possible to use more parameters with high correlation. It is well known that stock prices depend on many factors, such as government policies, company performance, investor interest, etc. News about any of these issues directly affects stock prices. This news can be categorized as “good”, “very good”, “bad” or “worse” and included in the analysis. Such a semi-supervised system would allow systems to be more robust and forecasts to be more accurate.

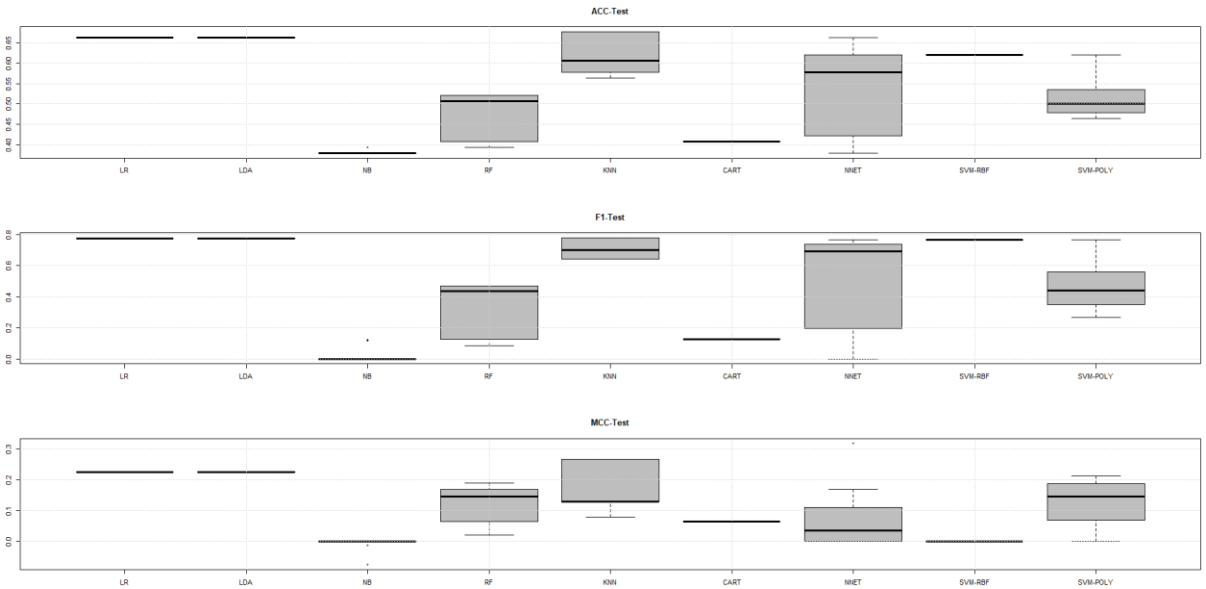
EK:

EK 1. Performans Ölçütleri Grafikleri

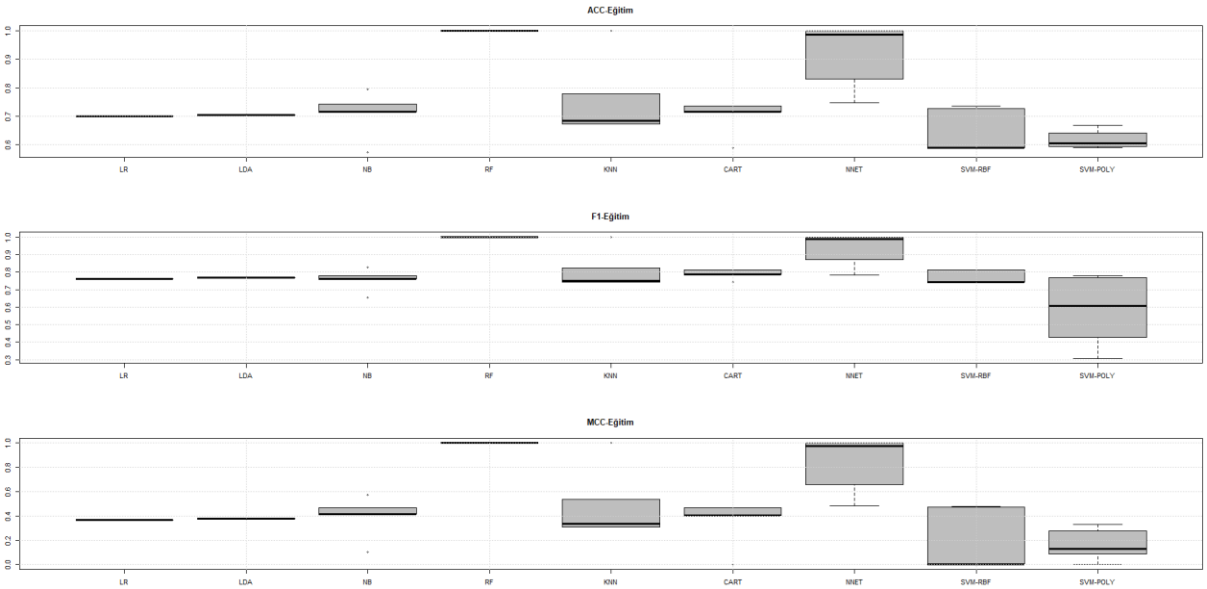
Şekil 1-6 arasında 10 tekrar için uzun dönem (%70-%30), orta dönem (%80-%20) ve kısa dönem (%90-%10) eğitim ve test verilerine dair performans ölçütleri sonuçlarının kutu grafikleri gösterilmektedir.



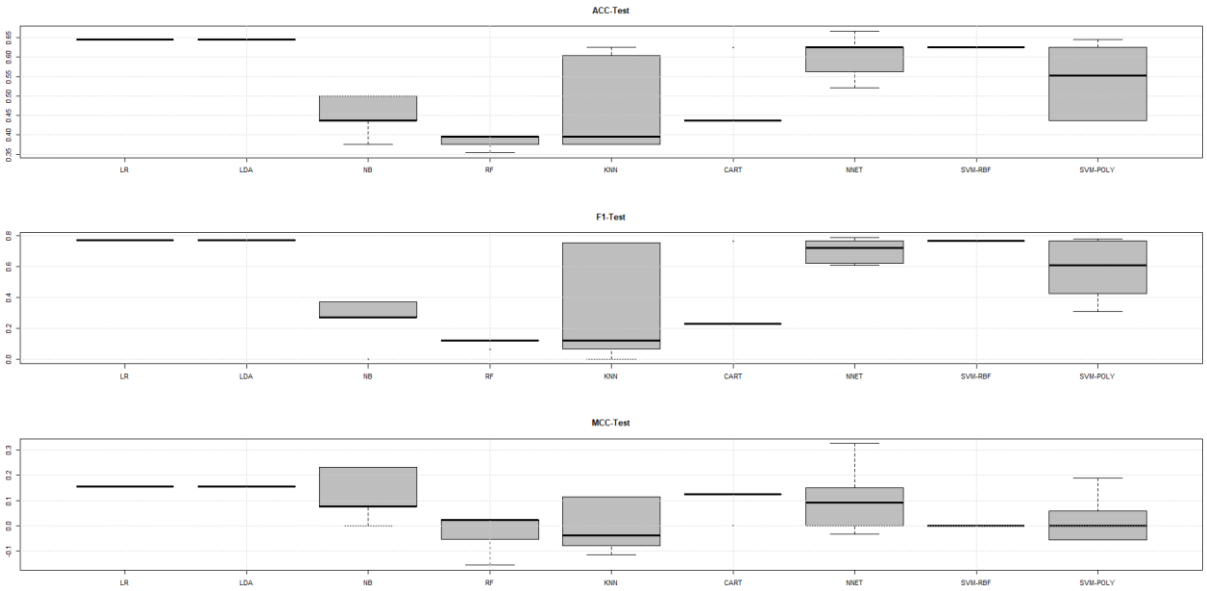
Şekil 1. Uzun Dönem Eğitim Verilerine Ait Performans Ölçütlerinin Grafiksel Sonuçları (%70-%30)



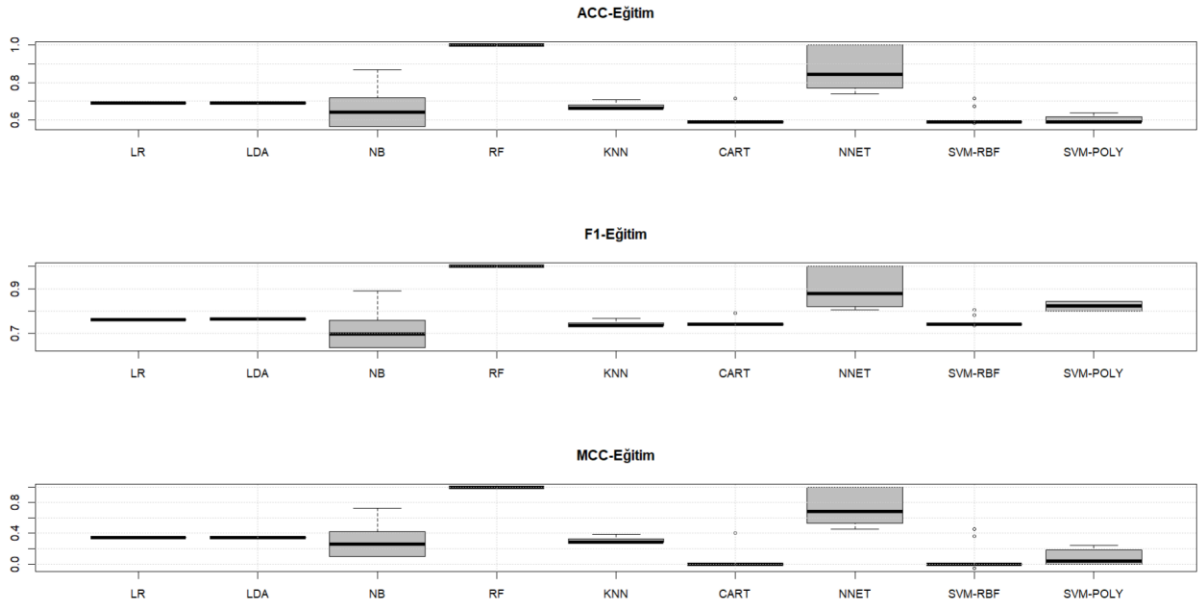
Şekil 2. Uzun Dönem Test Verilerine Ait Performans Ölçütlerinin Grafiksel Sonuçları (%70-%30)



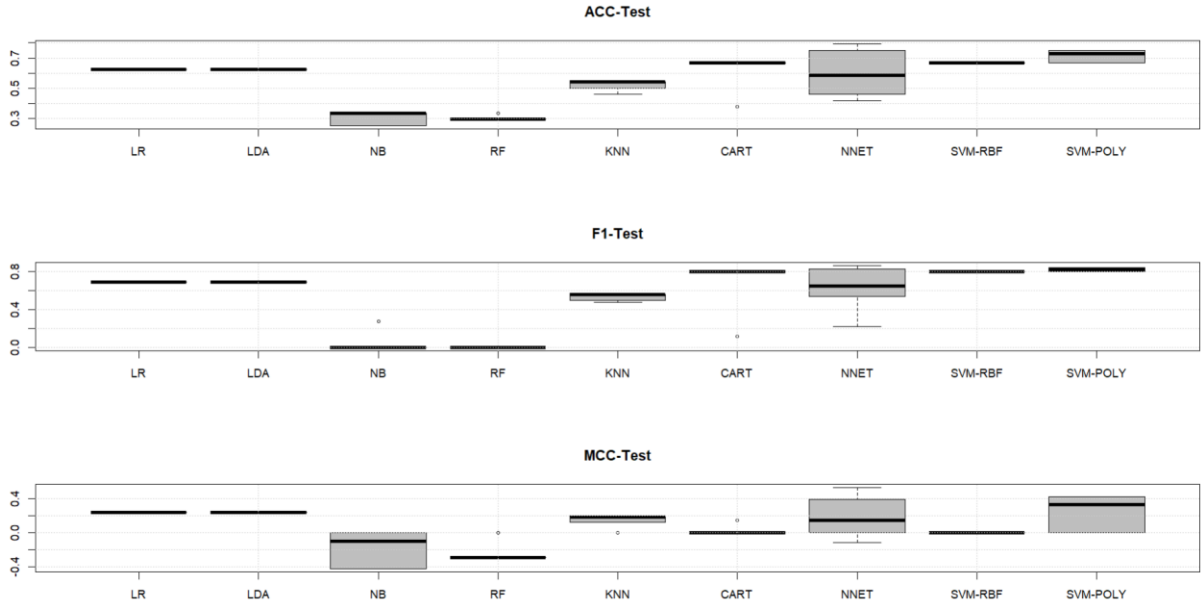
Şekil 3. Orta Dönem Eğitim Verilerine Ait Performans Ölçütlerinin Grafiksel Sonuçları (%80-%20)



Şekil 4. Orta Dönem Test Verilerine Ait Performans Ölçütlerinin Grafiksel Sonuçları (%80-%20)



Şekil 5. Kısa Dönem Eğitim Verilerine Ait Performans Ölçütlerinin Grafiksel Sonuçları (%90-%10)



Şekil 6. Kısa Dönem Test Verilerine Ait Performans Ölçütlerinin Grafiksel Sonuçları (%90-%10)