İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 19(38), Güz 2020, 224-233.

İSTANBUL TİCARET ÜNİVERSİTESİ
FEN BİLİMLERİ DERGİSİ

*Istanbul Commerce University Journal of Science*

http://dergipark.gov.tr/ticaretfbd

*Araştırma Makalesi / Research Article*

# İŞLETMELERİN İFLAS TAHMİNİNDE K- EN YAKIN KOMŞU ALGORİTMASI ÜZERİNDEN UZAKLIK ÖLÇÜTLERİNİN KARŞILAŞTIRILMASI*

## COMPARISON STUDY OF DISTANCE MEASURES USING K- NEAREST NEIGHBOR ALGORITHM ON BANKRUPTCY PREDICTION

**Gizem DİLKİ**[1] **Özlem DENİZ BAŞAR**[2]

*Sorumlu Yazar / Corresponding Author*
*gizemdilki2@gmail.com*

*Geliş Tarihi / Received*
*20.11.2020*

*Kabul Tarihi / Accepted*
*03.12.2020*

## Öz

Makine öğrenmesi biyoteknoloji alanından eğitim bilimlerine, doğal dil işlemeden duygu analizine kadar medikal, eğitim, işletme gibi birçok alanda aktif olarak kullanılan bir disiplindir. Kullanım alanı genişledikçe regresyon, sınıflama, kümeleme gibi farklı problemlere çözüm arayan makine öğrenmesi, iflas tahmini probleminde de kullanılmaya başlamıştır. Makine öğrenmesi disiplininde algoritma sayısı arttıkça, parametreler değiştikçe farklı doğruluk oranlarına ulaşmak mümkündür. Bu amaçla, çalışmada k En Yakın Komşu algoritmasına yer verilmiş; farklı uzaklık ölçütleri (Euclidean, Manhattan, Chebysev, Minkowski) kullanılarak yapılan sınıflandırma işlemi sonucunda en yüksek doğruluk oranına sahip uzaklık ölçütü belirlenmiştir. Veri seti %70 eğitim- %30 test seti olarak bölünmüş; çeşitli performans ölçütleri kullanılarak algoritmalar birbiriyle karşılaştırılmıştır.

**Anahtar Kelimeler:** İflas tahmini, k-en yakın komşu, makine öğrenmesi, uzaklık ölçütleri.

## Abstract

Machine learning is a discipline that is actively used in many areas such as medical, education and business management, from biotechnology to educational science, natural language processing to emotion analysis. As the area of use expanded, machine learning, which was looking for solutions to different problems such as regression, classing and clustering, also started to be used in the problem of bankruptcy prediction. As the number of algorithms increases in machine learning discipline, it is possible to achieve different accuracy rates as parameters change. For this purpose, the k Nearest Neighbor algorithm was involved in our study and the distance measure with the best accuracy were determined as a result of the classification process using different distance measures (Euclidean, Manhattan, Chebysev, Minkowski). The data set is divided into 70% training - 30% test; algorithms are compared using various performance criteria.

**Keywords:** Bankruptcy prediction, distance measures, k-nearest neighbor, machine learning.

*Bu çalışma, İstanbul Ticaret Üniversitesi Fen Bilimleri Enstitüsü'nde yapılan "İŞLETMELERİN İFLAS TAHMİNİNDE MAKİNE ÖĞRENMESİ ALGORİTMALARININ KARŞILAŞTIRMALI ANALİZİ" başlıklı yüksek lisans tezinden hazırlanmıştır.

[1]İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Anabilim Dalı, Küçükyalı, İstanbul, Türkiye.
gizemdilki2@gmail.com, Orcid.org/0000-0002-2316-8928.

[2]İstanbul Ticaret Üniversitesi, İnsan ve Toplum Bilimleri Fakültesi, İstatistik Bölümü, Sütlüce, İstanbul, Türkiye.
odeniz@ticaret.edu.tr, Orcid.org/0000-0002-9430-8975.

## 1. GİRİŞ

Teknolojinin hızla gelişmesi ve bilgisayar programlarının kullanımının yaygınlaşması ile birlikte birçok gerçek hayat problemine farklı çözümler aranmaya başlanmıştır. Bu problemleri çözmeye yardımcı dallardan biri makine öğrenmesidir. Makine öğrenmesi algoritmalarına, sağlık alanında çeşitli kanser hastalıklarının tanısında, Alzheimer gibi ölçümlemesi zor hastalıklarda, kronik rahatsızlarda ve kalp krizi gibi ani gerçekleşen vakalarda, pazarlama alanında müşteri sadakati ölçmede, bankacılık sektöründe kredi skoru belirlemede, sosyal medya alanında duygu analizleri çalışmalarında rastlanmaktadır. Algoritmalar, kullanıldığı veri tipine göre farklı sonuçlar üretmektedir. Bu nedenle, algoritmalar arası çeşitlilik sağlandıkça birden fazla yöntem kullanma, algoritma içi farklı metotlar kullanma ve bu yöntemleri karşılaştırma yoluna gidilmiştir.

Makine öğrenmesi algoritmalarının kullanıldığı problemlerden biri de iflas tahminidir. Şirketlerin iflas potansiyellerini ölçümlemeleri bu sürece girmelerini önlemeye yardımcı olabilir. Bu nedenle, tehditleri doğru zamanda fark etmek, değerlendirmek ve aksiyon almak için analitik faaliyetlerini düzenli olarak gerçekleştiren, riskliliklerini kontrol eden işletmelerde iflas süreçleri önceden tespit edilebilmekte ve iflas olasılığını azaltabilmek için makin öğrenmesi algoritmaları önem arz etmektedir.

## 2. MAKİNE ÖĞRENMESİ

Bilgisayarların hayatımıza girmesiyle birlikte işleri daha kısa sürede tamamlayabilen, objektif kararlar verebilen, beslenen fonksiyon dahilinde hata yapmayan, yorulmayan makineler ortaya çıkmıştır. Bu makinelerin, bahsi geçen özelliklere sahip olabilmeleri için arka planda bir program çalışması ve istenen fonksiyonların programa insan gücüyle öğretilmesi bilgisayar bilimcilerinin en temel uğraşlarından biri olmuştur. 1950'li yıllarda Alan Turing tarafından "düşünebilen makineler", "kendi öğrenebilen makineler" tanımları ortaya atılmış, böylelikle bugünkü yapay zeka kavramının temelleri oluşmuştur. Yapay zeka, en yalın haliyle, hedef odaklı akıllı makineler, özellikle de bilgisayar programları geliştirmek için türetilmiş bir bilim ve mühendislik dalı şeklinde tanımlanabilir. Zaman içerisinde birçok alt dala ayrılan yapay zekanın bilgisayar bilimi ve istatistiği birleştiren dalı makine öğrenmesi olarak anılır. Makine öğrenmesi, örnek verileri veya geçmiş deneyimleri kullanarak bir performans ölçütü optimize etmek için bilgisayarları programlamaktır (Alpaydin, 2010). Bilgisayar biliminin "sorunlara çözüm getirecek program üretme" güdüsüyle, istatistik biliminin "eldeki verilerden çıkarım yapma" güdüsünü birleştiren makine öğrenmesi, yeni bir disiplin oluşturarak daha etkin sistemler yaratmayı amaçlamaktadır (Mitchell, 2006).

Makine öğrenmesi kavramı, zaman içerisinde daha sık kullanılmaya başlanmasıyla birlikte literatüre bir takım yeni alt başlıklar katmıştır. Kendi içerisinde farklı öğrenme türlerine göre denetimli öğrenme, denetimsiz öğrenme, yarı-denetimli öğrenme ve takviyeli öğrenme olarak ayrılmaktadır. Denetimli öğrenme, girdi değişkenlerinin ve sonuç değişkenlerinin algoritmaya aynı anda öğretildiği öğrenme şekli olarak açıklanabilir. Algoritmanın yapması gereken sadece girdiden çıktıya ulaşmak için gereken adımlar veya süreç üzerinde çalışmaktır. Burada eğitim setinin görevi, algoritmayı doğru şekilde yönlendirmektir.

Denetimsiz öğrenmenin denetimli öğrenmeden en büyük farkı, denetimli öğrenmede her girdi karşılığında bir çıktı bulunması, denetimsiz öğrenmede ise sadece girdilerin bulunup sistemin bu girdileri kendi içinde yorumlayarak bir algoritma ortaya çıkarmaya çalışmasıdır. Denetimli öğrenme, eğitim setindeki farklı sınıflar arasında doğrudan ayrıştırma yapmaya odaklıdır ve yeterli miktarda veri varsa, eğitim setinde tahmin edilen sınıflandırma hatası, bilinmeyen verilerde de tahmin edilebilir (Hastie vd., 2008). Denetimli ve denetimsiz öğrenmenin amaçları

arasındaki temel fark, denetimsiz öğrenmenin, sadece, eğitim verilerini mümkün olduğunda temsil etmeye odaklanmasıdır. Denetimsiz bir şekilde kümeleme gerçekleştirildiyse ve daha sonra bulunan kümelere dayalı bir sınıflandırıcı oluşturulmak isteniyorsa, bu sınıflandırıcının iyi performans göstermesini beklenir. Kesin bir referansa erişim bulunmadığından, sınıflandırıcıyı doğrulanamaz ya da performansı başka bir şekilde tahmin edilemez. Bu nedenle, kümeleme işleminin sonuçlarına güvenilebilmesi için, eğitim verilerini temsil etmek üzere kullanılan modelin doğru olması çok önemlidir (Weber, 2000).

Denetimli öğrenme algoritmalarında, meydana gelebilecek tüm durumları içeren eğitim setine sahip olmak oldukça maliyetlidir. Denetimsiz öğrenme algoritmalarında ise veri seti genelde çok geniş ve yorumlanması zordur. Yarı denetimli öğrenme, bu iki algoritmayı birleştirerek, büyük ölçekli gözlemlerin yalnızca küçük bir alt kümesine karşılık gelen sınıf etiketleri olduğunda sınıflandırma sorununu göz önünde bulundurur. Yarı denetimli öğrenme, denetimsiz öğrenmeye göre daha az açıklamalı çabayla daha yüksek doğruluklar vaat eder. Takviyeli öğrenme ise, diğer öğrenme çeşitlerinden farklı olarak bir ödül mekanizması ile çalışır. Bu tip öğrenmede, bir ajan (agent) kullanılır ve ajan sisteme çeşitli girdilerde bulunur. Sistem gözlemlerini geri besleme mantığı ile makineye tekrar iletir ve bunun karşılığında bir ödül verir (Kotsiantis, 2007). Takviyeli öğrenme istatistiğin karar teorisi ile mühendisliğin kontrol teorisinin birleşimi bir öğrenme çeşididir. Ajanın amacı, aldığı ödülü maksimum düzeye çıkarmak; cezayı en aza indirmek üzerine kurulur ve bu amaçla aldığı her aksiyonu sisteme iletir. Denetimli öğrenmede ajana, verilen duruma nasıl tepki verileceğini öğretilir, takviyeli öğrenmede ise ajana nasıl tepki verileceğini öğretmez, daha çok "özgür seçim" yapma öğretilir.

## 2.1. İflas Tahmini

İflas, alacaklılara olan borçlarını ödenemediği durumlarda ortaya çıkan hukuki süreç olarak tanımlanabilir. İflas, gerçekleşme şeklinde göre ikiye ayrılır; ilk tip iflas, firmanın öz sermaye (toplam borç ile alacaklar arasındaki fark) durumuna göre değerlendirilir. İkinci tip iflas ise, şirketin resmi makamlara başvuru yapması üzerine, mal varlığını tasfiye etmek ya da bir kurtarma sürecine atılması olarak düşünülebilir (Altman ve Hotchkiss, 2006).

Firmalar, varlıklarını devam ettirebilmek ve piyasada uzun dönemler boyunca yer edinebilmek amacıyla birtakım stratejik çalışmalar yapmaktadır. Çalışmaların büyük bir bölümünü firmaların yaşam süresini direkt etkileyecek konulardan biri olan finansal konular oluşturmaktadır. Şirketlerin başarılı olmak ve uzun vadeli süreklilikleri sağlamak için mali durumlarını sürekli olarak değerlendirmesi gerekmektedir.

Organizasyonların iflas durumu ile ilgili ilk çalışmalar 1930'lu yıllarda başlamıştır. Bu dönemde yapılan çalışmalarda amaç, şirketlerin finansal oranlarını analiz ederek gidişat hakkında bilgi sahibi olmak; ilerleyen yıllarda ortaya çıkabilecek sorunları öngörebilmektedir. Teorik modeller olarak literatüre geçen bu çalışmalar genelde iflasın nedenini araştırmak yerine oranların şirketin finansal sağlığı konusunda verdiği cevaplara odaklanmaktır (Thian Cheng Lim vd., 2012).

Teorik modellerin ardından 1967'de Beaver'in öncü çalışması tek değişkenli diskriminant analizi ile istatistiksel tahmin yöntemleri kullanılmaya başlamıştır. 1968'de Altman'ın çok değişkenli diskriminant analizi çalışması ve geliştirdiği Altman Z Skor ile çalışmalar hız kazanmıştır (Balcaen ve Ooghe, 2006). 1980'de Ohlson'ın Logit modeli çalışmalarında kullanmasıyla model çeşitliliği artmıştır (Back vd., 1996).

İstatistiksel tahmin yöntemleri ile olumlu sonuçlar alınsa da bu tahmin yöntemlerinin doğrusallık, normallik, bağımsızlık gibi birtakım varsayımlarla ilgili dezavantajları bulunmaktadır. Bu nedenle son yirmi yılda iflas tahmin yöntemlerinin popülaritesi istatistiksel

yöntemlerden sinir ağları, genetik algoritmalar, destek vektör makineleri, bulanık mantık gibi makine öğrenmesine dayalı yöntemlere kaymıştır (Korol, 2019).

## 2.2. k En Yakın Komşu

k En Yakın Komşu (kNN) algoritması, ilk olarak 1950'lerin başında ortaya atılmıştır. Bu yöntemde, büyük eğitim setleri verildiğinde öğrenme işleminde oldukça zaman kaybedilmektedir. Bu nedenle, bilgi işlem gücü kullanılabilir hale gelene kadar popülerlik kazanmamıştır (Han vd., 2011). 1960'lardan sonra ise, 1965'te N.J. Nilsson tarafından hazırlanan minimum uzaklık sınıflayıcı üzerine çalışmalarla geliştirilmiş; 1967'de T. Cover ve P. Hart'ın sunduğu "Yakın Komşular Örüntü Sınıflama" çalışmalarıyla netlik kazanmıştır (Hu vd., 2016).

Denetimli öğrenme yöntemlerinden biri olan k En Yakın Komşu algoritması hem sınıflama hem de regresyon ayağında kullanılabilen çok yönlü bir algoritmadır. En basit haliyle tanımlayacak olursak, sınıfı bilinmeyen veri, eğitim setindeki diğer veriler ile karşılaştırılır ve bir uzaklık ölçümü yapılır. Hesaplanan uzaklığa göre henüz bir sınıfa atanamamış veriye en optimal sınıf bulunur.

Hemen hemen her sınıflandırma modeli kendi içinde bir artık sınıflayıcı oluşturur ve gelen her yeni veride bu sınıflayıcı kullanılır. kNN algoritmasında ise bu tür bir artık sınıflayıcı bulunmaz, bunun yerine gelen her yeni örnek için en yakın komşu kümesi tekrardan aranır. En Yakın Komşu sınıflandırma yönteminde, önceden hiçbir sınıflandırıcı model oluşturulmadığı ve her yeni verinin sınıflandırılmasında ham eğitim verilerine geri dönüldüğünden, eğitim kümesi tamamı sınıflandırıcı olarak değerlendirilir. Bu özelliği bakımından tembel öğrenici olarak nitelendirilen k En Yakın Komşu algoritması, her bir örnekte tek tek tarama yaptığı için sınıflama süreci uzun olan bir algoritmadır (Khan vd., 2002). kNN algoritması, yeni verilerin hızla geldiği ve eğitim kümesinin hızla değiştiği durumlarda diğer algoritmalara göre daha iyi bir sınıflandırıcı olarak değerlendirilebilir.

kNN algoritmasında en önemli hususlardan biri optimal k sınıf değerini bulmaktır. Sınıf değeri k, önceden belirlenir. En uygun k değeri verilerin boyutuna ve yapısına bağlıdır; k=1'den gözlem sayısı n'e kadar sınıf yaratmak mümkündür. Sınıf değerini olması gerekenden büyük kullanmak, çok benzer olmayan verileri aynı gruba alacağından, sınıflamada doğruluk değerini aşağı çekecektir. Tam tersi gereğinden küçük bir k değeri kullanmak ise bazı olası sınıfları saf dışı bırakacaktır; bu durumda yine sınıf doğruluğu aşağı yönlü ivme kazanır.

kNN algoritmasında ana fikir, benzer nokta ya da değişken gruplarının büyük olasılıkla aynı sınıfa ait olmasıdır. Bu noktada, önceden seçilmiş bir mesafe ölçütü kullanılarak sınıfı bilinmeyen verinin yakınlığı hesaplanır. Mesafe hesaplamada en çok kullanılan uzaklık ölçüsü, Euclidean uzaklığıdır (Hu vd., 2016).

Euclidean uzaklığı, iki nokta arasında, $x_1 = (x_{11}, x_{12} \dots x_{1n})$ ve $x_2 = (x_{21}, x_{22} \dots x_{2n})$ olmak üzere,

$$dist_{euclidean}(x_1 x_2) = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x_{1i} - x_{2i}})^2} \quad (1)$$

Euclidean uzaklığı dışında Manhattan, Minkowski, Chebyshev gibi farklı uzaklık hesaplama ölçütleri de kullanılabilir (Prasath vd., 2019). İlgili uzaklık değerlerine ilişkin fonksiyonlar aşağıdaki eşitliklerde gösterilmiştir.

$$dist_{minkowski}(x_1x_2) = \sqrt{\sum_{i=1}^{n}|x_{1i} - x_{2i}|^2} \quad (2)$$

$$dist_{manhattan}(X_1X_2) = \sum_{i=1}^{n}|x_{1i} - x_{2i}| \quad (3)$$

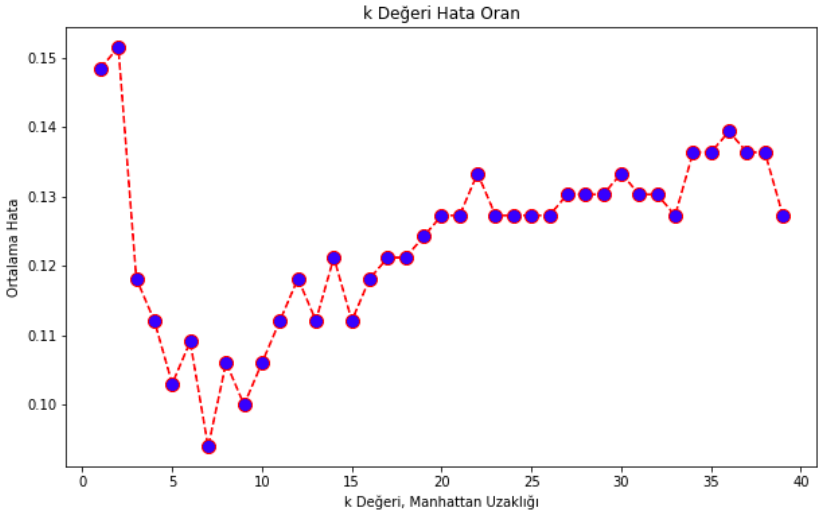
$$dist_{chebyshev}(X_1X_2) = \max_i|x_{1i} - x_{2i}| \quad (4)$$

İdeal olarak, kNN sınıflandırması için mesafe ölçümü çözülmekte olan soruna uyarlanmalıdır. Farklı uzaklık ölçütleri kullanılarak kNN algoritması özelinde daha doğru sınıflandırmalar yapılabildiğini gösteren çalışmalar mevcuttur (Weinberger vd.,2006) .

## 3. VERİ, ANALİZ ve BULGULAR

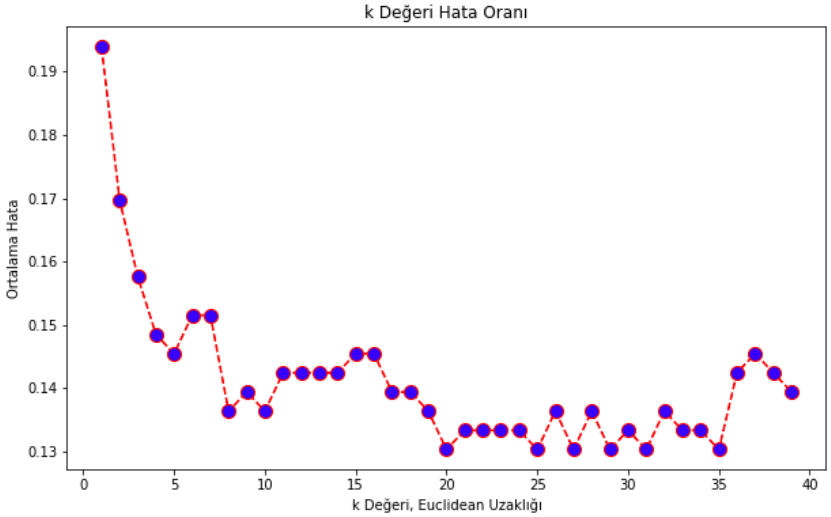
Sınıflandırma problemleri, denetimli öğrenme algoritmalarıyla çözülmektedir. Denetimli öğrenmede daha önce değinildiği gibi bir eğitim ve bir test sınıfı bulunmaktadır. Veri boyutu yeterliyse veri seti eğitim, test ve doğrulama setleri olarak ayrılabilir. Bu çalışmada, Kaliforniya Üniversitesi veri kütüphanesinde (UCI) bulunan Tayvan İflas Tahmini adlı veri setinden yararlanılmıştır. 1999-2009 yıllarını kapsayan veriler, Tayvan Ekonomi Dergisinden toplanmıştır. Yayınlanan verilerden elde edilen ana kütlede firmaların Tayvan Menkul Kıymetler Borsası'nda faaliyet gösteren firmalar olması kısıtı gözetilmiştir. İkinci olarak, firmaların iflas durumunun ortaya çıkmasından 3 yıl öncesine kadar kamuya açık şekilde finansal bilgilerini paylaşıyor olması gerekmektedir. Bu kısıtlara göre hazırlanan veri setinde benzer büyüklükte ve benzer sektörlerde (lojistik, hizmet, finans, imalat) faaliyet gösteren firmalar analize konu olmuştur (Liang vd., 2016). 1100 satır, 96 değişkenden oluşan veri seti içinde borçluluk, sermaye yeterliliği, karlılık, gelir, nakit akışı ve büyüme rasyoları bulunmaktadır. İflas etmiş sınıfa ait 220 veri bulunan veri seti, %70 eğitim verisi- %30 test verisi olarak ayrılmıştır.

Veriye kNN algoritması uygulanırken öncelikle optimum k değerinin bulunması gerekmektedir. Bu amaçla Python (3.7) üzerinde seçilen uzaklık ölçütüne göre k değerine karşılık ortalama hatayı hesaplayan bir döngü oluşturularak optimum k değerine aşağıdaki grafikler üzerinden karar verilmiştir.



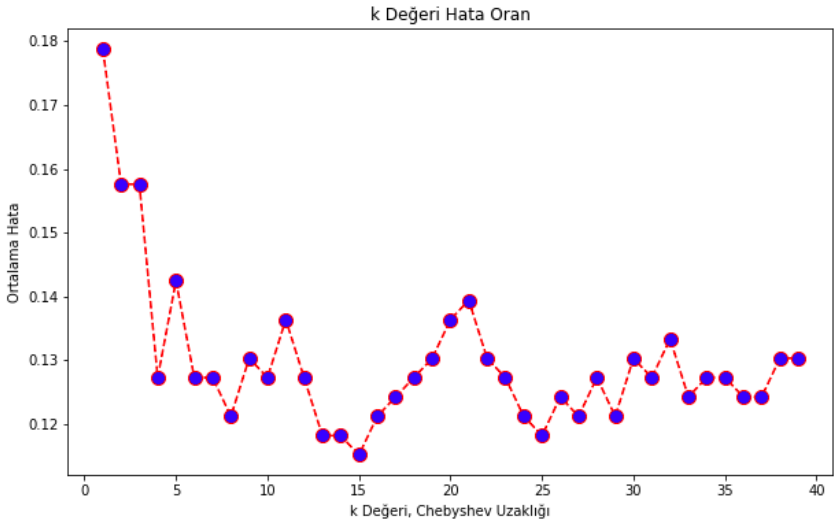

Şekil 1. Manhattan Uzaklığı İçin Optimal k Değeri

Şekil 1'de Manhattan uzaklığı için optimal k değeri verilmiştir. Gözlem sayısı 1'den n'e kadar iki nokta arasında farkın mutlak değerinin toplamı alınarak hesaplanan Manhattan uzaklığına göre optimal k değeri 7 olarak belirlenmiştir.



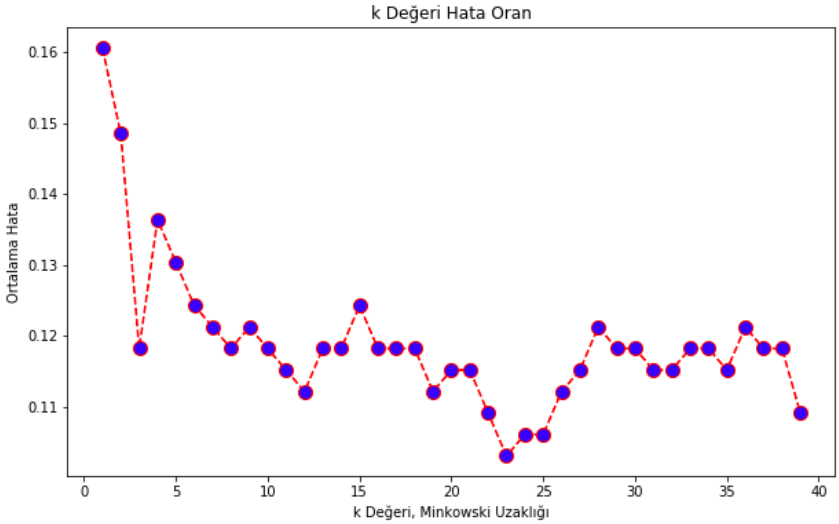

Şekil 2. Euclidean Uzaklığı İçin Optimal k Değeri

Şekil 2'de Euclidean uzaklığı için optimal k değeri belirlenmiştir. Kısaca iki nokta arasındaki farkın kareleri toplamının karekökü olarak ifade edilebilen Euclidean uzaklığına göre optimal k değeri çalışmaya göre 25 kabul edilir.




Şekil 3. Chebyshev Uzaklığı İçin Optimal k Değeri

Şekil 3'te Chebyshev uzaklığı için hesaplanan optimal k değeri verilmiştir. İki nokta arasındaki mutlak değer uzaklığın maksimum değerine göre hesaplanan bu uzaklık için optimal k değeri 15 olarak hesaplanmıştır.




Şekil 4. Minkowski Uzaklığı İçin Optimal k Değeri

Dördüncü uzaklık değeri olarak seçilen Minkowski uzaklığı için sonuçlar Şekil 4'te verilmiştir. Hesaplamada iki nokta arasındaki mutlak değer farkın kareleri toplamının karekökü dikkate alınan bu uzaklık ölçütü için optimal k değeri 23 olarak belirlenmiştir.

Algoritmaların performanslarını karşılaştırmak amacıyla Tablo 1'de verilen tahminlerin niteliklerini gösteren karmaşıklık matrisinden yararlanılarak Doğruluk, Keskinlik, Duyarlılık $F_1$ Puanı değerleri ölçüt olarak belirlenmiştir.

Tablo 1. Karmaşıklık Matrisi

| | | Tahmini Sınıf | |
|---|---|---|---|
| | | İflas Etti | İflas Etmedi |
| Gerçek Sınıf | İflas Etti | TP | FP |
| | İflas Etmedi | FN | TN |

Doğruluk, doğru tahmin edilen sınıfların tüm veriye oranıdır. Keskinlik, tüm pozitif tahminler içinde gerçek pozitif durumların tahmin edilme oranını göstermektedir. Hassaslık, tüm veri seti içinde pozitif durumların tahmin edilme oranını ifade eder. $F_1$ puanı ise Keskinlik ve Duyarlılık değerlerinin harmonik ortalamasıdır (Bulut ve Osmani, 2017).

$$\text{Keskinlik} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \quad (6)$$

$$F_1 \text{ Puanı} = 2\frac{P\ X\ R}{R+R} \quad (7)$$

Tablo 2: Model Performansları

| | **Euclidean Uzaklığı, k=25** | | | | **Manhattan uzaklığı, k=7** | | | |
|---|---|---|---|---|---|---|---|---|
| | Keskinlik | Duyalılık | $F_1$ Puanı | Doğruluk | Keskinlik | Duyalılık | $F_1$ Puanı | Doğruluk |
| **İflas Etti** | 0,89 | 0,96 | 0,92 | | 0,90 | 0,96 | 0,93 | |
| **İflas Etmedi** | 0,71 | 0,49 | 0,58 | | 0,80 | 0,62 | 0,70 | |
| **Ortalama** | 0,86 | 0,87 | 0,86 | 0,87 | 0,88 | 0,88 | 0,88 | 0,88 |
| | **Minkowski Uzaklığı, k=23** | | | | **Chebyshev Uzaklığı, k=15** | | | |
| | Keskinlik | Duyalılık | $F_1$ Puanı | Doğruluk | Keskinlik | Duyalılık | $F_1$ Puanı | Doğruluk |
| **İflas Etti** | 0,91 | 0,67 | 0,94 | | 0,85 | 0,97 | 0,9 | |
| **İflas Etmedi** | 0,83 | 0,64 | 0,72 | | 0,71 | 0,33 | 0,45 | |
| **Ortalama** | 0,89 | 0,90 | 0,89 | 0,90 | 0,82 | 0,84 | 0,81 | 0,84 |

Modellere ait performans sonuçlarına Tablo 2'de yer verilmiştir. Euclidean uzaklığı kullanılan kNN algoritmasına göre, iflas edecek firmaların %96'sı, iflas etmeyecek firmaların da %49'u doğru tahmin edilmiştir. Gerçekte iflas eden firmalar arasından, sınıflama sonucu iflas edeceği belirlenen firmalardan %89'u, iflas etmeyecek firmalar arasından %71'i doğru tahmin edilmiştir. Modelin geneli ele alındığında doğru sınıflandırma oranı %87 kabul edilir.

Manhattan uzaklığı kullanılarak gerçekleştirilen kNN sınıflamasında, tüm test setinde yer alan firmalardan iflas edeceği tahmin edilen firmalardan %96'sı gerçekten iflas etmiştir. Sınıflandırma işlemi sonucu, gerçekte iflas eden firmalar arasından %80'i doğru olarak tespit edilmiştir. Algoritmanın genel sınıflandırma doğruluğu ise %88 oranındadır.

Minkowski uzaklığı ile firmanın iflas ettiği yönünde yapılan tahmin %67 doğruluk oranına sahiptir. Test veri setinde ürünü satın almış kişilerin sınıflandırma işlemi sonucunda aynı şekilde ürünü satın aldığı %91'i doğru tahmin edilmiştir. Sınıflandırma işleminin genel doğruluk oranı ise %90 olarak belirlenmiştir.

Chebyshev uzaklığının uygulandığı kNN sınıflandırmasında ise iflas edeceği tahmin edilen firmalardan %97'si gerçekten iflas etmiştir. Chebyshev-kNN analizi sonucunda, gerçekte iflas eden firmalar arasından %71'i doğru olarak tespit edilmiştir. Algoritmanın genel sınıflandırma doğruluğu ise %84 oranındadır.

## 4. SONUÇ

İstatistik ve bilgisayar bilimlerinin birleşimiyle birçok alana katma değer sağlayan makine öğrenmesi, bu çalışmada firmaların iflas edip etmeyeceğini sınıflandırmak amacıyla kullanılmıştır. Firmaların finansal sağlıklarını düzenli olarak kontrol ederek iflas olasılığını ölçümlemeleri gelecek yıllarda düşecekleri zor durumu önceden görmeleri ve önem almalarına yardımcı olur. Aynı zamanda, iflas tehdidini önceden belirlemek sadece iflasın önüne geçmek için değil, aynı zamanda şirketin durumunu iyileştirmek amacıyla stratejik çözümler bulmak için de bir itici güç olabilmektedir.

Çalışmada iflas tahmini bir sınıflama problemi olarak ele alınmıştır. Sınıflandırma problemini çözümlemek amacıyla, denetimli öğrenme algoritmalarından o tembel öğrenici k En Yakın komşu algoritması kullanılmıştır. k En Yakın Komşu, kullanılan uzaklık ölçütüne göre farklı doğruluk oranları sağlayabilmektedir. Bu amaçla, Euclidean, Manhattan, Minkowski ve Chebysev uzaklıkları kullanılarak algoritmanın veri seti üzerinde sınıflama yapması sağlanmıştır. Uygulama sürecinde uzaklık ölçütlerine göre farklı optimal k değerleri elde edilmiştir.

Algoritmaların performansları karşılaştırıldığında, k En Yakın Komşu algoritması için en iyi tahminlemeyi yapan uzaklık ölçütü %90 doğrulu k ile Minkowski uzaklığı olmuştur. Seçilen dört uzaklık ölçütü arasından görece en düşük doğruluk oranına sahip uzaklık ölçütü ise %84 ile Chebyshev uzaklığı olmuştur. Finansal metriklerini doğru biçimde kayıt altına alan şirketler için iflas etme potansiyellerini ölçümlemek amacıyla yapılan analizde, ileriki yıllarda çalışma seti boyutu arttırılarak daha yüksek doğruluk oranları ile sınıflandırma yapmak mümkün olabilir.

## KAYNAKÇA

**Alpaydın, E.,** (2010), Introduction to Machine Learning (2nd ed). MIT Press, London.

**Altman, E.I., Hotchkiss, E.,** (2006), Corporate Financial Distress and Bankruptcy (Third Edition), John Wiley & Sons, New Jersey.

**Back, B., Latinen, T., Sere, K., Wezel, M.,** (1996), Choosing Bankruptcy Predictors Using Discriminant Analysis, Logit Analysis, and Genetic Algorithms. 20, Turku Centre for Computer Science Technical Report, No 40.

**Balcaen, S., Ooghe, H.,** (2006), "35 Years of Studies on Business Failure: An Overview of the Classic Statistical Methodologies and Their Related Problems". The British Accounting Review, 38(1), 63-93.

**Bulut, F., Osmani, S.,** (2017), "Scene Change Detection Using Different Color Pallets and Performance Comparison". Balkan Journal of Electrical and Computer Engineering, 66-72.

**Han, J., Kamber, M., Pei, J.,** (2011), Data Mining Concepts and Techniques Third Edition (Third Edition). Morgan Kaufmann, Massachusetts.

**Hastie, T., Friedman, J., Tibshirani, R.,** (2008), Unsupervised Learning. In: The Elements of Statistical Learning, Springer Series in Statistics. Springer, New York.

**Hu, L.Y., Huang, M.W., Ke, S.W., Tsai, C.F.,** (2016), "The Distance Function Effect on k-Nearest Neighbor Classification for Medical Datasets", Springer Plus, 5(1), 1-9.

**Khan, M., Ding, Q., Perrizo, W.,** (2002), "K-nearest Neighbor Classification on Spatial Data Streams Using P-trees", Advances in Knowledge Discovery and Data Mining, 2336, 517-528.

**Korol, T.,** (2019), "Dynamic Bankruptcy Prediction Models for European Enterprises". Journal of Risk and Financial Management, 12(4), 1-15.

**Kotsiantis, S. B.,** (2007), "Supervised Machine Learning: A Review of Classification Techniques", Informatica, 31, 249-268.

**Liang, D., Lu, C.C., Tsai, C.F., Shih, G.A.,** (2016), "Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A comprehensive Study", European Journal of Operational Research, 252(2), 561-572.

**Mitchell, T.M.,** (2006), The Discipline of Machine Learning, 10.11.2020, http://ra.adm.cs.cmu.edu/anon/usr0/ftp/anon/ml/CMU-ML-06-108.pdf

**Prasath, V.B.S., Alfeilat, H.A.A., Hassanat, A.B.A., Lasassmeh, O., Tarawneh, A.S., Alhasanat, M.B., Salman, H.S.E.,** (2019), "Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier—A Review", Big Data, 7(4), 221-248.

**Lim, T.C., Lim Xiu Yun, J., Siwei, G., Jiang, H.,** (2012), "Bankruptcy Prediction: Theoretical Framework Proposal", International Journal of Management Sciences and Business Research, 1(9), 69-74.

**UCI,** (2020), 25.08.2020, https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction.

**Weber, M.,** (2000), Unsupervised Learning of Models for Object Recognition, California Institute of Technology, Pasadena, California.

**Weinberger, K.Q., Blitzer, J., Saul, L.K.,** (2006), Distance Metric Learning for Large Margin Nearest Neighbor Classification, 20.11.2020, https://papers.nips.cc/paper/2005/file/a7f592cef8b130a6967a90617db5681b-Paper.pdf