Araştırma Makalesi / *Research Article*

Ege Tıp Dergisi / *Ege Journal of Medicine* 2021; 60 (4): 384-393

Rutin kan testleriyle COVID-19 tanı tahmininde makine öğrenmesi yöntemleriyle bir mobil uygulama geliştirilmesi

Development of a mobile application by using machine learning methods for the prediction of COVID-19 diagnosis with routine blood tests

Mert Demirarslan Aslı Suner

Ege Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, İzmir, Türkiye

ÖZ

Amaç: Tüm dünya Aralık 2019'dan bu yana SARS-CoV-2 virüsü ile başa çıkmaya çalışmaktadır. Hastalığın erken belirtileri, soğuk algınlığı ve grip gibi diğer yaygın durumlarla örtüştüğünden, hekimler için erken tanının önemi büyüktür. Bu çalışmada, genel kullanıma açık anonim bir veri seti kullanılarak, rutin kan testleri sonuçları üzerinden Yeni Koronavirüs Hastalığı (COVID-19) tanısının (pozitif/negatif) makine öğrenmesi algoritmaları yardımıyla tahmin edilmesine yönelik bir mobil uygulama geliştirilmesi amaçlanmaktadır.

Gereç ve Yöntem: Veri setinde yer alan, kayıp gözlem, sınıf dengesizliği, aykırı gözlem ve ilgisiz değişken problemleri giderildikten sonra makine öğrenmesi yöntemlerinin sınıflandırma performansları test edilmiş, ardından uygun değişkenlerle COVID-19 tanısı için lojistik regresyon modeli kurulmuştur. Bu model kullanılarak makine öğrenmesi tabanlı mobil uygulaması tasarlanmıştır.

Bulgular: Tanı koymada en iyi sonuç veren değişkenler, eozinofil, lökosit, trombosit, monosit, kırmızı kan hücresi, bazofildir. Veri ön işleme problemleri giderildikten sonra kullanılan algoritmaların sınıflandırma performansları, ham verideki performans değerlerine göre oldukça yükselmiştir.

Sonuç: Geliştirilen mobil uygulama ile rutin kan testi sonuçları kullanılarak, hızlı ve kolay bir şekilde Covid-19 tanısı tahmininde bulunulması mümkündür.

Anahtar Sözcükler: Covid-19, makine öğrenmesi yöntemleri, rutin kan testi, mobil uygulama, tanı.

ABSTRACT

***Objective**: The whole world has been dealing with the SARS-CoV-2 virus since December 2019. Early diagnosis is of great importance for physicians, as the early symptoms of the disease overlap with other common conditions such as cold and flu. In this study, we aimed to develop a mobile application to diagnose COVID-19 with machine learning algorithms that use anonymized publicly available routine blood tests results.*

***Materials and Methods**: After eliminating the missing observation, class imbalance, outlier observation, and unrelated variable problems in the data set, the classification performances of machine learning methods were tested, and then a logistic regression model was established for the detection of COVID-19 with appropriate variables. Using this model, a machine learning-based mobile application has been designed.*

***Results:** The variables that gave the best results in diagnosis were eosinophils, leukocytes, thrombocytes, monocytes, red blood cells, and basophils. After solving the data pre-processing problems, the classification performance of the algorithms used has increased considerably compared to the performance values in the raw data.*

***Conclusion**: With the developed mobile application, it is possible to estimate the diagnosis of Covid-19 quickly and easily by using routine blood test results.*

***Keywords:** Covid-19, machine learning methods, routine blood test, mobile application, diagnosis.*

Sorumlu yazar: Aslı Suner
Ege Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, İzmir, Türkiye
E-posta: *asli.suner@ege.edu.tr*
Başvuru tarihi: 28.04.2021 Kabul tarihi: 10.06.2021

GİRİŞ

Tüm dünya Aralık 2019'dan beri Yeni Koronavirüs Hastalığı (COVID-19) pandemisinin etkisi altında, SARS-CoV-2 virüsü ile baş etmeye çalışmaktadır. Çin'in Wuhan eyaletinde ilk gözlenen vakalardan bu yana (27 Nisan 2021 itibariyle) global olarak 147.377.159 kişi enfekte olmuş, 3.112.041 kişi ise hayatını kaybetmiştir (1). Toplam vaka sayısı açısından ilk üç ülke incelendiğinde; Amerika 1. sırada yer alırken (31.742.914 kişi), Hindistan 2. sırada (17.636.307 kişi), Brezilya ise 3. sırada (14.340.787 kişi) bulunmaktadır. Ülkemizin ise doğrulanmış 4.629.696 vaka ile Fransa ve Rusya'dan sonra 6. sırada yer aldığı belirtilmiştir. Dünyadaki ülkelerin virüsten etkilenmesi sonucunda, sokağa çıkma yasakları, seyahat kısıtlamaları ve tamamen kapanma gibi çeşitli tedbirler uygulanmaktadır. Artan COVID-19 vaka sayılarını yönetmek, dünya çapında sağlık hizmetleri için büyük zorluk oluşturmaktadır. Hastalık, yüksek bulaşıcılığı nedeniyle küresel bir salgın halini aldığından, farklı ülkelere ait veri tabanlarında araştırmacıların kullanımına sunulan ve sürekli güncellenen veriler ile virüse ilişkin farklı amaçlara yönelik çalışmaların yapılmasına olanak tanınmaktadır. SARS-CoV-2 enfeksiyonunun ortaya çıkmasından bu yana, çeşitli disiplinlerden araştırmacılar bu yeni virüsü araştırmaktadırlar.

Makine öğrenmesi, sağlık hizmetleri ve tıp bilişimi gibi alanlarda oldukça yaygın bir şekilde kullanılan bir yapay zeka dalı olarak, COVID-19 tanısının mortalitesi ve risk tahmini ile ilgili çok sayıda araştırmada kullanılmaya başlanmıştır (2). Bu araştırma alanı günümüzde oldukça aktif olduğundan ilgili yayınların sayısı hızla artmaktadır. Nisan 2021 itibariyle, PubMed taramasında makine öğrenmesi ve COVID-19 ile ilgili 2020 yılından bu yana yapılmış 1098 çalışma bulunmaktadır.

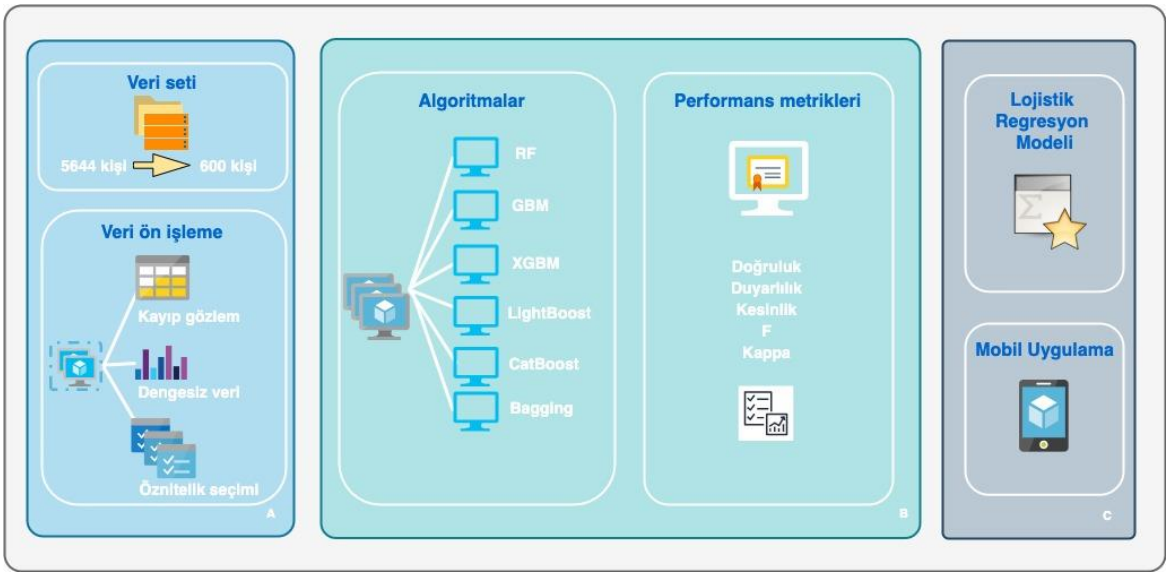
Makine öğrenmesi alanında pek çok algoritma mevcut olduğundan, bir kullanıcının verilen her sorun için en iyi algoritmayı bulması oldukça zordur. Bu nedenle bu çalışma kapsamında kullanılan topluluk öğrenme yöntemleri ile birden fazla modeli birleştirip, bir "meta" öğrenme şeması oluşturarak ve farklı hipotez alanlarını birleştirerek, en ideal çözüme "yakın" yaklaşımlar elde edilmiştir (3). Topluluk öğrenme algoritmalarını kullanırken en iyi yöntemler hangileri ve kaç tanesinin gerekli olduğu gibi yeni sorunlar da beraberinde geldiğinden, algoritmaların performansları kıyaslanarak en yüksek performans değerine sahip olan algoritma belirlenmektedir.

COVID-19'un yaygın semptomları genellikle hafif olsa da bazı hastalar için enfeksiyon ciddi ve bazen ölümcül komplikasyonlara neden olabilmektedir. SARS-CoV-2 enfeksiyonu ateş ve öksürük gibi en yaygın semptomları, başka bulaşıcı hastalıklara benzer semptomlar olduğundan, bu durum hekimlerin tanı aşamasında hızlı karar vermesini zorlaştırmaktadır (4). PCR testi şu anda en güvenilir tanı testi olarak kullanılsa da bazen doğru tanı koyma konusunda sıkıntı yaşanabilmekte ve testin sonuç vermesi zaman alabilmektedir. Bu çalışmada, şüpheli bir COVID-19 vakasına rutin kan sayımı verilerini kullanarak tanı konulmasına yardımcı olmak ve hastalık tanısına ilişkin kararların iyileştirilmesi için veri madenciliği algoritmaları, veri ön işleme yöntemleri sonrasında kullanılmıştır. Bu amaçla, ortak paylaşıma açık bir hastane verisi kullanılarak, yaygın olarak toplanan laboratuvar test sonuçları ile SARS-Cov-2 (pozitif/negatif) sonucunun makine öğrenmesi algoritmaları kullanılarak tahmin edilmesi için bir mobil uygulama geliştirilmiştir. Bu verilerde bulunan kayıp gözlem ve sınıf dengesizliği problemleri giderildikten sonra; ilgisiz değişkenler öznitelik seçim yöntemiyle veri setinden uzaklaştırılmış ve literatürde yaygın olarak kullanılan en güncel makine öğrenmesi sınıflandırma algoritmaları kullanılarak algoritmaların performanslarının arttırılması hedeflenmiştir.

GEREÇ ve YÖNTEM

1. Veri seti

Çalışmada kullanılan COVID-19 hasta verileri, Kaggle veri tabanında paylaşılan, Brezilya'nın São Paulo şehrinde Israelita Albert Einstein Hastanesi'nde 28 Mart 2020 ile 3 Nisan 2020 arasında toplanan 5644 kişiye ait anonim verilerden oluşmaktadır (5). Tüm veriler, en iyi uluslararası uygulamalar ve öneriler izlenerek hastane tarafından anonimleştirilip paylaşılmıştır. Cinsiyet dışındaki tüm değişkenler sayısal değerler içerdiğinden, ölçeklerdeki farklılığa bağlı aşırı büyük etkilerden kaçınmak için veriler normalize edilmiştir. Veri tabanındaki tüm klinik veriler, sıfır ortalamaya ve birim standart sapmaya sahip olacak şekilde standardize edilmiştir. İstatistiksel analizlerde 5644 sonuç, tam kan sayımı dışındaki verilerinin eksikliğinden dolayı kullanılmadığından, sadece 600 hastaya ait eksiksiz hasta verileri çalışmaya dahil edilmiştir



Şekil-1. Çalışmanın akış şeması.

Veri setinde yer alan 15 adet değişken arasında; yaş (yüzdelik grup), rt-PCR SARS-CoV-2 testinin sonucu ve standart tam kan sayımı: hematokrit, hemoglobin, trombosit, ortalama trombosit hacmi (MPV), kırmızı kan hücreleri (RBC), lenfosit, ortalama korpüsküler hemoglobin konsantrasyonu (MCHC), lökosit, bazofil, nötrofil, ortalama korpüsküler hemoglobin (MCH), eozinofil, ortalama korpüsküler hacim (MCV), monosit ve kırmızı kan hücresi dağılım genişliği (RBCDW) yer almaktadır.

Çalışmada kullanılan verilerin istatistiksel analizleri için IBM SPSS 25, makine öğrenmesi algoritmaları için Python, mobil uygulama için Android Studio'da Java dili kullanılmıştır. Şekil-1'de çalışmanın akış şeması özetlenmiştir.

2. Veri Ön İşleme

Yapay zekâ algoritmalarının doğru çalışmasını engelleyen ve düşük performans göstermesine neden olan kayıp gözlem, sınıf dengesizliği ilgisiz değişken ve aykırı gözlem gibi yaygın olarak karşılaşılan bazı problemler bulunmaktadır (6). Çalışmada incelenen veri setinde de bu problemlerle karşılaşılmıştır. Veride çok fazla kayıp gözlem olduğundan, bu problemi gidermek için atama işlemi yapılamamaktadır. Bu nedenle kayıp gözlem sayısı diğer değişkenlere göre az olan, tam kan sayımına ait değişkenler çalışmada kullanılmış, kayıp gözlem problemi ile baş etmek için satır bazında silme işlemi uygulanmıştır. Pozitif ve negatif COVID-19 tanısı almış hasta sayıları dengesiz olduğundan (dengesizlik oranı=6,22), veri madenciliği algoritmalarını uygulamadan önce dengesiz veri setini dengeli hale getirmek için veri ön işleme adımı olarak SMOTE yaklaşımı kullanılmıştır. Oluşturulan modelde, kan sonuçlarındaki tüm değişkenleri kullanmak yerine en ilgili olan minimum sayıdaki değişkenin seçilmesi amaçlanmaktadır. İlgisiz değişken probleminin giderilmesinde, aykırı gözlem problemini de çözümleyen istatistiksel tabanlı öznitelik seçim yöntemi olan OCtS methodu kullanılmıştır (7).

3. Kullanılan Veri Madenciliği Yöntemleri ve İstatistiksel Analizler

COVID-19'un pozitif tanısında tahmin performansını test etmek için 6 popüler topluluk öğrenme algoritması olan Random Forest (RF), Gradient Boosting Machine (GBM), Extreme Gradient Boosting Machine (XGBM), LightBoost, CatBoost ve Bagging yöntemleri (Tablo-1) 10 kat çapraz doğrulama ile kullanılmıştır (8-12). Algoritmalar kullanılırken, %80 eğitim verisi ve %20 test verisi olarak alınmıştır. Kullanılan topluluk öğrenme algoritmaları ile sınıflama yapılmış, veri ön işleme aşamasında dengeli hale getirilen verilerde, COVID-19'un pozitif tanısı için lojistik regresyon modeli kurulmuştur. Lojistik regresyon modeli, bağımsız özelliklerin değerine bağlı olarak belirli bir sınıfa ait veri noktalarının olasılığını modellemekte kullanılmaktadır. Bu çalışmada, model kurulurken ileriye doğru seçim (forward selection) yöntemi ile bağımsız değişkenlerin seçilmesi sağlanmıştır. Öznitelik seçimi yönteminde elde edilen değişkenlerle kurulan lojistik regresyon modeli sonuçları kıyaslanmıştır.

Tablo-1. Kullanılan makine öğrenmesi algoritmalarının özelliklerinin incelenmesi.

Algoritma	Özellikleri
RF	• Torbalama (bootstrap) adı verilen yeniden örnekleme süreciyle birden çok farklı ağacın ortak kararını yansıtır. • Eğitim aşamasında oldukça hızlıdır.
GBM	• Adaboost yönteminin sınıflandırma ve regresyon problemlerine kolayca uyarlanabilir genelleştirilmiş versiyonudur.
XGBM	• GBM algoritmasının hız ve tahmin performansını arttırmak üzere optimize edilmiş; ölçeklenebilir ve farklı platformlara entegre edilebilir halidir.
LightBoost	• XGBM algoritmasının eğitim süresi performansını arttırmaya yönelik geliştirilen bir diğer GBM algoritmasıdır.
CatBoost	• Çok fazla kategorik değişkene ve sınıf sayısına sahip verilerde GBM algoritmasında sınıflandırma performansı düştüğünden, CatBoost algoritması geliştirilmiştir.
Bagging	• İyi bir performansa sahip en eski, en sezgisel ve en basit topluluk tabanlı algoritmalardan birisidir.

Verilerin normal dağılım varsayımının kontrolü için Shapiro-Wilk normallik testi kullanılmıştır. Rutin kan testindeki değerlerin gruplar arasında farklı olup olmadığını test etmek için Mann Whitney-U testinden faydalanılmıştır. Son olarak, algoritmalar için en yüksek sınıflandırma performansını veren parametreler optimize edilmiştir.

4. Yöntemlerin Performans Karşılaştırmaları

Performans metriklerinin hesaplanmasında kullanılan hata matrisinde Tablo-2'de gösterilen doğru pozitif (DP), yanlış pozitif (YP), yanlış negatif (YN) ve doğru negatif (DN) ifadeleri yer almaktadır. Gerçekte pozitif olan bir durumun (örneğin hastalığın varlığı gibi) pozitif olarak tahminlenmesi DP olarak tanımlanırken, gerçekte negatif olan bir durum için pozitif olarak tahminde bulunulması sonucunda YP durumu ortaya çıkmaktadır. Gerçek durum pozitif olduğunda negatif tahmin yapılması durumunda YN, gerçek durumu negatif iken negatif tahmin yapıldığında da DN durumu oluşmaktadır (13). Algoritmaların sınıflandırma performanslarının karşılaştırılmasında; doğruluk, duyarlılık, kesinlik, F-ölçütü ve Kappa istatistiği ölçüm metrikleri kullanılmıştır (14-16).

Tablo-2. Hata matrisi.

		Gerçek Durum	
		Pozitif	Negatif
Tahmin Durumu	Pozitif	DP	YP
	Negatif	YN	DN

Performans ölçüm metrikleri 0 ile 1 arasında değerler almakta; 1'e yaklaşan değerlere sahip algoritmaları iyi performans göstermektedir

Doğruluk (Accuracy-ACC): Doğru pozitif ve doğru negatif değerlerinin tüm değerlere olan oranı ile bulunan doğruluk değeri (Formül 1) ile hesaplanmaktadır:

$$Doğruluk = \frac{DP+DN}{DP+YP+YN+DN} \quad (1)$$

Duyarlılık (Sensitivite-SEN): Gerçekte pozitif olan bir durumun pozitif olarak tahminlenmesi ile ilgilenen duyarlılık değerinde (Formül 2) gerçekte hasta olan birinin tahmin değerinin ya da test sonucunun da pozitif gelmesine ilişkin olasılık belirlenmektedir.

$$Duyarlılık = \frac{DP}{DP+YN} \quad (2)$$

Özgüllük (Specificity-SPE): Gerçekte negatif olan bir durumun negatif olarak tahminlenmesinin belirlendiği özgüllük değerinde (Formül 3) gerçekte hasta olmayan bir kişinin tahmin değerinin ya da test sonucunun da negatif gelmesine ilişkin olasılık hesaplanmaktadır.

$$Özgüllük = \frac{DN}{DN+YP} \quad (3)$$

Kesinlik: Pozitif tahmin edilen bir durumun gerçekten pozitif olduğu olasılığının belirlendiği değer kesinlik değeri (Formül 4), F-ölçütünün hesaplanmasında kullanılmaktadır.

$$Kesinlik = \frac{DP}{DP+YP} \quad (4)$$

F-ölçütü: Özgüllük ve kesinlik ölçümlerinin bir arada değerlendirilmesini sağlayan F-ölçütünün hesaplanmasında bu iki değerin harmonik ortalaması kullanılmaktadır (Formül 5).

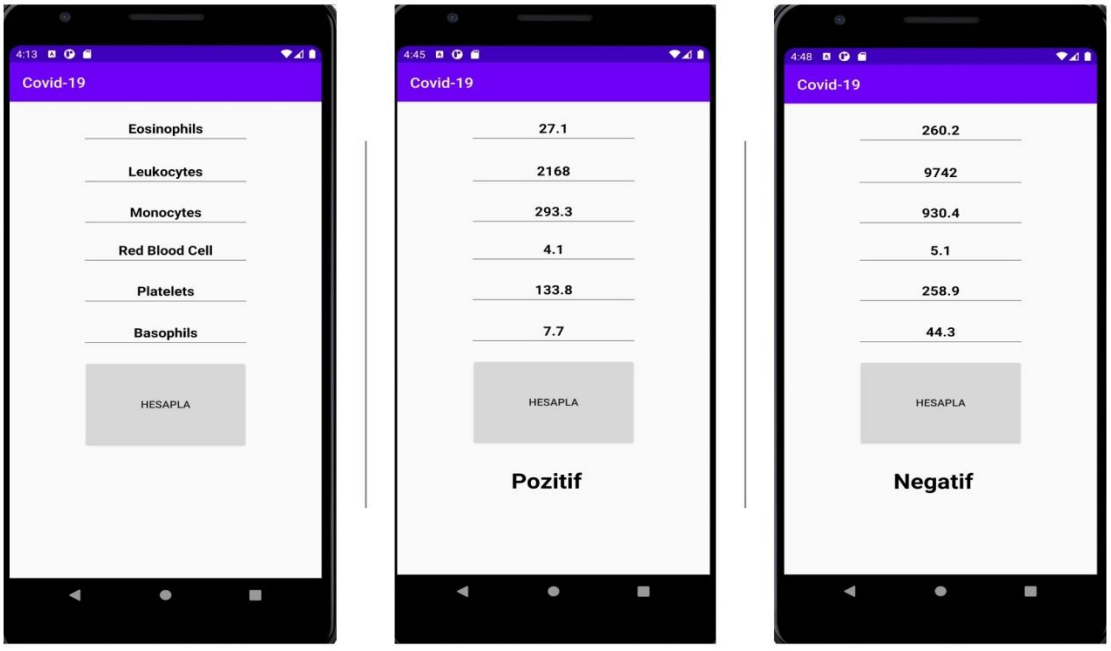
$$F\text{-}ölçütü = 2 * \frac{Özgüllük * Kesinlik}{Özgüllük + Kesinlik} \quad (5)$$

Kappa katsayısı: Gözlenen ve beklenen durumlar arasındaki uyumu ölçmede kullanılan Kappa istatistiği, bir sınıflandırıcının ne kadar iyi performans gösterdiğinin belirlenmesinde kullanılmaktadır. Formül 6'da p_0 gözlenen uyumu ve p_e beklenen uyumu belirtmektedir.

$$\textit{Kappa katsayısı}=\frac{p_o-p_e}{1-p_e} \quad (6)$$

5. Mobil Uygulamanın Geliştirilmesi

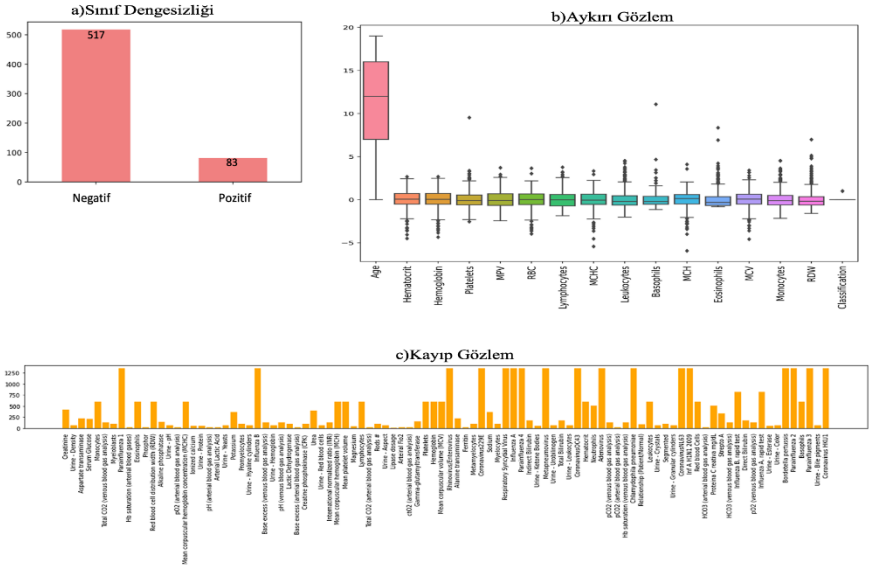
Hekimlerin COVID-19 tanısı konmasına yardımcı olmak üzere, hızlı ve pratik kullanım sağlayan makine öğrenmesi tabanlı mobil uygulama geliştirilmiştir. Uygulamada eozinofil, lökosit, monosit, RBC, trombosit ve bazofil olmak üzere 6 değişken yer almaktadır. Bu değişkenler, öznitelik seçiminin ardından elde edilen değişkenler ile kurulan ve istatistiksel olarak anlamlı çıkan lojistik regresyon modeli ile belirlenmiştir. Kullanıcı, hastanın ilgili kan sonuçlarını uygulamaya yazıp, hesapla butonuna tıkladığında, alt kısımda hastanın COVID-19 pozitif ya da negatif olduğuna ilişkin tanı tahmin sonucu ekranda yer almaktadır (Şekil-2).



Şekil-2. Geliştirilen mobil uygulamanın ekran görüntüsü.

BULGULAR

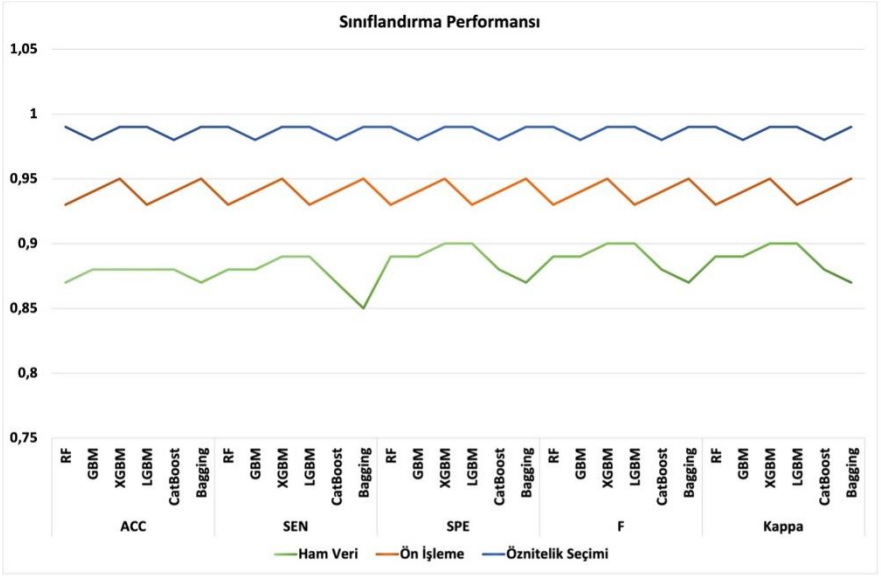
Araştırmada kullanılan COVID-19 hasta verisinde, sınıf dengesizliği (a), aykırı gözlemler (b) ve birçok kayıp gözlem (c) bulunmaktadır (Şekil-3). COVID-19 tanısı bağımlı değişkeninde pozitif olma durumu için kurulan modelde anlamlı bulunan bağımsız değişkenler eozinofil, lökosit, trombosit, monosit, RBC, bazofil şeklindedir. COVID-19 pozitif olanların negatif olanlara göre eozinofil için odds oranı 0,268, lökosit için odds oranı 0,291, trombosit için odds oranı 0,661, monosit için odds oranı 1,422, RBC için odds oranı 1,381 ve bazofil için odds oranı 0,798 olarak bulunmuştur. Öznitelik seçiminde, skor puanın yüksekliğine ve önem derecesine göre sıralanmıştır. Burada, değişkenlerin skorları bazofilin 1,08 değerinden sonra hızlı bir düşüş ile 0,50 altındaki değerlere indiğinden, eşik değeri 1,00 olarak belirlenmiştir. Seçilen ilk 6 değişken eozinofil, lökosit, trombosit, monosit, RBC ve bazofil ile kurulan lojistik modelinde de tüm değişkenler anlamlı çıkmıştır (Tablo-3). Seçilen değişkenlerle oluşturulan lojistik regresyon modeli için veri Logitboost (0,96 doğruluk) ve lojistik regresyon (0,86 doğruluk) algoritmaları ile eğitilip test edilmiştir. Rutin kan testindeki değerler olan hematokrit, hemoglobin, trombositler, MPV, RBC, MCHC, lökosit, bazofil, nötrofil, eozinofil, monosit ve RBCDW değişkenlerinin gruplar arasında istatistiksel olarak anlamlı derecede farklı olduğu tespit edilmiştir ($p<0,05$). Lenfositler ($p=0,250$), ortalama korpüsküler hemoglobin ($p=0,353$), ortalama korpüsküler hacim ($p=0,540$) değişkenlerinin ise gruplar arasında istatistiksel olarak farklı olmadığı belirlenmiştir.



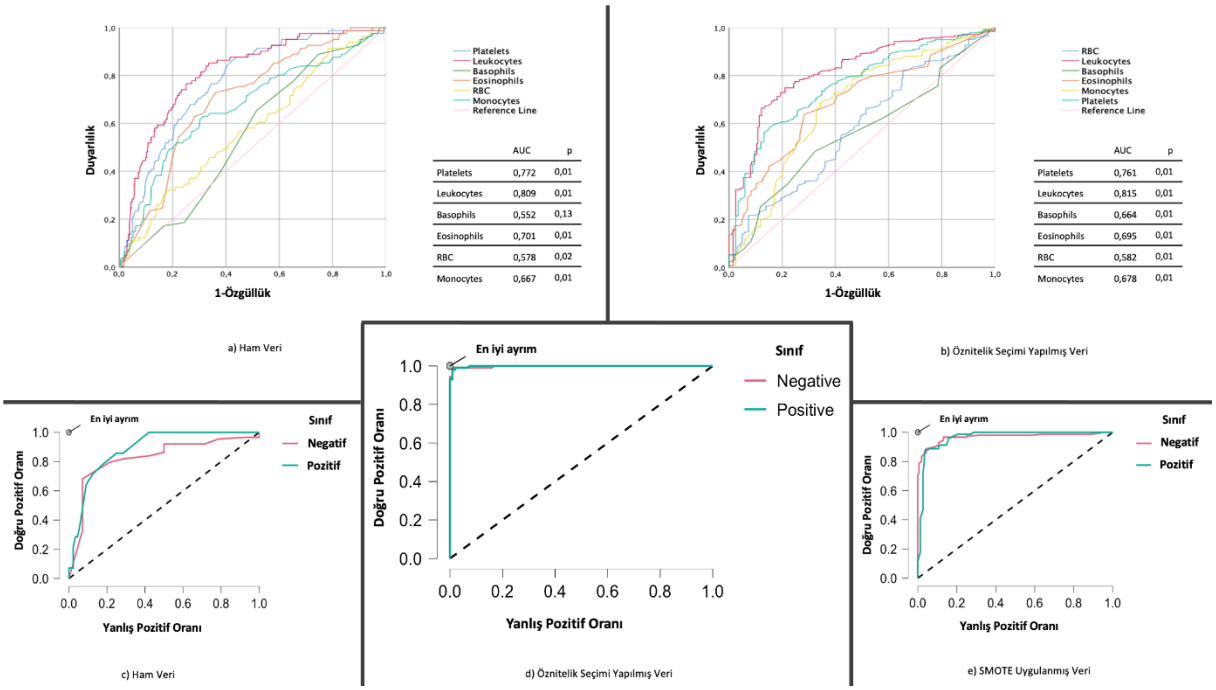
Şekil-3. Veri ön işleme problemlerinin incelenmesi: sınıf dengesizliği (a), aykırı gözlemler (b) ve kayıp gözlemler (c).

Tablo-3. Öznitelik seçimi sonrasında kurulan lojistik regresyon modelinde kullanılan değişkenler.

		Lojistik Regresyon			Öznitelik Seçimi
Değişken	**Odds Oranı**	**Katsayı**	**Güven Aralığı (%95)**	**p**	**Skor**
Eozinofil	0,268	1,315	0,225-0,420	0,001	4,31
Lökosit	0,291	1,234	0,178-0,336	0,001	2,95
Trombosit	0,661	0,413	0,554-0,879	0,001	2,11
Monosit	1,422	-0,352	1,202-1,693	0,001	1,81
RBC	1,381	-0,323	1,211-1,724	0,001	1,47
Bazofil	0,798	0,226	0,661-0,970	0,001	1,08
Sabit	0,340	1,078	-	-	-



Şekil-4. Algoritmalarının sınıflandırma performanslarının kıyaslanması.



Şekil-5. Değişkenlerin COVID-19 pozitif ayırt ediciliği için ROC eğrileri.

Tablo-4. İşlem akışına göre tüm topluluk öğrenmesi algoritmalarının sınıflandırma performanslarının değerlendirilmesi.

Veri Seti	Algoritma	Doğruluk	Duyarlılık	Kesinlik	F	Kappa
Ham Veri	**RF**	0,87	0,88	0,89	0,87	0,36
	GBM	0,88	0,88	0,89	0,87	0,45
	XGBM	0,88	0,89	0,90	0,88	0,49
	LightBoost	0,88	0,89	0,90	0,88	0,47
	CatBoost	0,88	0,87	0,88	0,86	0,34
	Bagging	0,87	0,85	0,87	0,85	0,28
SMOTE	**RF**	0,93	0,93	0,93	0,93	0,89
	GBM	0,94	0,94	0,94	0,94	0,90
	XGBM	0,95	0,95	0,95	0,95	0,91
	LightBoost	0,93	0,93	0,93	0,93	0,88
	CatBoost	0,94	0,94	0,94	0,94	0,90
	Bagging	0,95	0,95	0,95	0,95	0,91
Öznitelik Seçimi	**RF**	0,99	0,99	0,99	0,99	0,97
	GBM	0,98	0,98	0,98	0,98	0,96
	XGBM	0,99	0,99	0,99	0,99	0,97
	LightBoost	0,99	0,99	0,99	0,99	0,99
	CatBoost	0,98	0,98	0,98	0,98	0,96
	Bagging	0,99	0,99	0,99	0,99	0,97

Verinin ham halinin makine öğrenmesi algoritmalarındaki sınıflandırma performanslarına bakıldığında RF algoritmasında 0,87 doğruluk, 0,88 duyarlılık, 0,89 kesinlik, 0,87 F ölçütü ve 0,36 Kappa ölçütü değerleri hesaplanmıştır. SMOTE algoritması ile veriler dengeli hale getirildiğinde RF algoritmasının doğruluk, duyarlılık, kesinlik, F değerlerinin 0,94'e ve Kappa ölçütünün ise 0,90 değerine yükseldiği görülmektedir. Öznitelik seçimi sonrasında elde edilen değişkenlerle oluşturulan yeni veri setinin sınıflandırma performanslarına bakıldığında RF için doğruluk, duyarlılık, kesinlik, F değerleri 0,99'a; Kappa ölçütü de 0,97 değerine yükselmiştir. Ham veride en başarılı algoritmalar 0,88 doğruluk oranıyla; GBM, XGBM, LightBoost ve Catboost algoritmaları olmuştur. SMOTE işlemi uygulanan dengeli veride ise en yüksek doğruluk oranını 0,95 değeri ile XGBM algoritması göstermiştir. İlgili değişkenlerin seçiminin ardından GBM ve Catboost algoritmaları 0,98 doğruluk, RF, XGBM, LightBoost ve Bagging algoritmaları 0,99 doğruluk oranı sağlamıştır (Tablo-4). Genel olarak işlem akışına bakıldığında, tüm topluluk öğrenmesi algoritmalarının sınıflandırma performansları, verinin ham haline göre SMOTE yöntemi uygulandıktan sonra artış göstermiş; öznitelik seçimi yapıldıktan sonra ise daha yüksek performans değerleri elde edilmiştir (Şekil-4).

Verinin ham halindeki COVID-19 pozitif ayırt ediciliğine bakıldığında sadece bazofil (p=013) değişkeni istatistiksel olarak anlamlı bulunmamıştır (Şekil-5-a). Öznitelik seçimi yapılmış veride değişkenlerin COVID-19 pozitif ayırt ediciliği için ROC eğrisine bakıldığında, tüm değişkenler istatistiksel olarak anlamlı bulunmuştur ($p<0,05$). Ayırt edicilik gücünü anlamak için eğri altında kalan alana bakıldığında; trombosit 0,791 ile kabul edilir, lökosit 0,815 ile mükemmel, bazofil 0,664 ile kabul edilir, eozinofil 0,695 ile kabul edilir, RBC 0,582 ile zayıf ve monosit 0,678 ile kabul edilir ayrım gücüne sahiptir (Şekil 5-b). RF algoritmasının sınıf değişkeninde pozitif ve negatif ayırt ediciliğine bakıldığında, en iyi ayrım noktasına en yakın olan sırasıyla; öznitelik seçimi yapılmış veri, SMOTE yönteminin uygulandığı veri ve ham veri şeklindedir (Şekil-5, c-d-e).

Tüm veri ön işleme, öznitelik seçimi, makine öğrenmesi algoritmalarının performansları ve modelleme adımının uygulanmasının ardından, önemli bulunan değişkenlerle mobil uygulama oluşturulmuştur. Burada tam kan sayımından elde edilen en ilgili değişkenler ile COVID-19 tanı tahmini sağlamaktadır. Örneğin belirli bir hasta için eozinofil=27,1, lökosit=2168, monosit=293,3, RBC=4,1, trombosit=133,8, bazofil=7,7 şeklinde değerler tanımlandığında pozitif tanı sonucu tahminlenmektedir. Diğer bir örnekte ise, eozinofil=260,2, lökosit=9742, monosit=930,4, RBC=5,1, trombosit=258,9 ve bazofil=44,3 değerleri şeklinde girilen bir hasta için negatif tanı sonucu tahmini göstermektedir.

TARTIŞMA

Literatürde bu çalışmadaki veri setini kullanarak kayıp gözlem ve sınıf dengesizliği gibi problemlere veri ön işleme yapıldıktan sonra; RF, GBM, XGBM, LightBoost, Catboost ve Bagging algoritmalarının bir arada değerlendirildiği ve öznitelik seçimi yapılarak model kuran ve aynı zamanda algoritma performanslarını kıyaslayarak bu modelden geliştirmiş bir mobil uygulamaya rastlanmamıştır. Bunun yanı sıra bu çalışmada ham veri, işlenmiş veri ve öznitelik seçimi yapılmış olarak 3 durum da karşılaştırmalı olarak incelenmiştir.

Verinin ham halinde, algoritmaların sınıflandırma performansları oldukça düşük değerler göstermiştir. Bu durum veri setinde kayıp gözlem ve sınıf dengesizliği problemlerinin var olmasından kaynaklıdır. Bu problemlerin giderilmesinden sonra kullanılan algoritmaların sınıflandırma performansları, ham verideki performans değerlerine göre oldukça yükselmiştir.

Yavaş ve diğerlerinin (2020) yaptıkları çalışmada, aynı COVID-19 verisinde SMOTE kullanarak sınıflandırma yapmışlardır (17). Sınıflandırma yöntemlerinden derin öğrenme algoritmalarını kullanarak verinin ham halinde 0,86, dengelenmiş halinde 0,90 doğruluk değeri elde etmişlerdir. Derin öğrenme algoritmaları başarılı ancak açıklanabilir olmayan yapılardır. Ayrıca bu çalışmadaki doğruluk değeri bizim çalışmamızda kullanılan açıklanabilir ağaç tabanlı topluluk öğrenme algoritmalarına göre daha düşük performans göstermiştir. Yavaş ve diğerlerinin (2020) çalışmasında herhangi bir öznitelik seçim işlemi yapılmamış, optimal bir model önerilmemiş ve Kappa katsayısı ölçüm metriğini göz ardı edilmiştir (17). Banerji ve diğerleri (2020), tam kan sayımı ile COVID-19 tahmini yapmak için makine öğrenim ve derin öğrenme algoritmalarının performansını araştırmışlardır (18). Yapay sinir ağları (YSA) yöntemini

kullanarak %90 doğruluk değeri elde etmişler, ancak dengesiz veriler YSA'nın sınıflandırma performansını etkilediği için SMOTE yöntemi kullanarak verileri dengeli hale getirmişlerdir. Veriyi modellerken, veri setini yatan ve ayakta hasta olarak ikiye ayırarak incelemişlerdir. Dengeli verilerde, öznitelik seçimi yapmadan 14 değişkenle RF algoritmasını kullandıklarında yatan hastalarda COVID-19 pozitif tanısında %94 doğruluk elde ederken, ayakta hastalarda %86 doğruluk elde etmişlerdir. YSA'da yatan hastalarda %95, ayaktan hastalarda %80 doğruluk elde etmişlerdir. COVID-19 teşhisini pratik hale getirmek için; monosit, lökosit, eozinofil ve trombosit ile lojistik regresyon modeli kurmuşlardır. Model eğitildiğinde yatan hastalarda %81, ayaktan hastalarda %85 doğruluk göstermiştir. Modelde kullandıkların trombosit değerinin, COVID-19 tanısında influenza için ayırt edici olduğunu ve COVID-19 hastalarında pıhtılaşmanın arttığı belirtmişlerdir. Veride aynı zamanda MPV değişkeninde tespit ettikleri artışı, kemik iliği tarafından hızlı trombosit üretimi olarak yorumlamışlardır. Yazarlar bu çalışmada genel olarak, normalleştirilmiş verileri yorumlamanın zorluğunu vurgulamaktadırlar.

Yaşar ve Çolak (2020), çalışmalarında COVID-19 sınıflandırması için bir model önermişlerdir. Kayıp gözlem problemi için satır bazında silme uygulayarak 601 hasta verisi ve 10 değişken kullanmışlardır (19). Sınıflandırma algoritmaları dengesizlik probleminden olumsuz yönde etkilendiği için verileri dengelemiş ancak yöntemi paylaşmamışlardır. Çalışmada topluluk öğrenme algoritmalarından RF, basit öğrenme algoritmalarından CART, SVM, K-NN algoritmalarını kullanmışlardır. En yüksek %99 doğruluk oranıyla RF algoritması en yüksek başarı gösteren algoritma olmuştur. CART algoritması ise %81 doğruluk oranına sahiptir. Özniteliklerin sınıflandırılması için önem derecesinin sıralanmasında RF algoritmasını kullanmışlar ancak herhangi bir eşik değer belirleme ve öznitelik seçim işlemi yapmamışlardır.

Bizim bu çalışmamızda ise, kayıp gözlemler satır bazında silindikten sonra elde edilen 600 veri ve 14 adet değişken kullanılmıştır. Literatürde aynı veri setini kullanan çalışmalardan (17-19) farklı olarak, SMOTE yöntemiyle veri seti dengeli hale getirilmiş ve aykırı gözleme sahip değişkenler nedeniyle, aykırı gözlemlere duyarlı öznitelik seçim yöntemi tercih edilmiştir. Burada elde edilen değişkenlerle COVID-19 pozitif hastaları tespit etmek için lojistik regresyon modeli kurulmuş ve öznitelik seçiminden elde edilen her değişken istatistiksel olarak anlamlı bulunmuştur. Topluluk öğrenme algoritmaları genel olarak %99 sınıflandırma performansı göstermiştir. Analiz bulgularından elde edilen sonuçlarla geliştirilen mobil uygulama ile rutin kan testlerine ilişkin hasta bilgileri kullanılarak COVID-19 tanısının tahmininde kullanılacak bir model tasarlanmıştır.

COVID-19 tanısının konulmasında PCR testi ve radyolojik bulguların yanında rutin kan testlerinden de yararlanmaktadırlar. Tam kan sayımı sonuçlarındaki değişkenler arasındaki ilişkiyi kullanmaya yönelik klinik bilgi, sadeliği ve göreli kolay ölçülebilir değişkenleri nedeniyle oldukça önemlidir. Bu çalışmada geliştirilen model ile COVID-19 olma olasılığı en yüksek olan ve olmayan bireyleri ayırabilmeye karar desteği sağlamaya yardımcı bir tahmin modeli tasarlanmıştır. Sadece rutin kan testi değerlerine dayalı bir tahmin etme modeli, diğer viral hastalıklarda da görülebilen lenfopeni, lökopeni ve monositoz gibi bulgular ile de örtüşebileceğinden; bu bulguların da göz önünde bulundurulduğu durumlara ilişkin sonuçlar ile modelin sonuçları kıyaslanıp, COVID-19 ayırt ediciliğinin ilerleyen çalışmalarda çalışılması önerilmektedir. Çalışmada geliştirilen bu model bir klinik risk belirleme aracı olarak kullanılabilecek; makine öğrenmesi yöntemleriyle oluşturulan mobil uygulama, klinikteki karar vericilere mobil olarak karar desteği sağlama potansiyeline sahiptir.

SONUÇ

Kayıp gözlem, sınıf dengesizliği ve ilgisiz değişken problemleri giderildikten sonra topluluk öğrenme algoritmalarının sınıflandırma performanslarının arttığı görülmüştür. COVID-19 tanısının tahminlenmesi için oluşturulan model ile hızlı ve pratik bir ara yüze sahip mobil uygulama tasarlanmıştır. Hastanın tam kan sayımı sonuçlarındaki ilgili değişkenlere ait kan değerlerinin girilmesinin ardından COVID-19 tanısı tahminlenebilmektedir. Gelecek çalışmalarda, örneklem sayısının fazla olması, radyolojik bulgu sonuçları PCR testi sonuçları gibi başka değişkenlerin de modele eklenmesi ve farklı hastanelerden elde edilecek hasta sonuçları ile bölgesel çeşitliliğin sağlanması sınıflandırma algoritmalarının eğitilmesi açısından önemli fayda sağlayacaktır.

Çıkar Çatışması: Yazarlar çıkar çatışması beyan etmemişlerdir.

KAYNAKLAR

1. WHO Coronavirus (COVID-19) Dashboard Website [cited 27 April 2021]. Available from: https://covid19.who.int/
2. Alballa, N., & Al-Turaiki, I. Machine Learning Approaches in COVID-19 Diagnosis, Mortality, and Severity Risk Prediction: A Review. Informatics in Medicine Unlocked 2021; 100564.
3. Zhou, Z. H. Ensemble methods: Foundations and algorithms. In Ensemble Methods: Foundations and Algorithms. 1st Edition. New York: Chapman and Hall/CRC. 2012..
4. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. The Lancet 2020; 395(10229):1054-62.
5. Open Datasets and Machine Learning Projects|Kaggle [Internet]. Available from: https://www.kaggle.com/datasets
6. García, Salvador, Julián Luengo, and Francisco Herrera. Data preprocessing in data mining. Vol. 72. Cham, Switzerland: Springer International Publishing, 2015.
7. Demirarslan, M., & Suner, A. A Proposal of New Feature Selection Method Sensitive to Outliers and Correlation 2021; bioRxiv 2021.03.11.434934; doi: https://doi.org/10.1101/2021.03.11.434934
8. Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. Random Forests for land cover classification. Pattern Recognit Lett. 2005; 27 (4): 294-300. https://doi.org/10.1016/j.patrec.2005.08.011
9. Ke, G., Meng, Q., Finley, T., et al. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017; 30: 3146-54.
10. Chen, T., & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–94). New York, NY, USA: ACM; 2016 https://doi.org/10.1145/2939672.2939785
11. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst. 2018; 31.
12. Breiman,L. Bagging predictors. Machine Learning 1996; 24 (2): 123–40. https://doi.org/10.1007/bf00058655.
13. Ian Goodfellow, Yoshua Bengio, A. C. Deep Learning Book. Deep Learning 2015 https://doi.org/10.1016/B978-0-12-391420-0.09987-X.
14. Powers D. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. J of Machine Learn Tech 2011; 2 (1): 37-63.
15. Delgado R & Tibau X-A. Why Cohen's Kappa should be avoided as performance measure in classification. PLoS ONE 2019; 14 (9): e0222916. https://doi.org/10.1371/journal.pone.0222916
16. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas. 1960; 20 (1): 37-46. https://doi.org/10.1177/001316446002000104
17. Yavaş M, Güran A, ve Uysal M. Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması. Avrupa Bilim ve Teknoloji Dergisi. 2020:258-64. https://doi.org/10.31590/ejosat.779952
18. Banerjee A, Ray S, Vorselaars B, et al. Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. Int Immunopharmacol 2020; 86: 106705. https://doi.org/10.1016/j.intimp.2020.106705
19. Yaşar, Ş. & Çolak, C. A Proposed Model Can Classify the Covid-19 Pandemic Based on the Laboratory Test Results. Journal of Cognitive Syst 2020; 5 (2): 60-3.