GUFBD / *GUJS* (2022) 12(1): 370-379
DOI: 10.17714/gumusfenbil.892253 Araştırma Makalesi / *Research Article*

Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasına yönelik tahmininin destek vektör makineleri ve k-en yakın komşu algoritmaları ile gerçekleştirilmesi

Classification in the change of estimated number of Covid-19 daily cases by using support vector machine and k-nearest neighbor algorithm

Enes FİLİZ*[1,a]
[1]*Fırat Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 23000, Elazığ*

• Geliş tarihi / *Received*: 06.03.2021 • Düzeltilerek geliş tarihi / *Received in revised form*: 25.12.2021 • Kabul tarihi / *Accepted*: 01.01.2022

Öz
Covid-19 virüsü hayatımıza girdiği Aralık 2019'dan bu yana etkinliğini kaybetmeden tüm dünyayı etkilemeye devam etmektedir. Dünya sağlık örgütünün önerileri, ülkelerin kendi bünyelerinde aldıkları tedbirler ve aşı çalışmaları virüsün üstesinden gelmek için büyük önem arz etmektedir. Bu bağlamda birçok bilimsel çalışma virüsün geleceği için değerli bilgiler ortaya koymuştur. Çalışmada Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasına yönelik tahminler destek vektör makinesi ve k-en yakın komşu algoritmaları ile yapılmıştır. Günlük hasta sayısının sınıflandırılmasının tahmininde etkin rol oynayan öznitelikler 'pozitif çıkma oranı', 'filyasyon oranı', 'işyerleri hareketliliği' ve 'parklardaki hareketlilik' olarak belirlenmiştir. Bu etkin öznitelikler yardımıyla yapılan günlük hasta sayısının sınıflandırılması tahmininde de k-en yakın komşu algoritmasının (%84.7) en başarılı algoritma olduğu gözlemlenmiştir.

Anahtar kelimeler: Covid-19, Hasta sayısı, Makine öğrenmesi, Öznitelik seçimi, Sınıflandırma

Abstract
Since December 2019, the Covid-19 virus afftected our lives and continues to affect the whole world significantly. The investigistion of the indicators of the Covid-19 virus and vaccination studies are of great interest to overcome the Covid-19 pandemic based on the World health organization recommendations. In this context, many scientific studies have revealed valuable information for the future of the virus. In this study, estimation of the cOvid-19 cases and Classification of changes in the daily number of cases in Turkey was carried out by using support vector machine and k-nearest neighbor algorithms. The indicators that play a critical role in the estimation of the daily patient number classification have been determined as "positivity rate", "fillation rate", "workplace mobility" and "mobility in parks". It has been observed that the k-nearest neighbor algorithm (84.7%) is the most successful algorithm in the estimation of the daily number of cases when considering the highlighted features.

Keywords*: Covid-19, Daily case, Machine learning, Feature selection, Classification*

[*a] Enes FİLİZ; efiliz@firat.edu.tr, Tel: (0424) 237 00 00, orcid.org/ 0000-0002-8006-9467

1. Giriş
1. Introduction

Koronavirüs ailesinin bir üyesi olarak 2019 yılı Aralık ayında Çin'in Wuhan eyaletinden hayatımıza giren Covid-19 virüsü Dünya Sağlık örgütü tarafından 11 Mart 2020 tarihinde salgın olarak ilan edilmiştir. Bu virüs geçen süre zarfına rağmen etkinliğini kaybetmeden insan hayatına hasar vermeye devam etmektedir. Son hızla süren aşı ve ilaç çalışmaları, en kısa sürede Covid-19 virüsünün üstesinden gelmeyi amaçlamaktadır. Bilimin ışığı ile bu virüsün yenileceği aşikârdır. Elde edilen başarılı denemeler insanlığın umudunu arttırmaktadır. Şu anda Covid-19 virüsüne karşı net bir çözüm olmamasına rağmen en önemli korunma yolu ülkelerin kendi durumları için aldıkları tedbirlere uymak, temizliğe dikkat etmek ve insan hareketliliğini en aza indirmektir. Bu doğrultuda bilimsel çalışmalar ön plana çıkmaktadır. Ülkelerin ve Dünya Sağlık Örgütünün ortaya koyduğu tablolardan yola çıkarak önemli bilgiye ulaşmak, analizler yapmak çözüme ulaşmada yardımcı olacaktır. Makine öğrenmesi yöntemleri de veriden anlamlı bilgiyi ortaya çıkarmak için kullanılan yöntemlerden biridir. Makine öğrenmesi, veri madenciliğinde kullanılan yöntemlerden birisidir. Büyük verilerden anlamlı bilgiler ortaya çıkarmak için kullanılan makine öğrenmesi algoritmaları, klasik istatistiksel yöntemlerden farklı olarak varsayım istemediğinden tercih edilmektedir. Makine öğrenmesi algoritmaları kümeleme, sınıflandırma ve tahmin çalışmalarında tatminkâr sonuçlar vermektedir. Ayrıca bünyesinde barındırdığı öznitelik seçim algoritmaları ile sınıflandırma başarısı için etkin değişkenleri belirlemeye yardımcı olmaktadır. Makine öğrenmesi algoritmaları sağlık, ekonomi ve birçok alanda uygulanabilirliği ile ön plana çıkmaktadır. İçinde bulunduğumuz salgın dönemi dolayısıyla ülkelerin sağlık sistemleri, alınan tedbirler ve bu durumlara yönelik ortaya çıkan sonuçlar ile ilgili yapılan çalışmalar salgının geçmişi, bugünü ve geleceği açısından büyük önem arz etmektedir.

Literatürde Covid-19 salgının başladığı ilk günden itibaren veriler analiz edilmeye başlanmış ve salgının gidişatı ile ilgili önemli sonuçlar elde edilmiştir. Ülkemizde ve Dünya genelinde birçok Covid-19 çalışması yapılmıştır. Bu çalışmalar genellikle epidemik modeller ve istatistiksel modeller üzerine olmuştur. De Felice ve Polimeni (2020), çalışmalarında Covid-19 araştırma eğilimlerini belirlemek için makine öğrenmesi yardımıyla bibliyometri analizi yapmışlardır. Kushwaha vd., (2020) Covid-19 salgınında makine öğrenmesi ile ilgili makaleleri incelemişler ve bu salgın krizini çözmek için makine öğrenmesinin önemini araştırmışlardır. Ülkemizde yapılan bir istatistiksel çalışmada ise Ayaz (2021) makine öğrenimi algoritmalarını kullanarak pozitif hastaların tespit edilebilmesi için tam kan sayımı sonuçlarından yararlanmıştır. Hatalığın daha önceden tespit edilebilmesi için tam kan sayım sonuçlarının kullanılabileceğini makine öğrenmesi algoritmaları ile göstermiştir. Ulaş (2021) virüsün yaklaşık üreme hızı tahmin ederek, belirli bir tarih aralığında bu virüsten kaç kişini etkilenebileceğini ve bu kişilerden ne kadarının aktif vaka olabileceğini tahmin etmiştir. Punn vd. (2020) Covid-19'un uluslararası toplumun refahı ile ilgili gelecekteki durumunu tahmin etmek için makine öğrenmesi ve derin öğrenme yöntemlerini kullanmışlardır. Barstugan vd. (2020) Covid-19'un erken teşhisindeki önemini belirlemek için farklı sınıflandırma kriterleri yardımıyla destek vektör makinelerinin performansını incelemişlerdir. Ardabili vd. (2020) Covid-19'un ülkeler arasındaki çeşitliliği de göz önünde bulundurarak salgını modellemek için makine öğrenmesinin uygun bir yöntem olduğunu belirtmişlerdir. Yadav vd. (2020) çalışmalarında Covid-19 salgınına yönelik belirledikleri 5 farklı durumu analiz etmek için destek vektör makineleri regresyon yönteminden yararlanmışlardır. Benzer şekilde Malki vd. (2020) farklı hava şartları faktörleri ile Covid-19'un yayılımı arasındaki ilişkiyi incelemek için makine öğrenmesi regresyon modellerini incelemişlerdir. Bazı çalışmalarda Covid-19'un epidemik gelişimin sınıflandırılması ve tahmini için makine öğrenmesi algoritmaları kullanılmıştır (Fanelli & Piazza, 2020; Wang vd., 2020). Literatürde makine öğrenmesi yardımıyla Covid-19 salgınının tarama, izleme, tahmin gibi durumları için kullanıldığı incelenmiştir (Lalmuanawma vd., 2020; Tuli vd., 2020).

Türkiye Cumhuriyeti Sağlık Bakanlığı'nın açıkladığı günlük koronavirüs bilgileri (T.C. Sağlık Bakanlığı Web Sayfası) ve Google Covid-19 Topluluk Hareketliliği Raporları (Google Haberler Web Sayfası-Google Covid-19 Community Mobility Reports) tarafından açıklanan insanların hareketliliğine dair veriler kullanılarak Türkiye'deki günlük hasta sayısındaki değişimlerin sınıflandırılmasının tahminini yapılacaktır. Çalışmadaki ilk amaç bu veri seti yardımıyla Türkiye'deki Covid-19 günlük hasta sayılarının değişimlerinin sınıflandırılmasının tahmin başarısında destek vektör makinesi puk çekirdeği (DVM-puk) ve k en yakın komşu (knn) algoritmasının performanslarının belirlenmesidir. İkinci olarak günlük hasta sayısının tahmininin sınıflandırma başarısında etkin rol oynayan

değişkenleri ortaya çıkararak algoritmaların sınıflandırma başarılarında değişiklik olup olmadığını incelemektir.

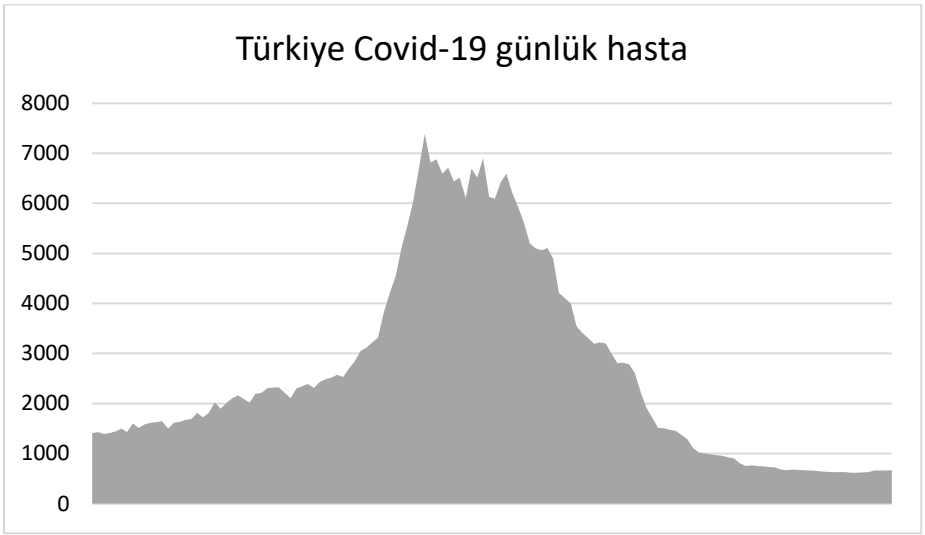
2. Materyal ve metot
2. Material and method

2.1. Veri seti
2.1. Data set

Çalışmada Türkiye Cumhuriyeti Sağlık Bakanlığı'nın açıkladığı günlük koronavirüs bilgileri ve Google Covid-19 Topluluk Hareketliliği Raporları tarafından açıklanan insanların hareketliliğine dair 28.09.2020 – 12.02.2021 tarihleri arasındaki veriler kullanılmıştır. Türkiye'de Covid-19'a bağlı günlük hasta sayısı bağımlı değişken olarak belirlenmiştir. Bağımsız değişkenler olarak, zatürre oranı, ağır hasta sayısı, günlük vefat sayısı, günlük test sayısı, günlük iyileşen hasta sayısı, yatak doluluk oranı, erişkin yoğun bakım doluluk oranı, ventilatör doluluk oranı, ortalama temaslı tespit süresi, filyasyon oranı, yapılan test sayısına göre pozitif çıkma oranından yararlanılmıştır. Ayrıca Google Covid-19 Topluluk Hareketliliği Raporları tarafından Türkiye için belirlenen insanların hareketliliği ile ilgili perakende ve rekrasyon (cafe, restoran, avm, müze, sinema vb.), market ve eczane (market, eczane, gıda deposu, pazar vb.), parklar (milli parklar, halk plajları, yat limanları, plazalar, halka açık parklar vb.), toplu taşıma istasyonları (metro, otobüs, tren istasyonları vb.), işyerleri ve konutlar değişkenleri belirlenen referans değerine göre çalışmaya dahil edilmiştir. Referans değeri ise Google Covid-19 Topluluk Hareketliliği Raporları tarafından 3 Ocak 2020 ile 6 Şubat 2020 tarihleri arasındaki 5 haftalık döneme ait ortanca değer olarak belirlenmiştir. Çalışmada kullanılan değişkenler Tablo 1'de verilmiştir. Ayrıca belirlenen tarihler arasındaki günlük hasta sayısındaki değişim Şekil 1'de gösterilmiştir.

Tablo 1. Çalışmada kullanılan değişkenler
Table 1. *Variables used in the study*

Bağımlı değişken	***Bağımsız değişkenler***	
• *Günlük hasta sayısı*	• *Zatürre oranı*	• *Filyasyon oranı*
	• *Ağır hasta sayısı*	• *Pozitif çıkma oranı*
	• *Günlük vefat sayısı*	• *Perakende ve rekrasyon*
	• *Günlük test sayısı*	• *Market ve eczane*
	• *Günlük iyileşen hasta sayısı*	• *Parklar*
	• *Yatak doluluk oranı*	• *Toplu taşıma istasyonları*
	• *Erişkin yoğun bakım doluluk oranı*	• *İşyerleri*
	• *Ventilatör doluluk oranı*	• *Konutlar*
	• *Ortalama temaslı tespit süresi*	



Şekil 1. Türkiye Covid-19 28.09.2020-12.02.2021 arasında günlük hasta sayıları grafiği
Figure 1. *The number of Daily Covid-19 patients between 28th September 2020 and 12th January 2021 in Turkey*

2.2. Sınıflandırma algoritmaları
2.2. Classification algorithms

Çalışmada Türkiye Cumhuriyeti Sağlık Bakanlığı'nın açıkladığı günlük koronavirüs bilgileri ve Google Covid-19 Topluluk Hareketliliği Raporları tarafından açıklanan insanların hareketliliğine dair veriler kullanılarak Türkiye'deki günlük hasta sayısının değişimin sınıflandırılmasının tahmini yapılacaktır. Bunun için literatürde sıklıkla kullanılan makine öğrenmesi algoritmalarından DVM-puk ve knn algoritması kullanılacaktır. Belirlenen algoritmalar ile ilgili açıklamalar aşağıda verilmiştir.

DVM-puk, veri madenciliği ve makine öğrenmesi algoritmaları arasında yer alan önemli bir yöntemdir. İlk olarak Cortez ve Vapnik tarafından geliştirilen destek vektör makineleri verileri iki kategoriye ayırmak için n boyutlu bir hiperdüzlem oluşturarak işlem yapmaktadır (Cortes & Vapnik, 1995; Haykin, 1999). Denetimli bir öğrenme olan bu algoritmada veriler doğrusal ayrılmışsa doğrusal destek vektör makineleri, doğrusal olarak ayrılmamışsa doğrusal olmayan destek vektör makineleri kullanılmaktadır (Shahiri vd., 2015; Alpaydın, 2004). Doğrusal olmayan destek vektör makineleri kullanımında farklı çekirdek fonksiyonlarından yararlanılmaktadır. Kullanılan çekirdek fonksiyonu seçimine göre farklı sonuçlar elde edilmektedir (Shawe-Taylor vd., 1998). Çalışmada Pearson VII fonksiyon tabanlı evrensel (DVM-puk) çekirdek fonksiyonu kullanılacaktır. Literatür incelendiğinde DVM-puk çekirdek fonksiyonun diğer çekirdek fonksiyonlarına göre daha başarılı sonuçlar verdiği ortaya konulmuştur (Kavzoğlu & Çölkesen, 2010; Abakar & Yu, 2014; Tuncer & Bolat). Ayrıca, DVM-puk çekirdeği, boyutlar arasında esnek bir geçişe sahiptir ve bu nedenle genel bir çekirdek fonksiyonu olarak kullanmak mümkün olmaktadır (Abakar & Yu, 2014). Bu durumlar göz önünde bulundurularak çalışmada DVM-puk çekirdek fonksiyonu tercih edilmiştir. DVM-puk çekirdek fonksiyonunun genel formülü aşağıdaki gibidir.

$$\frac{1}{\left[1+\left(\frac{2*\sqrt{\|x-y\|^2\sqrt{2^{\left(\frac{1}{\omega}\right)}-1}}}{\sigma}\right)^2\right]^{\omega}} \quad (1)$$

knn algoritması, bir veri setinde en yakın komşuları bulmayı amaçlar ve bu komşuları bulmak için farklı uzaklık ölçüleri kullanarak analiz yapar. Algoritmanın performansı bu uzaklık ölçülerine ve k parametresi ile ilgilidir (Liv d., 2003; Xia vd, 2015). Sınıflandırma, tahmin etme gibi analizlerde kullanılmaktadır. Kullanımı kolay, sınıflandırma performansı yüksek ve popüler bir algoritmadır. Sınıflandırma algoritmalarının performanslarının karşılaştırılmasında ön plana çıkmaktadır (Horton & Nakai, 1997; Zhang vd., 2017). Bu durumlar göz önünde bulundurularak çalışmada knn algoritması tercih edilmiştir. Genel formülü aşağıdaki gibidir;

$$y(x_i) = \begin{cases} \infty & , \quad eğer\ d(x_i, q) = 0\ ise \\ \frac{1}{d(x_i,q)} & , \quad aksi\ takdirde \end{cases} \quad (2)$$

Denklem de, xi uzaydan alınan rasgele bir x ∈ X örneğini, d komşular arasındaki uzaklığı, q ise komşu ile sınıfı belirlenmek istenen nokta arasındaki mesafenin tersini göstermektedir (Mitchell, 1997).

2.3. Öznitelik seçimi
2.3. Feature selection

Çalışmanın amacına yönelik olarak yapılacak işlemlerden biriside öznitelik seçimidir. Öznitelik seçimi, algoritmaların sınıflandırma başarısında etkin rol oynayan özniteliklerin belirlenmesidir. Bu yöntem ile daha az sayıda öznitelik yardımıyla sınıflandırma başarısından ödün vermemek amaçlanmaktadır. Yapılan öznitelik seçimi, zaman tasarrufu ve işlem kolaylığı sağlamaktadır. Analizlerde Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde etkin rol oynayan öznitelikler belirlenecektir. Bu öznitelikleri belirlemek için ReliefF öznitelik seçim algoritması kullanılacaktır. Geniş bir kullanım alanı olan, başarılı bir öznitelik seçim algoritmasıdır (Kononenko vd., 1997). Bu durum göz önünde bulundurularak çalışmada ReliefF öznitelik seçim algoritması tercih edilmiştir.

ReliefF öznitelik seçim algoritması, Kira ve Rendell (Kira & Rendell, 1992) tarafından formüle edilen bu öznitelik seçim algoritması özniteliklerin aralarında bağımlılık olup olmadığını incelemektedir (Urbanowicz vd., 2018). Özniteliklerin ait olduğu ve ağırlıklarının belirlenmediği sınıflardaki en yakın örnekleri karşılaştırır. ReliefF öznitelik seçim algoritması ilk olarak ikili sınıf problemleri için kullanılmış daha sonrasında ise çoklu sınıf problemleri için uygulanmıştır. Ayrıca bu öznitelik seçim algoritması Relief istatistik modelinin temeline dayanmaktadır (Kononenko, 1994). ReliefF öznitelik seçim algoritmasının formülü aşağıdaki gibidir;

$$S_i = \frac{\sum_{j=1}^{m} -fark(x_{ij,en\ yakın\ aynı_{ij}}) + fark(x_{ij,en\ yakın\ farklı_{ij}})}{m} \quad (3)$$

Denklemde m, verideki tüm örnek sayısını; $fark$ (x_{ij},$en\ yakın\ farklı_{ij}$) değeri, j. örnekteki i. değerin farklı sınıfa mensup en yakın örneğe olan mesafesini, $fark$(x_{ij},$en\ yakın\ aynı_{ij}$) değeri, j. örnekteki i. özniteliğin aynı sınıfa mensup en yakın örneğe olan mesafesini belirtmektedir (Gümüşçü vd., 2016).

2.4. Sınıflandırma performans kriterleri
2.4. Classification performance criteria

Çalışmanın amacına yönelik olarak Türkiye Covid-19 günlük hasta sayısının değişiminin sınıflandırılmasının tahmininde kullanılacak algoritmaların performanslarının belirlenmesi için bazı sınıflandırma kriterlerinden yararlanılacaktır. Doğruluk (Acc), kappa istatistiği (κ), hassasiyet, ortalama mutlak hata (Mean absolute error - Mae), Matthews korelasyon katsayısı (Mcc) kriterleri kullanılacaktır.

Acc, sınıflandırma performansının ölçümünde önemli rol oynamaktadır. Doğru sınıflandırılmış tüm örneklerin toplam örnek sayısına oranı ile bulunmaktadır. Algoritmaların sınıflandırma performansında etkin rol oynayan bir diğer değer olan kappa değeri, 1'e ne kadar yakınsa o kadar iyi sonuç vermektedir. Algoritmalar karşılaştırılırken kappa değeri yüksek olan algoritmanın daha başarılı sınıflandırma yaptığı söylenmektedir. Hassasiyet değeri ise doğru sınıflandırılmış pozitif örneklerin sayısının, toplam pozitif örneklerin sayısına oranı ile bulunmaktadır. Hassasiyet değeri yüksek olan algoritma sınıflandırmada daha iyi sonuç vermektedir. Mae, tahmin edilenler ile gözlenen değerler arasındaki farkı göstermektedir. Bu değer hangi algoritmada daha düşükse o algoritmanın sınıflandırmada daha iyi performans gösterdiği söylenmektedir (Willmott & Matsuura, 2005). Mcc kriteri, -1 ile 1 arasında değer alan sınıflandırma performanslarının ölçümünde başarılı sonuçlar veren bir korelasyon katsayısıdır. Algoritmalar arasında Mcc kriteri yüksek olan daha başarılı sınıflandırma performansı göstermektedir (Kılıç-Depren vd., 2017).

Makine öğrenmesi algoritmaları uygulamasında önemli noktalardan birisi de eğitim ve test verilerinin belirlenmesidir. Veri seti öncelikle eğitim ve test veri seti olarak iki gruba ayrılarak işlem yapılır. k-katlı çapraz doğrulama yönteminde veri setini k eşit parçaya bölünür. Bunların k-1 tanesi eğitim veri seti, diğer kısmı ise test verisi olarak belirlenir. Ayrı ayrı her bir parça, test kümesi olarak alınarak işlem k kez tekrarlanır. Tüm sonuçların ortalaması hesaplanarak sınıflandırma değerleri belirlenmiş olur. Çalışmada 10-katlı çapraz doğrulama uygulanmıştır (Filiz & Öz, 2019).

2.5. Uygulama
2.5. Application

Türkiye Cumhuriyeti Sağlık Bakanlığı'nın açıkladığı günlük koronavirüs bilgileri ve Google Covid-19 Topluluk Hareketliliği Raporları tarafından açıklanan insanların hareketliliğine dair verilerden yararlanılacaktır. Veri setini zaman serisinden kurtarmak ve denetimli öğrenme haline getirmek için günlük hasta sayısı değişkenine 1 gün sonrasına denk gelecek şekilde kaydırma işlemi uygulanmıştır (Kemalbay & Alkış, 2020; Bontempi vd., 2012). Çalışmada bağımsız değişkenlerden günlük ağır hasta sayısı, günlük test sayısı, günlük iyileşen hasta sayısı, günlük vefat sayısı, pozitif çıkma oranı, perakende ve rekreasyon, market ve eczane, parklar, toplu taşıma istasyonları, işyerleri ve konutlar günlük değişimleri göz önünde bulundurularak bir önceki güne göre artış göstermişse 1 azalış göstermişse 0 olarak kodlanmıştır. Ayrıca bağımlı değişken günlük hasta sayısı da günlük değişimler göz önünde bulundurularak bir önceki güne göre artış göstermişse 1 azalış göstermişse 0 olarak işleme dahi edilmiştir. Bu değişkenler dışındaki kalan değişkenler haftalık açıklandığından haftalık değerleri kullanılmıştır. Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde DVM-puk ve knn algoritması kullanılacaktır. Bu algoritmalar arasından en başarılı tahmin sınıflandırmasını yapan algoritma belirlenecektir. Ardından reliefF öznitelik seçim algoritması kullanılarak Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde etkin öznitelikler bulunacaktır. Bulunan öznitelikler yardımıyla Türkiye günlük hasta sayısındaki değişimin sınıflandırılmasının tahmin başarısında düşüş olup olmadığı incelenecektir. Analizler makine öğrenmesi ve öznitelik seçim algoritmalarını bünyesinde bulunduran weka programı ile yapılacaktır.

3. Analiz sonuçları
3. Results

Çalışmanın amacı doğrultusunda Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde DVM-puk ve knn

algoritması kullanılmıştır. Bu iki makine öğrenmesi algoritması arasından hangi algoritmanın daha başarılı tahmin sınıflandırması yaptığı sınıflandırma kriterleri yardımıyla incelenmiştir. Ayrıca reliefF öznitelik seçim algoritması kullanılarak günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde hangi özniteliklerin etkin rol oynadığı belirlenmiştir.

Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde tüm değişkenler kullanılarak elde edilen sonuçlar Tablo 2'de gösterilmiştir.

Tablo 2. Tüm değişkenler kullanılarak algoritmaların sınıflandırma tahmininde başarı performansları
Table 2. *Classfication success rate of algorithms in estimation by using all variables*

	Acc	**Kappa**	**Hassasiyet**	**Mae**	**Mcc**
DVM-puk	**0.6930**	0.3785	0.693	0.3066	0.3800
knn	0.6350	0.2619	0.634	0.3669	0.2620

Tablo 2 incelendiğinde Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde DVM-puk (0.6930) algoritmasının başarılı sonuç verdiği belirlenmiştir. Bu sonucu kappa (0.3485), hassasiyet (0.693), mae (0.3066) ve mcc (0.3800) kriterlerinin desteklediği görülmüştür.

Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde hangi öznitelikleri etkin rol oynadığını belirlemek için reliefF öznitelik seçim algoritması yardımıyla analizler yapılmış ve etkin rol oynayan öznitelikler ile önem düzeyleri Tablo 3'te verilmiştir.

Tablo 3. ReliefF öznitelik seçim algoritması ile elde edilen tahminlerin sınıflandırılmasına etkin rol oynayan öznitelikler ve önem düzeyleri
Table 3. *The features and significance levels that play an critical role in the classification of the predictions obtained by the ReliefF feature selection algorithm*

Öznitelikler	**Önem düzeyleri**
Pozitif çıkma oranı	*0.10657*
Filyasyon oranı	*0.09197*
İşyerleri hareketliliği	*0.07226*
Parklardaki hareketlilik	*0.06058*

Tablo 3'te görüldüğü gibi Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde 4 özniteliğin etkin rol oynadığı belirlenmiştir. Bu öznitelikler arasından en etkin olan öznitelik olarak pozitif çıkma oranı (0.10657) belirlenmiştir. Filyasyon oranı (0.09197), işyerleri hareketliliği (0.07226) ve parklardaki hareketlilik (0.06058) özniteliklerinin pozitif çıkma oranına yakın sonuçlar verdiği görülmüştür.

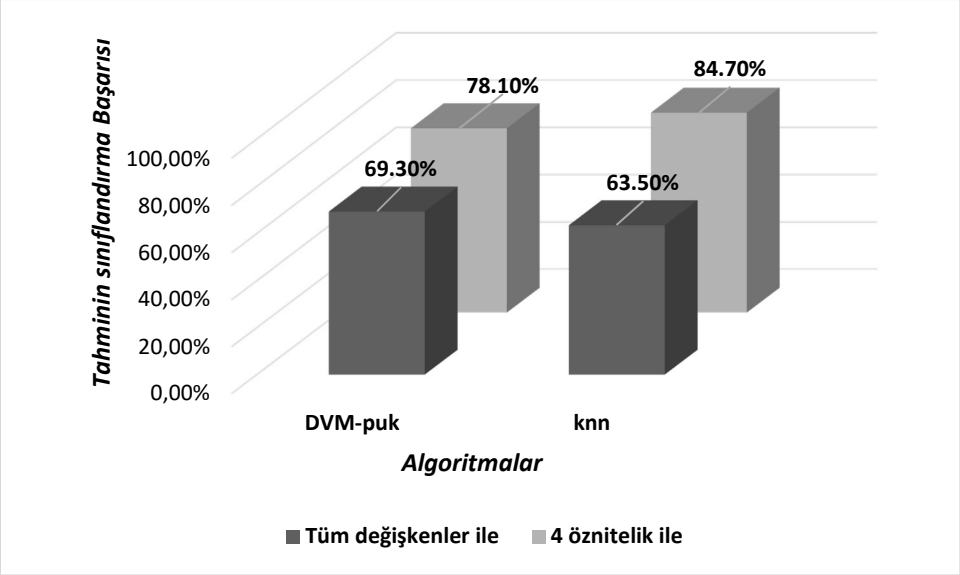
Belirlenen 4 etkin öznitelik yardımıyla DVM-puk ve knn algoritması için Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmin işlemi yinelenmiş ve sonuçlar Tablo 4'te gösterilmiştir.

Tablo 4. ReliefF öznitelik seçim algoritması ile elde edilen öznitelikler yardımıyla algoritmaların tahminde sınıflandırma performansları
Table 4. *Classification performances of algorithms by using features that obtained from ReliefF feature selection algorithm*

	Acc	**Kappa**	**Hassasiyet**	**Mae**	**Mcc**
DVM-puk	0.7810	0.5550	0.7830	0.2190	0.5590
knn	**0.8470**	0.6896	0.8480	0.2182	0.6920

Tablo 4 incelendiğinde reliefF öznitelik seçim algoritması yardımıyla belirlenen 4 etkin öznitelik ile yapılan Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde knn (0.8470) en başarılı algoritma olarak belirlenmiştir. Bu sonucu kappa (0.6896), hassasiyet (0.8480), mae (0.2182) ve mcc (0.6920) kriterleri desteklemektedir.

Çalışmada hem tüm değişkenler hem de reliefF öznitelik seçim algoritması ile Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmini yapılmıştır. Bunun için iki farklı makine öğrenmesi algoritması kullanılmıştır. Genel karşılaştırma Şekil 2'de verilmiştir.



Şekil 2. Tüm değişkenler ve etkin öznitelikler kullanılarak elde edilen algoritmaların tahminde sınıflandırma performanslarının karşılaştırılması
Figure 2. *Comparison of classification performances of algorithms obtained by using all variables and active features.*

4. Tartışma ve değerlendirme
4. Discussion and conclusion

Çalışmanın amacı doğrultusunda Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde DVM-puk ve knn algoritmasının performansı sınıflandırma kriterleri çerçevesinde incelenmiştir. Bunun yanında reliefF öznitelik seçim algoritması yardımıyla Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde etkin rol oynayan öznitelikler belirlenmiştir. Bu belirlenen etkin öznitelikler yardımıyla sınıflandırma algoritmalarının performansları yeniden incelenerek tahminde sınıflandırma başarılarında bir değişim olup olmadığı araştırılmıştır.

Uygulama aşamasında ilk olarak, Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmini probleminde tüm değişkenler kullanılarak yapılan analizler sonucunda DVM-puk %69.3 oranında başarı göstermiştir. Günlük hasta sayısındaki değişimi etkileyen öznitelikler bilindiğinde Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmini %69.3 oranında doğru olarak yapılabileceğini söylemek mümkündür.

Çalışmadaki Covid-19 hastalığının günlük hasta sayısındaki değişimi, gelişimi ve çözümü için önemli noktalardan biri de öznitelik seçimidir. Makine öğrenmesi algoritmaları bünyesinde bulunan ReliefF öznitelik seçim algoritması kullanılarak Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde başarıdan ödün vermeden daha az değişkenle başarı elde etmek amaçlanmıştır. Bu çerçevede 'pozitif çıkma oranı', 'filyasyon oranı', 'işyerleri hareketliliği' ve 'parklardaki hareketlilik' öznitelikleri etkin öznitelikler olarak belirlenmiştir. Covid-19 virüsü insandan insana ve temas ile bulaşabilen bir virüstür. Bu duruma bağlı olarak insanların işyerlerindeki ve parklardaki hareketliliğinin Covid-19 hasta sayısı üzerine etkisi olduğu bu çalışmada belirlenmiştir. Afacan & Avcı (2020), yaptıkları çalışmada insanların işleri ile ilgili yüz yüze görüşmelerden ziyade interaktif olarak görüşmeler yapılabileceğini söylemişlerdir. Benzer şekilde yapılan bir çalışma da Covid-19 salgınının insan hareketliliğinin belirlenen tedbirler kapsamında, günlük hareketliliklerinin de sınırlanmasını meydana getirdiği belirtilmiştir (Sirkeci vd., 2020). Türkiye'de Covid-19 salgınının başladığı ilk günden itibaren gerek işyerlerinde seyreltme veya uzaktan çalışma gerekse belli yaş altında ya da üstünde kişilere getirilen kısıtlamalar (parklar vs.) çalışmada

belirlenen etkin özniteliklerin destekleyicisi niteliğindedir. Bir diğer etkin öznitelik olarak belirlenen filyasyon oranı, temaslı kişilerin belirlenip; çevre ile ilişkileri kesilen kişilerin oranı olarak söylenebilir. Literatür de incelenen çalışmalarda filyasyonun ve temaslı kişilerin belirlenmesinin öneminden bahsedilmiştir (Demirtas & Tekiner, 2020; Durusoy vd., 2020; Şimşek vd., 2020). Kişilerin yapılan Covid-19 testleri sonrasında pozitif çıkma oranları da etkin değişkenlerden biri olarak bulunmuştur. Bu durumda yapılan test sayısı sonucu pozitif çıkan kişilerin günlük hasta sayısı için önemli bir gösterge olduğu ortaya konmuştur.

ReliefF öznitelik seçimi ile elde edilen etkin özniteliklerin belirlenmesi, Türkiye için Covid-19 salgınının gelişimi için önemli noktaları işaret etmektedir. Bu özniteliklere karşı gerekli tedbirler alınarak günlük hasta sayısındaki değişimin durumu ile ilgili bilgilere ulaşılabilir. Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde etkin rol oynayan 4 öznitelik belirlenmiş ve bu etkin 4 öznitelik yardımıyla yapılan analizlere göre %84.7 ile knn algoritmasının en başarılı algoritma olduğu görülmüştür. Günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde etkin rol oynayan 4 özniteliğin bilinmesi durumunda %84.7 oranında Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmini doğru olarak yapılabilecektir. Ayrıca çalışmada kullanılan iki algoritmada etkin 4 öznitelik yardımıyla daha başarılı sonuçlar verdiği belirlenmiştir.

Tüm dünyada etkisini sürdüren Covid-19 salgını için yapılan çalışmada bu salgına yönelik önemli sonuçlar elde edilmiştir. Son yıllarda popüler hale gelen makine öğrenmesi algoritmaları bu çalışma için de başarılı bilgiler elde edilmesine yardımcı olmuştur. Ayrıca yapılan öznitelik seçimi ile daha az öznitelik ile daha yüksek başarı elde edilebileceği gösterilmiştir. Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde tüm değişkenleri kullanmak yerine etkin 4 özniteliğin kullanılmasının daha uygun olacağı ortaya konulmuştur.

Çalışmada bazı sınırlamalar vardır. Türkiye Covid-19 günlük hasta sayısındaki değişimin sınıflandırılmasının tahmininde Türkiye Cumhuriyeti Sağlık Bakanlığı'nın açıkladığı günlük Covid-19 bilgileri ve Google Covid-19 Topluluk Hareketliliği Raporları tarafından açıklanan insanların hareketliliğine dair veriler kullanılmıştır. 28.09.2020 – 12.02.2021 tarihleri arasındaki verilerden yararlanılmıştır. Etkin öznitelikler yalnızca Türkiye için belirlenmiştir.

Yazar katkısı
Author contribution

Enes FİLİZ çalışmayı gerçekleştirmiştir.

Etik beyanı
Declaration of ethical code

Bu makalenin yazarları, bu çalışmada kullanılan materyal ve yöntemlerin etik kurul izni ve / veya yasal-özel izin gerektirmediğini beyan etmektedir.

Çıkar çatışması beyanı
Conflicts of interest

Herhangi bir çıkar çatışması bulunmamaktadır.

Kaynaklar
References

Abakar, K. A. A., & Yu, C. (2014). Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity. *Indian Journal of Fibre and Textile Research*, *39*, 55-59.

Afacan, E., & Avcı, N. (2020). Koronavirüs (Covid-19) Örneği Üzerinden Salgın Hastalıklara Sosyolojik Bir Bakış. *Avrasya Sosyal ve Ekonomi Araştırmaları Dergisi, 7*(5), 1-14.

Alpaydın, E. (2004). *Introduction to machine learning.*

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., & Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. *Algorithms*, *13*(10), 249. https://doi.org/10.3390/a13100249

Ayaz, M. (2021). *Makine öğrenmesi algoritmaları ile covid-19 hastalarının belirlenmesi* [Yüksek Lisans Tezi, Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü].

Barstugan, M., Ozkaya, U., & Ozturk, S. (2020). Coronavirus (covid-19) classification using ct images by machine learning methods. *arXiv preprint arXiv:2003.09424*.

Bontempi, G., Taieb, S. B., & Le Borgne, Y. A. (2012, July). Machine learning strategies for time series forecasting. In *European business intelligence summer school* (ss. 62-77). Springer, Berlin, Heidelberg.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297. https://doi.org/10.1007/BF00994018

De Felice, F., & Polimeni, A. (2020). Coronavirus Disease (COVID-19): A Machine learning bibliometric analysis. *In vivo*, *34(3 suppl)*, 1613-1617. https://doi.org/10.21873/invivo.11951

Demirtas, T., & Tekiner, H. (2020). Filiation: a historical term the COVID-19 outbreak recalled in Turkey. *Erciyes Medical Journal, 42*(3), 354-359.

Depren, S. K., Aşkın, Ö. E., & Öz, E. (2017). Identifying the classification performances of educational data mining methods: a case study for TIMSS. *Educational Sciences: Theory & Practice, 17*(5), 1605-1623. https://doi.org/10.12738/estp.2017.5.0634

Durusoy, R., Teneler, A. A., Geçim, C., Özbay, N. F., Küçük, E. F., Şimşek, S., & Ersel, M. (2020). Ege Üniversitesi Tıp Fakültesi Hastanesi'nde COVID-19 vakalarının sürveyansı, filyasyonu ve temaslılarının belirlenmesi. *Turkish Journal of Public Health*, *18(COVID-19 Special)*, 25-39. https://doi.org/10.20518/tjph.771286

Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons & Fractals*, 134, 109761. https://doi.org/10.1016/j.chaos.2020.109761

Filiz, E., & Öz, E. (2019). Finding The Best Algorithms And Effective Factors In Classification Of Turkish Science Student Success. *Journal of Baltic Science Education, 18*(2), 239. https://doi.org/10.33225/jbse/19.18.239

Google Web Sayfası - Google Covid-19 Topluluk Hareketliliği Raporları, https://www.google.com/covid19/mobility/ (Erişim tarihi: 16.02.2021).

Gümüşçü, A., AydileK, İ.B., & Taşaltın, R. (2016). Mikro-dizilim Veri Sınıflandırmasında Öznitelik Seçme Algoritmalarının Karşılaştırılması. *Harran Üniversitesi Mühendislik Dergisi*, *1*(1), 1-7.

Haykin, S. (1999). *Neural Networks: A comprehensive Foundation*.

Horton, P., & Nakai, K. (1997, June). Better Prediction of Protein Cellular Localization Sites with the it k Nearest Neighbors Classifier. *In Ismb, 5*, 147-152.

Kavzoğlu, T., & Çölkesen, İ. (2010). Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. *Harita Dergisi, 144*(7), 73-82.

Kemalbay G., & Alkiş B. N. (2020). Borsa endeks hareket yönünün çoklu lojistik regresyon ve k-en yakın komşu algoritması ile tahmini. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, *26* (8). https://doi.org/10.5505/pajes.2020.57383

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. *In Machine learning proceedings*, (ss. 249-256). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-247-2.50037-1

Kononenko, I. (1994, April). Estimating attributes: Analysis and extensions of RELIEF. *In European conference on machine learning*, (ss. 171-182). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-57868-4_57

Kononenko, I., Šimec, E., & Robnik-Šikonja, M. (1997). Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, *7*(1), 39-55. https://doi.org/10.1023/A:1008280620621

Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., & Singh, R. P. (2020). Significant applications of machine learning for COVID-19 pandemic. *Journal of Industrial Integration and Management*, *5*(4). https://doi.org/10.1142/S2424862220500268

Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, Solitons & Fractal*s, 110059. https://doi.org/10.1016/j.chaos.2020.110059

Li, B., Yu, S., & Lu, Q. (2003). An improved k-nearest neighbor algorithm for text categorization. *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages*. https://arxiv.org/ftp/cs/papers/0306/0306099.pdf .

Malki, Z., Atlam, E. S., Hassanien, A. E., Dagnew, G., Elhosseini, M. A., & Gad, I. (2020). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos, Solitons & Fractals*, *138*, 110137. https://doi.org/10.1016/j.chaos.2020.110137

Mitchell, T. M. (1997). Machine Learning. *Burr Ridge, IL: McGraw Hill, 45*(37), 870-877.

Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2020). COVID-19 epidemic analysis using machine learning and deep learning algorithms. *MedRxiv*. https://doi.org/10.1101/2020.04.08.20057679

Shahiri, A.M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer*

Science, *72*, 414-422. https://doi.org/10.1016/j.procs.2015.12.157

Shawe-Taylor, J., Bartlett, P. L., Williamson, R.C., & Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, *44*(5), 1926-1940. https://doi.org/10.1109/18.705570

Şimşek, A. Ç., Kara, A., Baran-Aksakal, F. N., Gülüm, M., Ilter, B., Ender, L., & Demirkasimoğlu, M. (2020). Contact tracing management of the COVID-19 pandemic. *Türk Hijyen ve Deneysel Biyoloji Dergisi*, 269. https://doi.org/10.5505/TurkHijyen.2020.80688

Sirkeci, I., Özerim, M. G., & Bilecen, T. (2020). Editörden: kovid-19'un uluslararası hareketlilik ve göçmenliğe ilişkin etkisi üzerine. *Göç Dergisi*, *7*(1), 1-8. https://doi.org/10.33182/gd.v7i1.688

T.C. Sağlık Bakanlığı Web Sayfası, https://covid19.saglik.gov.tr/ (Erişim tarihi: 15.02.2021).

Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things, 11*, 100222. https://doi.org/10.1016/j.iot.2020.100222

Tuncer, E., & Bolat, E. D. Destek Vektör Makinaları ile EEG Sinyallerinden Epileptik Nöbet Sınıflandırması. *Politeknik Dergisi*, 1-1.https://doi.org/10.2339/politeknik.672077

Ulaş, E. (2021). Prediction of COVID-19 Pandemic Before The Latest Restrictions in Turkey by Using SIR Model. *Suleyman Demirel University Journal of Science*, *16*(1), 77-85. https://doi.org/10.29233/sdufeffd.852222

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, *85*, 189-203. https://doi.org/10.1016/j.jbi.2018.07.014

Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons & Fractals*, *139*, 110058. https://doi.org/10.1016/j.chaos.2020.110058

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, *30*(1), 79-82. https://doi.org/10.3354/cr030079

Xia, S., Xiong, Z., Luo, Y., Dong, L., & Zhang, G. (2015). Location difference of multiple distances based k-nearest neighbors algorithm. *Knowledge-Based Systems*, *90*, 99-110. https://doi.org/10.1016/j.knosys.2015.09.028

Yadav, M., Perumal, M., & Srinivas, M. (2020). Analysis on novel coronavirus (COVID-19) using machine learning methods. *Chaos, Solitons & Fractals*, *139*, 110050. https://doi.org/10.1016/j.chaos.2020.110050

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2017). Efficient kNN classification with different numbers of nearest neighbors. *IEEE transactions on neural networks and learning systems, 29*(5), 1774-1785. https://doi.org/10.1109/TNNLS.2017.2673241