Pamukkale Univ Muh Bilim Derg, 27(4), 556-569, 2021

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Borsa endeks hareket yönünün çoklu lojistik regresyon ve k-en yakın komşu algoritması ile tahmini

Prediction of stock index movement direction with multiple logistic regression and k-nearest neighbors algorithm

Gülder KEMALBAY[1*], *Begüm Nur ALKIŞ*[2]

[1]İstatistik Bölümü, Fen Edebiyat Fakültesi, Yıldız Teknik Üniversitesi, İstanbul, Türkiye.
kemalbay@yildiz.edu.tr
[2]Fen Bilimleri Enstitüsü, Yıldız Teknik Üniversitesi, İstanbul, Türkiye.
begumalkis@gmail.com

Geliş Tarihi/Received: 08.02.2020 Düzeltme Tarihi/Revision: 10.08.2020 doi: 10.5505/pajes.2020.57383
Kabul Tarihi/Accepted: 02.09.2020 Araştırma Makalesi/Research Article

Öz

Hisse senedi piyasası birçok makroekonomik değişkenler ve politik faktörlerden etkilendiği için finansal veri madenciliğinde, hisse senedi endeksi hareket yönü tahmini zor bir sınıflandırma problemidir. Bu problemin doğru tahmini kısa vadeli yatırımcılara erken öneri sistemi olarak hizmet verebileceği için birçok araştırmacının ilgisini çekmektedir. Bu çalışma, sınıflamaya dayalı denetimli makine öğrenmesi algoritmaları yardımı ile Borsa İstanbul 100 (BIST100) endeksinin günlük aşağı veya yukarı hareket yönünü tahmin etmeyi amaçlar. İlgilendiğimiz problem, belirli bir günde BIST100 endeksinin yükseleceğini veya düşeceğini tahmin etmektir. Bu amaç doğrultusunda, BIST100 endeks hareket yönü üzerindeki etkisi istatistiksel olarak anlamlı bulunan bağımsız değişkenler kullanılarak çoklu lojistik regresyon ve K-en yakın komşu algoritması modelleri kurulmuştur. Son olarak, örneklem dışı tahminler borsadaki gerçek hareketlerle karşılaştırılmıştır. Performanslar sadece doğruluk ile değil, diğer istatistiksel metrikler ile de ölçülmüştür. Elde edilen sonuçlara göre, lojistik regresyon analizi verilen zaman dilimi içinde BIST100 verileri üzerinde K-en yakın komşu algoritmasına karşı %81 doğruluk oranı ile daha iyi tahmin performansı elde etmiştir.

Anahtar kelimeler: Endeks hareket yönü, K-En yakın komşu algoritması, Lojistik regresyon, Denetimli öğrenme.

Abstract

In financial data mining, stock index movement direction prediction is a challenging classification problem, since stock index is affected by many economic and political factors. The accurate prediction of this problem is of interest to many researchers as it can serve as an early recommender system for short-term financiers. This study aims to predict daily upward or downward movement direction of Borsa Istanbul 100 (XU100) index with the aid of supervised machine learning algorithms based on classification. Problem we deal with includes whether on a specific day the XU100 index fall into up bucket or fall into down bucket. For this purpose, the multiple logistic regression and K-nearest neighbors algorithm models are fitted using independent variables whose effect on XU100 index movement direction was statistically significant. Lastly, the out-of sample predictions are compared with the actual movements in the stock market. Performances are measured not only with accuracy but also other statistical metrics. According to the results obtained, logistic regression analysis achieves better predict performance with 81% accuracy opposed to K-nearest neighbors algorithm on XU100 data over the given time period.

Keywords: Index movement direction, K-Nearest neighbors algorithm, Logistic regression, Supervised learning.

1 Giriş

Finansal zaman serileri içerisinde önemli bir yere sahip olan borsa endeks verileri üzerine yapılacak fiyat ya da hareket yönü tahmini çalışmaları hem yatırımcıların hem de araştırmacıların ilgisini çekmektedir. Pay piyasalarındaki belirsizlikler, fiyatlarda gözlemlenen oynaklık yanı sıra hisse senedi fiyatlarının ekonomik, politik ve spekülatif birçok etmenden etkilenmesi sebebi ile yatırımcılar için geleceğin öngörülmesi zorlaşmaktadır. Gelişen teknoloji ile beraber ortaya çıkan veri madenciliği ve makine öğrenmesi temelli istatistiksel teknikler ile finansal tahmin problemlerinin doğru kestirilmesi yatırımcılar için bir erken uyarı sistemi olarak kullanılması bakımından önem taşımaktadır.

Bu çalışmada, sınıflama temelli denetimli makine öğrenmesi algoritmaları kullanarak BIST100 endeks hareket yönünü tahmin etmek amaçlanmıştır. BIST100 endeksi, Türkiye borsasının en temel göstergesi olup Borsa İstanbul'da işlem gören, işlem hacmi ve piyasa değeri bakımından en yüksek olan 100 adet hisse senedinin performansını ölçmek için kullanılır. Çalışmada, lineer sınıflayıcı olan lojistik regresyon analizi ile lineer olmayan sınıflayıcı olan K-en yakın komşu (K-nn) algoritması BIST100 endeks hareket yönü verisi üzerinde başarılı bir tahmin performansına sahip midir sorusuna cevap aranmaktadır. Bu amaç doğrultusunda, hem literatürdeki çalışmalar takip edilmiş hem de finans alanında uzman kişi görüşü alınarak Borsa İstanbul'u etkileyeceği düşünülen çeşitli finansal göstergeler kullanılmıştır. Literatürde yapılan çalışmaların çoğunun endeks fiyat tahmini olması yanında Borsa İstanbul'a ait çeşitli endekslerin hareket yönünü tahmin eden çalışmalar incelendiğinde bilgimiz dahilinde BIST100 endeksi için lojistik regresyon analizi ile K-nn algoritmasını kıyaslayan çalışmaya rastlanmamıştır. Bu çalışma, BIST100 endeks hareket yönünün öngörülebilirliği için kıyaslanan metodolojinin farklı oluşu ve elde edilen bulguların yatırımcılar tarafından strateji oluşturma aşamasında kullanılabilirliği açısından özgünlük sunmaktadır ve literatüre katkı sağlamaktadır.

*Yazışılan yazar/Corresponding author

Bu kapsam dahilinde çalışmanın kalan kısmı literatür, metodoloji, analiz ve bulgular ile sonuçlar olmak üzere dört bölümden oluşmaktadır.

2 Literatür

Bu bölümde, başta BIST endeksleri olmak üzere çeşitli ülkelerin borsa endeksleri üzerine yapılan endeks hareket yönü tahminine yönelik bir takım çalışmalara kronolojik sıralamaya göre yer verilmiştir. Hisse senedi fiyatı yön tahmini için yapılan çalışmaların çoğunluğu lojistik regresyon, yapay sinir ağları (YSA), K-nn ve destek vektör makineleri (DVM) üzerine yoğunlaşmıştır [1]. Son yıllarda yapılan çalışmalarda ise derin yapay sinir ağları kullanıldığı görülmüştür. Literatürde konu ile ilgili yapılan çalışmalar ve sonuçları izleyen şekilde özetlenmiştir.

Diler [2], İstanbul Menkul Kıymetler Borsası (İMKB) 100 endeksinin bir sonraki günkü yönünü YSA ile tahmin etmeye çalışmıştır. Yedi adet teknik gösterge değişkeni ile oluşturduğu modelin test verilerindeki doğru sınıflama oranı %60.81'dir.

Huang ve diğ. [3], 1 Ocak 1990-31 Aralık 2002 tarihleri arasında işlem gören NIKKEI 225 endeksinin haftalık hareket yönünü DVM yöntemine ek olarak geri yayılımlı YSA, lineer ve karesel diskriminant analizi ile tahmin etmiştir. Sonuç olarak, DVM'yi diğer sınıflandırma yöntemleriyle entegre ederek önerdiği bileşik model (%75) en başarılı sınıflama performansına sahip olmuştur.

Kutlu ve Badur [4], IMKB endeksi fiyat hareket yönünü tahmin etmek için denetimli öğrenme modelleri arasından ileri beslemeli YSA modeli oluşturmuştur. 02.07.2001 ve 13.07.2006 tarihleri arasındaki 1270 iş gününe ait endeks değerleri kullanılmıştır. Elde edilen sonuçlar hareketli ortalamalar yöntemi ile karşılaştırılmış ve sonuçlara göre YSA %55.1 doğruluk oranı ile yön tahmini için hareketli ortalamalar yöntemine göre daha başarılı olmuştur.

Kara ve diğ. [5], BIST100 endeksinin günlük hareket yönü tahmini için 02.01.1997-31.12.2007 tarihleri arasındaki endeks kapanış fiyatları ile on adet teknik göstergeyi kullanarak YSA ve DVM modelleri kurmuşlar ve sonuçları kıyaslamışlardır. Her iki model için de en iyi parametreler ayarlanmış ve sonuç olarak YSA (%75.74) tahmin performansı ile DVM yöntemine göre (%71.52) daha başarılı olmuştur.

Özdemir ve diğ. [6], IMKB 100 endeksi getiri yönünü tahmin etmek için ikili sınıflandırma teknikleri arasından lojistik regresyon analizi ve DVM yöntemini kullanmışlardır. Her iki yöntem ile yaklaşık olarak eğitim kümesinde %75, test kümesinde ise %86 genel doğruluk oranı elde edilmiştir. Ayrıca lojistik regresyon negatif getiri yön tahmini için daha başarılı iken DVM ise pozitif getiri yön tahmininde daha başarılı olmuştur.

Dutta ve diğ. [7], Hint borsasında 2005-2008 yılları arasında işlem gören hisse senetlerinin performans sınıflaması için lojistik regresyon analizi uygulamıştır. 8 adet finansal oranın bağımsız değişken olarak kullanıldığı çalışma, %74.6 doğruluk oranı elde etmiştir. Benzer olarak; Subha ve Nambi [8], Hint borsasına ait BSE-SENSEX ve NSE-NIFTY endekslerinin Ocak 2006-Mayıs 2011 tarihleri arasındaki günlük hareket yönünü tahmin etmek için K-nn algoritmasını kullanmışlar ve elde edilen sonuçları lojistik regresyon analizi ile kıyaslamışlardır. Sonuç olarak K-nn algoritması (%79.65), lojistik regresyon analizine (%54.11) göre daha başarılı sınıflandırma performansı sergilemiştir.

Tayyar ve Tekin [9], İMKB 100 endeksi hareket yönünü teknik analiz göstergeleri kullanarak DVM ile tahmin etmiş ve elde edilen sonuçları lojistik regresyon analizi ile kıyaslamıştır. Tahmin edilen sonuçlara göre iki yöntem benzer sonuçlar elde etse de DVM haftalık verilerin yön tahmini için %70 sınıflama başarısı ile en iyi model seçilmiştir.

Türkmen ve Cemgil [10], ABD borsasına ait NASDAQ endeksinde 01.01. 2010 ile 30.06.2014 tarihleri arasında işlem gören seçilmiş hisse senetlerinin hareket tahmini için DVM, çok katmanlı algılayıcı, rassal ormanlar modellerinin yanı sıra derin öğrenme algoritmalarından biri olan yığılmış oto-kodlayıcı modellerini kullanmışlardır. Tahmin sonuçları kesinlik ve F_1 skora göre değerlendirilmiş, her iki değere göre DVM en başarılı sınıflayıcı seçilmiştir.

Patel ve diğ. [11], Hint borsasına işlem gören CNX Nifty ve S&P BSE Sensex endekslerinin hareket yönü tahmini için YSA, DVM, rassal orman algoritması ve naive-Bayes yöntemlerini kullanmıştır. Ocak 2003-Aralık 2012 tarihleri arasındaki endeks değerleri ve 10 adet teknik göstergenin kullanıldığı çalışma sonucu rassal ormanlar algoritması diğer üç yönteme göre daha iyi performans sağlamıştır.

Gündüz ve diğ. [12], Borsa İstanbul'da en çok işlem gören üç hisse senedinin günlük hareket yönünü tahmin etmeyi amaçladıkları çalışmalarında Ocak 2011-Aralık 2015 tarihlerini kapsayan veri seti ile derin sinir ağlarının bir türü olan evrişimsel sinir ağları modeli oluşturmuşlardır. Tahmin için öznitelik olarak hem hisse senedi fiyatı hem de dolar-altın fiyatı ile hesaplanan teknik göstergeler kullanılmıştır. Sonuç olarak, hem fiyat hem de dolar-altın özniteliklerinin birlikte kullanıldığı model, %61 ile en yüksek doğruluk oranına ulaşmıştır.

Yakut ve Gemici [13], BIST100 endeksinde 2009-2014 yılları arasında işlem gören 18 adet hisse senedi verileri ile C5.0 algoritması, lojistik regresyon, CART algoritması ve DVM yöntemlerini kullanarak getiri yönünü sınıflamaya çalışmışlardır. Analiz aşamasında Tüfe ve 20 adet finansal oran bağımsız değişken olarak kullanılmış ve sonuç olarak CART algoritması %89.8 doğru sınıflama oranı ile en başarılı olmuştur.

Kara ve Ecer [14], BIST banka endeksinin hareket yönünü sınıflamak için 10 adet teknik gösterge ile beraber YSA, DVM, lojistik regresyon ve lineer diskriminant analizi kullanmıştır. Sonuçlara göre YSA %81.74 ile en yüksek doğru sınıflama oranı elde etmiştir.

Oğuz ve diğ. [15], BIST100 endeks fiyat yönü hareket sınıflaması için teknik göstergeleri kullanarak doğrusal regresyon, lojistik regresyon ve naive-Bayes algoritmaları ile tahmin yapmışlardır. F_1 skoruna göre her üç yöntem de yaklaşık %71 doğru sınıflama oranı ile benzer sonuçlar vermiştir.

Livieris ve diğ. [16], borsa endeks hareket yönünü tahmin etmek için ağırlık kısıtlamalı derin sinir ağları modeli önermişlerdir. Önerilen modelin performansını ölçmek için DJIA, NASDAQ ve S&P 500 endekslerine ait 3 ayrı veri seti üzerinde yön tahmini yapılmış ve modelin sınıflandırma performansı üzerinde etkinliği gösterilmiştir.

3 Metodoloji

Çalışmada ilgilendiğimiz problem, zaman serisi verisinin denetimli makine öğrenme algoritmaları ile sınıflandırılması problemidir. Makine öğrenmesi problemleri genellikle statiktir,

bu nedenle denetimli öğrenme algoritmaları uygulanmadan önce zaman serisi verisi üzerinden öznitelik çıkarımı yapılması gereklidir. Böylece zaman serisi verisi, özellik vektörü ile temsil edilerek zamansal problem statik bir probleme dönüşür [17], [18].

Zaman serisi verisini özellik vektörü ile temsil ederken genellikle serinin gecikme (lag) değerleri kullanılır. T bir indis olmak üzere $\{Y_t, t \in T\}$ bir zaman serisi olsun. L gecikme operatörü Y_t'nin keyfi bir t anındaki değerini $t-k, k \in Z$ anındaki değerine dönüştürür ve $L^k Y_t = Y_{t-k}$ ile tanımlanır. Öznitelik çıkarımı aşamasında girdi değişkenleri olarak gecikme değerleri yanı sıra dış veriler de kullanılabilir [19].

Sınıflandırma, hedef sınıf nitelik değeri bilinen örneklerin incelenmesi yolu ile sınıf niteliği bilinmeyen yeni bir örneğin mevcut sınıflardan hangisine ait olacağını tahmin eden bir denetimli öğrenme algoritmasıdır [20]. Çalışmada kullanılan sınıflandırma yöntemleri mevcut avantajları nedeni ile tercih edilmiştir ve izleyen şekilde açıklanabilir. Lojistik regresyon analizi, çıktı olarak yeni bir örneğin sınıf niteliklerinden birine atanma olasılığını veren bir sınıflandırma yöntemi olup sonuçlarının kolaylıkla yorumlanabilir olması, mevcut varsayımlarının az olması, diskriminant analizi gibi alternatif sınıflandırma yöntemlerinde var olan varyansların homojenliği ve çok değişkenli normallik varsayımlarını içermemesi nedeni ile çalışmada tercih edilmiştir [21]. K-nn algoritması yeni bir örneğin sınıfını tahmin ederken eğitim örneklerine olan benzerliklerini göz önüne alan örnek tabanlı sınıflandırıcı olup yaygın kullanımı, kolay uygulanması, ayrıca parametrik olmayan bir sınıflandırıcı olması dolayısı ile değişkenlerin dağılım varsayımı gerektirmemesi nedeni ile çalışmada kullanılmıştır [22].

İzleyen alt bölümlerde, BIST100 endeks hareket yönünü sınıflamak amacı ile kullanılan yöntemlerin yanı sıra hata matrisi ve model performans ölçütleri ile uygun kesim noktasının belirlenmesine yer verilmiştir.

3.1 Çoklu lojistik regresyon analizi

Y bağımlı değişken ve p adet bağımsız değişkenler vektörü $\mathbf{X}^T = (X_1, \dots, X_p)$ olmak üzere bağımlı değişkenin iki kategorili olduğunu varsayalım. Y'nin sahip olduğu kategori genellikle Y=0 ve Y=1 olarak kodlanır. $\mathbf{X}$ vektörünün aldığı değer $\mathbf{x}^T = (x_1, \dots, x_p)$ bilindiğinde Y'nin 1 değerini alma olasılığı $\pi(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x})$ ile gösterilir. O halde çoklu lojistik regresyon modeli Denklem (1)'de tanımlanmıştır:

$$\pi(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}} \quad (1)$$

$\pi(\mathbf{x})$ olasılığı ile $\mathbf{X}$ bağımsız değişkenleri arasındaki ilişki doğrusal olmayıp ilişkinin grafiği S biçiminde eğri şeklindedir. Denklem (1)'deki çoklu lojistik regresyon modeline $lojit[\pi(\mathbf{x})] = g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right)$ ile gösterilen lojit dönüşüm uygulandığında Denklem (1)'deki model, Denklem (2)'de verilen parametrelerine göre doğrusal bir modele dönüşür:

$$\mathrm{g}(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (2)$$

$\mathbf{x}$'in değer kümesine bağlı olacak şekilde $g(\mathbf{x})$ sürekli olabilir ve $g(\mathbf{x})$'in alacağı değerler $-\infty$ ile $+\infty$ arasında değişebilir. Modelin $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \dots, \beta_p)$ parametreleri, $L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}(1 - \pi(\mathbf{x}_i))^{1-y_i}$ olabilirlik fonksiyonu kullanılarak en çok olabilirlik yöntemi ile tahmin edilir [23]. Model kurulduktan sonra, $\beta_1, \dots, \beta_p$ katsayılarının aynı anda sıfıra eşit olmadığı olabilirlik oran testi ile test edilir ve sıfırdan farklı olduğu sonucuna ulaşılırsa her bir değişkenin anlamlılığı Wald testi ile araştırılır. Model uyumunu değerlendirmek için Cox-Snell R^2 ve Nagelkerke R^2 istatistikleri kullanılabilir. Modelin uyum iyiliğini araştırmak için Hosmer ve Lemeshow testi uygulanır [23].

Çoklu lojistik regresyon analizinde, bağımlı değişken binom dağılımına olup varyansı $\pi(\mathbf{x})$'in bir fonksiyonudur; dolayısı ile x değerlerine bağlı olarak değişeceği için varyansların homojenliği varsayımını içermemektedir. Ayrıca, bağımsız değişkenlerin dağılımı ile ilgili bir varsayımı yoktur ancak bağımsız değişkenler arasında çoklu doğrusal bağlantı problemi olup olmadığı incelenmelidir. Problem, genellikle bağımsız değişkenler arasında yüksek korelasyon olması, ya da gözlem sayısının değişken sayısından küçük olması (n<p) durumlarında gözlemlenir. Bu durumda lojistik regresyon katsayılarının standart hataları olduğundan büyük kestirildiği için Wald istatistiği küçülür ve aslında modele katkısı önemli olan bir değişken istatistiksel olarak anlamsız bulunur. Problemin tespiti için lineer regresyonda olduğu gibi tolerans değeri ya da varyans şişirme faktörü (VIF) değeri incelenir. Tolerans değeri 0.10'dan küçük ya da VIF değeri 10'dan büyük ise çoklu doğrusal bağlantı problemi vardır; öyleyse ilgili değişkenler modelden çıkartılabilir veya gözlem sayısı arttırılabilir. Lojistik regresyon analizi ile iyi uyuma sahip model elde edebilmek için bağımlı değişken ile ilgisiz olan bağımsız değişkenlerin incelenerek modele alınmaması önerilmektedir. Bu durumlar, yöntemin kısıtlamaları olarak sayılabilir ancak göz ardı edilmemesi gerekir [21],[24].

Sınıflandırma yapılırken her bir gözlem için lojistik regresyon ile kestirilen $\hat{\pi}(x_i) = P(Y = 1|x_i)$ olasılıklar genellikle 0.5 kesim noktası ile kıyaslanır. Eğer $\hat{\pi}(x_i) > 0.5$ ise bu gözleme karşılık gelen y_i değeri 1; aksi halde 0 olarak sınıflandırılır. Ancak, farklı kesim noktaları için daha iyi sınıflandırma performansı elde edilebilir. Alt başlık 3.4'te uygun kesim noktasının nasıl hesaplanacağına değinilecektir.

3.2 K-en yakın komşu algoritması

K-nn algoritması, sınıf kategorisi bilinen eğitim veri setindeki örneklerin uzaklık ölçüsüne dayalı olan örnek-tabanlı sınıflama tekniklerinden biridir. p adet nitelik ile tanımlanan n tane $\mathbf{x}_1^T = (x_{11}, \dots, x_{1p}), \dots, \mathbf{x}_n^T = (x_{n1}, \dots, x_{np})$ veri örneğinin olduğunu varsayalım. Her örnek, p boyutlu vektör uzayında bir noktayı temsil etmektedir. *i.* ve *j.* örnek noktaları arasındaki uzaklık $d(\mathbf{x}_i, \mathbf{x}_j)$ ile gösterilir ve niteliklerin hepsi nümerik olduğunda genellikle Denklem (3)'te verilen Euclidean uzaklığı ile hesaplanır:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - y_{jk})^2} \quad (3)$$

Yeni bir örneğin hangi sınıfa ait olacağını belirleyebilmek için bu noktadan eğitim veri setindeki tüm noktalarına olan uzaklık hesaplanır ve hesaplanan uzaklıklar küçükten büyüğe sıralanır. Önceden belirlenmiş olan K komşu sayısı parametresi dikkate alınarak K adet en yakın uzaklığa sahip noktalar seçilir. Seçilen K tane sınıf kategorisi bilinen eğitim veri noktaları içerisinde yeni gözlemin sınıfını belirlemek için genellikle çoğunluk oylaması veya ağırlıklı oylama yöntemlerine başvurulur. Çoğunluk oylamasına göre, yeni örneğin sınıf ataması

yapılırken K adet komşu arasında en çok tekrar eden sınıf seçilir. Ağırlıklı oylama yönteminde ise yeni örnek noktasına daha yakın olan komşuların sınıflama yaparken söz hakkının yüksek olması amaçlanır. Eğitim veri setindeki K tane $\mathbf{x}_m; m = 1, \dots, K$ en yakın komşunun oyu, yeni örnek noktası $\mathbf{x}_Z$'ye olan mesafe ile ters orantılı olduğu Denklem (4)'te verilmiştir:

$$oy(\mathbf{x}_m) = \begin{cases} \infty, & d(\mathbf{x}_m, \mathbf{x}_Z) = 0 \; ise \\ \dfrac{1}{d(\mathbf{x}_m, \mathbf{x}_Z)}, & diğer \; durumda \end{cases} \quad (4)$$

Denklem (4)'te yer alan tüm oylar toplanır ve yeni örneğin sınıfı, en yüksek ağırlıklı oylamaya sahip olan sınıf kategorisine atanır. K-nn algoritmasında genellikle $K = \sqrt{n}$ ile seçilir veya çapraz doğrulama yöntemi ile K parametresi deneme yapılarak bulunur [20],[25],[26].

K-nn algoritması bahsedilen avantajları yanı sıra bazı kısıtlamalara sahiptir. Algoritmanın sınıflandırma başarısı özellikle K parametresinin seçimine bağlıdır. K parametresinin çok küçük seçilmesi durumunda algoritma gürültü içeren veriye karşı daha hassas çalışacaktır. Tersi durumda ise, başka sınıflara ait olan veriler komşuluğa dahil olacaktır. Diğer önemli bir nokta ise uygun uzaklık ölçütünün seçilmesi ile ilgilidir. Uzaklık ölçütü uygulamadan önce farklı birime sahip veriler var ise standardize edilmesi gereklidir. Bağımsız değişkenlerin sürekli olması halinde uygun olan Euclidean uzaklık ölçütünün, değişken sayısının çok fazla olduğu yüksek boyutlu verilerde iyi sınıflama yapamadığı bilinmektedir. Ayrıca, yüksek boyutlu verilerde K-nn algoritması yüksek bellek kapasitesine ihtiyaç duyar. Bu nedenle daha az sayıda ve ilgili değişkenler ile sınıflandırılma yapılması önerilir [22],[27].

3.3 Hata matrisi ve model performans ölçütleri

Oluşturulan sınıflandırma modelinin başarı performansını ölçmek için hata matrisi ve hata matrisi yardımı ile elde edilen çeşitli performans ölçütleri kullanılır. Hata matrisi, sütunda gözlemlenen çıktı değerleri, satırda ise model tarafından tahmin edilen değerleri gösterir. İkili sınıflandırma yöntemi için hata matrisi Tablo 1'de verilmektedir.

Tablo 1. İkili sınıflandırma yöntemi için hata matrisi.

Table 1. Confusion matrix for binary classification method.

	Gözlemlenen		
Tahmin Edilen	Pozitif	Negatif	Toplam
Pozitif	doğru pozitif (dp)	yanlış pozitif (yp)	$\hat{n}^+$ tahmin pozitif sayısı (tpoz)
Negatif	yanlış negatif (yn)	doğru negatif (dn)	$\hat{n}^-$ tahmin negatif sayısı (tneg)
Toplam	n^+ gerçek pozitif sayısı (poz)	n^- gerçek negatif sayısı (neg)	n toplam örnek sayısı

Hata matrisinden elde edilen ve aşağıda verilen ölçütler sınıflandırma yöntemlerinin performansını değerlendirmek için kullanılır [25],[26].

Doğruluk, doğru sınıflandırılmış örnek sayısının toplam örnek sayısına bölünmesi ile hesaplanır: $\frac{dp+dn}{n}$.

Hata oranı, yanlış sınıflandırılmış örnek sayısının toplam örnek sayısına bölünmesi ile ya da 1-doğruluk ile hesaplanır: $\frac{yp+yn}{n} = 1 - \frac{dp+dn}{n}$.

Duyarlılık ya da doğru pozitif oranı, doğru sınıflandırılmış pozitif örneklerin gerçek pozitif sayısına oranıdır: $\frac{dp}{poz} = \frac{dp}{dp+yn}$.

Belirleyicilik ya da doğru negatif oranı, doğru sınıflandırılmış negatif örneklerin gerçek negatif sayısına oranıdır: $\frac{dn}{neg} = \frac{dn}{dn+yp}$.

Doğruluk, duyarlılık ve belirleyicilik türünden ifade edilebilir: $duyarlılık\,\frac{dp+yn}{n} + belirleyicilik\,\frac{dn+yp}{n}$. Bu ifadede $\frac{dp+yn}{n}$ yaygınlık oranı (pr), $\frac{dn+yp}{n}$ ise 1-yaygınlık oranı (1-pr) ölçütüdür.

Yanlış pozitif oranı, 1-belirleyicilik ile hesaplanır: $\frac{yp}{neg} = \frac{yp}{yp+dn}$.

Yanlış negatif oranı, 1-duyarlılık ile hesaplanır: $\frac{yn}{poz} = \frac{yn}{yn+dp}$.

Pozitif öngörü değeri ya da kesinlik, doğru sınıflandırılan pozitif örneklerin tahmin pozitif sayısına oranıdır: $\frac{dp}{tpoz} = \frac{dp}{dp+yp}$.

Negatif öngörü değeri, doğru sınıflandırılan negatif örneklerin tahmin negatif sayısına oranıdır: $\frac{dn}{tneg} = \frac{dn}{dn+yn}$.

F-ölçüsü, duyarlılık ve kesinlik ölçütlerinin harmonik ortalamasıdır: $\frac{2*kesinlik*duyarlılık}{kesinlik+duyarlılık}$.

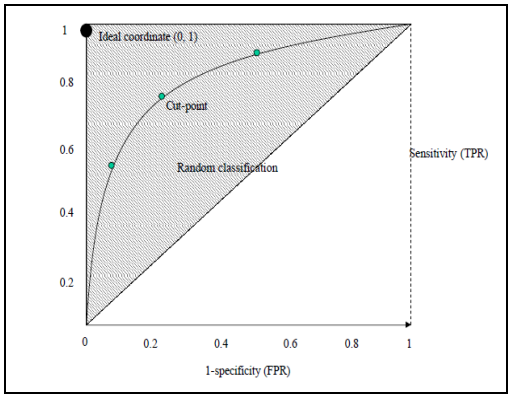
Pozitif olabilirlik oranı: $\frac{doğru\;pozitif\;oranı}{yanlış\;pozitif\;oranı} = \frac{duyarlılık}{1-belirleyicilik}$.

Negatif olabilirlik oranı: $\frac{yanlış\;negatif\;oranı}{doğru\;negatif\;oranı} = \frac{1-duyarlılık}{belirleyicilik}$.

Tanısal üstünlük oranı: $\frac{pozitif\;olabilirlik\;oranı}{negatif\;olabilirlik\;oranı}$.

Modelin iyi uyumu için doğruluk, duyarlılık, belirleyicilik, pozitif öngörü değeri, negatif öngörü değeri, F-ölçüsü, pozitif olabilirlik oranı, tanısal üstünlük oranı değerlerinin yüksek; hata oranı, yanlış pozitif oranı, yanlış negatif oranı, negatif olabilirlik oranı değerlerinin düşük çıkması beklenmektedir.

Sınıflandırıcıların performanslarını kıyaslamak için kullanılan diğer bir ölçüt, ROC (Receiver Operator Characteristic) eğrisi altında kalan ve AUC (Area Under Curve) ile gösterilen alan hesabıdır. ROC eğrisi, farklı kesim noktaları için dikey eksende duyarlılık yatay eksende ise (1-belirleyicilik) değerlerinin yer aldığı eğridir [21]. Şekil 1'de ROC uzayı verilmektedir.



Şekil 1. ROC uzayı [28].

Figure 1. ROC space [28].

ROC uzayında (0,0) noktası tüm örneklerin negatif; (1,1) noktası tüm örneklerin pozitif sınıflandırıldığını gösterir. (0,1) noktası ideal sınıflandırmayı verirken, (0,0) ile (1,1) noktalarını birleştiren köşegen çizgisi, sınıflandırmanın rastgele olarak yapıldığını göstermektedir. ROC eğrisi altındaki alan yani AUC değeri, pozitif ve negatif örnekleri doğru sınıflandırma oranını vermektedir. AUC değeri 1'e ne kadar yakınsa modelin uyumunun o kadar iyi olduğu söylenir. AUC değerinin 0.5'e eşit olması, köşegen çizgisi altında kalan alanı yani sınıflandırmanın rastgele yapıldığını gösterir. Bu nedenle, hiçbir gerçekçi sınıflandırıcının AUC değeri 0.5'ten küçük olmamalıdır [29].

ROC eğrisinin kullanım amaçlarından biri de sınıflandırıcı için uygun kesim noktasının hesaplanmasına imkan sağlamaktır.

3.4 Uygun kesim noktasının hesaplanması

Hata matrisi yardımı ile model performans ölçütleri değerlendirilirken seçilen kesim noktasına bağlı kalınmaktadır. Uygun kesim noktası, ROC eğrisi yardımı ile duyarlılık ve belirleyicilik ölçütlerini eş zamanlı değerlendiren çeşitli kriterler ile belirlenebilir. Bu çalışmada, Youden indeksi, Minimax kriteri ve duyarlılığı belirleyiciliğe eşit kılan yöntem kullanılmıştır.

Youden indeksi, J ile gösterilir ve c keyfi bir kesim noktası olmak üzere Denklem (5) ile ifade edilir:

$$J(c) = [duyarlılık(c) + belirleyicilik(c) - 1] \quad (5)$$

Denklem (5)'te verilen $J(c)$ fonksiyonunu maksimum yapan c değeri, uygun kesim noktası olarak seçilir ve c^*_{Youden} ile gösterilir. Youden indeksi, duyarlılık ve belirleyicilik ölçütlerine eşit önem verir. Youden indeksi doğru pozitif oranı ile yanlış pozitif oranı arasındaki farkı maksimum yapmayı amaçlar. Youden indeksine göre uygun kesim noktası, ROC eğrisi ile rastgele sınıflandırmayı gösteren köşegen çizgisi arasındaki dikey mesafenin maksimum olduğu noktadır.

Minimax kriteri, M ile gösterilir ve c keyfi bir kesim noktası olmak üzere Denklem (6) ile ifade edilir:

$$M(c) = max\begin{bmatrix} pr\big(1 - duyarlılık(c)\big), \\ (1-pr)\big(1 - belirleyicilik(c)\big) \end{bmatrix} \quad (6)$$

Denklem (6) ile verilen $M(c)$ fonksiyonunu minimum yapan c değeri, uygun kesim noktası olarak seçilir ve $c^*_{Minimax}$ ile temsil edilir. Minimax kriteri, en sık rastlanan hatayı minimize etmeyi sağlayan uygun kesim noktası bulmayı amaçlar.

Bir diğer kritere göre ROC eğrisi üzerinde duyarlılığın belirleyiciliğe eşit olduğu değere karşılık gelen kesim noktası, en uygun kesim noktası olarak seçilir ve $c^*_{SE=SPC}$ ile gösterilir.

Eğitim veri seti üzerinde farklı yöntemler ile elde edilen uygun kesim noktaları arasından seçim yapabilmek için test veri seti üzerinde duyarlılık ve belirleyicilik dengesini sağlayacak şekilde performans değerlendirmesi yapılabilir [30]-[32].

4 Analiz ve bulgular

Bu bölümde, çalışmada kullanılan veri seti, analiz öncesi veri seti hazırlık aşaması ve elde edilen analiz sonuçları açıklanmıştır. Veri seti, R programı 3.6.2 versiyonu kullanılarak analiz edilmiştir.

4.1 Veri seti

Bu araştırma, BIST100 endeksinin 11 Ocak 2010 tarihinden 13 Ekim 2016 tarihine kadar geçen 1700 işlem gününe ait yüzde getiri verileri üzerinde gerçekleştirilmiştir. Finansal bir zaman serisi olan BIST100 endeksi verisi üzerinden öznitelik çıkarımı için serinin gecikme değerleri kullanılmıştır. Günlük veri ile çalışıldığı için, t gününün endeks değerini beş işlem günü önceki değerlerinin etkileyebileceği göz önüne alınmıştır. Girdi değişkenleri olarak BIST100 endeks getiri serisinin gecikmeleri yanı sıra endeksi etkileyebilecek diğer finansal veriler de kullanılmış ve bu veriler modele ilave edilirken bir gün önceki gecikme değerleri alınmıştır. Böylece, zaman serisi verisi denetimli makine öğrenmesi problemi ile kullanılabilir hale getirilmiştir.

Araştırmanın bağımlı değişkeni BIST100 endeks hareket yönü olup iki kategorili nominal ölçekli nitel bir değişkendir. Endeks getirisinin bir önceki günkü değeri pozitif ise hareket yukarı yönlü 1; değilse 0 şeklinde kodlanmıştır. BIST100 endeks hareket yönüne ait 1700 adet gözlemin 795 tanesi aşağı, 905 tanesi ise yukarı yönlüdür. Endeks hareket yönünü tahmin etmek için işlem hacmi; önceki beş işlem gününün kapanış getirileri; ABD Doları kapanış getirisi; gösterge niteliğindeki Türkiye Cumhuriyeti Merkez Bankası (TCMB) Euro, İngiliz Sterlini, Rus Rublesi kurları; alternatif yatırım araçları olarak Bono, Altın, Brent Petrol, Aktif Tahvil kapanış getirisi; küresel borsa endeksi olarak ABD Borsası Dow Jones endeksi kapanış getirisi; büyük Dünya endeksleri olarak Brezilya Borsası Bovespa endeksi, Çin Borsası Shanghai endeksi kapanış getirileri; diğer borsa endeksleri olarak Belarus ve Merval endeksi kapanış getirileri olmak üzere toplam 19 adet nicel bağımsız değişken kullanılmıştır.

Veri seti, Finnet Elektronik Yayıncılık Data İletişim San.Tic.Ltd.Şti.'ne ait lisanslı yazılım tarafından sağlanmıştır; bu nedenle verilere erişim serbest değildir. Verilere erişim talebi için https://www.finnet.com.tr adresi ile iletişime geçilebilir [33].

Çalışmada, BIST100 endeks hareket yönünün belirli bir t gününde yükseleceğini veya düşeceğini tahmin etmek için girdi olarak kullandığımız bağımlı değişkenin gecikmeleri $t-1, \ldots, t-5$ (bir gün önce,...,beş gün önce) periyodunda ve diğer bağımsız değişkenler $t-1$ (bir gün önce) periyodundadır. Analizde kullanılan bağımsız değişken isimleri ve açıklamaları Tablo 2'de sunulmaktadır.

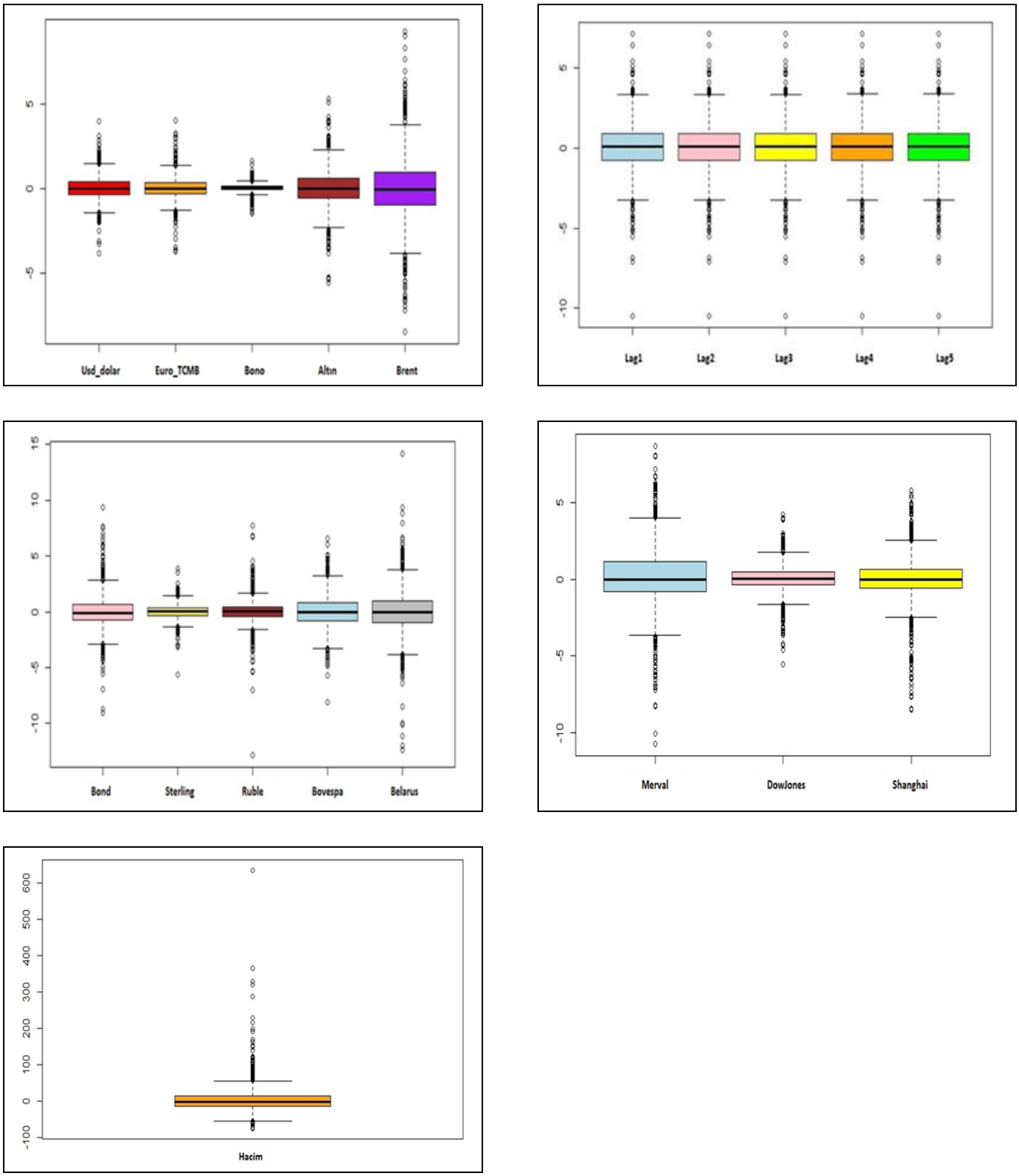
Tablo 2. Analizde kullanılan bağımsız değişkenler.

Table 2. Independent variables used in the analysis.

Değişken İsimleri	Değişken Açıklamaları
Hacim	Toplam işlem hacminin günlük getirileri
Lag1	BIST100 kapanışının bir gün önceki getirileri
Lag2	BIST100 kapanışının iki gün önceki getirileri
Lag3	BIST100 kapanışının üç gün önceki getirileri
Lag4	BIST100 kapanışının dört gün önceki getirileri
Lag5	BIST100 kapanışının beş gün önceki getirileri
Usd.dolar	ABD doları günlük kapanış getirisi
Euro.TCMB	TCMB Euro kuru günlük getirisi
Bono	Bono günlük kapanış getirisi
Altın	Altın (Ons/Dolar) fiyatının günlük getirisi
Brent	Brent petrolün günlük kapanış getirisi
Bond	Aktif tahvilin günlük kapanış getirisi
Sterling	TCMB İngiliz Sterlini kuru günlük getirisi
Ruble	TCMB Rus Rublesi kuru günlük getirisi
Bovespa	Bovespa endeksinin günlük kapanış getirisi
Belarus	Belarus endeksinin günlük kapanış getirisi
Merval	Merval endeksinin günlük kapanış getirisi
DowJones	Dow Jones endeksinin günlük kapanış getirisi
Shanghai	Shanghai endeksinin günlük kapanış getirisi

4.2 Veri setinin hazırlanması

Ham veri setini analize hazır hale getirebilmek için aykırı değerler, çoklu doğrusal bağlantı, yinelenen ve eksik değerler kontrol edilmiştir. Veri önişleme sonucuna göre eksik değer gözlenmemiştir, veri tekrarlarını ortadan kaldırıp verinin tutarlılığını arttırmak için normalizasyon incelemesi yapılmıştır. Aykırı değer analizi için tüm nümerik değişkenlerin kutu grafikleri çizdirilmiştir ve Şekil 2'de sunulmuştur.



Şekil 2. Bağımsız değişkenlere ait kutu grafikleri.

Figure 2. Box plot of independent variables.

Şekil 2 incelendiğinde veri setindeki tüm değişkenlerde uç değerlerin varlığı ve değişkenlerin aldığı değerlerin farklı aralıklarda olduğu gözlenmiştir. Finansal piyasalarda getirilerin uç değerlerde gerçekleşmesi bazı ekonomik veya spekülatif faktörlerden kaynaklanabileceği için veri setinden uç değerler çıkartılmayarak analizlere devam edilmiştir. İki nümerik bağımsız değişken arasındaki doğrusal ilişkinin incelenmesi için Pearson korelasyon katsayı değerleri hesaplanmıştır. Ancak ikiden fazla bağımsız değişken arasında yakın doğrusal bağımlılığının incelenmesi için korelasyon katsayısı yeterli değildir. Bu nedenle, bağımsız değişkenler arasında çoklu doğrusal bağlantı probleminin olup olmadığını incelemek amacı ile VIF ve tolerans değerleri hesaplanarak Tablo 3'te sunulmuştur.

Tablo 3. Bağımsız değişkenlerin tolerans ve VIF değerleri.

Table 3. Tolerance and VIF values of independent variables.

Bağımsız Değişkenler	Tolerans	VIF
Hacim	0.984	1.016
Lag1	0.762	1.312
Lag2	0.921	1.086
Lag3	0.992	1.008
Lag4	0.993	1.007
Lag5	0.988	1.012
Usd.dolar	0.448	2.234
Euro.TCMB	0.462	2.166
Bono	0.362	2.763
Altın	0.963	1.039
Brent	0.735	1.360
Bond	0.367	2.727
Sterling	0.404	2.473
Ruble	0.910	1.099
Bovespa	0.583	1.714
Belarus	0.603	1.659
Merval	0.649	1.540
DowJones	0.551	1.815
Shanghai	0.946	1.057

Tablo 3'e göre tüm değişkenlerin tolerans değeri 0.10'dan büyük ya da VIF değeri 10'dan küçük çıktığı için bağımsız değişkenler arasında çoklu doğrusal bağlantı problemi bulunmamaktadır.

Analize başlamadan önce veri seti hold-out yöntemi ile (%95) eğitim ve (%5) test olmak üzere ikiye parçalanmıştır. 2016 yılından önceki 1615 adet gözleme ait veriler eğitim veri setini, 2016 yılını kapsayan 85 adet gözleme ait veriler ise test veri setini oluşturmaktadır. Tüm analizler, eğitim veri setinde gerçekleştirilmiştir ve modellerin tahmin performansları test veri setinde kıyaslanmıştır. Çalışma boyunca istatistiksel anlamlılık düzeyi %5 olarak belirlenmiştir.

4.3 Lojistik regresyon analizi bulguları

Analize dahil edilen tüm bağımsız değişkenlerin yer aldığı çoklu lojistik regresyon modeline ilişkin olmak üzere bağımsız değişkenlerin tahmin edilen eğim katsayısı $\hat{\beta}$, eğim katsayısına ait standart hatası S.E($\hat{\beta}$), -2Log-olabilirlik değeri, Wald istatistiği, p değeri, odds oranı Exp($\hat{\beta}$) ve odds oranı için %95 güven sınırları Tablo 4'te verilmiştir.

Tablo 4'e göre, Lag1 (Wald=6.195, p=0.013), Bono (Wald=15.770, p=0.000), Brent (Wald=5.128, p=0.024), Bovespa (Wald=6.964, p=0.008), Belarus (Wald=50.686, p=0.000) ve DowJones (Wald=11.485, p=0.001) değişkenlerinin endeks hareket yönü üzerinde etkilerinin önemli olduğu bulunmuştur. Örneğin, Bono günlük kapanış getirilerindeki bir birimlik artış, endeks hareketinin yukarı yönlü olma odds değerini 4.053 (%95 güven aralığı, 2.031-8.085) katına çıkarmaktadır.

Elde edilen lojistik regresyon modelinin uyum iyiliğini değerlendirmek için pseudo R^2 istatistikleri ve Hosmer-Lemeshow testi Tablo 5'te verilmiştir.

Tablo 4. Çoklu lojistik regresyon analiz sonuçları.

Table 4. Multiple logistic regression analysis results

Değişkenler	$\hat{\beta}$	S.E($\hat{\beta}$)	-2 Log-olabilirlik	Wald	p değeri	Exp($\hat{\beta}$)	%95 Güven Sınırları	
							Alt	Üst
Sabit	0.082	0.055	0.082	2.218	0.136	1.085		
Hacim	0.003	0.002	0.003	2.149	0.143	1.003	0.999	1.007
Lag1	-0.104	0.042	-0.104	6.195	0.013*	0.901	0.819	0.983
Lag2	0.023	0.039	0.023	0.353	0.553	1.023	0.947	1.099
Lag3	-0.034	0.036	-0.034	0.882	0.348	0.966	0.899	1.037
Lag4	0.019	0.036	0.019	0.288	0.591	1.020	0.949	1.094
Lag5	0.067	0.036	0.067	3.428	0.064	1.069	0.996	1.147
Usd.dolar	-0.144	0.122	-0.144	1.393	0.238	0.866	0.681	1.099
Euro.TCMB	-0.002	0.121	-0.002	0.000	0.989	0.998	0.787	1.265
Bono	1.399	0.352	1.399	15.770	0.000*	4.053	2.031	8.085
Altın	0.015	0.055	0.015	0.077	0.782	1.015	0.912	1.129
Brent	-0.076	0.034	-0.076	5.128	0.024*	0.927	0.867	0.989
Bond	-0.014	0.060	-0.014	0.057	0.811	0.986	0.876	1.108
Sterling	0.119	0.131	0.119	0.824	0.364	1.126	0.871	1.455
Ruble	0.033	0.061	0.033	0.288	0.591	1.033	0.917	1.163
Bovespa	0.129	0.049	0.129	6.964	0.008*	1.138	1.033	1.253
Belarus	0.303	0.043	0.303	50.686	0.000*	1.354	1.245	1.472
Merval	0.046	0.034	0.046	1.817	0.178	1.047	0.979	1.118
DowJones	0.284	0.084	0.284	11.485	0.001*	1.329	1.127	1.566
Shanghai	0.020	0.038	0.020	0.283	0.595	1.021	0.946	1.100

Tablo 5. Modelin uyum iyiliği.

Table 5. Goodness of fit of the model.

Model Özeti		
-2 Log-olabilirlik	Cox-Snell R^2	Nagelkerke R^2
2067.616	0.153	0.204
Hosmer-Lemeshow testi		
χ^2	Serbestlik derecesi	p değeri
5.825	8	0.670

Tablo 5'te yer alan Cox-Snell R^2 ve Nagelkerke R^2 değerleri genel olarak küçük çıkma eğilimdedir. Elde edilen sonuçlara göre, Cox-Snell R^2 (0.153) ve Nagelkerke R^2 (0.204) istatistiklerinin 0.20'ye yakın ve üzerinde olması nedeniyle elde edilen modelin uyumlu olduğu söylenebilir.

Modelin veriye uyumunun iyi olup olmadığını incelemenin bir diğer yolu da Ki-kare uyum iyiliği testi olarak da bilinen Hosmer-Lemeshow testidir. Tablo 5'te χ^2 istatistiğine ilişkin p=0,670>0.05 olduğu için %5 anlamlılık düzeyinde için modelin veriye uyumunun yeterli olduğuna karar verilir.

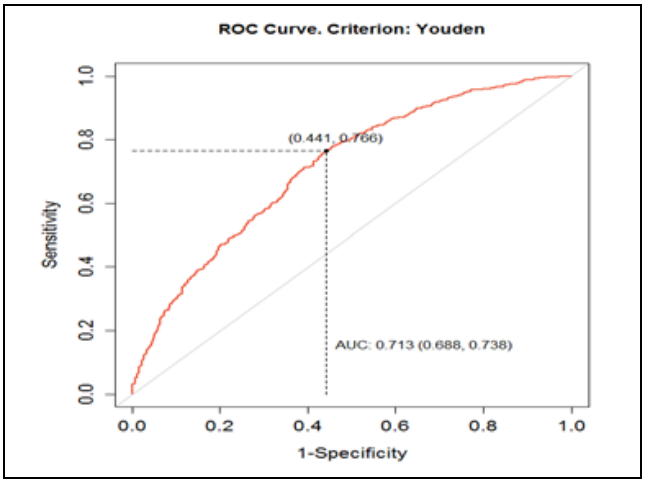
Hata oranını düşürüp modelin sınıflama performansını arttırmak için tüm bağımsız değişkenleri kullanmak yerine yalnızca endeks hareket yönü üzerinde etkisi anlamlı olan değişkenler (Lag1, Bono, Brent, Bovespa, Belarus, DowJones) ile lojistik regresyon analizine devam edilecektir.

Sınıflandırma yapılırken her bir gözlem için lojistik regresyon analizi ile kestirilen $\hat{\pi}(x_i) = P(Y = 1|x_i)$ olasılıklar, uygun bir kesim noktası ile kıyaslanmalıdır. c^* uygun kesim noktası olmak üzere $\hat{\pi}(x_i) > c^*$ ise bu gözleme karşılık gelen y_i değeri 1; aksi halde 0 olarak sınıflandırılır.

1615 adet gözlemden oluşan eğitim veri seti kullanılarak tüm mümkün kesim noktalarına karşılık gelen duyarlılık ve (1-belirleyicilik) değerlerini gösteren ROC eğrisi çizdirilmiştir.

ROC eğrisi üzerinde uygun kesim noktasını seçmek için Youden indeksi, Minimax kriteri ve duyarlılığın belirleyiciliğe eşit olduğu nokta esas alınmış; aynı zamanda literatürde varsayılan değer olarak alınan 0.5 kesim noktası da kıyaslama yapmak için analize dahil edilmiştir.

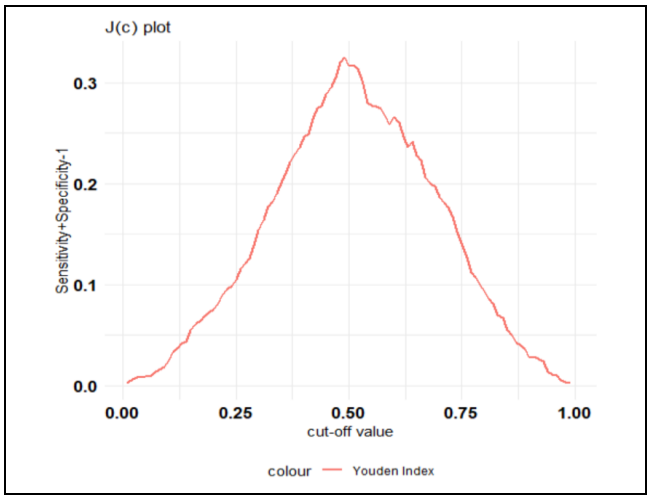
Youden indeksine göre ROC eğrisi üzerinde belirlenen uygun kesim noktası Şekil 3'te verilmiştir.



Şekil 3. Youden indeksine göre kesim noktası.

Figure 3. Cut-point according to Youden index.

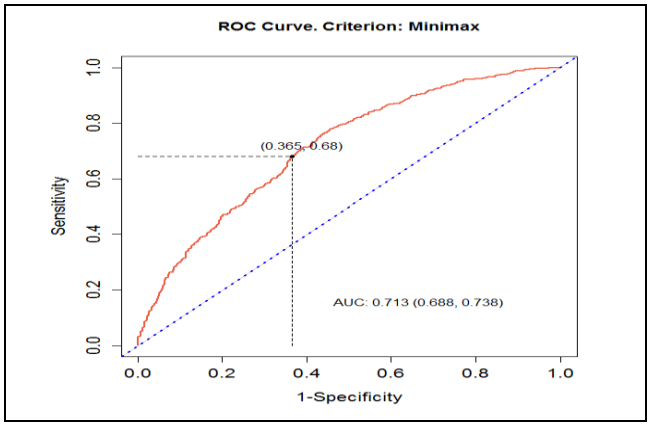
Şekil 3'e göre $\forall c \in [0,1]$ olmak üzere, $J(c) = [duyarlılık(c) + belirleyicilik(c) - 1]$ fonksiyonunu maksimum yapan nokta, yani uygun kesim noktası $c^*_{Youden} = 0.488$ bulunmuştur. Bu nokta, aynı zamanda duyarlılık ve belirleyicilik toplamını maksimum yapar, yanlış sınıflandırma oranını ise minimize eder. Bulunan noktaya karşılık gelen duyarlılık değeri 0.766, belirleyicilik değeri ise 0.559'dur. Şekil 4'te $J(c)$ fonksiyonunun, $\forall c \in [0,1]$ için çizdirilen grafiği sunulmaktadır.



Şekil 4. $J(c)$ fonksiyonunun kesim noktalarına göre grafiği.

Figure 4. Plot of $J(c)$function according to cut-points.

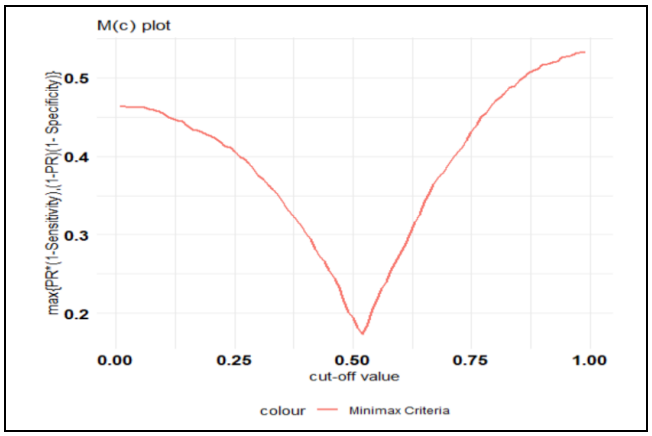
Şekil 4'e göre Youden indeksini maksimum yapan nokta, grafikten de gözükmektedir. Bir diğer kriter Minimax olup, ROC eğrisi Şekil 5'te verilmiştir.



Şekil 5. Minimax kriterine göre kesim noktası.

Figure 5. Cut-point according to Minimax criterion.

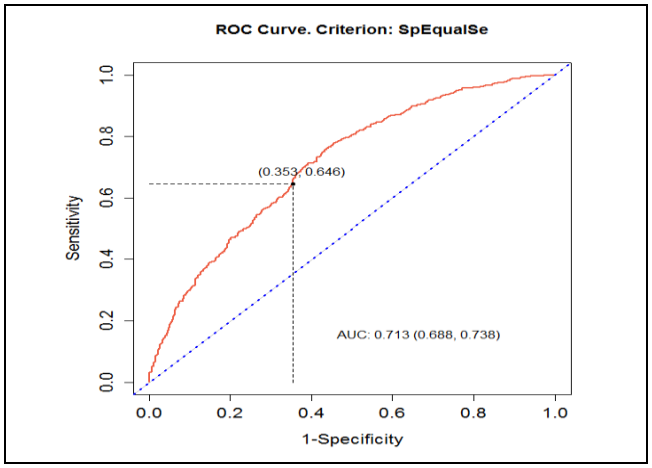
Minimax kriteri, duyarlılık ve belirleyiciliğin yanı sıra yaygınlık oranından etkilenmektedir. Eğitim veri setindeki 1615 adet gözlem değerinin 803 tanesi aşağı, 862 tanesi yukarı yönlü olup pr oranı 0.534'tür. Şekil 5'e göre $\forall c \in [0,1]$ olmak üzere Denklem (6)'da verilen $M(c)$ fonksiyonunu minimum yapan nokta, $c^*_{Minimax} = 0.523$; bu noktaya karşılık gelen duyarlılık değeri 0.679 ve belirleyicilik değeri 0.635 bulunmuştur. Bu nokta, görsel olarak ta Şekil 6'da verilen $M(c)$ fonksiyonunun tüm mümkün kesim noktalarına karşılık gelen grafiğinde görülmektedir. $M(c)$ fonksiyonu, yanlış pozitif oranı ve yanlış negatif oranı fonksiyonlarının maksimumu ile oluşturulmuştur, bu nedenle $c^*_{Minimax}$ noktası en sık gözlemlenen hatayı minimum yapan noktadır.



Şekil 6. $M(c)$ fonksiyonunun kesim noktalarına göre grafiği.

Figure 6. Plot of $M(c)$function according to cut-points.

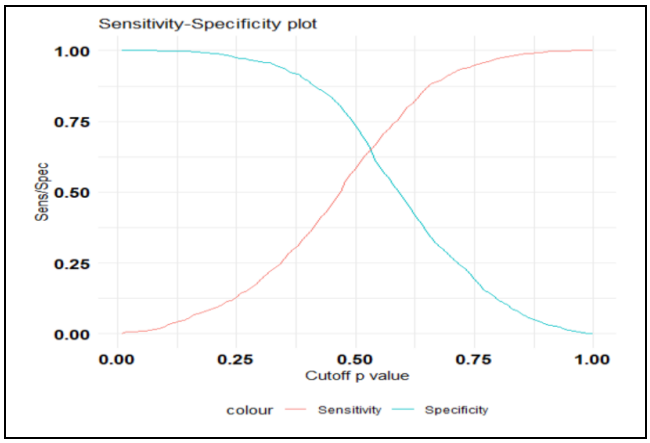
Kesim noktası değerleri arttıkça, duyarlılık artarken belirleyicilik azalır. Bu yüzden, duyarlılık ve belirleyicilik değerlerinin kesiştiği nokta, uygun kesim noktası olarak alınabilir. Matematiksel olarak bu nokta, ROC uzayında (1,0) ve (0,1) noktalarını birleştiren doğru ile ROC eğrisinin kesiştiği noktadır. Şekil 7'de ROC eğrisi üzerinde duyarlılığın belirleyiciliğe eşit olduğu kesim noktası verilmiştir.



Şekil 7. Duyarlılığın belirleyiciliğe eşit olduğu kesim noktası.

Figure 7. Cut-point where sensitivity equals specificity.

Şekil 7'ye göre uygun kesim noktası $c^*_{SE=SPC} = 0.532$; bu noktaya karşılık duyarlılık ile belirleyicilik değeri 0.646 elde edilmiştir. Şekil 8'de duyarlılık ve belirleyiciliğin kesiştiği değere karşılık gelen kesim noktası görülmektedir.



Şekil 8. Duyarlılık ve belirleyicilik fonksiyonlarının kesim noktalarına göre grafiği.

Figure 8. Plot of sensitivity and specificity functions according to cut-points.

Farklı yöntemler ile elde edilmiş uygun kesim noktalarının test veri seti üzerinde performansları karşılaştırılmış ve Tablo 6'da sunulmuştur. İdeal kesim noktasının duyarlılık, belirleyicilik ve doğruluk değerlerini en yüksek yapması beklenir. Tablo 6'ya göre Youden indeksi ile elde edilen kesim noktası en yüksek duyarlılığı (0.907) verirken, belirleyicilik değeri (0.667) en düşüktür. Benzer sonuç, varsayılan 0.5 kesim noktası için de geçerlidir. Minimax kriteri ile elde edilen ve duyarlılığın belirleyiciliğe eşit olduğu değere karşılık gelen kesim noktaları, aynı belirleyicilik değerine (0.738) sahiptir. Buna karşılık, Minimax kriteri ile elde kesim noktasının duyarlılık (0.884) ve doğruluk (0.812) değerleri daha yüksek olduğu için en uygun kesim noktası değeri $c^*_{Minimax} = 0.523$ bulunmuştur.

Lojistik regresyon analizi ile sınıflama yapılırken kestirilen $\hat{\pi}(x_i)$ olasılıkları, $c^*_{Minimax} = 0.523$ ile kıyaslanmış ve Tablo 7'de test seti hata matrisi verilmiştir. Tablo 7'den görüleceği üzere BIST100 endeks hareket yönü %81.176 oranında doğru olarak tahmin edilmiştir. Test hata oranı ise %18.824'tür.

Tablo 6. Farklı yöntemler ile bulunan uygun kesim noktalarının karşılaştırılması.

Table 6. Comparison of feasible cut-points found by different methods.

		Gözlemlenen							
		Youden indeksi $c^*_{Youden} = 0.488$		Minimax $c^*_{Minimax} = 0.523$		Duyarlılık= Belirleyicilik $c^*_{SE=SPC} = 0.532$		Varsayılan $c^* = 0.5$	
		Endeks Hareket Yönü		Endeks Hareket Yönü		Endeks Hareket Yönü		Endeks Hareket Yönü	
Tahmin Edilen		Yukarı	Aşağı	Yukarı	Aşağı	Yukarı	Aşağı	Yukarı	Aşağı
Endeks Hareket Yönü	Yukarı	39	14	38	11	36	11	39	12
	Aşağı	4	28	5	31	7	31	4	30
Doğruluk		0.788		0.812		0.788		0.812	
Duyarlılık		0.907		0.884		0.837		0.907	
Belirleyicilik		0.667		0.738		0.738		0.714	

Lineer sınıflama yöntemlerinden biri olan çoklu lojistik regresyon analizi ile elde edilen sonuçları kıyaslayabilmek için lineer olmayan sınıflama teknikleri içerisinden K-en yakın komşu algoritması aynı veri seti üzerinden denenecektir.

4.4 K-nn algoritması bulguları

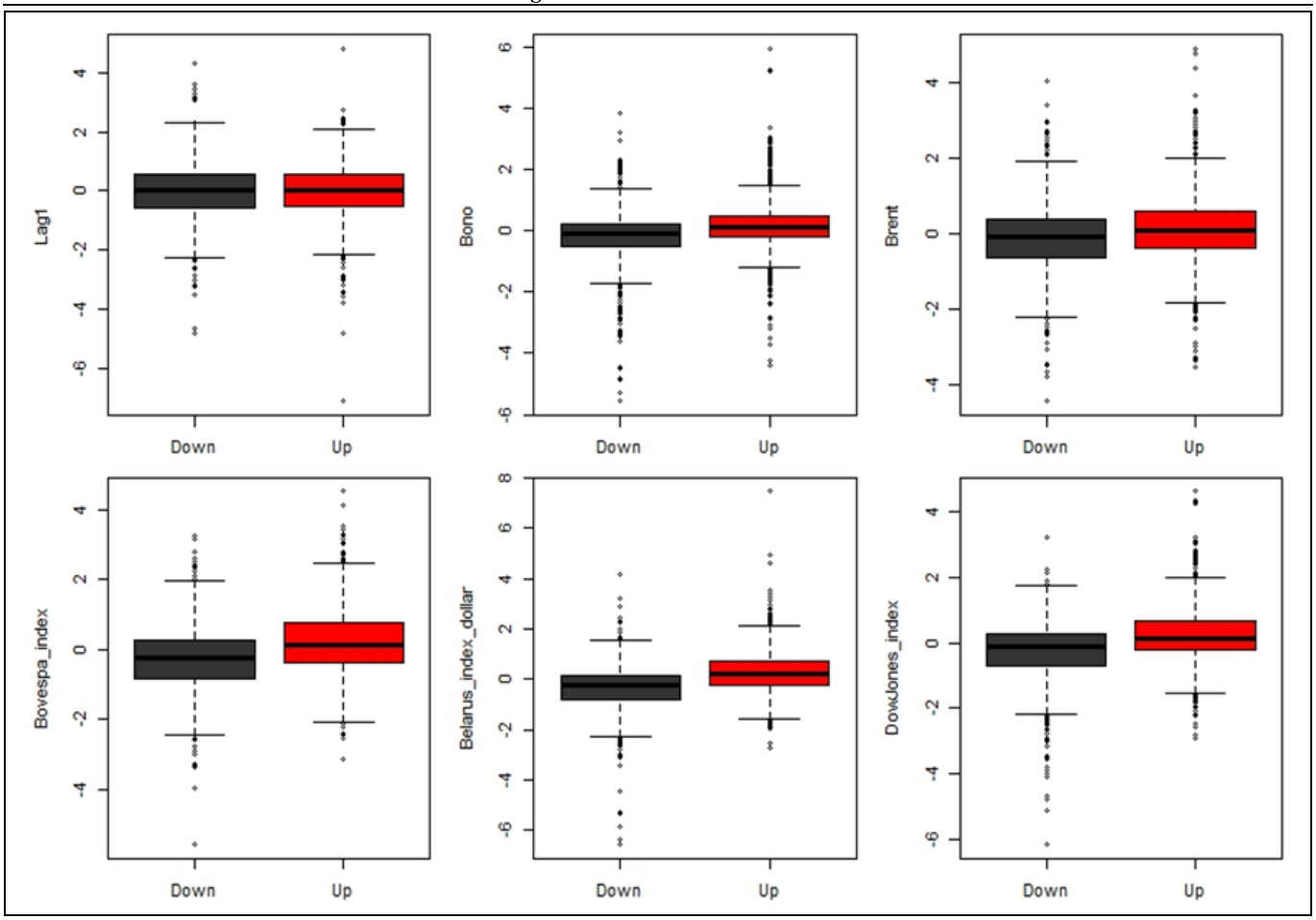
K-nn sınıflama algoritması ile yapılacak analize, bir önceki bölümde BIST100 endeks hareket yönü üzerinde anlamlı etkisi olduğu tespit edilen değişkenler (Lag1, Bono, Brent, Bovespa, Belarus, DowJones) ile devam edilmiştir. Tüm bağımsız değişkenler sürekli olduğu için Euclidean uzaklık ölçütü kullanılmıştır. Analize başlamadan önce farklı birimlere sahip değişkenler olması nedeni ile tüm veriler standardize edilmiştir. Bağımlı değişkene göre, analize dahil edilen değişkenlerin kutu grafikleri Şekil 9'da verilmektedir.

Şekil 9, endeks hareket yönüne göre aşağı ve yukarı olarak gruplandırılmış 6 bağımsız değişkenin kutu diyagramları olup, özelliklerin hareket yönü açısından nasıl değiştiğini göstermektedir. K-nn algoritmasının sınıflama performansını etkileyen önemli nokta K komşu sayısının değeridir. Araştırmada, optimal K değerinin seçimi için birini dışarıda bırak çapraz doğrulama yöntemi uygulanmıştır. K=1,2,...,15 değerleri ayrı ayrı denenen çapraz doğrulama sonucunda elde edilen doğru sınıflama yüzdeleri küçükten büyüğe sıralanarak en iyi sınıflama performansı veren K parametresi belirlenmiştir.

Tablo 7. Lojistik regresyon için test seti hata matrisi.

Table 7. Test set confusion matrix for logistic regression.

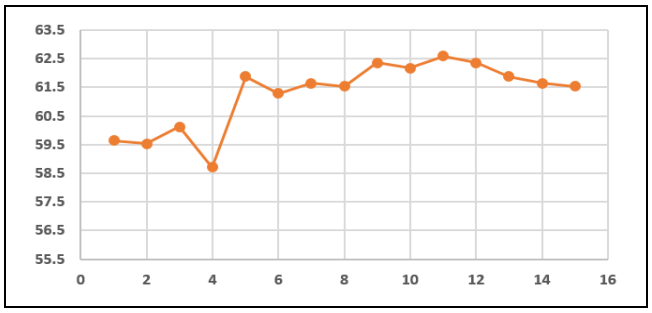
		Gözlemlenen		
		Endeks Hareket Yönü		Doğruluk %
Tahmin Edilen		Yukarı	Aşağı	
Endeks Hareket Yönü	Yukarı	38	11	77.551
	Aşağı	5	31	86.111
Genel Doğruluk %				81.176



Şekil 9. Endeks hareket yönüne göre bağımsız değişkenlerin kutu grafikleri.

Figure 9. Box plots of independent variables according to the index movement direction.

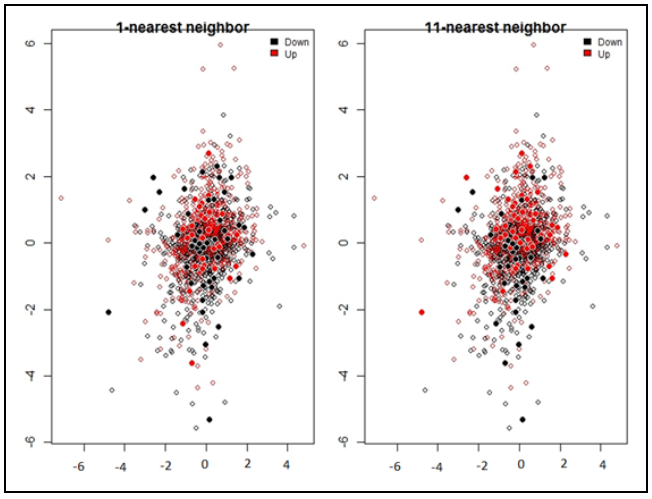
Farklı K değerlerine için her bir çapraz doğrulamaya ait doğru sınıflandırma yüzdelerinin grafiği Şekil 10'da verilmiştir.



Şekil 10. K değerlerine karşılık gelen doğru sınıflandırma oranları.

Figure 10. Correct classification rates corresponding to K values.

Şekil 10 incelendiğinde, farklı K değerleri arasında K=11 değeri %62.6 doğru sınıflama yüzdesi ile optimal komşu sayısı olarak belirlenmiştir. Optimal komşu sayısı ile yapılan K-nn analizi sonuçlarını kıyaslama bakımından K=1 komşu seçildiğinde oluşturulan test seti hata matrisi Tablo 8'de verilmiştir. Tablo 8'e göre komşu sayısı 1 alındığında K-nn algoritması test veri setindeki 85 gözlemin %57.647'sini doğru şekilde tahmin etmiştir. Model performansını arttırmak amacı ile optimal K değeri için analiz yinelenmiştir ve Tablo 9'da sonuçlar verilmiştir. Tablo 9'da verilen hata matrisine göre test veri setinde endeks hareket yönünü doğru sınıflama oranı K=11 için %70.588 bulunmuş ve birden fazla komşuluğa bakılması model performansını arttırmıştır. Komşu sayısı K=1 ve K=11 için elde edilen en yakın komşu grafikleri Şekil 11'de verilmiştir.



Şekil 11. K=1 ve K=11 için en yakın komşu grafikleri.

Figure 11. Nearest neighbor plots for K = 1 and K = 11.

Tablo 8. K-nn algoritması (K =1) için test seti hata matrisi.

Table 8. Test set confusion matrix for K-nn algorithm (K=1).

		Gözlemlenen		
		Endeks Hareket Yönü		Doğruluk %
Tahmin Edilen		Yukarı	Aşağı	
Endeks Hareket Yönü	Yukarı	22	15	59.459
	Aşağı	21	27	56.250
Genel Doğruluk %				57.647

Tablo 9. K-nn algoritması (K =11) için test seti hata matrisi.

Table 9. Test set confusion matrix for K-nn algorithm (K=11).

		Gözlemlenen		
		Endeks Hareket Yönü		Doğruluk %
Tahmin Edilen		Yukarı	Aşağı	
Endeks Hareket Yönü	Yukarı	34	16	68.000
	Aşağı	9	26	74.286
Genel Doğruluk %				70.588

Sonuç olarak, lineer olmayan sınıflayıcılardan biri olan K-nn algoritması ile n=85 adet gözlemden oluşan test veri setinde BIST100 endeks hareket yönünü %70.588 doğru sınıflama yüzdesi ile tahmin etmiştir.

4.5 Model performanslarının karşılaştırılması

Çoklu lojistik regresyon ve K-nn algoritması ile elde edilen modellerin BIST100 endeks hareket yönünü tahmin etmedeki performanslarını kıyaslamak için test veri setindeki hata matrisleri ve hata matrisinden elde edilen model performans değerlendirme ölçütleri Tablo 10'da sunulmuştur.

Tablo 10. Çoklu lojistik regresyon ve K-nn algoritması model performans değerlendirmesi.

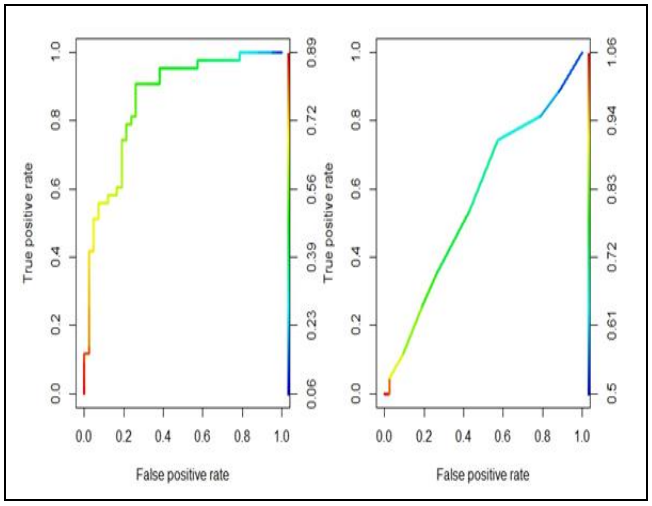
Table 10. Model performance evaluation of multiple logistic regression and K-nn algorithm.

		Gözlemlenen			
		Çoklu Lojistik Regresyon		K-En Yakın Komşu	
		Endeks Hareket Yönü		Endeks Hareket Yönü	
Tahmin Edilen		Yukarı	Aşağı	Yukarı	Aşağı
Endeks Hareket Yönü	Yukarı	38	11	34	16
	Aşağı	5	31	9	26
Doğruluk		0.812		0.706	
Hata Oranı		0.188		0.294	
Duyarlılık		0.884		0.791	
Belirleyicilik		0.738		0.619	
Yanlış Pozitif Oranı		0.262		0.381	
Yanlış Negatif Oranı		0.116		0.209	
Pozitif Öngörü Değeri		0.776		0.680	
Negatif Öngörü Değeri		0.861		0.743	
F-Ölçüsü		0.826		0.731	
Pozitif Olabilirlik Oranı		3.374		2.076	
Negatif Olabilirlik Oranı		0.157		0.338	
Tanısal Üstünlük Oranı		21.490		6.142	

Elde edilen ölçütlerden doğruluk, duyarlılık, belirleyicilik, pozitif öngörü değeri, negatif öngörü değeri, F-ölçüsü, pozitif olabilirlik oranı, tanısal üstünlük oranı değerlerinin yüksek; hata oranı, yanlış pozitif oranı, yanlış negatif oranı, negatif olabilirlik oranı değerlerinin ise düşük çıkması beklenir ve bu durumda modelin uyumunun iyi olduğu söylenir. Tablo 10'a göre çoklu lojistik regresyon tüm ölçütlere göre K-nn algoritmasına kıyasla daha iyi sınıflama performansı sergilemiştir. BIST100 endeks hareket yönünü verilen zaman diliminde tahmin etmede lineer sınıflayıcı tekniklerinden biri

olan lojistik regresyon modeli, lineer olmayan sınıflayıcı alternatifi olan K-nn algoritmasına göre oldukça başarılıdır.

Hata matrisinden hareketle modelin veriye uyumu incelenirken seçilen kesim noktasına bağlı olacak şekilde doğru sınıflama oranları değişebilir. Bu nedenle istenen kesim noktasına karşılık yatay eksende yanlış pozitif oranı, dikey eksende ise eksende duyarlılık olacak biçimde her iki model için ROC eğrisi çizdirilmiş ve Şekil 12'de verilmiştir.



Şekil 12. Lojistik regresyon ve K-nn algoritması için ROC eğrisi.

Figure 12. ROC curve for logistic regression and K-nn algorithm.

Şekil 12 incelendiğinde sol tarafta çoklu lojistik regresyon, sağ tarafta ise K-nn algoritması için ROC eğrileri çizdirilmiştir. ROC eğrisi altında kalan alan AUC değeri olup bu değer, pozitif ve negatif gözlemleri doğru sınıflama oranını göstermektedir. Şekil12'ye göre lojistik regresyon analizine ait ROC eğrisinin (0,1) sol üst köşe noktasına yakın olması bu yöntem ile yanlış pozitif oranının düştüğünü gösterirken K-nn algoritmasına ait ROC eğrisinin sınıflandırmanın rastgele yapıldığını gösteren köşegen çizgisine yakın olması bu yöntem ile yanlış sınıflandırılan örnek sayısının arttığını gösterir. ROC eğrisi altında kalan alan yani AUC değeri çoklu lojistik regresyon analizi ve K-nn algoritması için sırası ile 0.86 (p=0.000) ve 0.57 (p=0.009) olarak hesaplanmış ve istatistiksel olarak anlamlı bulunmuştur. AUC değerinin 1'e yakın olması yöntemin sınıflama gücünün yüksek; AUC değerinin 0.5'e yakın olması ise sınıflama gücünün zayıf olduğunu gösterir. Buna göre, çoklu lojistik regresyon analizi ile elde edilen modelin sınıflama performansının K-nn algoritmasına göre yüksektir.

5 Sonuçlar

Finansal veri madenciliğinde hisse senedi endeks hareket yönünü tahmin etmek zor bir problemdir. Bu çalışma, BIST100 endeks hareket yönünü sınıflandırmaya dayalı teknikler ile tahmin etmeyi amaçlamıştır. Bu amaç doğrultusunda lineer sınıflandırıcı olarak çoklu lojistik regresyon, lineer olmayan sınıflandırıcı olarak K-en yakın komşu algoritması tercih edilmiştir. Araştırmanın kategorik bağımlı değişkenini oluşturulurken, BIST100 endeksinin bir önceki işlem günü değerine bakılmış ve değer pozitif ise endeks hareket yönü yukarı temsilen 1, aksi durumda ise 0 olarak kodlanmıştır. Endeks hareketinin yukarı veya aşağı yön tahmini için toplam 19 adet nicel bağımsız değişken kullanılmıştır. Analiz öncesi; verilerde çoklu doğrusal bağlantı problemi, tekrar eden, eksik ve aşırı değerler kontrol edilmiştir. Toplam 1700 işlem gününden oluşan veri seti hold-out yöntemi ile eğitim (%95, n=1615) ve test (%5, n=85) olmak üzere ikiye parçalanmıştır.

Çoklu lojistik regresyon analizi sonuçlarına göre; BIST100 endeksi günlük kapanışının bir gün önceki getirisini gösteren Lag1 ile Bono, Brent petrol, Bovespa endeksi, Belarus endeksi ve Dow Jones endeksi günlük kapanış getirisi değişkenleri BIST100 endeks hareket yönünü tahmin etmede istatistiksel olarak anlamlı bulunmuştur. Uyum iyiliği testi sonuçlarına göre kurulan modelin veriye uyumu yeterlidir.

Çoklu lojistik regresyon analizi ile kestirilen olasılıklar, uygun kesim noktası ile karşılaştırılmış ve verilen gözlem değerlerine karşılık gelen endeks hareket yönü 1 (yukarı) veya 0 (aşağı) olarak sınıflandırılmıştır. Uygun kesim noktasını belirlemek için ROC eğrileri çizdirilmiştir. Duyarlılık ve belirleyiciliği dengeleyecek şekilde ROC eğrisi üzerinde uygun kesim noktasını seçmek için Youden indeksi, Minimax kriteri ve duyarlılık ile belirleyiciliğin eşit olduğu nokta esas alınmıştır. Farklı yöntemlere göre elde edilen uygun kesim noktalarının test veri seti üzerindeki performansları karşılaştırılmış ve sonuç olarak Minimax kriteri ile elde edilen 0.523 değeri en uygun kesim noktası seçilmiştir. En uygun kesim noktası göz önüne alınarak, anlamlı değişkenler ile çoklu lojistik regresyon modeli kurulmuş ve test kümesinde oluşturulan hata matrisine ait doğru sınıflama performansı %81.176 bulunmuştur.

Çoklu lojistik regresyon analizinin performansını kıyaslamak üzere, anlamlı değişkenler ile aynı veri seti üzerinde K-en yakın komşu algoritması uygulanmıştır. K-nn algoritmasındaki optimal *K* değeri birini dışarıda bırak çapraz doğrulama ile belirlenmiş ve çapraz doğrulama sonucuna göre %62.6 doğru sınıflama oranı ile K değeri 11 bulunmuştur. Optimal *K* değeri ile oluşturulan K-nn modelinin test veri setindeki performansına göre, endeks hareket yönü %70.588 ile doğru tahmin edilmiştir.

Bu çalışma, BIST100 endeks hareket yönünün sınıflamaya dayalı denetimli makine öğrenmesi algoritmaları aracılığı ile başarılı bir şekilde tahmin edilebileceğini göstermektedir. Analizlere göre, lineer sınıflayıcı olan lojistik regresyon analizi, lineer olmayan K-nn sınıflayıcısına göre daha iyi sınıflama performansına sahiptir. Çalışmada elde edilen bulgular ile modelin endeks hareket yönünü yukarı yani artan bir pazar olarak tahmin ettiği günlerde yatırımcılar için satın alma stratejisinin geliştirilebileceği söylenebilir.

Kullanılan sınıflandırıcıların performansının seçilen bağımsız değişkenlere bağlı olduğu kadar algoritmaya özgü parametrelerin belirlenmesinde seçilen yönteme de bağlı olması, çalışmanın kısıtlamaları olarak gösterilebilir. Gelecekte, BIST100 endeks hareket yönünü tahmin etmek için farklı veri setleri ile farklı öğrenme teknikleri denenecektir.

6 Conclusions

In financial data mining, stock index movement direction prediction is a challenging problem. This study aimed to predict movement direction of XU100 index with supervised machine learning algorithms based on classification. For this purpose, multiple logistic regression as a linear classifier and K-nearest neighbor algorithm as a nonlinear classifier were preferred. If XU100 index value on the previous trading day was positive, then the dependent variable of the study was coded as 1 representing the direction of movement upwards; otherwise as 0. A total of 19 quantitative independent variables were used for upward or downward prediction of the index movement direction. Before data analysis; multicollinearity problem,

repetitive, missing and extreme values in the data were checked. The dataset consisting of 1700 trading days in total, was divided into two subsets as training (95%, n = 1615) and test (5%, n = 85) by hold-out method.

According to the results of the multiple logistic regression analysis; Lag1, which shows the closing return of the BIST100 index on the previous trading day and the daily closing return of Bond, Brent oil, Bovespa index, Belarus index and Dow Jones index variables were found to be statistically significant in predicting the movement direction of XU100 index. According to the goodness of fit test results, the model fit the data well.

The probabilities estimated by multiple logistic regression analysis were compared with the feasible cut-off point and the index movement direction corresponding to the given observation values was classified as 1 (up) or 0 (down). The feasible cut-off point has been determined with the help of ROC curves. To select the feasible cut-off point on the ROC curve that balance sensitivity and specificity, the Youden index, the Minimax criterion, and the point where sensitivity and specificity are equal were taken as basis. The performances of the feasible cut-off points obtained according to different methods on the test data set were compared and as a result, the 0.523 value obtained with the Minimax criterion was chosen as the optimal cut-off point. Considering the optimal cut-off point, a multiple logistic regression model was fitted with significant variables and the correct classification rate of the test confusion matrix was found to be 81.176%.

To compare the performance of multiple logistic regression analysis, the K-nearest neighbor algorithm was applied on the same data set with significant variables. The optimal *K* value in the K-nn algorithm was determined by leave-one-out cross validation and as a result of cross validation, *K* value was found to be 11 with 62.6% correct classification rate. According to the performance of the K-nn model fitted with the optimal *K* value in the test data set, the index movement direction was correctly predicted with 70.588%.

This study shows that XU100 index movement direction can be successfully predicted by using supervised machine learning algorithms based on classification. According to the analysis, the performance of logistic regression, which is a linear classifier, is better than the K-nn algorithm, which is a nonlinear classifier. With the findings obtained in the study, it can be said that the buying strategy for investors can be developed in the days when the model predicts the direction of the index movement as an upward market trend.

The fact that the performance of the classifiers used depends not only on the selected independent variables but also on the method chosen in determining the algorithm-specific parameters, can be shown as the limitations of the study. In the future, different learning techniques will be tried with different datasets to predict the BIST100 index movement direction.

7 Yazar katkı beyanı

Gerçekleştirilen çalışmada Gülder KEMALBAY, fikrin oluşması, tasarımın yapılması, metodoloji yazımı, veri analizinin yapılması, sonuçların incelenmesi, elde edilen sonuçların değerlendirilmesi, yazım denetimi ve içerik açısından makalenin kontrol edilmesi başlıklarında; *Begüm Nur ALKIŞ*, literatür taraması, metodoloji yazımı, veri toplanması, veri analizinin yapılması başlıklarında katkı sunmuşlardır.

8 Etik kurul onayı ve çıkar çatışması beyanı

Hazırlanan makalede etik kurul izni alınmasına gerek yoktur. Hazırlanan makalede herhangi bir kişi/kurum ile çıkar çatışması bulunmamaktadır.

9 Kaynaklar

[1] Ballings M, Van den Poel D, Hespeels N, Gryp, R. "Evaluating multiple classifiers for stock price direction prediction". *Expert Systems with Applications*, 42(20), 7046-7056, 2015.

[2] Diler Aİ. "İMKB Ulusal-100 endeksinin yönünün yapay sinir ağları hata geriye yayma yöntemi ile tahmin edilmesi". *İMKB Dergisi*, 7(25-26), 65-81, 2003.

[3] Huang W, Nakamori Y, Wang SY. "Forecasting stock market movement direction with support vector machine". *Computers & Operations Research*, 32(10), 2513-2522, 2005.

[4] Kutlu B, Badur B. "Yapay sinir ağları ile borsa endeksi tahmini". *Yönetim*, 63, 25-40, 2009.

[5] Kara Y, Boyacıoglu MA, Baykan ÖK. "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, 2011.

[6] Özdemir K, Tolun S, Demirci E. "Endeks getirisi yönünün ikili sınıflandırma yöntemiyle tahmin edilmesi: İMKB 100 endeksi örneği". *Niğde Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 4(2), 45-59, 2011.

[7] Dutta A, Bandopadhyay G, Sengupta S. "Prediction of stock performance in Indian stock market using logistic regression". *International Journal of Business and Information*, 7(1), 105-136, 2012.

[8] Subha MV, Nambi ST. "Classification of stock index movement using k-nearest neighbours (k-nn) algorithm". *WSEAS Transactions on Information Science and Applications*, 9(9), 261-270, 2012.

[9] Tayyar N, Tekin S. "İMKB-100 endeksinin destek vektör makineleri ile günlük, haftalık ve aylık veriler kullanarak tahmin edilmesi". *Abant İzzet Baysal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 13(1), 189-217, 2013.

[10] Türkmen AC, Cemgil AT. "An application of deep learning for trade signal prediction in financial markets". *IEEE 2015 23nd Signal Processing and Communications Applications Conference*, Malatya, Turkey, 16-19 May 2015.

[11] Patel J, Shah S, Thakkar P, Kotecha, K. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques". *Expert Systems with Applications*, 42(1), 259-268, 2015.

[12] Gündüz H, Çataltepe Z, Yaslan Y. "Stock market direction prediction using deep neural networks". *IEEE 2017 25th Signal Processing and Communications Applications Conference*, Antalya, Turkey, 15-18 May 2017.

[13] Yakut E, Gemici E. "Predicting stock return classification through LR, C5.0, CART and SVM methods, and comparing the methods used: an application at BIST in Turkey". *Ege Academic Review*, 17(4), 461-479, 2017.

[14] Kara, İ, Ecer F. "BİST endeks hareket yönünün tahmininde sınıflandırma yöntemlerinin performanslarının karşılaştırılması". *The Journal of Academic Social Science*, 6(83), 514-524, 2018.

[15] Oğuz RF, Uygun Y, Aktaş MS, Aykurt İ. "On the use of technical analysis indicators for stock market price movement direction prediction". IEEE *27th Signal Processing and Communications Applications Conference,* Sivas, Turkey, 24-26 April 2019.
[16] Livieris IE, Kotsilieris T, Stavroyiannis S, Pintelas P. "Forecasting stock price index movement using a constrained deep neural network training algorithm". *Intelligent Decision Technologies*, 14(3), 1-14, 2019.
[17] Bontempi G, Taieb SB, Le Borgne YA. *Machine Learning Strategies for Time Series Forecasting.* 1st ed. Berlin, Germany, Springer, 2012.
[18] Fulcher BD, Jones NS. "Highly comparative feature-based time-series classification". *IEEE Transactions on Knowledge and Data Engineering*, 26(12), 3026-3037, 2014.
[19] Jiao Y, Jakubowicz J. "Predicting stock movement direction with machine learning: an extensive study on S&P 500 stocks". *IEEE 2017 International Conference on Big Data,* Boston, USA, 11-14 December 2017.
[20] Balaban ME, Kartal E. *Veri Madenciliği ve Makine Öğrenmesi*. 2. baskı. İstanbul, Türkiye, Çağlayan, 2018.
[21] Alpar R. *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*. 5. Baskı. Ankara, Türkiye, Detay, 2017.
[22] Lantz B. *Machine Learning with R: Expert Techniques for Predictive Modeling.* 3rd ed. Birmingham, UK, Packt Publishing Ltd. 2019.
[23] Hosmer JR, David W, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. New Jersey, USA, John Wiley & Sons, 2013.
[24] Tabachnick BG, Fidell LS. *Çok Değişkenli İstatistiklerin Kullanımı*. 6. baskı. Ankara, Türkiye, Nobel, 2015.
[25] Mitchell, TM. *Machine Learning*. 1st ed. New York, USA, McGraw-Hill, 1997.
[26] Han J, Pei J, Kamber M. *Data Mining: Concepts and Techniques*. 3rd ed. Waltham, USA, Elsevier, 2011.
[27] Steinbach M, Tan PN. *kNN: k-Nearest Neighbors*. Editors: Wu X, Kumar V. The Top Ten Algorithms in Data Mining, 151-162, Boca Raton, FL, USA, CRC Press, 2009.
[28] Zhu W, Zeng N, Wang N. "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations". *NESUG Proceedings: Health Care and Life Sciences*, Baltimore, Maryland, 14-17 November 2010.
[29] Fawcett T. "An introduction to ROC analysis". *Pattern Recognition Letters*, 27(8), 861-874, 2006.
[30] Unal, I. "Defining an optimal cut-point value in ROC analysis: an alternative approach". *Computational and Mathematical Methods in Medicine*, 2017, 1-14, 2017.
[31] Kelly MJ, Dunstan FD, Lloyd K, Fone DL. "Evaluating cutpoints for the MHI-5 and MCS using the GHQ-12: a comparison of five different methods". *BMC Psychiatry*, 8(10), 1-9, 2008.
[32] Habibzadeh F, Habibzadeh P, Yadollahie M. "On determining the most appropriate test cut-off value: the case of tests with continuous results". *Biochemia Medica*, 26(3), 297-307, 2016.
[33] Finnet Elektronik Yayıncılık Data İletişim San. Tic. Ltd. Şti. "Finnet Kurumsal Web Sitesi". https://www.finnet.com.tr (26.10.2017).