Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 5(2) (2022) 127-136
DOI: 10.38016/jista.1082310

Covid-19 Hastalarının Ölüm Oranlarının ve Yüksek Ölüm Riskine Sahip Hastaların Belirlenmesi için Temel Bileşen Analizinin Kullanılması

Ebru Efeoğlu[1*]

[1] Dumlupınar Üniversitesi, Yazılım Mühendisliği, Kütahya, Türkiye

ebru.efeoglu@dpu.edu.tr

Öz

Covid-19 virüsü 2019 yılında ortaya çıktı ve kısa bir sürede tüm dünyaya yayıldı. Milyonlarca insanın enfekte olmasına ve yüz binlerce insanın ölümüne neden oldu. Her geçen gün vaka sayısı artmakta ve virüsün yeni varyantlar meydana gelmektedir. Bu hastalığa sahip kişileri tespit etmek için Polimeraz Zincir Reaksiyonu (PCR) testleri uygulanmaktadır. Hastalığı tespit edilen kişilerin durumlarının incelenmesi yoğun bakım ve ölüm oranlarının önceden tespiti oldukça önemlidir. Bu çalışmada Covid-19 hastalarından ölüm oranlarının tespitinde özellik çıkarımı yöntemi olarak Temel Bileşen Analizi (PCA) kullanılmış ve yöntemin başarılı sonuçları en popüler makine öğrenmesi teknikleri ile gösterilmiştir. Çalışmada kullanılan makine öğrenmesi teknikleri K-En Yakın Komşu (KNN), Doğrusal Ayrımcılık Analizi (LDA), Extra Ağaçlar, Random Tree, Rep Tree ve Naive Bayes algoritmalarıdır. Bu tekniklerin performans değerlendirmesinde Doğruluk, Kesinlik, Duyarlılık, Rms, F-skoru değerleri hesaplanmıştır. Ayrıca ROC Eğrileri ve Karışıklık matrisleri incelenerek sonuçlar karşılaştırılmıştır. Sonuç olarak, en iyi performansın Temel bileşenler analizi uygulandıktan sonra Doğrusal Ayrım Analizi (PCA+LDA) kullanımı ile elde edildiği görülmüştür. PCA+LDA uygulaması ile %96,39 Doğruluk oranı elde edilmiştir. Makalede ayrıca özellik çıkarımının kullanılmasıyla Covid-19 virüsünden Zatürre, Şeker, KOAH ve Astım hastalarının, hamile, yaşlı ve entrube insanların daha çok etkilendiği ve ölüm riskinin daha yüksek olduğu ortaya çıkmıştır. Virüsün varyantlarının ölümcüllüğünün incelenmesi, riskli hastaların tedavisi, ölüm riski bulunan hastaların izolasyonu için gereken önlemlerin alınması ve hastane kapasite planlamasının iyileştirilmesi açısından bu çalışma önem arz etmektedir.

Anahtar kelimeler: Covid-19, Performans analizi, Sınıflandırma algoritmaları, Temel Bileşen Analizi.

Using Principal Component Analysis to Identify Mortality Rates of Covid-19 Patients and Patients at High Risk of Death

Abstract

The Covid-19 virus emerged in 2019 and spread all over the world in a short time. It caused millions of people to be infected and hundreds of thousands to die. The number of cases is increasing day by day and new variants of the virus are emerging. Polymerase Chain Reaction (PCR) tests are used to detect people with this disease. It is very important to examine the conditions of the people with the disease and to determine the intensive care and mortality rates in advance. In this study, Principal Component Analysis (PCA) was used as a feature extraction method to determine mortality rates from Covid-19 patients, and the successful results of the method were demonstrated with the most popular machine learning techniques. Machine learning techniques used in the study are K-Nearest Neighbor (KNN), Linear Discrimination Analysis (LDA), Extra Trees, Random Tree, Rep Tree and Naive Bayes algorithms. In the performance evaluation of these techniques, Accuracy, Precision, Sensitivity, Rms, F-score values were calculated. In addition, ROC Curves and Confusion matrices were examined and the results were compared. As a result, it was seen that the best performance was obtained with the use of Linear Discrimination Analysis (PCA+LDA) after applying Principal component analysis. With the PCA+LDA application, an accuracy rate of 96.39% was obtained. In the article, it has also been revealed that Pneumonia, Diabetes, COPD and Asthma patients, Pregnant, Elderly and Intubated people are more affected and the risk of death is higher from the Covid-19 virus by using feature extraction. This study is important in terms of examining the lethality of virus variants, taking the necessary precautions for the treatment of risky patients isolation of patients at risk of death, and improving hospital capacity planning.

Keywords: Covid-19, Performance Analysis, Classifications, Principal Component Analysis.

* Sorumlu yazar.
E-posta adresi: ebru.efeoglu@dpu.edu.tr

Alındı : 3 Mart 2022
Revizyon : 13 Mayıs 2022
Kabul : 28 Mayıs 2022

1. Giriş (Introduction)

Çin'de ortaya çıkan koronavirüs hastalığı (Covid-19) dünya çapında bir pandemi haline geldi (Velavan & Meyer, 2020). Bu pandeminin en başından itibaren, vaka tespiti konusu her zaman bilimsel ve kamusal söylemin merkezinde yer aldı. Salgının kontrol altına alınabilmesi için Popülasyonda gerçekten kaç enfeksiyonun bulunduğunu bilmek büyük önem taşımaktadır. Ancak hastalığa dair belirti göstermeyip taşıyıcı olarak hastalığı başkasına bulaştıran, koronavirüs testi pozitif çıkan asemptomatik bireyler ve ilk önce belirti göstermeyip ilerleyen süreçlerde hastalık belirtisi gösteren presemptomatik bireylerin, özellikle genç popülasyonda bir pandemik hastalığın yayılmasında önemli bir etkiye sahiptir (Stella, Martínez, Bauso, & Colaneri, 2020). Hastalığın yayılmasının dinamiklerini tahmin etmek için epidemiyolojik modeller (de León, Pérez, & Avila-Vales, 2020) (Chinazzi et al., 2020), enfekte hastaları ve yeni vakaları izlemek için mobil cihaz uygulamaları geliştirilmiştir (Zens, Brammertz, Herpich, Südkamp, & Hinterseer, 2020) (Drew et al., 2020).

Ayrıca bu vakalar gerçek ölüm oranını ortaya çıkarma sorunuyla sıkı sıkıya iç içedir. Covid-19 enfeksiyonlarının gerçek sayılarını tahmin etme sorunu, ölüm oranıyla ilişkili olduğu tamamen istatistiksel bir bakış açısıyla da tartışılmıştır (Manski & Molinari, 2021). Bu bağlamda, farklı ulusal eksik raporlama oranları karşılaştırılmış (Rahmandad, Lim, & Sterman, 2020; Jagodnik, Ray, Giorgi, & Lachmann, 2020) ve enfeksiyon ölüm oranının değerlendirilmesine ilişkin genel bir tartışma ve anket yapılmıştır (Levin, Cochran, & Walsh, 2020).

Son zamanlarda Covid-19 hastalığı ile ilgili farklı amaçlar için çeşitli makine öğrenimi yaygın olarak uygulanmıştır (Albahri et al., 2020). Örneğin Röntgen görüntülerinden, (Kassania, Kassanib, Wesolowskic, Schneidera, & Detersa, 2021), tam kan sayımından makine öğrenimi algoritmaları kullanılarak hastalık teşhisi yapılmıştır (Akhtar et al., 2021). Çin'de Covid-19 vaka ölüm oranının erken tahmini için veriye dayalı bir analiz yapılmıştır (Yang et al., 2020). Makine öğrenimi yaklaşımını kullanarak COVİD-19 bulaşması ve ölümle ilişkili yeni faktörlerin belirlenmesi ile ilgili çalışmalar yapılmıştır (M. Li et al., 2021). Hastanede yatan Covid-19 hastalarının ölümcül risk tahmini için klinik ve inflamatuar özelliklere dayalı makine öğrenimi kullanılmış (Guan et al., 2021) (Quiroz-Juárez, Torres-Gómez, Hoyo-Ulloa, León-Montiel, & U'Ren, 2021) ve risk faktörlerini değerlendirilmiştir (Gansevoort & Hilbrands, 2020), (Parra-Bracamonte, Lopez-Villalobos, & Parra-Bracamonte, 2020). Ayrıca pozitif ve negatif Covid-19 vakaları için epidemiyoloji etiketli veri seti kullanılarak lojistik regresyon, karar ağacı, destek vektör makinesi, Naive Bayes ve yapay sinir ağları içeren öğrenme algoritmaları ile Covid-19 enfeksiyonu için denetimli makine öğrenimi modelleri geliştirilmiştir (Muhammad et al., 2021). Hipertansiyon, diyabet, koroner kalp hastalığı, kronik obstrüktif akciğer hastalığı, kronik böbrek hastalığı (Escobedo-de la Peña et al., 2021), obezite (Bello-Chavolla et al., 2020) ve Hamilelik durumunun (Ríos-Silva, Murillo-Zamora, Mendoza-Cano, Trujillo, & Huerta, 2020) Covid-19 mortalitesi üzerindeki etkisi incelenmiştir.

Ölüm riski yüksek hastaların tespiti onlara gereken özenin gösterilmesi ve önlemlerin alınması ölüm oranlarının düşmesine önemli bir katkı sağlayacaktır. Ayrıca bu hastaların tespiti hastanelerdeki kaynakları ve kapasiteleri yönetmek (Singh et al., 2021), (Zawiah et al., 2020) hastalara zamanında tedavi sağlamak (Nemati, Ansary, & Nemati, 2020) için oldukça önemlidir.

Çalışmanın amacı, Covid-19 hastalarının ölüm oranı tespitini en yüksek başarı oranı ile en kısa işlem süresine sahip bir yöntem önermek ve ölüm riski yüksek olan hastaların hangi özelliklere sahip olduklarının tespiti ile hem bu hastalara gereken izolasyon sağlanarak ölüm oranlarının düşürülmesine hem de hastane kapasite planlamasının iyileştirilmesine yardımcı olmaktır. Bu hastaların tespiti için özellik seçiminde sıklıkla kullanılan ve başarılı bir algoritma olan PCA yöntemi tercih edilmiştir. Bu yöntem sayesinde hem yüksek ölüm riskine sahip hastaların özellikleri tespit edilebilecek hem de daha kısa sürede ve daha yüksek başarıyla ölüm oranları tespiti yapılabilecektir. PCA yöntemine ek olarak ölüm oranları tespitinde sınıflandırma algoritmalarından yararlanılmıştır. En başarılı algoritmanın belirlenebilmesi için Algoritmaların performans analizi yapılmıştır. Analizde PCA yöntemi kullanılmadan ve PCA yöntemi kullanıldıktan sonra veri setinden ölüm oranı tahmin etme başarıları dikkate alınmıştır. Analiz sonucunda elde edilen sonuçlar karşılaştırılmıştır.

2. Materyal ve Yöntem (Material and Method)

2.1. Veri seti (Dataset)

Makalede kullanılan veri seti, Kaggle sitesinde bulunan ve "COVID-19 Mexico Patient Health Dataset" isimli veri setidir. Bu veri setindeki Covid-19 hastalığına ilişkin veriler Meksika hükümeti tarafından 15 Ocak 2020 ile 3 Mayıs 2020 tarihleri arasında kaydedilmiştir. Bu veri seti daha önce ölüm oranı tespiti için (Yavuz & Dudak, 2020) makalesinde kullanıldı. Bu çalışmada hem ölüm oranları tahmin edilmiş hem de hangi özelliklerin hastada ölüm riskinin artmasında daha etkili olduğu incelenmiştir.

Veri seti 19 özellikten oluşan 95805 örnekten oluşmaktadır. (Kaggle.com, 2020). Veri setindeki özellikler hastanın cinsiyeti, hastalığın tipi, entübe olma durumu, zatürre olma durumu, hastanın yaşı, hamilelik durumu, diyabet olma durumu, Kronik Obstrüktif Akciğer Hastalığı olması durumu (KOAH), astım olma durumu, Bağışıklık sistemi baskılanması durumu (İmmünosupresyon), Hipertansiyon durumu, diğer

hastalık durumu, Kardiyovasküler durum, Obezite olma durumu, Kronik böbrek yetmezliği durumu (Chronic_kidneyfailure), sigara içme durumu, başka bir vaka durumu, yoğun bakım durumu (icu), ölüm tarihi bulunmaktadır. Çalışmada ölüm tarihi özelliğinin yerine ölü tarihi yazan hastanın öldüğü, tarih yoksa hastanın yaşadığı kabul edilmiştir. Veri seti sayısal değerlerden oluşmaktadır. Hastalarda sayılan özelliklerin bulunması durumu 1, bulunmama durumu 2 ile bu özelliğe ait veri yoksa 99 ile gösterilmiştir. 98/97 değeri ise bu özellik için kişinin uygun olmadığını göstermektedir.

2.2. Sınıflandırma algoritmaları (Classification algorithms)

K-En Yakın Komşu Algoritması: Bu algoritmanın amacı, bir veri setinde en yakın komşuları bulmaktır. Bu komşuları bulmak için farklı mesafe metrikleri kullanır. Algoritmanın başarısı bu mesafe metrikleri ve k ile gösterilen komşu sayısına göre değişkenlik gösterir (B. Li, Yu, & Lu, 2003; Xia, Xiong, Luo, Dong, & Zhang, 2015).

Naive Bayes : Bayes teoremine dayanır. Sınıfı belli olan örnek verileri kullanarak yeni verinin sınıflara ait olma olasılığını hesaplar. Bulunan değerlerden en yüksek olasılık değerine sahip sınıf, örneğinin sınıfıdır. (Bermejo, Gámez, & Puerta, 2011).

Doğrusal Ayrımcılık Analizi (LDA): Bu algoritma ilk olarak 1936 yılında ikili sınıflandırmalar için R. A. Fisher tarafınca geliştirilmiştir. Daha sonra C. R. Rao tarafından ikiden fazla sınıflandırmalar için formülüze edilmiştir. Diskriminant analizinde ilk olarak sınıfları birbirinden ayırmayı sağlayan diskriminant fonksiyonları bulunur. Daha sonra bulunan fonksiyonlar kullanılarak yeni örneğin hangi sınıfa dahil edilmesi gerektiğine karar verilir (Ünsal, Bileşenler, _Faktür, & Mali, 1996).

Random tree: Her düğümde belirli sayıda özellik kullanılır ve ağaç oluşturulur. Seçilen bu özellikler rasgele seçilmiş özelliklerdir. Budama işlemi yapılmaz. Ayrıca, tutulan veri kümesine dayalı olarak sınıf olasılıklarının tahmin edilmesini sağlayan bir seçeneğe de sahiptir.

Rep Tree algoritması: İlk olarak Quinlan tarafından (Quinlan, 1999) önerildi. Eğitim örneklerinde gürültünün etkilerini azaltmak istenir ve bunun için budama işlemi yapılarak bir karar ağacı oluşturulur. Hızlı makine öğrenmesi algoritmalarındandır (J. Li, Zhang, Lu, & Yan, 2008). Rep Tree algoritması varyanstan kaynaklanan hatayı en aza indirme ilkesine ve entropi (Amasyali & Ersoy, 2009) ile bilgi edinme ilkesine dayanır ve sadece sayısal verilerle çalışır.

Ekstra Ağaçlar algoritması: Klasik yukarıdan aşağıya prosedüre göre budanmamış bir karar ağaçları topluluğu oluşturur. Diğer ağaç tabanlı topluluk yöntemleriyle arasındaki iki temel fark, kesme noktalarını tamamen rastgele seçmeleri, düğümleri ayırmaları ve ağaçları büyütmek için tüm öğrenme örneğini kullanmalarıdır (Freund & Mason, 1999).

Hoefding tree: Hoeffding ağacı, veri dağılımının zaman içinde değişmediğini varsayan büyük veri akışı için artan bir karar ağacı öğrenicisidir. Hoeffding sınırına dayanan bir karar ağacı aşamalı olarak büyür (Zhang, Ding, & Wang, 2011).

Random Forest: Torbalama yöntemine rastgelelik özelliği eklenerek oluşturulan bir algoritmadır (Breiman 2001). Regresyon ve sınıflandırma yöntemi olarak kullanılabilir.

2.3. Özellik Seçimi ve PCA Yöntemi (Feature Selection and PCA Method)

Özellik seçimi kısaca veri setini temsil edebilecek en iyi altkümenin seçilmesidir. Bu işlem, çözülmek istenen problem için en faydalı ve en önemli özelliklerin seçilmesiyle veri kümesindeki özellik sayısının azaltılmasıdır. Birçok özellik seçim yöntemi bulunmaktadır. Bir örneği tekrar tekrar örnekleyerek ve aynı ve farklı sınıfın en yakın örneği için verilen özniteliğin değerini göz önünde bulundurarak bir özniteliğin değerini değerlendiren Relief yöntemi (Kira & Rendell, 1992), Sınıfa göre simetrik belirsizliği ölçerek bir özelliğin değerini değerlendiren simetrik belirsizlik katsayısı (Novaković, Strbac, & Bulatović, 2011). Sınıfa göre kazanç oranını ölçerek bir özelliğin değerini değerlendiren Kazanç oranı yöntemi, Bir öğrenme şeması kullanarak öznitelik kümelerini değerlendiren Sarmalayıcı Alt Kümes (Wrapper Subset Eval) yöntemi. (Kohavi & John, 1997).Veri setinde negatif olmayan bir şekilde lineer olarak temsil edilmesini sağlayan ve Negatif olmayan sinyallerin bulunduğu alanlarda başarılı olan Negatif olmayan matris yöntemi bu yöntemler arasında yer almaktadırlar.

PCA yöntemi, veri içindeki etkin özellikleri tespit ederek verinin boyutunu azaltır (Abdi & Williams). Oldukça popüler ortogonal linear dönüşümdür. Genel olarak PCA yöntemi verideki maksimum varyansa sahip verinin az boyuta sahip uzayda gösterimi şeklinde ifade edilebilir.

PCA dönüşümü denklem (1) de gösterildiği gibi yapılır.

$$\mu^T = X^T W \quad (1)$$

Burada,

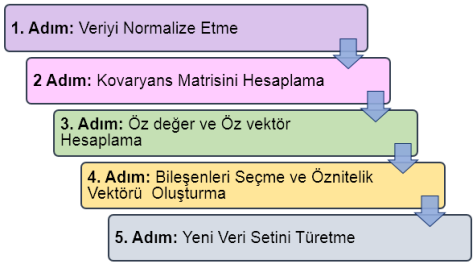
X ortogonal matrisi, μ^T linear dönüşümünü, W ise özvektörleri ifade eder.

Verinin PCA yöntemi ile ayrılmış şekli denklem (2) de verildiği gibidir (Maglaveras, Stamkopoulos, Diamantaras, Pappas, & Strintzis, 1998) .

$$X_i = \sum_{j=1}^{p} w_{ij} Q_j \quad (2)$$

Q_j, j=1, .., p, faktör veya özellik adı verilen gizli gösterim parametreleridir.

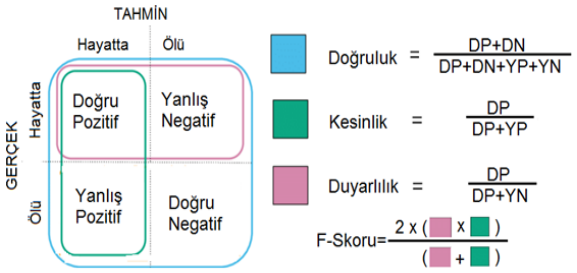
PCA yönteminin uygulanması için izlenmesi gereken adımlar Şekil 2' de verilmiştir.



Şekil 1. PCA yönteminin işlem adımları (Process steps of the PCA method)

2.4. Performans metrikleri (Performance metrics)

Algoritmaların sınıflandırma performansı hakkında en fazla bilgi içeren metrik karışıklık matrisi olduğundan performans karşılaştırmada en sık kullanılan metriktir. Karışıklık matrisinde 4 farklı değer bulunmaktadır. Gerçek durum pozitifken test sonucunun pozitif olması durumunda DP (Doğru Pozitif) değeri, gerçek durum negatifken test sonucunun pozitif olması durumunda YP (Yanlış Pozitif) değeri, gerçek durum negatifken test sonucunun negatif olması durumunda DN (Doğru Negatif) değeri ve gerçek durum pozitifken test sonucunun negatif olması durumunda ise YN (Yanlış Negatif) hesaplanır. Bu değerlerin karışıklık matrisinde gösterimi ve bu matristen hesaplanan diğer metrikler Şekil 2' de verilmiştir.

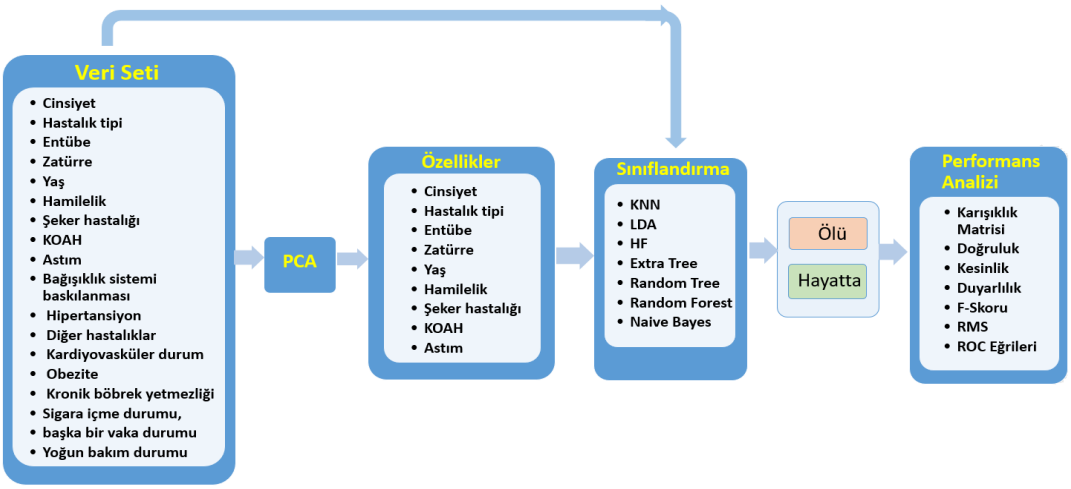


Şekil 2. Karışıklık matrisi ve diğer performans metrikleri (Confusion matrix and other performance metrics)

Covid-19 hastalarından ölüm oranlarının tespiti için yapılan çalışmayı özetleyen akış diyagramı Şekil 3'de verilmiştir. Akış diyagramından da anlaşılacağı gibi ölüm oranı tahmini için önce veri setine PCA yöntemi kullanılmadan bir sınıflandırma işlemi uygulanmış daha sonra veri setine PCA yöntemi uygulanarak veri setinin boyutu azaltılmış ve özellik seçimi yapılmıştır. Seçilen özellikler kullanılarak bir sınıflandırma yapılmıştır. Daha sonra PCA kullanılmadan ve PCA kullanılarak yapılan sınıflandırma işlemlerine performans analizi uygulanmıştır. Yapılan performans analizinde kullanılan performans metrikleri akış diyagramında belirtilmiştir. Son olarak en iyi performansa sahip teknik belirlenmiştir.

PCA yöntemi kullanılmadan ve PCA yöntemi kullanıldıktan sonra yapılan sınıflandırmalardan elde edilen karışıklık matrisi Şekil 4'te verilmiştir. Şekilde mavi ile gösterilen yerler algoritmaların doğru tahmin ettiği örnek sayısını ifade etmektedir. Görüldüğü gibi PCA yönteminin kullanılması ile algoritmaların doğru tahmin ettiği örnek sayılarının arttırmıştır. Örneğin KNN algoritması ile yapılan sınıflandırmada algoritma 91381 örneği doğru tahmin etmiştir. Bu sayı PCA yöntemi kullanılması ile 92202'e yükselmiştir.

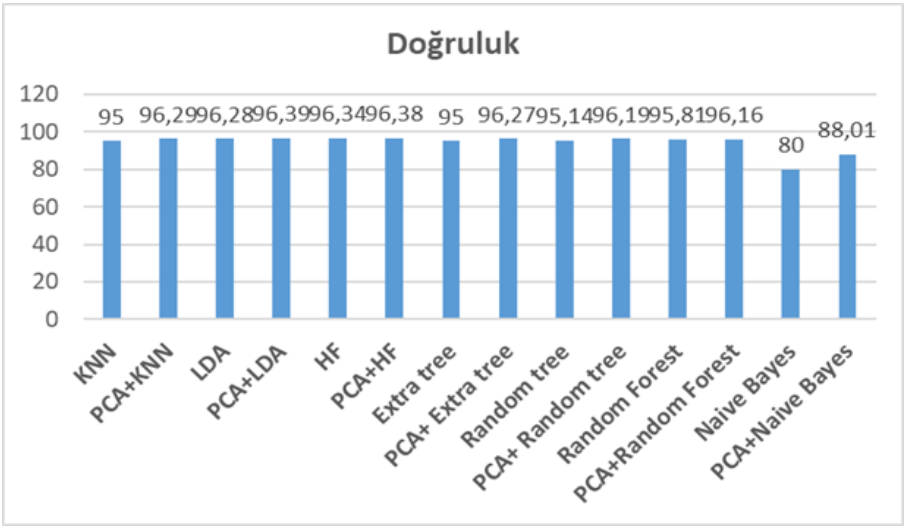
Algoritmaların yüzde cinsinden doğruluk değerlerini gösteren grafik Şekil 5' de diğer performans metrikleri de Tablo 2'de verilmiştir. PCA yönteminin algoritmaların veri seti üzerinde model oluşturulma süreleri ve toplam işlem süreleri karşılaştırmalı olarak gösterilmektedir. Tablodan PCA yönteminin işlem sürelerini azalttığı görülmektedir. Tüm işlemler Intel(R) Core(TM) i7-4600U CPU@ 2.10 Ghz işlemci, 8 GB Ram özelliklerine sahip Windows 10 işletim sistemi kurulu bir bilgisayar ile gerçekleştirilmiştir.



Şekil 3. Çalışmanın akış diyagramı (Flow chart of the study)

		Tahmin edilen Sınıf					*Tahmin edilen Sınıf*	
K-EYK	*Gerçek Sınıf*		*Hayatta*	*Ölü*	**PCA+ KNN**	*Gerçek Sınıf*	*Hayatta*	*Ölü*
		Hayatta	*90854*	*1519*		*Hayatta*	91991	382
		Ölü	*2905*	*527*		*Ölü*	3221	211
LDA	*Gerçek Sınıf*		*Hayatta*	*Ölü*	**PCA+ LDA**	*Gerçek Sınıf*	*Hayatta*	*Ölü*
		Hayatta	92030	343		*Hayatta*	92338	35
		Ölü	3215	217		*Ölü*	3426	6
HF	*Gerçek Sınıf*		*Hayatta*	*Ölü*	**PCA+ HF**	*Gerçek Sınıf*	*Hayatta*	*Ölü*
		Hayatta	92142	231		*Hayatta*	92291	82
		Ölü	3275	157		*Ölü*	3424	8
Extra tree	*Gerçek Sınıf*		*Hayatta*	*Ölü*	**PCA+ Extra Tree**	*Gerçek Sınıf*	*Hayatta*	*Ölü*
		Hayatta	90325	2048		*Hayatta*	91927	446
		Ölü	2703	729		*Ölü*	3200	232
Random Tree	*Gerçek Sınıf*		*Hayatta*	*Ölü*	**PCA+ Random Tree**	*Gerçek Sınıf*	*Hayatta*	*Ölü*
		Hayatta	90441	1932		*Hayatta*	91941	432
		Ölü	2724	708		*Ölü*	3210	222
Random Forest	*Gerçek Sınıf*		*Hayatta*	*Ölü*	**PCA+ Random Forest**	*Gerçek Sınıf*	*Hayatta*	*Ölü*
		Hayatta	91225	1148		*Hayatta*	91879	494
		Ölü	2858	574		*Ölü*	3176	256
Naive Bayes	*Gerçek Sınıf*		*Hayatta*	*Ölü*	**PCA+ Naive Bayes**	*Gerçek Sınıf*	*Hayatta*	*Ölü*
		Hayatta	74460	17913		*Hayatta*	82279	10094
		Ölü	444	2988		*Ölü*	1387	2045

Şekil 4. Karışıklık matrisi (Confusion matrix)

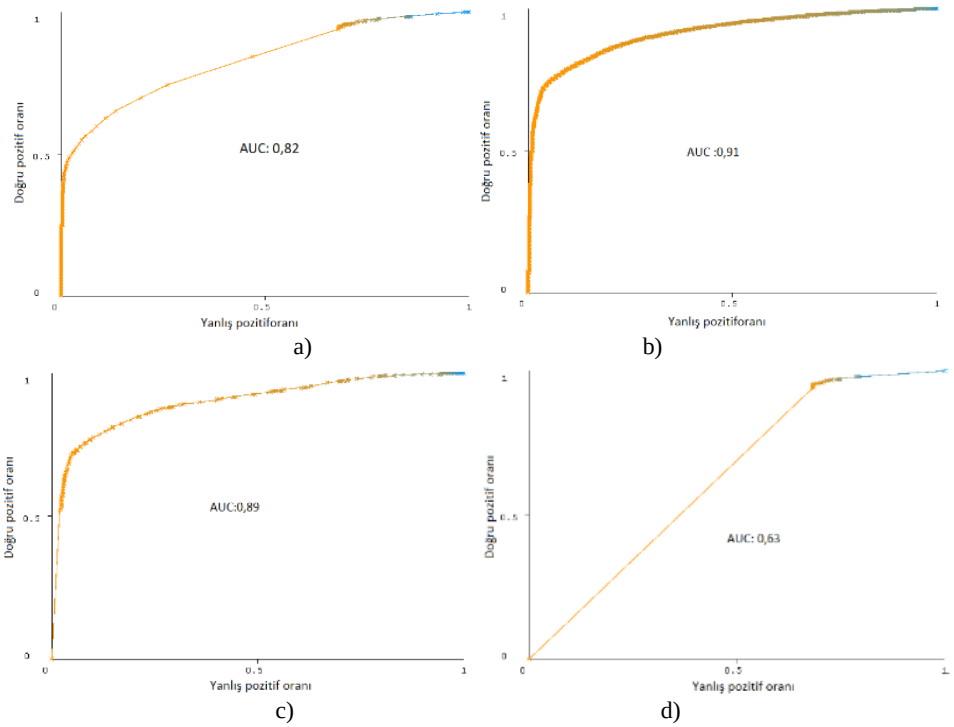


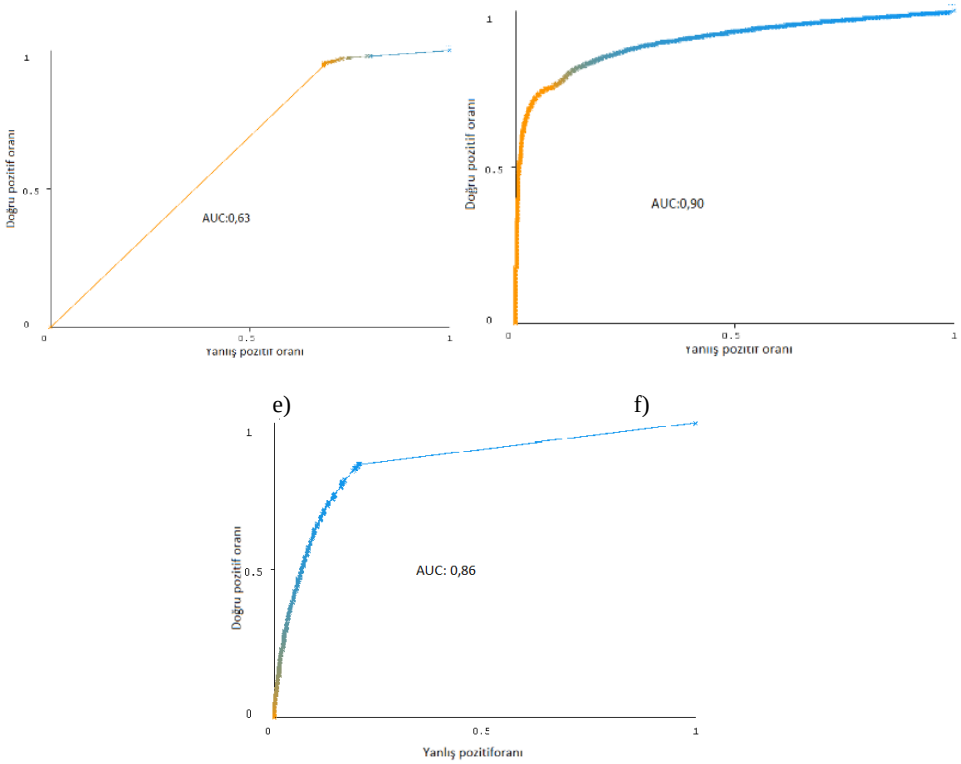
Şekil 5. Doğruluk değerleri (Accuracy values)

Tablo 2. Performans metrikleri (Performance metrics)

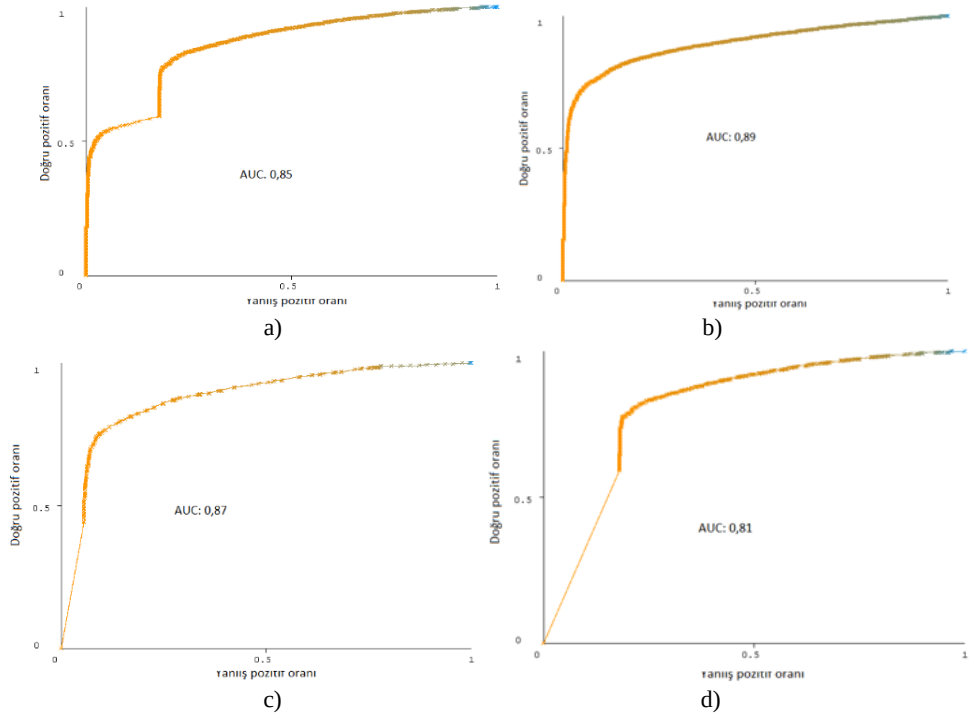
Yöntem	Kesinlik	Duyarlılık	F-Skoru	Rms	Özellik sayısı	Doğru sınıflandırılan örnek sayısı	Modelin Oluşturulma Süresi (sn)	Toplam İşlem Süresi (sn)
KNN	0,94	0,95	0,94	0,21	19	91381	0,11	1440
PCA+ KNN	0,94	0,96	0,94	0,18	9	92202	0,02	1320
LDA	0,94	0,96	0,95	0,17	19	92247	0,22	5
PCA+LDA	0,93	0,96	0,94	0,17	9	92344	0,12	3
HF	0,94	0,96	0,94	0,17	19	92299	0,94	10
PCA+HF	0.93	0.96	0.94	0,17	9	92299	0,71	6
Extra tree	0,94	0,95	0,94	0,22	19	91054	0,39	9
PCA+ Extra tree	0,94	0,96	0,94	0,18	9	92159	0,33	8
Random tree	0,94	0,95	0,94	0,22	19	91149	1,86	18
PCA+ Random tree	0,94	0,96	0,94	0,18	9	92163	1,16	13
Random Forest	0,94	0,95	0,95	0,18	19	91799	89,7	1200
PCA+Random Forest	0,94	0,96	0,95	0,17	9	92135	69,96	840
Naive Bayes	0,96	0,80	0,86	0,37	19	77448	0,3	5
PCA+Naive Bayes	0,95	0,88	0,91	0,27	9	86369	0,23	4

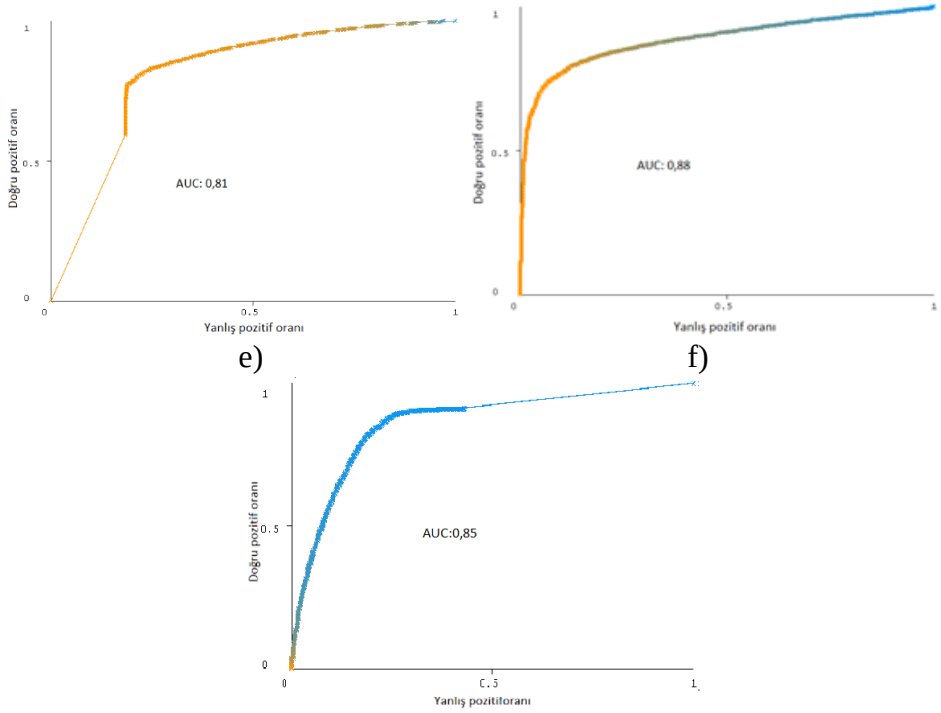
Temel bileşen analizinin uygulanması ile bütün algoritmaların işlem süreleri azalmış ve doğruluk oranları artmıştır. En düşük doğruluk oranı Naive bayes algoritmasına aittir. Naive bayes algoritmasının doğruluk oranı %80 dir. PCA uygulaması ile bu oran %88,01'e çıkmıştır. En yüksek doğruluk oranı ise PCA ve LDA algoritması uygulanması ile elde edilmiştir. Bu metrikler dışında kullanılan diğer metrik ROC eğrileridir. Yatay eksen yanlış pozitif oranını düşey eksen ise doğru pozitif oranını gösterecek şekilde çizilen eğrilere ROC eğrileri denir. Bu eğrilerin altında kalan alan da algoritmanın performansının değerlendirilmesinde kullanılır. Bu alan kısaca AUC ile gösterilir. Algoritmaların PCA yöntemi kullanılmadan ve PCA yöntemi kullanılması ile ölüm oranı tahminlerinden elde edilen ROC eğrileri Şekil 6 ve Şekil 7'te verilmiştir.





Şekil 6. ROC eğrileri a) KNN b) LDA c) HF d) Extra tree e) Random tree f) Naivebayes g)Random Forest





Şekil 7. ROC eğrileri a) PCA+KNN b) PCA+LDA c) PCA+HF d)PCA+Extra tree e) PCA+Random tree f) PCA+Naive Bayes g) PCA+Random Forest

İyi bir sınıflandırma için performans metriklerinin 1'e yakın değerler almaları gereklidir. Duyarlılık, Pozitif durumların ne kadar başarılı tahmin edildiğini gösteren bir metriktir. Algoritmaların genel olarak duyarlılık değerleri 0,95 ve üstü olması nedeniyle ölüm oranı tahminlerinde başarılı oldukları söylenebilir. Kesinlik, algoritmanın pozitif olarak tahmin ettiği değerlerin gerçekte kaç tanesinin pozitif olduğunu gösteren bir metriktir. F-skoru daha çok dengesiz veri setlerinin sınıflandırılmasında kullanılır. Yapılan tahminlerdeki hata oranını belirten RMS değeri de PCA yöntemi kullanılması ile azalmıştır.

4. Sonuçlar (Conclusions)

Bütün dünyayı etkileyen ve milyonlarca insanın ölümüne sebep olan Covid 19 virüsünden kaynaklanan ölüm oranlarının tahmin edilmesi için çalışmada PCA yöntemi ve sınıflandırma algoritmalarından yararlanılmıştır. Başarı oranını yükseltmek ve hangi özelliklere sahip olan hastanın ölüm riskinin yüksek olduğunun tespitinin yapılabilmesi için PCA yöntemi kullanılmıştır. Veri setinden PCA yöntemi kullanılmadan ve PCA yöntemi kullanıldıktan sonra yapılan sınıflandırma olmak üzere 2 farklı sınıflandırma uygulaması yapılmıştır. Sonuçların karşılaştırılmasında ise algoritmaların performans değerlendirmesinin yapılmasında kullanılan performans metrikleri incelenmiştir. Bu metrikler göz önüne alınarak yapılan değerlendirmede PCA yöntemi uygulandıktan sonra sınıflandırma işleminin yapılması durumunda bütün algoritmalarda performans metriklerinde bir iyileşme olduğu sınıflandırma başarısının arttığı görülmüştür. Bununla birlikte PCA uygulamasının yapılması ile hastanın cinsiyeti, Entübe, Zatürre olma durumu, yaşı, hamile olması ayrıca Şeker, KOAH ve Astım hastalığının bulunması durumu ölüm riskinin artmasında etkin rol oynadığı anlaşılmıştır. Bu hastalıklara sahip yaşlı insanların bu hastalıktan korunmak için daha dikkatli olmaları gerektiği, yakalanan insanların ise tedavisinde daha özen gösterilmesi gerekmektedir.

Kaynaklar (References)

Abdi, H., & Williams, L. J. 2010. Principal component analysis. Computational Statistics.

Akhtar, A., Akhtar, S., Bakhtawar, B., Kashif, A. A., Aziz, N., & Javeid, M. S. 2021. COVID-19 Detection from CBC using Machine Learning Techniques. International Journal of Technology, Innovation and Management (IJTIM), 1(2), 65-78.

Albahri, A. S., Hamid, R. A., Alwan, J. K., Al-Qays, Z., Zaidan, A., Zaidan, B., . . . Almahdi, E. 2020. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. Journal of medical systems, 44, 1-11.

Amasyali, M. F., & Ersoy, O. 2009. Evaluation of regression ensembles on drug design datasets.

Bello-Chavolla, O. Y., Bahena-López, J. P., Antonio-Villa, N. E., Vargas-Vázquez, A., González-Díaz, A., Márquez-Salinas, A., . . . Aguilar-Salinas, C. A. (2020). Predicting mortality due to SARS-CoV-2: a mechanistic score relating obesity and diabetes to COVID-19 outcomes in

Mexico. *The Journal of Clinical Endocrinology & Metabolism, 105*(8), 2752-2761.

Bermejo, P., Gámez, J. A., & Puerta, J. M. 2011. Improving the performance of Naive Bayes multinomial in e-mail foldering by introducing distribution-based balance of datasets. Expert Systems with Applications, 38(3), 2072-2080.

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

COVID-19 Mexico Patient Health Dataset. (2020, 05 19). Retrieved from Kaggle.com: https://www.kaggle.com/datasets/riteshahlawat/covid19-mexico-patient-health-dataset

Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., . . . Sun, K. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science, 368*(6489), 395-400.

de León, U. A.-P., Pérez, Á. G., & Avila-Vales, E. (2020). An SEIARD epidemic model for COVID-19 in Mexico: mathematical analysis and state-level forecast. *Chaos, Solitons & Fractals, 140*, 110165.

Drew, D. A., Nguyen, L. H., Steves, C. J., Menni, C., Freydin, M., Varsavsky, T., . . . Wolf, J. (2020). Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science, 368*(6497), 1362-1367.

Escobedo-de la Peña, J., Rascón-Pacheco, R. A., de Jesús Ascencio-Montiel, I., González-Figueroa, E., Fernández-Gárate, J. E., Medina-Gómez, O. S., . . . Borja-Aburto, V. H. (2021). Hypertension, diabetes and obesity, major risk factors for death in patients with COVID-19 in Mexico. *Archives of medical research, 52*(4), 443-449.

Freund, Y., & Mason, L. 1999. The alternating decision tree learning algorithm. Paper presented at the icml.

Gansevoort, **R.** T., & Hilbrands, L. B. (2020). CKD is a key risk factor for COVID-19 mortality. *Nature Reviews Nephrology, 16*(12), 705-706.

Guan, X., Zhang, B., Fu, M., Li, M., Yuan, X., Zhu, Y., . . . Lu, Y. 2021. Clinical and inflammatory features-based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study. Annals of Medicine, 53(1), 257-266.

Jagodnik, K. M., Ray, F., Giorgi, F. M., & Lachmann, A. 2020. Correcting under-reported COVID-19 case numbers: estimating the true scale of the pandemic. medRxiv.

Kassania, S. H., Kassanib, P. H., Wesolowskic, M. J., Schneidera, K. A., & Detersa, R. 2021. Automatic detection of coronavirus disease (COVID-19) in X-ray and CT images: a machine learning based approach. Biocybernetics and Biomedical Engineering, 41(3), 867-879.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992* (pp. 249-256): Elsevier.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence, 97*(1-2), 273-324.

Levin, A. T., Cochran, K., & Walsh, S. 2020. Assessing the age specificity of infection fatality rates for COVID-19: Meta-analysis & public policy implications. NBER Working Paper(w27597).

Li, B., Yu, S., & Lu, Q. 2003. An improved k-nearest neighbor algorithm for text categorization. arXiv preprint cs/0306099.

Li, J., Zhang, S., Lu, Y., & Yan, J. 2008. Real-time P2P traffic identification. Paper presented at the IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference.

Li, M., Zhang, Z., Cao, W., Liu, Y., Du, B., Chen, C., . . . Chen, C. 2021. Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach. Science of the Total Environment, 764, 142810.

Maglaveras, N., Stamkopoulos, T., Diamantaras, K., Pappas, C., & Strintzis, M. 1998. ECG pattern recognition and classification using non-linear transformations and neural networks: A review. International journal of medical informatics, 52(1-3), 191-208.

Manski, C. F., & Molinari, F. 2021. Estimating the COVID-19 infection rate: Anatomy of an inference problem. Journal of Econometrics, 220(1), 181-192.

Muhammad, L., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN computer science, 2*(1), 1-13.

Nemati**,** M., Ansary, J., & Nemati, N. (2020). Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data. *Patterns, 1*(5), 100074.

Novaković, J., Strbac, P., & Bulatović, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of operations research, 21*(1), 119-135.

Parra-Bracamonte**,** G. M., Lopez-Villalobos, N., & Parra-Bracamonte, F. E. (2020). Clinical characteristics and risk factors for mortality of patients with COVID-19 in a large data set from Mexico. *Annals of epidemiology, 52*, 93-98. e92.

Quiroz-Juárez, M. A., Torres-Gómez, A., Hoyo-Ulloa, I., León-Montiel, R. d. J., & U'Ren, A. B. (2021). Identification of high-risk COVID-19 patients using machine learning. *Plos one, 16*(9), e0257234.

Quinlan, J. R. 1999. Simplifying decision trees. International Journal of Human-Computer Studies, 51(2), 497-510.

Rahmandad, H., Lim, T. Y., & Sterman, J. 2020. Estimating COVID-19 under-reporting across 86 nations: implications for projections and control. medRxiv.

Ríos-Silva, M., Murillo-Zamora, E., Mendoza-Cano, O., Trujillo, X., & Huerta, M. (2020). COVID-19 mortality among pregnant women in Mexico: a retrospective cohort study. *Journal of Global Health, 10*(2).

Singh, J., Green, M. B., Lindblom, S., Reif, M. S., Thakkar, N. P., & Papali, A. (2021). Telecritical care clinical and operational strategies in response to COVID-19. *Telemedicine and e-Health, 27*(3), 261-268.

Stella, L., Martínez, A. P., Bauso, D., & Colaneri, P. 2020. The role of asymptomatic individuals in the Covid-19 pandemic via complex networks. arXiv preprint arXiv:2009.03649.

Ünsal, A., Bileşenler, Ö., _Faktür, M., & Mali, D. A. Y. I. Ş. 1996. Başarılarının Analizi. In: Ankara.

Velavan, T. P., & Meyer, C. G. 2020. The COVID-19 epidemic. Tropical medicine & international health, 25(3), 278.

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.

Yang, S., Cao, P., Du, P., Wu, Z., Zhuang, Z., Yang, L., . . . Wang, X. 2020. Early estimation of the case fatality rate

of COVID-19 in mainland China: a data-driven analysis. Annals of translational medicine, 8(4).

Yavuz, Ü., & Dudak, M. N. 2020. Classification of covid-19 dataset with some machine learning methods. journal of amasya university the institute of sciences and technology, 1(1), 30-37.

Zawiah, M., Al-Ashwal, F. Y., Saeed, R. M., Kubas, M., Saeed, S., Khan, A. H., . . . Abduljabbar, R. (2020). Assessment of healthcare system capabilities and preparedness in Yemen to Confront the novel coronavirus 2019 (COVID-19) outbreak: a perspective of healthcare workers. *Frontiers in public health*, 419.

Zens, M., Brammertz, A., Herpich, J., Südkamp, N., & Hinterseer, M. (2020). App-based tracking of self-reported COVID-19 symptoms: analysis of questionnaire data. *Journal of medical Internet research, 22*(9), e21956.

Zhang, Y., Ding, L., & Wang, Y. 2011. Research and design of ID3 algorithm rules-based anti-spam email filtering. Paper presented at the 2011 IEEE 2nd International Conference on Software Engineering and Service Science.