Kaynar, T., Yiğit, Ö. E. / Journal of Yasar University, 2021, 16/64, 1741-1762

Öznitelik Mühendisliği ile Makine Öğrenmesi Yöntemleri Kullanılarak BIST 100 Endeksi Değişiminin Tahminine Yönelik Bir Yaklaşım

An Approach for Estimating BIST 100 Index Change Using Machine Learning Methods with Feature Engineering

Tamara KAYNAR, Milli Savunma Üniversitesi, Türkiye, tamarakaynar@gmail.com

Orcid No: 0000-0002-3792-0295

Öyküm Esra YİĞİT, Yıldız Teknik Üniversitesi, Türkiye, oeyigit@yildiz.edu.tr

Orcid No: 0000-0001-7805-3979

Öz: Finansal piyasaların ana çıktısı bir zaman serisi problemidir ve zaman serileri doğaları gereği gürültülü, durağan olmayan ve karmaşık bir yapı sergilemektedirler. Bu karmaşık yapı sebebiyle zaman serilerinin gelecekteki davranışlarını öngörme süreci araştırmacılar açısından hayli zorlu bir çalışma alanı olmaktadır. Bu çalışmada BIST 100 endeksi günlük getiri yönünün tahmin edilmesinde kapsamlı bir öznitelik mühendisliği işlemi uygulanmış ve farklı makine öğrenmesi algoritmaları kullanılarak modellemeler gerçekleştirilmiştir. Modellere girdi olarak alınacak öznitelikler, serinin özetleyici istatistiklerine, örnekleme dağılımının ek karakteristiklerine ve serinin lineer olmayan/karmaşık yapısını yansıtan gözlenen dinamiklerine bağlı olarak çıkartılmış ve dışsal değişken kullanmadan da sınıflandırma performanslarının oldukça yüksek olduğu gösterilmiştir. Ayrıca farklı eğitim-test oranları kullanılarak tahminlerin dayanıklılığı araştırılmıştır.

Anahtar Sözcükler: Öznitelik Mühendisliği, Makine Öğrenmesi, Finansal Zaman Serileri, Sınıflandırma

JEL Sınıflandırması: C51, C52, F31, G17

Abstract: The main output of financial markets is a time series problem and a time series exhibit noisy, non-linear and chaotic structure by nature. Due to this complex structure, the process of predicting the future behavior of time series is a very challenging field for researchers. In this study, a comprehensive feature engineering process was applied to estimate the daily return direction of the BIST 100 index and models were carried out using different machine learning algorithms. The features to be taken as input to the models were extracted depending on the summative statistics of the series, the additional characteristics of the sampling distribution, and the observed dynamics reflecting the non-linear/complex structure of the series and it was shown that the classification performances are quite high without using exogenous variables. In addition the durability of the predictions performances was investigated using different training-test ratios.

Keywords: Feature Engineering, Machine Learning, Financial Time Series, Classification

JEL Classification: C51, C52, F31, G17

1. Giriş

Finansal piyasalar ve bu piyasaların etkinliği, sahip olduğu yüksek getiri potansiyeli göz önünde bulundurulduğunda yatırımcılar açısından oldukça önemlidir (Başak vd., 2019). Finansal piyasalarda "Etkin Piyasalar Hipotezi" ve "Rastgele Yürüyüş Hipotezi" olarak bilinen ve finans biliminin temel taşları olarak kabul edilen kurallar yüzünden oldukça uzun bir süre hisse senedi fiyatlarındaki değişikliklerin öngörülemeyeceğine inanılıyordu (Başak vd., 2019). Etkin Piyasalar Hipotezi, mevcut fiyatların ilgili tüm bilgileri tam olarak yansıttığı ve piyasalar üstü bir kazanımın elde edilmesinin olanaksız olduğunu savunurken Rastgele Yürüyüş Hipotezi,

Makale Geçmişi / Article History
Başvuru Tarihi / Date of Application : 4 Haziran / June 2021
Kabul Tarihi / Acceptance Date : 16 Eylül / September 2021

piyasa fiyatlarının rastgele bir yürüyüş süreci olduğunu ve bu sebeple geçmiş piyasa hareketlerine ve davranışlarına dayalı olarak tahmin yapılmasının olanaklı olmadığını savunmaktadır. Finansal piyasalardaki rastlantısallık kavramına dayanan Rastgele Yürüyüş Hipotezi genel olarak kabul görmesine rağmen, finansal sistemin daha karmaşık dinamiklerini modelleyebilen algoritmaları kullanarak bu teorinin geçerliliğini sorgulayan birçok araştırmacı bulunmaktadır (Khaidem vd., 2016).

Zaman serisi verilerinin karmaşık yapılandırılmış davranışları nedeniyle gelecek tahmini oldukça zor bir iştir. Yatırımcıların riskini en aza indirmek ancak ve ancak gelecekteki fiyat hareketleri hakkında doğru tahminler yapmak ile mümkündür. Yatırımcı doğal olarak değerinin artmasını beklediği bir varlığa yatırım yapma, düşmesini beklediği bir varlıktan ise kaçınma eğiliminde olacaktır. Kazancını maksimize etmek isteyen bir yatırımcının trendleri doğru bir şekilde takip etmesi gerekmektedir. Piyasanın gidişatını tahmin etmek için farklı metodolojiler kullanılmaktadır. Çoğu çalışma, makine öğrenmesi tekniklerinin borsa fiyat hareketlerinin tahmininde oldukça tatmin edici sonuçlar ürettiğini göstermiştir.

Makine öğrenmesi, eldeki bilgiye dayanarak öğrenilen sistem üzerinde yapılan tahminlere odaklanmaktadır. Eldeki verinin tipine göre kullanılan makine öğrenmesi yöntemleri temelde gözetimli ve gözetimsiz olmak üzere 2 çeşittir. Gözetimli öğrenmede etiketlenmemiş veri ile etiketlenmiş veri arasındaki ilişkiyi tasvir eden bir fonksiyon üretilir (Alacan, 2020). Amaç, çıktısı belli olan etiketli veriler yardımıyla kurulan fonksiyon kullanılarak çıktıları bilinmeyen veri setleri için tahminlerde bulunmaktır. Sınıflandırma problemlerinde sıkça kullanılan algoritmalar Karar Ağaçları (KA), Rasgele Orman (RO), k-En Yakın Komşu (k-NN), Naive Bayes (NB), Lojistik Regresyon (LR) ve Destek Vektör Makineleri (DVM) olmaktadır.

Öte yandan finansal varlıkların gelecekteki yönünü tahmin etmede kullanılan sınıflandırma yöntemlerinde çoğunlukla girdi olarak makroekonomik ve mikroekonomik faktörler ele alınmaktadır. Mikroekonomik faktörler işletme düzeyinde ortaya çıkarken, makroekonomik faktörler olarak petrol fiyatı, altın fiyatı, döviz kuru, faiz oranı, para arzı, tüketici fiyat endeksi, cari açık ve ödemeler dengesi gibi göstergeler sayılabilmektedir. Girdi(ler) ve çıktı(lar) arasındaki ilişkinin matematiksel formunun belirlenmesi ve sonrasında etkili tahminlerin elde edilmesi, şüphesiz ki istatistiksel açıdan modele anlamlı katkı sağlayan ve kuramsal teori ile çelişmeyen girdilerin analize dâhil edilmesiyle gerçekleşir. Bu sebeple yüksek oranda doğru tahminler üreten bir modelleme sürecinde girdinin, diğer bir ifade ile özniteliklerin araştırmacı tarafından doğru ve eksiksiz bir şekilde belirlenmesi oldukça önem arz etmektedir. Ancak gerçek hayat problemlerinde çıktıyı etkileyen girdileri eksiksiz bir şekilde belirlemek ve bunları model sürecine dâhil etmek çoğu zaman araştırmacı açısından zor olabilmektedir. Özellikle

araştırmacı, kuramsal ilişkiler konusunda uzman değilse ya da ilgili girdi ile çıktı arasındaki ilişkiyi bilmiyorsa modelleme süreci daha da karmaşık hale gelmektedir. Bu sebeple makine öğrenme yöntemlerinin uygulanışında veri boyutunu azaltarak modelin karmaşıklığını düşürmek, bozuk ve gürültülü bilgi sorununu çözmek gibi sınıflandırıcının performansını doğrudan etkileyen konularda özellikle son yıllarda adından sıkça bahsedilen "öznitelik mühendisliği (feature engineering)" kavramı araştırmacıların dikkatini çekmektedir. En genel ifade ile öznitelik mühendisliği; tahmine dayalı öğrenme performansını iyileştirmek amacıyla öznitelik dönüşümü (feature transformation), öznitelik üretimi (feature generation), öznitelik çıkarımı (feature extraction), öznitelik seçimi (feature selection), öznitelik analizi ve değerlendirmesi, genel otomatik öznitelik mühendisliği metodolojisi (general automatic feature engineering) ve öznitelik mühendislik uygulamaları (feature engineering applications) olmak üzere farklı ana başlıkları içeren bir yöntemler topluluğudur (Dong ve Liu, 2018).

Bu çalışma kapsamında Türkiye ekonomisinin önemli göstergelerinden biri olan ve BIST tarafından hesaplanan BIST 100 endeksinin gelecekteki davranışının tahminine yönelik bir sınıflandırma süreci önerilmiştir. Önerilen bu süreç, öznitelik mühendisliği yöntemlerine kapsamlı bir bakış açısı sunmaktadır. Literatürdeki diğer çalışmalardan farklı olarak, BIST 100 endeksinin gelecekteki davranışını belirlemede herhangi bir mikro veya makro ekonomik açıklayıcı kullanılmamış, modelleme sürecine sadece verinin iç yapısından elde edilen öznitelikler dahil edilmiştir. Bu öznitelikler 69 farklı matematiksel fonksiyon kullanarak hesaplanmış ve özellikle zaman serisinin özetleyici istatistiklerine (merkezi eğilim/yayılım ölçüleri, çarpıklık vb.), seriye ait örnekleme dağılımının ek karakteristiklerine (mutlak enerji, binned entropi vb.) ve serinin gözlenen dinamiklerine (ortalama mutlak değişim, otokorelasyon vb.) dayanmaktadır. Veri ön işleme aşaması sonrasında sınıflandırma sürecinde yaşanabilecek muhtemel bir aşırı uyumdan (overfitting) kaçınmak adına, çok boyutlu öznitelik uzayını en iyi temsil yeteneğine sahip olan, daha az boyutlu öznitelik uzayı oluşturmada temel bileşenler (TB) analizinden faydalanılmıştır. Sonrasında Benjamini-Hochberg (BH) öznitelik seçim prosedürü sonuçlarına dayanan bir öznitelik seçim süreci gerçekleştirilerek nihai öznitelik vektörü oluşturulmuş ve BIST 100 endeksi değişiminin (artan/azalan) tahmin edilmesinde 8 farklı sınıflandırma algoritmasından yararlanılmıştır. NB dışındaki tüm sınıflandırıcılar için (k-NN, DVM, KA, RO ve LR) parametre optimizasyonları Grid Search (GS) yöntemi ile yapılmış ve sınıflandırıcılara ait sınıflama performansları farklı performans metrikleri baz alınarak karşılaştırılmıştır. Sınıflandırmalar, farklı oranlarda eğitim verisi ile (%80, %70, %60, %50) tekrarlanmış ve tahmin performanslarının tutarlılığı araştırılmıştır.

2. Finansal Literatürde Makine Öğrenmesi

Yatırımcıların en az risk ile kârlarını maksimize etmek için borsa hareketlerine ait eğilimleri doğru bir şekilde tahmin etmesi gerekmektedir. Borsa fiyat hareketlerine ait davranışın tahmin edilmesinde kullanılan farklı metodolojiler bulunmaktadır. Bunlar arasında en sık kullanılanlar, teknik analiz, zaman serisi tahmini, makine öğrenimi ve veri madenciliğidir (Khaidem vd., 2016). Bu çalışma kapsamında makine öğrenmesi ve veri madenciliği yaklaşımlarına odaklanan çalışmalar ele alınmıştır.

Özellikle ulusal literatürde finansal piyasaların hareketini tahmin etme konusunda oldukça önemli çalışmalar bulunmaktadır. Gündüz (2019), yatırımcıların kârlarını maksimize etmesi için finansal zaman serileri tahmininde derin öğrenme modellerinin ne oranda iyi performans sergilediğini araştırmıştır. Çalışmasında kullandığı derin öğrenme modellerinin tahmin performanslarının literatürde sıklıkla kullanılan diğer yapay öğrenme modellerine (DVM, LR) güçlü bir alternatif teşkil ettiğini göstermiştir. Alkış (2017), BIST 100 endeksi hisse senedi getiri hareketlerinin (aşağı/yukarı) veri madenciliği yöntemleriyle tahmin edilmesi üzerine yaptığı çalışmada LR ve k-NN algoritmasını kullanmış ve LR sınıflandırıcısının borsa tahmininde oldukça başarılı sonuçlar ortaya koyduğunu göstermiştir. Filiz, Karaboğa ve Akoğul (2017), BIST 50 endeksini etkileyen faktörlerden yararlanarak makine öğrenmesi yöntemleri ile sınıflandırma modelleri oluşturmuşlardır. Öznitelik seçim yöntemleri ile endeksin değişim yönünü etkileyen faktörlerin önem sırasını belirleyerek çalışmalarında girdi olarak kullanmış ve k-NN, NB, C4.5 ve YSA sınıflandırıcıları ile gerçekleştirdikleri analizler sonucunda C4.5 algoritmasının %92.71 doğruluk oranı ile en iyi performansı gösterdiğini ortaya koymuşlardır. Özdemir, Tolun ve Demirci (2011), BIST 100 endeksinin getiri yönünü tahmin etmeye çalıştıkları çalışmada LR ve DVM'yi kullanmış olup her iki yöntemin de yatırımcılar tarafından getiri tahmininde etkin bir şekilde kullanılabileceğini göstermişlerdir. Kara ve Ecer (2018), BIST Banka Endeksi hareket yönünün tahmininde çeşitli makine öğrenmesi sınıflandırma yöntemlerinin performanslarını karşılaştırmak için girdi değişkeni olarak BIST'de hesaplanan 10 adet teknik göstergeyi, çıktı değişkeni olarak ise bir sonraki günün borsa endeksi kapanış değerini almışlardır. Analizler sırasında kullandıkları YSA, DVM, LR ve LDA (lineer diskriminant analizi) modellerinin doğru sınıflandırma performanslarını sırasıyla %81.74, %60.87, %76.70, %76.87 olarak bulmuşlardır. Kartal (2020), BIST 100 ve başka ülkelere ait çeşitli borsa endekslerinin yukarı ve aşağı eğilim yönünü modellemek amacıyla çalışmasında DVM'den yararlanmış ve makroekonomik faktörlerin borsa endeksleri üzerindeki etkilerini de incelemiştir. Pabuçcu (2019), BIST 100 endeksinin hareket yönünü tahminlemek için üç farklı yönteme (YSA, DVM, NB) dayalı 3 farklı model oluşturarak bu

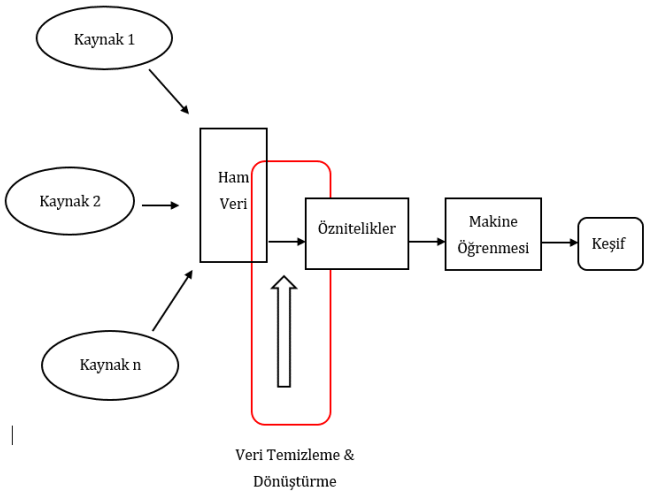
modellerin tahmin performanslarını karşılaştırmıştır. 10 faklı teknik göstergenin modellere girdi olarak alındığı çalışmada YSA sınıflandırıcısı diğer iki sınıflandırıcıdan daha yüksek performans sunduğunu göstermiştir. Aksoy (2021), BIST 30 endeksinde yer alan beş şirketin gelecek üç aylık ortalamaları ile hesaplanan hisse senedi fiyatlarını, 2010-2020 yılları arasındaki verileri ve beş makroekonomik değişkeni kullanarak tahmin etmeye çalışmıştır. Öztekin vd. (2016), çalışmalarında üç tahmin modelini (uyarlanabilir nöro-bulanık çıkarım sistemleri, YSA ve DVM) baz alarak günlük hisse senedi fiyat hareketlerini tahmin etmek için bir metodoloji geliştirmişler ve DVM'nin endeks tahmininde diğer modellere kıyasla daha üstün geldiğini göstermişlerdir. Gündüz ve Çataltepe (2015), BIST 100 endeksi açılış fiyatlarının yönünü tahmin etmek için bir gün önce yayınlanan haberleri ve fiyat verilerini kullanarak yeni bir metot geliştirmişlerdir. Çalışmalarında sınıflandırma performansını iyileştirebilecek öznitelikleri seçmek için öznitelik seçim yöntemlerini birleştirerek azınlık sınıflarının aşırı örneklenmesinden dolayı ortaya çıkan sınıf dengesizliği sorunuyla baş eden ve seçilen öznitelikleri bir bütün olarak ele alabilen bir öznitelik seçim yöntemi önermişlerdir. Koç-Ustalı, Tosun ve Tosun (2021), BIST 30 endeksinde yer alan firmaların hisse senedi fiyatlarını tahmin etmeyi amaçladıkları çalışmada YSA, RO ve XGBoost (Extreme Gradient Boosting) algoritmalarını kullanarak firmalara ait hisse senetlerinin gelecekteki fiyatlarını tahmin etmişledir. Filiz ve Öz (2017), BIST 100 endeks değerini etkileyerek değişimine sebep olan çeşitli makro ekonomik faktörleri ele aldıkları çalışmalarında tahmin yöntemi olarak makine öğrenimi sınıflandırma algoritmalarından k-NN, NB, LR ve C4.5 sınıflandırıcılarını kullanmışlardır. Weka üzerinden gerçekleştirdikleri analizde en başarılı yöntemi C4.5 Algoritması (%66.2) ve LR (%65.9) olarak bulmuşlardır. Kara, Boyacioglu ve Baykan (2011), BIST 100 endeksinin hareket yönünü tahmin etmek amacıyla çalışmalarında makine öğrenmesi yöntemlerinden YSA ve DVM algoritmalarını kullanarak bu algoritmaların tahmin performanslarını karşılaştırmışlardır. Basak vd. (2018), tahmin hatasını en aza indirmenin yatırım riskini en aza indirmek olacağını belirterek hisse senedi fiyatlarının n gün önce geçerli olan fiyata göre artıp azalmayacağını öngören sınıflandırma probleminin çözümü için RO ve XGBoost sınıflandırıcılarını kullanmışlardır.

Ulusal literatürün yanısıra uluslararası literatürde de finansal piyasaların hareketini tahmin etmede makine öğrenmesi yöntemleri sıkça kullanılmaktadır (Dai ve Zhang, 2013; Khan vd., 2016; Di, 2014; Nakagawa, Uchida ve Aoshima, 2018; Kakushadze ve Yu, 2019; Cao ve Tay, 2001; Nti, Adekoya ve Weyori, 2019; Teixeira ve De Oliveira, 2010; Lohrmann ve Luukka, 2019; Ou ve Wang, 2009; Kumar ve Thenmozhi, 2006; Huang, Nakamori ve Wang, 2005;

Creamer ve Freund, 2004). Bu yöntemler kıyaslandığında çoğunlukla LR, RO ve DVM sınıflandırıcılarının öne çıktığı görülmektedir.

3. Öznitelik Mühendisliği

Karar vericiler ellerindeki büyük veriyi keşfetmenin getirdiği kârın ve bu verilerin işlenmesiyle elde edilen öngörülerin oldukça güçlü bir rekabet avantajı doğurduğunu görmeye başladıklarından bu yana veri kavramı oldukça önem kazanmıştır. Verilerin hacmi ve çeşitliliği, manuel analiz kapasitesinin çok ötesine geçmiş ve çoğu durumda geleneksel veri tabanlarının kapasitesini aşmıştır (Provost ve Fawcett, 2013). Bilgisayar teknolojisinin güçlenmesi ve veri kümelerini birbirine bağlayan algoritmaların geliştirilmesi ile veri bilimi kavramı artık sıkça adından bahsettirmeye başlamıştır. Bilginin bilim ve endüstriyi yönlendirebilecek nitelikteki yararlı bilgiye dönüştürülmesine yardımcı olan ve ileriye dönük öngörüleri elde etmede bilimsel süreçleri kullanan veri bilimi; matematik, istatistik ve bilgisayar gibi farklı disiplinleri entegre kullanan çok disiplinli bir bilim alanıdır ve genel olarak ardışık düzeni takip eden dört temel adımdan oluşur: (1) problemin tanımlanması, (2) verilerin toplanması, (3) öznitelik mühendisliği ve (4) makine öğrenme modellerinin seçimi (Song, 2018). Makine öğrenmesinde öznitelikler girdi olarak ele alınır ve ham verilerin sayısal bir temsilini gösterir (Zheng ve Casari, 2018). Bir problemde makine öğrenmesi yöntemleri ile ele alınan süreç genellikle Şekil 1'deki gibi ilerler ve bu sürecin bir parçası olan öznitelik mühendisliği, sürecin performansını oldukça etkileyen bir kavram olarak araştırmacının karşısına çıkar. İş akışında sürecin her bir parçası bir sonrakini etkiler. Veriyi iyi temsil eden bir öznitelik uzayı, modelleme adımını kolaylaştırır ve nihai model, istenen görevi tamamlama konusunda daha yetenekli hale getirir. Eksik tanımlanmış veya verinin iyi bir temsili olmayan öznitelikler, aynı performans düzeyine ulaşmada çok daha karmaşık bir modelleme çabası gerektirebilir (Zheng ve Casari, 2018).



Şekil 1. Öznitelik Mühendisliğinin Makine Öğrenmesi İş Akışındaki Yeri

Kaynak: Zheng ve Casari, 2018.

Öznitelik mühendisliği, önceden de bahsedildiği gibi makine öğrenimi iş akışında en temel ve en fazla emek isteyen bir bileşendir. Çoğu makine öğrenme modellerinin performansı büyük ölçüde hangi özniteliklerin modele girdi olarak alınacağına bağlıdır. Bu sebeple veri bilimi ile uğraşan araştırmacılar proje sürecinin çoğunu ön işleme aşaması alarak da tanımlanan öznitelik vektörünü tasarlamaya ayırmaktadırlar. Gerçek hayat uygulamalarında öznitelik mühendisliği, deneme yanılma yoluyla model performansını sürekli gözlemlemeye ve buna tepki vermeye dayanan, sezgi ve uzmanlık alan bilgisini kullanarak modeli en iyilemeyi amaçlayan bir çabalar bütünüdür (Khurana, Samulowitz ve Turaga, 2018). Öznitelik mühendisliği modeli en iyilemede aşağıda verilen farklı ana başlıkları içeren yöntemlerin bir bütünüdür.

3.1. Öznitelik Ölçeklemesi

Model başarısı modele girdi olarak alınacak ve modele seçilecek özniteliklere bağlı olarak performans gösterecektir. Özniteliklerin iyi belirlenmesinin yanısıra bu özniteliklerin kullanılan modelin varsayımlarına uygun hale getirilmesi gerekmektedir. Bu sebeple model varsayımlarını göz önüne alarak girdinin ölçeklemesine ihtiyaç duyulabilir. Ölçekleme, farklı aralıklara veya birimlere sahip özniteliklerin karşılaştırılabilir değerlere dönüştürülmesine yardımcı olur. Uygulamada sıklıkla standardizasyon ve normalizasyon olmak üzere iki ölçekleme yöntemi kullanılır.

3.2. Öznitellik Üretimi

Modelleme işleminde kullanmak için ham verileri alma ve öznitelikleri tanımlama sürecidir. Faiz oranı, sanayi üretim endeksi, enflasyon oranı ve döviz kuru gibi göstergeler finansal serilerin hareketinde etkili olduğu bilinmekte ve çoğu modelleme çalışmalarında bu göstergelerin öznitelik olarak modele alındığı bilinmektedir.

3.3. Öznitelik Çıkartma ve Seçme

Öznitelik çıkartma ve öznitelik seçim sürecini birbirinden ayırmak çok da mümkün değildir. İki sürecin de birlikte kullanımının en temel amaçlarından biri, modelin aşırı uyumunu önlemektir. Bu iki süreç, ayrı ayrı veya aynı anda kullanılabilmektedir (Khalid, Khalil ve Nasreen, 2014).

3.3.1. Öznitelik Çıkartma

Kesit verilerinde farklı birimlere ait girdiler ve çıktı(lar) tek bir zaman noktasından toplanır ve araştırmacının temel olarak odaklandığı problem birimler arasındaki farklılıkların belirlenmesidir. Dolayısıyla farklılığın zaman içindeki değişimi araştırma kapsamı dâhilinde değildir. Bu sebeple modelleme sürecinde kesit verileri için öznitelik çıkartma işlemi genel akışa uygun bir biçimde gerçekleştirilir. Ancak araştırmacı zaman serisi verileri ile ilgileniyorsa, her bir çıktı, farklı zaman noktalarındaki serilerle ilişkili olacağı için öznitelik çıkarma işlemi biraz daha karmaşıklaşacaktır. Zaman serilerinin iç dinamiklerine dayalı özniteliklerin çıkartılması işlemini otomatik olarak gerçekleştiren ve Python kütüphanesi olarak gerçek hayat problemlerine entegre edebilecek yöntemler bulunmaktadır. Bu kütüphanelerden en sıklıkla kullanılanları FATS (Nun vd., 2015), CESIUM (Naul vd., 2016), TSFRESH (Christ vd., 2018) ve TSFEL (Barandas vd., 2020) sayılabilir.

TSFRESH (Time Series Feature Extraction Based on Scalable Hypothesis Tests) kütüphanesinde zaman serilerine ait yüzlerce öznitelik otomatik olarak çıkartılabilmektedir. Güncel sürümünde 69 farklı fonksiyon ele alınarak 700'den fazla öznitelik üretme olanağı sağlayan kütüphanede öznitelik çıkartma işleminin gerçekleşmesi, her zaman dilimine ait çıktıya karşılık gelen bir girdi üretmede kaydırma işlemi (rolling) sayesinde yapılabilmektedir. Zaman serisinin belirli bir alt kesitini oluşturmak için bir kesme penceresini zaman serisi üzerinde kaydırma işlemi ile zaman serisi tek boyutludan çok boyutlu hale gelir ve araştırmacının kullandığı matematiksel fonksiyonun ürettiği değer her zaman dilimi için bulunmuş olunur (Christ vd., 2018).

3.3.2. Öznitelik Seçme

Modele girdi olarak alınması muhtemel öznitelikler (1) modele katkı sağlayacak, modelde yer almadığında bilgi kaybını arttıracak öznitelikler, (2) modele katkı sağlamayacak, modele alınmadığında bilgi kaybını etkilemeyecek öznitelikler olmak üzere temelde iki kategoride değerlendirilir. Öznitelik seçim teknikleri ile kullanılacak modelin karmaşıklığını azaltmak için modele katkı sağlamayacak özellikler kapsam dışı kalır. Araştırmacının nihai hedefi, yüksek doğrulukta en az bilgi kaybıyla, daha hızlı sonuç veren ve en az sayıda özniteliğe sahip basit bir model elde etmektir. Böyle bir modeli elde etmek için bazı öznitelik seçme tekniklerinde birden fazla aday modelin eğitim sürecinden geçmesi gerekmektedir. Öznitelik seçme işlemi modelin eğitim sürecini kısaltmayı hedeflemez hatta bazı tekniklerde bu süre oldukça uzar (Zheng ve Casari, 2018).

Öznitelik seçimi süreci temelde boyut indirgeme veya iteratif seçim yöntemleri (filtreleyici teknikler, sarmal tekbikler, gömülü teknikler) kullanılarak gerçekleştirilir. Yüksek boyutlu öznitelik uzayında boyut indirgeme işlemi (dimensionality reduction) dönüştürülmüş öznitelikleri elde etmek için orijinal özniteliklere uygulanır. Bu tekniklerle elde edilen dönüştürülmüş öznitelikler modele girdi olarak alınır. Boyut indirgemenin en bilinen dezavantajı, dönüştürülmüş özniteliklerin orijinal özniteliklere kıyasla bilgi kaybı içermesidir. Ancak yüksek boyutlu öznitelik uzayı ile çalışmak, boyutsallığın laneti (curse of dimensionality) olarak adlandırılan problemi ortaya çıkaracaktır. Özellikle veri sayısının öznitelik sayısından büyük olduğu durumlarda boyut sorunu ile karşılaşmak oldukça olasıdır. Temel bileşenler analizi (TB) ve bağımsız bileşenler analizi boyut indirgemede kullanılan gözetimli yöntemler arasında sayılabilirken lineer ayırma analizi gözetimsiz yöntemler arasında yer alır.

Filtreleyici tekniklerin temel amacı ise modele katkısının olmadığı veya az olduğu özniteliklerin filtrelenerek modele dahil edilmesini engellemektedir. BH prosedürü (Benjamini ve Hochberg, 1995) filtreleyici teknikler arasında kullanılan ve TSFRESH kütüphanesinde yer alan bir öznitelik seçim yöntemidir. Bu prosedür, çoklu hipotez testi için sıralı değiştirilmiş Bonferroni düzeltmesini kullanarak yanlış keşif oranını (false discovery rate-FDR) kontrol eder. Bonferroni düzeltmesi FWER'e (Family Wise Error Rate) dayanırken BH birden fazla hipotez testini kontrol etmede FDR'ye dayalı bir yöntem olarak tasarlanmıştır (Haynes, 2013). BH prosedüründe her bir özniteliğin çıktı üzerindeki etkisi tek değişkenli testlerle değerlendirilir ve p değerleri hesaplanır. p değerini hesaplayan yöntemlere öznitelik seçiciler denir. Daha sonra çoklu test prosedürü olan BH prosedürü, hangi özelliklerin alınacağına ve hangilerinin model dışı kalacağına karar verir.

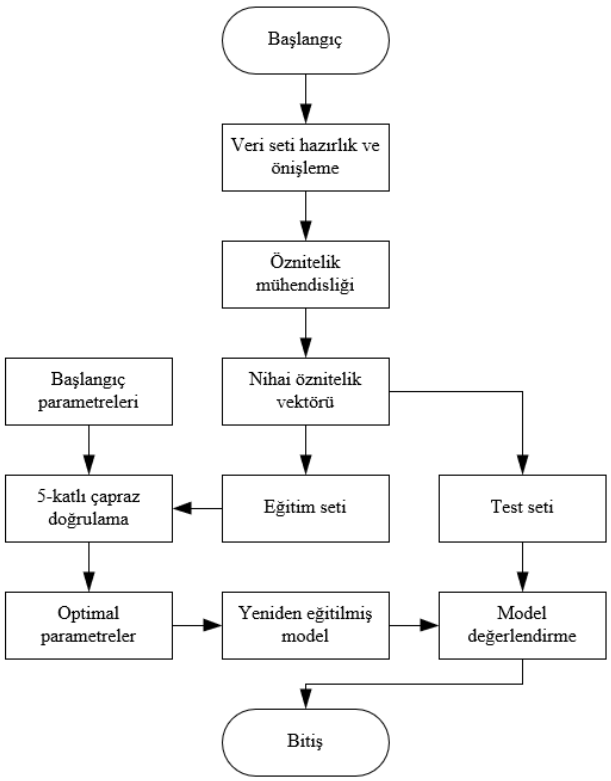
3.4. Öznitelik Değerlendirme

Öznitelik değerlendirmesi, özniteliklerin modeldeki faydasını değerlendirmekle ilgilidir. Bu süreç bazen öznitelik seçim sürecinin bir parçası olarak da kabul edilir.

4. Yöntem

4.1. Deneysel Tasarım

Çalışmanın tasarımı birkaç bağımsız sürecin bir bileşimidir. Şekil 2, önerilen sistemin genel çerçevesini göstermektedir. Bu çerçevenin ilk adımı veri setinin hazırlanmassı olup ikinci adımı öznitelik mühendisliği uygulamasıdır. Üçüncü ve son adımında sınıflandırma süreci yer almaktadır.

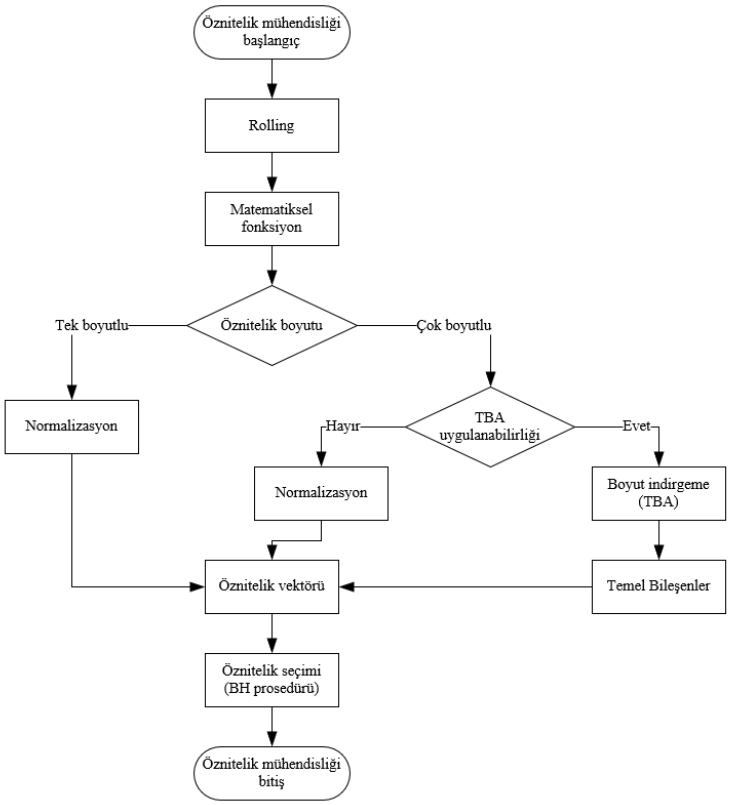


Şekil 2. Deneysel Tasarım

İlk aşamada kullanılacak veri seti www.investing.com adlı internet sitesinden temin edilmiştir. Ham veri setinde BIST 100 endeksine ait günlük fiyat bilgisi bulunmaktadır. Bu çalışmadaki temel amaç endeks değişiminin yönünü tahmin etmek olduğu için çıktı değişkenini oluşturmada önce günlük getiri hesaplanmış daha sonra getirisi negatif olan değerler (-1) pozitif olan değerler ise (+1) olarak etiketlenmiştir.

İkinci aşama kapsamlı bir öznitelik mühendisliği sürecinden oluşmaktadır. Bu süreç Şekil 3'de ayrıntılı olarak gösterilmiştir. Bir zaman serisinin tipik bir makine öğrenmesi probleminde

kullanılabilmesi için tek boyutlu halden çok boyutlu getirilmesinde rolling mekanizması kullanılmış ve seçilen rolling parametreleri ile oluşturulan her bir alt kesit için TSFRESH kütüphanesinde farklı matematiksel fonksiyonun değeri hesaplanarak öznitelikler çıkartılmıştır. Fonksiyonların bazıları tek bir öznitelik sütunu bazıları aynı fonksiyonun farklı parametreleri ile birbirinden farklı boyutlarda öznitelikler üretmektedir. Eğer bir fonksiyon tek bir öznitelik üretmiş ise ilgili sayısal değerlere normalizasyon uygulanmış ve dönüşüm sonrası elde edilen değerler tahmin algoritmasına girdi olacak şekilde saklanmıştır. Çok boyutlu öznitelik üreten fonksiyonlar ise ayrı ayrı ele alınmıştır. Öncelikle her bir sütun için normalizasyon uygulanmıştır. İlişkili değişkenleri modele aynı anda girdi olarak almaktan kaçınmak adına özniteliklerin istatistiksel açıdan birbirleriyle anlamlı ilişkiye sahip olup olmadığı incelenmiştir. Farklı parametreler ile aynı fonksiyondan üretilen özniteliklerin çoğunun kendi aralarında oldukça yüksek korelasyona sahip olduğu görülmüştür. Sonraki adım için temel amaç çok boyutlu öznitelik uzayını en iyi temsil yeteneğine sahip az boyutlu öznitelik uzayı oluşturmaktır. Bu amaç doğrultusunda TB analizinden faydalanılmış ve boyut indirgemenin mümkün olduğu durumda, çok boyutlu öznitelik uzayı, analizin önerdiği kadar boyutta öznitelik uzayına indirgenmiştir. TB analizine uygun olmayan çok boyutlu ve aynı matematiksel fonksiyonla hesaplanmış öznitelik uzayı için ise herhangi bir boyut indirgeme işlemi uygulanmamış, TSFRESH kütüphanesi tarafından üretilen orjinal değerlerin normalizasyon sonrasındaki ölçeklenmiş değerleri tahmin algoritmasına girdi olacak şekilde saklanmıştır. Sonrasında tek ve çok boyutlu öznitelik vektörleri birleştirilerek bir öznitelik vektörü oluşturulmuştur. Birleştirilen öznitelik vektörüne BH öznitelik seçim prosedürü yardımıyla bir öznitelik seçim süreci uygulanmış nihai öznitelik vektörü oluşturulmuştur.



Şekil 3. Öznitelik Mühendisliği Süreci

Nihai öznitelik vektörünün oluşturulması sonraki aşama üçüncü ve son adım olan sınıflandırma sürecidir. İlk olarak nihai öznitelik vektörü eğitim ve test setlerine farklı oranlar kullanılarak bölünmüştür. Farklı oranlar kullanılmasındaki temel amaç, elde edilen sınıflandırma başarılarının gittikçe azalan eğitim oranları ile değişip değişmediğinin gösterilmesi ve sınıflandırma sonuçlarının tutarlılığının araştırılmasıdır. Kullanılan her bir eğitim-test veri seti için BIST 100 endeks hareketi değişiminin tahmin edilmesinde farklı sınıflandırma algoritmaları kullanılmıştır. NB dışındaki tüm sınıflandırıcılar ile modellemelerde parametre optimizasyonu ızgara arama tekniği ile yapılmış ve optimize edilmiş parametrelerle sınıflandırma sonuçları elde edilmiştir. Bu amaç doğrultusunda çapraz doğrulama katsayısı tüm sınıflandırma işlemlerinde k=5 olarak seçilmiştir. Sınıflandırma algoritmalarının performanslarının değerlendirilmesinde temel metrik olarak seçilen doğruluk (accuracy) dışında ayrıca özgüllük (specificity), duyarlılık (sensitivity), F1-skor ve ROC (Receiver Operating Characteristic) eğrisi altında kalan alan değeri (AUC) hesaplanmıştır. Bahsi geçen performans metrikleri, çıktı değişkenin gerçek değeri ile tahmin edilen değerinin bir görselini sunan karmaşıklık matrisinden (confusion matrix) elde edilmiştir.

4.2. Sınıflandırma Algoritmaları

Çalışmada KA, RO, k-NN, NB, LR ve DVM'ye ait 3 farklı çekirdek fonksiyonu olan DVM_Lineer, DVM_RBF ve DVM_Sigmoid olmak üzere 8 farklı sınıflandırma algoritması kullanılmıştır. Bu algoritmalar aşağıda kısaca verilmiştir.

4.2.1. Karar Ağaçları

Temel amaç çok miktarda kayıt içeren bir veri setini böl-fethet metoduyla belli kurallar uygulayarak daha anlaşılır, küçük alt gruplara ayırmak ve veri setinin özelliklerinden ortaya çıkarılan basit karar kurallarını öğrenerek çıktı değişkenin değerini tahminleyecek şekilde bir model oluşturmaktır (Şahin, 2019; Alp ve Öz, 2019). Algoritma çıktısı bir ağaç görünümüne benzediği için KA adını alan yapı; düğümler, dallar ve yapraklardan (terminal düğüm) meydana gelmektedir. Kök düğümde başlayan sınıflandırma süreci terminal düğümde son bulur. Gözlemler homojen bir yapıya eriştiklerinde dallanma sona ererek sınıflandırma süreci sonlanmaktadır. Dallanmanın sonlandığı terminal düğümler, verilerin ait oldukları sınıfları göstermektedir (Alp ve Öz, 2019; Özekeş, 2003).

4.2.2. Rasgele Orman

Leo Breiman (2001) tarafından geliştirilen RO algoritması, ağaç tabanlı sınıflandırma algoritmalarının bütünü olarak tanımlanmaktadır. Algoritmanın asıl amacı farklı farklı öğrenmelerin gerçekleştiği veri setlerinden oluşturulan ağaç yapılarını birleştirerek bir karar ormanı oluşturmak ve bu sayede tahmin başarısını arttırmaktır (Elesan, 2019). Ormandaki ağaç sayısı ne kadar fazla olursa elde edilecek sonuç da o oranda iyileşir.

4.2.3. k-En Yakın Komşu

k-NN algoritması, 1951 yılında örüntü tanımada kullanılmak amacıyla Fix ve Hodges tarafından önerilmiş olup 1967 yılında ise Cover ve Hart tarafından geliştirilmiştir. Oldukça basit ve yorumlanması kolay bu algoritmanın işleyişi gözlemler arasındaki mesafeye dayanmaktadır (Elesan, 2019; Cunningham ve Delanay, 2007). Algoritma temelde bir sınıfa atanacak olan herhangi bir gözlemi hangi sınıfa atayacağına karar vermek için eğitim veri kümesi içerisinde bu gözleme en yakın k adet komşunun sınıflarını dikkate alarak bu komşular arasında en fazla karşılaşılan sınıfı, atanacak gözlemin sınıfı olarak tayin eder. Gerçekleştirilen sınıflandırma işlemi sonucunda sınıf içi benzerliğin maksimum, sınıflar arası benzerliğin ise minimum olması sağlanır. Yüksek doğruluğa sahip başarılı bir tahmin için veri örnekleri arasındaki mesafe ölçümünün nasıl yapılacağı ve dikkate alınacak komşu sayısı k'nın ne olacağı

son derece önem arzetmektedir (Elesan, 2019; Mucherino, Papajorgji ve Pardalos, 2009; Çalışkan ve Soğukpınar, 2008)

4.2.4. Naive Bayes

Temeli Bayes teoremine dayanan NB sınıflandırıcısı önceden sınıflanmış gözlemlerin sınıf bilgilerini kullanarak sisteme yeni giren bir verinin hâlihazırdaki sınıflardan hangisine ait olduğunun olasılığını hesaplayan bir algoritmadır (Silahtaroğlu, 2008)

4.2.5. Lojistik Regresyon

Regresyon analizi ile temelde benzer yapıda olan ve bir sınıflandırma yöntemi olarak da sıklıkla kullanılan LR, çıktı değişkeninin kategorik olması durumunda tercih edilmektedir. Kategorik bağımlı değişken değerinin tahmin edilmeye çalışıldığı LR analizinde temelde yapılmaya çalışılan şey diğer makine öğrenmesi yöntemlerinde olduğu gibi gruplara ilişkin üyelik tahminidir (Mertler ve Vannatta, 2002). Analizde yeni gözlemlerin mevcut bir sınıfa ait olma olasılıkları hesaplanmakta ve hangi sınıfın olasılık hesaplaması yüksek ise gözlem o sınıfa atanmaktadır (Alp ve Öz, 2019) Amaçlarından biri sınıflandırma olan yöntem ayrıca bağımlı ve bağımsız değişkenler arasındaki istatistiksel ilişkiyi de ortaya çıkarmaktır (Mertler ve Vannatta, 2002).

4.2.6. Destek Vektör Makineleri

Boser ve arkadaşları tarafından 1992 yılında sunulan ve daha sonra Cortes ve Vapnik tarafından 1995 yılında geliştirilen DVM hem sınıflandırma hem de regresyon problemlerinde kullanılabilen, verinin sınıflara ayrılmasında bu ayrımı en iyileyen hiperdüzlemi bulmayı amaçlayan bir yöntemdir (Yakut, Elmas ve Yavuz, 2014). Parametrik olmayan bir yaklaşım olması, veri dağılımına ilişkin herhangi bir varsayım gerektirmemesi ve basit ve kolay uygulanabilirliği algoritmanın avantajlarındandır (Hearst vd, 1998). Doğrusal DVM'lerde veri örneklerinin sınıflara ayrılması işlemi direkt olarak doğrusal bir hiperdüzlem ile gerçekleşmektedir. Ancak pratikte veri örnekleri orijinal giriş uzayında doğrusal bir düzlem ile ayrılamazlar. Bu durum çeşitli çekirdek fonksiyonları kullanılarak çözümlenebilir (Yakut, Elmas ve Yavuz, 2014). Doğrusal olmayan DVM'lerde çekirdek fonksiyonun seçimi oldukça önem arz etmektedir. Analizde hangi fonksiyonun kullanılacağı; problemin tipine, veri adedine ve çeşidine, veri setinin gürültü içerip içermemesine göre değişiklik göstermektedir. Fonksiyon seçiminde tek bir formül olmadığı için farklı fonksiyonlar denenmeli ve model için en uygun olan/en yüksek performans gösteren fonksiyon belirlenmelidir (Scholkopf, 2001). Çalışmada

sıkça kullanılan 3 farklı çekirdek fonksiyonu ele alınmıştır. Bunlar radial tabanlı (RBF), lineer ve sigmoid fonksiyonlarıdır.

5. Uygulama

Çalışmanın uygulama kısmı adımsal olarak verilmiştir.

5.1. Adım 1: Veri Setinin Hazırlanması

Çalışmada kullanılan veri seti, 18.06.1993 tarihinden 26.02.2021 tarihine kadar olan BIST 100 endeksine ait günlük kapanış fiyatlarıdır. Toplamda 6930 güne ait gözlem değerlerini içeren veri setine ait bazı özetleyici istatistikler Tablo 1'de verilmiştir. Çarpıklık ve basıklık testleri, serinin sağa çarpık ve platikurtik bir dağılım sergilediğini göstermektedir.

Tablo 1. BIST 100 Endeksi Günlük Fiyatlarına Ait Özetleyici İstatistikler

	N	*Ortalama*	*Std.Sapma*	*Medyan*	*Çarpıklık*	*Basıklık*
BIST 100	6930	437.26	376.14	384.14	0.49***	-0.85***

5.2. Adım 2:Öznitelik Mühendisliği

Zaman serisine ait öznitelik üretmek için seriyi çok boyutlu hale getirmede kullanılan rolling mekanizmasında ilgili parametreler *max_timeshift=20* ve *min_timeshift=5* olarak seçilmiştir. Oluşturulan her bir alt kesit için özniteliklerin çıkartılmasında TSFRESH kütüphanesinde yer alan 69 farklı matematiksel fonksiyonun tamamı kullanılmıştır. Fonksiyonların 39'u tek bir öznitelik sütunu üretirken 30'u aynı fonksiyonun farklı parametreleri ile birbirinden farklı boyutlarda öznitelikler üretmektedir. Farklı parametreler ile aynı fonksiyondan üretilen çok boyutlu özniteliklerin yüksek korelasyon yapısına sahip olması sebebiyle boyut indirgemede TB analizi kullanılmış ve öznitelik vektörünün TB analizine uygunluğunun değerlendirmesinde Keiser-Meyer-Olkin (KMO) ve Bartlett Küresellik testinden faydalanılmıştır. KMO örneklem yeterliliği ölçüsü için eşik değer 0.5 iken, Bartlett'in küresellik testi için hesaplanan p değeri ile karşılaştırılacak olan anlamlılık seviyesi 0.05 olarak alınmıştır. Tek ve çok boyutlu öznitelikler için yukarıda bahsedilen işlemler sonucunda 101 boyutlu bir öznitelik vektörü elde edilmiştir. Nihai öznitelik vektörünü oluşturmada kullanılan BH prosedüründe FDR=0.001 olarak alınmış ve 101 boyutlu öznitelikten 22 tanesinin istatistiksel olarak anlamlı olduğu ve modele girdi olarak alınması gerektiği sonucuna varılmıştır. Tablo 2'de sınıflandırıcılara girdi olarak alınacak öznitelikler (önem sırasına göre) ve açıklamaları verilmiştir. Bu öznitelikler zaman serisinin özetleyici istatistiklerine, örnekleme dağılımının ek karakteristiklerine ve gözlenen dinamiklerine göre farklılık göstermektedirler.

Tablo 2. Nihai Öznitelik Vektöründe Bulunan Öznitelikler

*Öznitelik Adı**	*Matematiksel Fonksiyon*	*Açıklama*
Mean second_derivate _central	$\frac{1}{2(n-2)}\sum_{i=1}^{n-1}\frac{1}{2}(x_{i+2}-2x_{i+1}+x_i)$	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. Tek boyutlu bir özniteliktir.
last_location_of _maximum	$\frac{max(arg\,max\,X)}{n}$	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. x'in maksimum değerinin bulunduğu son konumu verir. Tek boyutlu bir özniteliktir.
first_location_of _maximum	$\frac{min(arg\,max\,x)}{n}$	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. x'in maksimum değerinin bulunduğu ilk konumu verir. Tek boyutlu bir özniteliktir.
agg_linear_tren d	Dizi için doğrusal bir en küçük kareler regresyonunu hesaplar. Bu özellik, sinyalin tekdüze (uniformly) olarak örneklendiğini varsayar.	Zaman serisinin özetleyici istatistiklerine dayalı bir özniteliktir. İlgili fonksiyon için farklı parametrelerle parçacık boyutunun (chunk_len) 5 ve 10 olduğu 32 adet öznitelik üretilmiştir. Çıkartılan özniteliklere uygulanan TB analizinde (Bartlett p=0.00; KMO=0.77) ilk 5 temel bileşenin özdeğerlerinin 1'den büyük olduğu ve bu ilk beş bileşenin toplam varyasyonun yaklaşık %87'sini açıkladığı görülmüştür. Böylece 32-boyutlu öznitelik uzayı 5-boyuta indirgenmiş ve nihai öznitelik vektörüne alınmıştır.
mean_change	$\frac{1}{n}\sum_{i=1}^{n-1}\lvert x_{i+1}-x_i\rvert$	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. Ardışık x değerleri arasındaki mutlak değişimin ortalamasını hesaplar. Tek boyutlu bir özniteliktir.
number_cwt_ peaks	Bu özellik hesaplayıcı, x zaman serisinin Ricker dalgacık dönüşümünün katsayılarını inceleyerek zaman serilerindeki zirveleri tespit eder.	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. x'deki farklı zirvelerin (peak) sayısını verir. Bu fonksiyona ait farklı parameter değerleri için 6 farklı öznitelik çıkartılmıştır. Çıkartılan özniteliklere uygulanan TB analizinde (Bartlett p=0.00; KMO=0.56) ilk üç temel bileşenin toplam varyasyonun yaklaşık %58'ini açıkladığı görülmüştür. 6-boyutlu öznitelik uzayı 3-boyuta indirgenmiş ve nihai öznitelik vektörüne alınmıştır.
change_ quantiles__f_agg	$\frac{\sum_{i=1}^{n-1}\lvert x_{i+1}-x_i\rvert\,1_{Q_{ql}}\le x_i\le Q_{qh}\,1_{Q_{ql}}\le x_{i+1}\le Q_{qh}}{\sum_{i=1}^{n-1}1_{Q_{ql}}\le x_i\le Q_{qh}\,1_{Q_{ql}}\le x_{i+1}\le Q_{qh}}$	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. Farklı q parametreleri ile toplamda 60 adet öznitelik üretilmiştir. Çıkartılan özniteliklere uygulanan TB analizinde (Bartlett p=0.00; KMO=0.88) ilk 9 temel bileşenin özdeğerlerinin 1'den büyük olduğu bu ilk bileşenlerin toplam varyasyonun yaklaşık %84'ünü açıkladığı görülmüştür. Böylece 60-boyutlu öznitelik uzayı 9-boyuta indirgenmiş ve nihai öznitelik vektörüne alınmıştır.
fft_coefficient	$\sum_{m=0}^{n-1}a_m\,exp\left\{-2\pi i\frac{mk}{n}\right\}$	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. Tek boyutlu ayrık fourier dönüşümünün fourier katsayılarını hesaplar. Kütüphanede farklı parametrelerle toplamda 11 öznitelik üretilmiştir. Çıkartılan özniteliklere uygulanan TB analizinde (Bartlett p=0.00; KMO=0.95) ilk 2 temel bileşenin özdeğerlerinin 1'den büyük olduğu bu ilk bileşenlerin toplam varyasyonun yaklaşık %57'sini açıkladığı görülmüştür. Böylece 11-boyutlu öznitelik uzayı 2-boyuta indirgenmiş ve nihai öznitelik vektörüne alınmıştır.

skewness	$\frac{n^2}{(n-1)(n-2)} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2)^{\frac{3}{2}}}$	Zaman serisinin özetleyici istatistiklerine dayalı bir özniteliktir. Serinin çarpıklık katsayısıdır. Tek boyutlu bir özniteliktir.
last_location_of_minimum	$\frac{\max(\arg\min x)}{n}$	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. x'in minimum değerinin bulunduğu son konumu verir. Tek boyutlu bir özniteliktir.
first_location_of_minimum	$\frac{\min(\arg\min x)}{n}$	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. x'in minumum değerinin bulunduğu ilk konumu verir. Tek boyutlu bir özniteliktir.
longest_strike_below_mean	x'in ortalama değerine eşit ya da daha küçük verilerden oluşan en uzun ardışık alt dizinin uzunluğunu hesaplar.	Zaman serisinin gözlenen dinamiklerine dayalı bir özniteliktir. Tek boyutlu bir özniteliktir
index_mass_quantile	Kartil (yüzdelik) fonksiyonudur. Bir veri kümesinin q'uncu kartili, q'ya eşit veya daha küçük olan bir özetleme değerini temsil eder ($0 \leq q \leq 1$).	Zaman serisinin örnekleme dağılımının ek karakteristiklerine dayalı bir özniteliktir. Farklı q parametreleri ile toplamda 8 adet öznitelik üretilmiştir. Çıkartılan özniteliklere uygulanan TB analizinde (Bartlett p=0.00; KMO=0.81) ilk 2 temel bileşenin özdeğerlerinin 1'den büyük olduğu bu ilk bileşenlerin toplam varyasyonun yaklaşık %80'nini açıkladığı görülmüştür. Böylece 8-boyutlu öznitelik uzayı 2-boyuta indirgenmiş ve nihai öznitelik vektörüne alınmıştır.
Not: Matematiksel fonksiyonlarda verilen x, ilgili zaman serisini göstermektedir. *(TSFRESH kütüphanesindeki orjinal isimleri kullanılmışır.)		

5.3. Adım 3:Sınıflandırma

Nihai öznitelik vektörü 4 farklı oranda (%80, %70, %60 ve %50) eğitim ve test setlerine bölünmüştür. Kullanılan her bir eğitim veri seti için KA, RO, k-NN, LR, DVM_Lineer, DVM_RBF ve DVM_Sigmoid sınıflandırıcılarına ait parametre optimizasyonu sonucu elde edilen optimal patametreler kullanılarak test veri setleri için sınıflandırmalar gerçekleştirilmiş ve bu sınıflandırıcılara ait performans metrikleri hesaplanmıştır. NB sınıflandırıcısı için parametre optimizasyonu olmadan model eğitim veri setleri ile eğitilmiştir. Tablo 3'de her bir sınıflandırıcı ve her bir eğitim oranı için elde edilen sınıflandırma sonuçları verilmiştir.

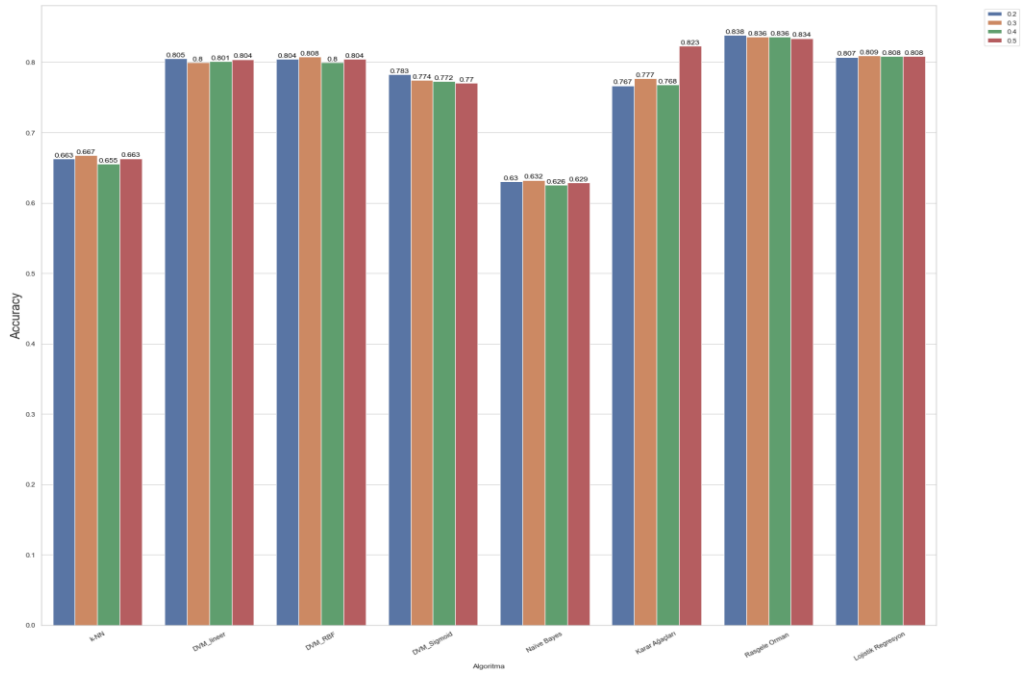
Tablo 3. Sınıflandırma Sonuçları

	%80 Eğitim	*%70 Eğitim*	*%60 Eğitim*	*%50 Eğitim*
Karar Ağacı Sınıflandırıcısı				
DO	0.7667	0.7771	0.7682	0.8229
Duyarlılık	0.7799	0.8085	0.7964	0.8438
Özgüllük	0.7526	0.7432	0.7363	0.7995
F1-skor	0.7761	0.7905	0.7847	0.8345
AUC	0.8580	0.8560	0.8770	0.9060
Rasgele Orman Sınıflandırıcısı				
DO	0.8383	0.8358	0.8357	0.8339
Duyarlılık	0.8398	0.8418	0.8366	0.8504
Özgüllük	0.8365	0.8294	0.8347	0.8154

F1-skor	0.8433	0.8422	0.8347	0.8442
AUC	0.9200	0.9210	0.9220	0.9230
k-NN Sınıflandırıcısı				
DO	0.6628	0.6674	0.6552	0.6630
Duyarlılık	0.7116	0.7150	0.6936	0.7139
Özgüllük	0.6101	0.6158	0.6118	0.6057
F1-skor	0.6863	0.6911	0.6809	0.6915
AUC	0.7270	0.7260	0.7200	0.7170
Naive Bayes Sınıflandırıcısı				
DO	0.6303	0.6323	0.6256	0.6289
Duyarlılık	0.6727	0.6799	0.6140	0.6277
Özgüllük	0.5847	0.5807	0.6387	0.6303
F1-skor	0.6535	0.658	0.6349	0.6415
AUC	0.6940	0.6960	0.7010	0.7030
Lojistik Regresyon Sınıflandırıcısı				
DO	0.8072	0.8094	0.8014	0.8081
Duyarlılık	0.8342	0.8473	0.8420	0.8438
Özgüllük	0.7781	0.7683	0.7555	0.7682
F1-skor	0.8177	0.8222	0.8181	0.8232
AUC	0.8910	0.8890	0.8900	0.8930
Destek Vektör Makineleri (Lineer Çekirdek)				
DO	0.8050	0.7998	0.8014	0.8039
Duyarlılık	0.8426	0.8482	0.8420	0.8498
Özgüllük	0.7646	0.7472	0.7555	0.7522
F1-skor	0.8175	0.8151	0.8181	0.8209
AUC	0.8900	0.8870	0.8900	0.8910
Destek Vektör Makineleri (RBF Çekirdek)				
DO	0.8043	0.8079	0.7996	0.8042
Duyarlılık	0.8398	0.851	0.8250	0.8356
Özgüllük	0.7661	0.7612	0.7709	0.7688
F1-skor	0.8165	0.8217	0.8136	0.8187
AUC	0.9000	0.8990	0.8940	0.8930
Destek Vektör Makineleri (Sigmoid Çekirdek)				
DO	0.7826	0.7743	0.7725	0.7704
Duyarlılık	0.8300	0.8334	0.8318	0.8335
Özgüllük	0.7316	0.7101	0.7056	0.6995
F1-skor	0.7983	0.7934	0.7950	0.7934
AUC	0.8670	0.8640	0.8600	0.8610

Sürecin performansı incelendiğinde tüm eğitim oranları için RO sınıflandırıcısının kullanılan performans metriklerine göre diğer sınıflandırıcılara kıyasla daha iyi sonuç verdiği görülmektedir. Eğitim oranı %80'den %50'ye düştüğünde RO sınıflandırıcısının başarısında sadece yaklaşık %0.4'lük bir kayıp görülmektedir. %50 eğitim verisi ile çalışıldığında dahi sınıflandırma başarısının %83.3'ün altına inmemesi, RO sınıflandırıcısının tahmin başarısının oldukça tutarlı olduğunu göstermektedir. RO sınıflandırıcısının ardından tahmin performansı DO metriğine göre en iyi sınıflandırıcının LR olduğu görülmektedir. Eğitim oranı %80'den

%50'ye doğru gittikçe düştüğünde DO değerinin %80.7 seviyelerinde kaldığı görülmektedir. Farklı çekirdek fonksiyonları ile kurulan DVM modellerinde DO metriğine göre en düşük performansı sigmoid çekirdek fonksiyonu vermiştir. LR sınıflandırıcısı ile DVM_lineer ve DVM_RBF sınıflandırıcıları DO metriği bakımından hemen hemen benzer sonuçlar üretmektedir. Sınıflandırma başarısının en düşük olduğu ilk 2 algoritma NB ve k-NN olduğu dikkat çekmektedir. DO metriği baz alınırsa, tüm eğitim oranlarında performansı en düşük algoritma NB olmaktadır. Ayrıca bu algoritmanın duyarlılık ve özgüllük değerlerinin yaklaşık olarak birbirlerinin tamamlayıcı olması, özellikle yüksek eğitim oranlarında aşırı uyum probleminin var olabileceğini göstermektedir. Farklı test oranlarında tüm sınıflandırıcıların DO metriğine göre performansları Şekil 4'de verilmiştir.



Şekil 4. Test Oranlarına Göre Sınıflandırıcıların Başarı Değişimi

6. Sonuç ve Tartışma

Hisse senedi piyasaları yıllardır ekonomilerin merkezinde yer almıştır. Bu piyasalarda meydana gelen herhangi bir kriz, ekonomi üzerinde etkili değişimlere sebep olur. Ekonomilerin yöneticileri beklenmedik bir kriz durumunda etkili ve hızlı şekilde önlem almak için borsaların davranışlarını özenle izler (Demir, 2019). Piyasa hareketlerini takip eden karar vericilere/yatırımcılara faydalı rehber olması amacıyla BIST tarafından farklı endeksler hesaplanmaktadır. Bu endeksler, piyasanın mevcut durumunu anlamayı kolaylaştırır ve karşılaştırma ölçütü olarak kullanılabilirler. Karar vermede yatırımcılar minimum risk/maksimum kazanç ilkesini benimser ve ilgili varlığın gelecekteki davranışını tahmin etmek

isterler. Geleneksel veya makine öğrenimi yöntemleri tahmin doğruluğunu etkili bir şekilde artırabilir ve bu yöntemlerden elde edilen çıktılar yatırımcıların risklerden kaçınmasına ve faydayı artırmasına yardımcı olabilir (Luo ve Wang, 2016; Yigit, Alp ve Öz, 2020).

BIST 100 endeksi BIST'de işlem gören 100 hisse senedinin (piyasa ve işlem hacmi en yüksek) performansını ölçmek için hesaplanan ve Türkiye borsasının temel göstergelerinden sayılan bir endekstir. Bu endeksin gelecekteki davranışını tahmin etmeye yönelik çabalar literatürde genişçe yer bulmuştur. Bu çalışmada makine öğrenmesi yöntemleri yardımı ile BIST 100 endeks getirilerine ait artış/azalış yönünü yüksek doğrulukla tahmin etmede ardışık düzeni takip eden bir süreç önerilmiştir. Bu amaç doğrultusunda kapsamlı bir öznitelik mühendisliği işlemleri yürütülmüş ve parametre optimizasyonları yapılarak tahmin modelleri arasında karşılaştırmalar yapılmıştır. Çalışmanın iki temel araştırma sorusu mevcuttur: (1) Sınıflandırma modellerine girdi olarak seçilecek öznitelikler verinin kendi iç dinamiğinden üretildiğinde sınıflandırma performansları hangi oranda olur? (2) Sadece verinin kendisi kullanılarak, diğer bir ifade ile modele herhangi bir mikro/makro ekonomik gösterge kullanmadan oluşturulan öznitelik uzayı ile gerçekleştirilen sınıflandırmalarda en iyi performansı hangi algoritma üretir? Veri ön işleme aşamasında, literatürdeki diğer çalışmalardan farklı olarak BIST 100 endeksinin gelecekteki davranışını belirlemede herhangi bir mikro veya makro ekonomik açıklayıcı kullanılmamış, modelleme sürecine sadece verinin iç yapısından elde edilen öznitelikler dahil edilmiştir. Bu öznitelikler TSFRESH kütüphanesinde yer alan farklı matematiksel fonksiyonlar ile hesaplanmıştır.

Bu çalışma, benzer amaç taşıyan gelecek çalışmalara bir temel oluşturmakla birlikte diğer çalışmalarda kullanılan makine öğrenmesi algoritmaları için karşılaştırma yapılabilmesini sağlayacak çıktılar sunmaktadır. Çalışmanın bazı kısıtları bulunmaktadır. Verinin meta bilgisinden yararlanılarak elde edilen tahmin sonuçlarının oldukça yüksek olduğu görülmekle birlikte, ileriki çalışmalarda ele alınan öznitelikler haricinde önemli bazı makro veya mikro göstergeler kullanılarak tahminin gücü arttırılabilir. Ayrıca sınıflandırma algoritmalarının çeşitliliği arttırılarak daha genelleştirilebilir sonuçlar elde edilebilir.

KAYNAKÇA

Aksoy, B. 2021. "Pay senedi fiyat yönünün makine öğrenmesi yöntemleri ile tahmini: Borsa İstanbul örneği." *Business and Economics Research Journal*, 12(1), 89-110.

Alacan, S. 2020. "*Makine öğrenmesi ile teknoloji perakende sektöründe ek garanti satış modellemesi*." Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü: İstanbul.

Alkış, B. N. 2017. "*Çoklu lojistik regresyon ve k-en yakın komşu yöntemleri ile BIST 100 endeks getiri yönünün tahmini*." Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü: İstanbul.

Alp, S. ve Öz, E. 2019. *Makine Öğrenmesinde Sınıflandırma Yöntemleri ve R Uygulamaları,* İstanbul: Nobel Akademik Yayıncılık:

Barandas, M., Folgado, D., Fernandes, L., Santos, S., Abreu, M., Bota, P., ... ve Gamboa, H. 2020. "TSFEL: time series feature extraction library," *Softwarex*,11, 100456.

Basak, S., Kar, S., Saha, S., Khaidem, L. ve Dey, S. R. 2019. "Predicting the direction of stock market prices using tree-based classifiers." *The North American Journal of Economics and Finance*, 47, 552-567.

Benjamini, Y. ve Hochberg, Y. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of The Royal Statistical Society Series B*, 57(1), 289–300.

Breiman, L. 2001. "Random forests machine learning," *Kluwer Academic Publishers*, 45(1), 5-32.

Çalışkan, S. K. ve Soğukpınar, İ. 2008. "Kxknn: K-means ve k en yakin komşu yöntemleri ile ağlarda nüfuz tespiti," *EMO Yayınları,*120-124.

Cao, L. ve Tay, F.E. 2001. "Financial forecasting using support vector machines." *Neural Computing and Applications*, 10(2), 184-192.

Christ, M., Braun, N., Neuffer, J., ve Kempa-Liehr, A. W. 2018. "Time series feature extraction on basis of scalable hypothesis tests (Tsfresh–a python package)." *Neurocomputing*, 307, 72-77.

Creamer, G. G. ve Freund, Y. 2004. "Predicting performance and quantifying corporate governance risk for latin american adrs and banks." *Financial Engineering and Applications*, MIT, Cambridge.

Cunningham P. ve Delanay, S.J. 2007 "K-neighbour classifiers technical report." https://www.researchgate.net/publication/228686398_k-Nearest_neighbour_classifiers. Erişim Tarihi: 4.02.2021.

Dai, Y. ve Zhang, Y. 2013. "Machine learning in stock price trend forecasting." Stanford University, http://cs229.stanford.edu/proj2013/DaiZhang-MachineLearningInStockPriceTrendForecasting.pdf . Erişim Tarihi: 21.05.2021.

Demir, C. 2019. "Macroeconomic determinants of stock market fluctuations: the case of BIST 100." *Economies*, 7(1), 1-14.

Di, X. 2014. "Stock trend prediction with technical indicators using SVM." *Independent Work Report,* Stanford Univ, Stanford.

Dong, G. ve Liu, H. 2018. *Feature Engineering for Machine Learning and Data Analytics*. CRC Press.

Elesan, S. 2019. "*Veri madenciliğinde farklı karar ağaçları ve k-en yakın komşuluk yöntemlerinin incelenmesi: kadın hastalıkları ve doğum verisinde bir uygulama*." Doktora Tezi, Van Yüzüncü Yıl Üniversitesi Sağlık Bilimleri Enstitüsü: Van.

Filiz, E. ve Öz, E. 2017. "Classification of BIST 100 ındex changes via machine learning methods." *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 39(1), 117-129.

Filiz, E., Karaboğa, H. ve Akoğul, S. 2017. "BIST-50 endeksi değişim değerlerinin sınıflandırılmasında makine öğrenmesi yöntemleri ve yapay sinir ağları kullanımı." *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 26(1), 231-241.

Gündüz, H. 2019. "*Derin öğrenme yöntemleri ile zaman serisi tahmini*." Doktora Tezi, İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü: İstanbul.

Gunduz, H. ve Cataltepe, Z. 2015. "Borsa Istanbul (BIST) daily prediction using financial news and balanced feature selection." *Expert Systems with Applications*, 42(22), 9001-9011.

Haynes W. 2013. Benjamini–Hochberg Method. W. Dubitzky, O. Wolkenhauer, K. H. Cho ve H. Yokota içinde, *Encyclopedia of Systems Biology*, Springer: New York.

Hearst, M. A., Dumais, S.T., Osuna, E., Platt, J. ve Scholkopf, B. 1998. "Support vector machines." *IEEE Intelligent Systems and Their Applications*, 13(4), 18-28.

Huang, W., Nakamori, Y. ve Wang, S.Y. 2005. "Forecasting stock market movement direction with support vector machine." *Journal of Computers and Operational Research*, 32(10), 2513-2522.

Kakushadze, Z. ve Yu, W. 2019. "Machine learning risk models." *Journal of Risk and Control*, 6(1), 37-64.

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

Kara, Y., Boyacıoğlu, M.A. ve Baykan, Ö.K. 2011. "Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the İstanbul Stock Exchange," *Expert Systems with Applications*, 38(5), 5311-5319.

Kartal, C. 2020. "Destek vektör makineleri ile borsa endekslerinin tahmini." *Itobiad: Journal of The Human & Social Science Researches,* 9(2), 1394-1418.

Khaidem, L., Saha, S., ve Dey, S. R. 2016. "Predicting the direction of stockmarket prices using random forest." https://arxiv.org/abs/1605.00003. Erişim Tarihi: 21.05.2021.

Khalil, T. ve Nasreen, S. (2014). "A survey of feature selection and feature extraction techniques in machine learning." Science and İnformation Conference, https://ieeexplore.ieee.org/document/6918213?denied= Erişim Tarihi: 10.05.2020.

Khan, W., Ghazanfar, M. A., Asam, M., Iqbal, A., Ahmad, S., & Khan, J. A. 2016. "Predıctıng trend ın stock market exchange usıng machıne learnıng classıfıers." *Science International*, 28(2), 1363-1367.

Khurana, U., Samulowitz, H. ve Turaga, D. 2018. "Feature engineering for predictive modeling using reinforcement learning." *Proc. AAAI Conference on Artificial Intelligence*, 32(1), 3407-3414.

Koç-Ustalı, N., Tosun, N. ve Tosun, Ö. 2021. "Makine öğrenmesi teknikleri ile hisse senedi fiyat tahmini." *Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 16(1), 1-16.

Kumar, M. ve Thenmozhi, M. 2006. "Forecasting stock index movement: A comparison of support vector machines and random forest." *In Indian Institute of Capital Markets 9th Capital Markets Conference Paper,* Vashi: India.

Lohrmann, C. ve Luukka, P. 2019. "Classification of intraday S&P500 returns with a random forest." *International Journal of Forecasting*, 35(1), 390-407.

Luo, H ve Wang, S. 2016. "Based on the pca-arıma-bp hybrid model of stock price prediction research." *Anzıam Journal*, 58, 162-178.

Mertler, C.A. ve Vannatta, R.A. 2002. *Advanced and multivariate statistical methods: Practical application and interpretation,* California: Pyrczak Publishing.

Mucherino, A., Papajorgji, P. J. ve Pardalos, P. M. 2009. "K-nearest neighbor classification in data mining in agriculture." *Springer Optimization and Its Applications*, 34, 83-106.

Nakagawa, K., Uchida, T., ve Aoshima, T. 2018. *Deep Factor Model*. ECML PKDD Workshops.

Naul, B., van der Walt, S., Crellin-Quick, A., Bloom, J. S., ve Pérez, F. 2016. "Cesium: open-source platform for time-series inference." https://arxiv.org/abs/1609.04504. Erişim Tarihi: 15.03.2020.

Nti, K. O., Adekoya, A. ve Weyori, B. 2019. "Random forest based feature selection of macroeconomic variables for stock market prediction." *American Journal Of Applied Sciences*, 16(7), 200-212.

Nun, I., Protopapas, P., Sim, B., Zhu, M., Dave, R., Castro, N., ve Pichara, K. 2015. "FATS: feature analysis for time series." https://arxiv.org/abs/1506.00010. Erişim Tarihi: 21.05.2021.

Ou, P. ve Wang, H. 2009. "Prediction of stock market index movement by ten data mining techniques." *Modern Applied Science*, 3(12), 28-42.

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

Özekeş, S. 2003. "Veri madenciliği modelleri ve uygulama alanları." *İstanbul Ticaret Üniversitesi Dergisi,* 3, 65-82.

Öztekin, A., Kizilaslan, R., Freund, S., ve Iseri, A. A. 2016. "A data analytic approach to forecasting daily stock returns in an emerging market." *European Journal of Operational Research,* 253(3), 697-710.

Pabuçcu, H. 2019. "Borsa endeksi hareketlerinin makine öğrenme algoritmaları ile tahmini." *Uluslararası İktisadi ve İdari İncelemeler Dergisi,* 23, 179-190.

Provost, F. ve Fawcett, T. 2013. "Data science and its relationship to big data and data-driven decision making." *Big Data*, 1(1), 51-59.

Şahin, N. 2019. "*Yapay sinir ağları ve karar ağaçları modelleri ile işletmelerin finansal başarısızlıklarının tahminlenmesi*." Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü: İstanbul.

Scholkopf, B. 2001. "The kernel trick for distances." In *Advances in Neural Information Processing Systems*, 13, 301-307.

Silahtaroğlu, G. 2008. *Veri Madenciliği*, İstanbul: Papatya Yayınları.

Song, H. 2018. *Autofe: efficient and robust automated feature engineering*. Doctoral Dissertation, Massachusetts Institute of Technology, Massachusetts.

Teixeira, L. A. ve De Oliveira, A. L. I. 2010. "A method for automatic stock trading combining technical analysis and nearest neighbor classification." *Expert Systems with Applications*, 37(10), 6885-6890.

Yakut, E., Elmas, D. ve Yavuz, Y. 2014 "Yapay sinir ağları ve destek vektör makineleri yöntemleriyle borsa endeksi tahmini." *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 19(1), 139-157.

Yiğit, Ö. E., Alp, S. ve Öz, E. 2020. "Prediction of bist price indices: a comparative study between traditional and deep learning methods". *Sigma Journal of Engineering and Natural Sciences,* 38(4), 1693-1704.

Zheng, A. ve Casari, A. 2018. *Feature engineering for machine learning: principles and techniques for data scientists.* O'Reilly Media, Inc.,USA.