Türkiye'de Zaman Serisi, Makine Öğrenmesi ve Yapay Sinir Ağı ile Covid-19 Vaka Tahmini

Covid-19 Case Prediction in Turkey using Time Series, Machine Learning, Artifical Neural Network

Erdem Demir
Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği
İstanbul Sabahattin Zaim Üniversitesi
İstanbul, Türkiye
demir_.erdem@hotmail.com

Muhammed Nafiz Canıtez
Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği
İstanbul Sabahattin Zaim Üniversitesi
İstanbul, Türkiye
nafizcntz@hotmail.com

Mohamed Elazab
Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği
İstanbul Sabahattin Zaim Üniversitesi
İstanbul, Türkiye
elazabmuhamed@gmail.com

***Öz*—Koronavirüs (Covid-19) hastalığı, 2019 yılının sonunda Çin'in Wuhan eyaletinde ortaya çıkan birçok insanın ve hayvanın hastalığına sebep olan ve hızla yayılan bir virüs türüdür. Bu çalışmanın yapıldığı dönemde hala yayılmaya devam ve kontrol altına alınamayan koronavirüs ile ilgili başta tıp alanı olmak üzere farklı disiplinlerde de yoğun çalışmalar sürmektedir. Bilgisayar bilimlerinde yer alan Zaman Serisi, Makine Öğrenmesi ve Yapay Sinir Ağı, mevcut veriler ile tahminler yapılabilmektedir. Bu amaçlarla geliştirilen SARIMAX, Ekstrem Gradyan Arttırma (XGBoost), Lineer Regresyon (Linear Regression), Karar Ağacı (Desicion Tree), Gradyan Arttırma (Gradient Boosting), Yapay Sinir Ağı (Artificial Neural Network) modelleri kullanılarak Türkiye Cumhuriyeti Sağlık Bakanlığının internet sitesinde açıklanan güncel veriler alınarak gerçek vakalarla karşılaştırılarak bu modellerin tutarlı sonuç verdikleri saptanmıştır.**

Anahtar Kelimeler*—*Covid-19, Makine Öğrenmesi, Vaka Tahminleme, SARIMAX, Ekstrem Gradyan Arttırma, Lineer Regresyon, Karar Ağacı, Gradyan Arttırma, Yapay Sinir Ağı

***Abstract*—Coronavirus (Covid-19) disease is a rapidly spreading type of virus that was discovered in Wuhan, China and emerged towards the end of 2019. In the period this study was conducted, intensive studies are continuing in different disciplines, especially in the field of medicine regarding coronavirus, which continues to spread and is yet to be controlled. Time Series, Machine Learning and Artificial Neural Network are a known field branched from computer science that has the ability of making predictions using existing data. Among these models that was developed for this purpose there are SARIMAX, XGBoost, *Linear Regression*, Decision Tree, Gradient Boosting, Artificial Neural Network. Predictions were made applying these models on the daily updated data announced on the website of the Ministry of Health of Turkey and it has been determined that these models produced consistent results by comparing the results with the real data.**

Keywords*—*Covid-19, Machine Learning, Case prediction, SARIMAX, XGBoost (eXtreme Gradient Boosting), Linear Regression, Decision Tree, Gradient Boosting, Artificial Neural Network.

I. GİRİŞ

Koronavirüs hastalığı, 21.yüzyılda ortaya çıkan insan solunum sistemini hedefleyen başlıca patojenlerden biridir [1]. Tarihsel sürece bakıldığında 2003 yılında Çin'in Guangdong eyaletinde, Şiddetli Akut Solunum Sendromu-CoV (Severe Acute Respiratuary Sendrom-CoV, [SARS-CoV]) [2] ve 2012 yılında ise Arap Yarımadası'nda ortaya çıkan Orta Doğu Solunum Sendromu-CoV (Middle East Respiratory Sendrome-CoV, [MERS-CoV]) [3] birer koronavirüs türleridir. Bu iki virüs türünden sonra Covid-19'da 21.yüzyılın üçüncü büyük salgını olarak görülmektedir [4].

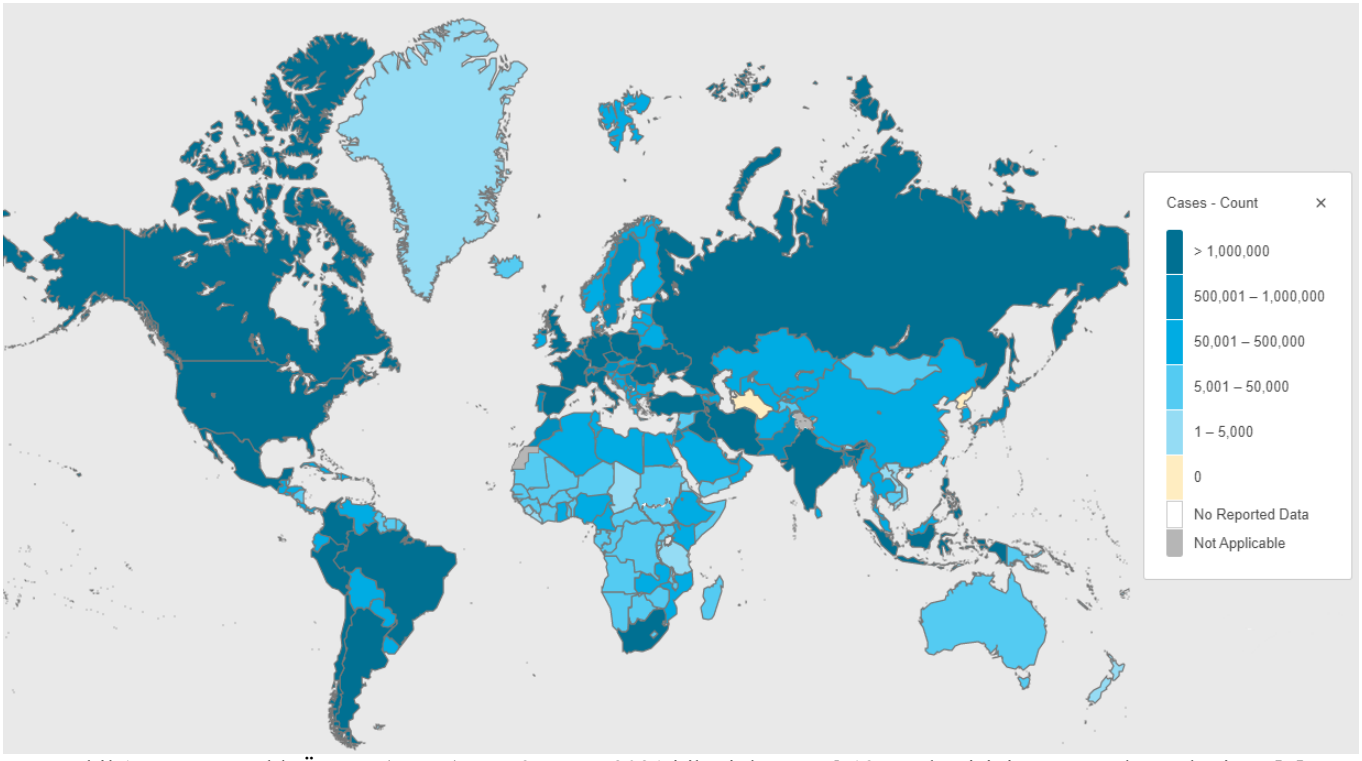
Çin Hastalık Kontrol ve Önleme Merkezi (Centers for Disease Control and Prevention, [CDC]) tarafından CoV ailesine ait olduğu saptanmış ve Covid-19 olarak adlandırılmıştır. Dünya Sağlık Örgütü (WHO) tarafından da pandemik bir hastalık olarak ilan edilmiştir.

Covid-19'un Çin'in Wuhan eyaletinin Huanan deniz ürünleri pazarında bir insana bulaşmış olduğu tahmin edilmekte ve günümüzde de hızla yayılmaktadır. Covid-19 insan vücudunda; öksürük, ateş, halsizlik, nefes darlığı ve boğaz ağrısı gibi belirtiler göstermektedir. Covid-19 teşhisi ise bu virüse özgü geliştirilmiştir test kitleri ile yapılabilmektedir.

Covid-19'a maske, dezenfektan ve sosyal mesafeyi koruyarak önlem alınmaya çalışılmaktadır. Aynı zamanda ABD'li ilaç firması Pfizer ile Alman BioNTech'in geliştirdiği aşı başarı oranı %95, Rusya tarafından geliştirilen Sputnik V aşısı başarı oranı %92 ve Çinli bilim insanları tarafından viral vektör ve RNA tabanlı teknik ile geliştirilen Sinovac (CoronaVac) aşısının başarı oranı %50'nin üstündedir.

Dünya Sağlık Örgütü (WHO) tarafından açıklanan güncel verilere göre 9 Mayıs 2021 itibariyle dünya genelinde doğrulanmış vaka sayısı 157 milyon, virüsten kaynaklı vefat sayısı 3 milyon ve aşı olanların sayısı ise 1 milyar 170 milyon civarındadır [5]. Bu araştırmanın yapıldığı esnada Türkiye'deki doğrulanmış vaka sayısı 5 milyon, virüsten kaynaklı vefat sayısı 43 bin ve toplam yapılan aşı sayısı 25 milyon civarındadır [6]. Hâlâ dünya genelindeki virüs vaka ve vefat sayıları artmaya devam etmekte ve aşı uygulanan insan sayıları artmaktadır.

Bu makalenin I. Giriş bölümünde koronavirüs genel olarak nasıl bir virüs olduğundan, nerede başladığından, çeşitlerinden, nasıl önlem almamız gerektiğinden, şu ana kadar çıkmış aşıların başarı oranlarından, dünya ve Türkiyedeki vakalardan bahsedilmiştir. II. Materyal ve metot bölümü ise veri setinin nereden alındığından, modele verilmeden önce nasıl işlediğimizden, Zaman Serisi, Makine Öğrenmesi ve Yapay Sinir Ağı modellerinin ne olduğundan, projede kullandığımız modellerin nasıl çalıştığından ve



Şekil 1. Dünya Sağlık Örgütü (WHO)'nün 9 Mayıs 2021 itibariyle Covid-19 pandemisini gösteren dünya haritası [7]

modellerin sonuçlarında kullandığımız hata analizi ölçütlerinin ne olduğundan bahsettik. III. Araştırma Sonuçları ve Tartışma bölümünde modele verdiğimiz verilerden çıkan sonuçları gerçek verilerle görselleştirip ve bu çıkan sonuçları hata analiziyle ölçüp tüm modelleri karşılaştırdığımızdan bahsettik. IV. Sonuç bölümünde ise elde edilen sonuçları tartıştık.

II. MATERYAL VE METOT

Türkiye Cumhuriyeti Sağlık Bakanlığı tarafından düzenlenen, 11 Mart 2020 ile 9 Mayıs 2021 tarih aralığındaki verileri kapsayan veri seti üzerinde, SARIMAX, Ekstrem Gradyan Arttırma, Lineer Regresyon, Karar Ağacı, Gradyan Arttırma, Yapay Sinir Ağı modelleri kullanılarak gerçek veriler üzerinden vaka tahminleri gerçekleştirilmiştir.

Yapılan çalışmada kullanılan modeller Python dili ile kodlanmıştır. Çalışmada 8GB RAM, Intel Core i7-7700H işlemci, NVIDIA GeForce GTX 1050 ekran kartın sahip bir donanım üzerinden gerçekleştirilmiştir.

Veri seti %80 eğitim ve %20 test seti şeklinde ayrılmıştır. Eğitim verileri kullanılarak vaka sayılarına yönelik olarak tahminlemeler yapılmıştır. Modelin tahmin ettiği değerler gerçek vakalar ile karşılaştırılmış, Kök Ortalama Kare Hata (Root Mean Square Error (RMSE)) ve Ortalama Mutlak Hata (Mean Absolute Error (MAE)) değerleri üzerinden model perfonması yorumlanmıştır.

A. *Veri Seti*

1) *Veri Seti Toplama*

Türkiye Cumhuriyeti Sağlık Bakanlığı tarafından açıklanan, Türkiye genelindeki Covid-19 vakalarını içeren, sürekli güncellenen ve kamunun erişimine açık bir şekilde paylaşılan "Türkiye Cumhuriyeti Sağlık Bakanlığı Genel Koronavirüs Tablosu" veri seti kullanılmıştır [8]. Bu veri seti Selenium kütüphanesi kullanılarak çalışmaya dahil edilmiştir. Veri setine Türkiye genelindeki günlük vakalar sürekli eklenmekte ve veri miktarı sürekli artmaktadır. Veri setinde yer alan ilk vaka 11 Mart 2020 tarihine aittir ve bu çalışmanın gerçekleştirildiği 9 Mayıs 2021 tarihine kadar olan güncel veriler kullanılmıştır. Çalışmanın yapıldığı tarih itibariyle veri seti içerisinde 425 adet satır yer almaktadır.

Veri setinde bulunan ve bu çalışmada kullanılan özellikler şunlardır (Tablo 1):

2) *Veri Ön İşleme*

Veri ön işleme kısmında kullandığımız veri setindeki değişkenler, tiplerinin uygunluğuna göre tam sayı (integer) ve kesirli sayı (float) olarak değiştirilmiştir.

Tarih değişkeninde "23 Mart 2020" formatından uluslararası format olan "2020-03-23" formatına dönüştürülmüştür.

Günlük Vaka Sayısı değişkenindeki eksik değerlere, Toplam Vaka Sayısı değişkeninde tarihsel olarak art arda bulunan değerleri, birbirinden çıkararak o günkü tarihe çıkarılan değerler atanmıştır.

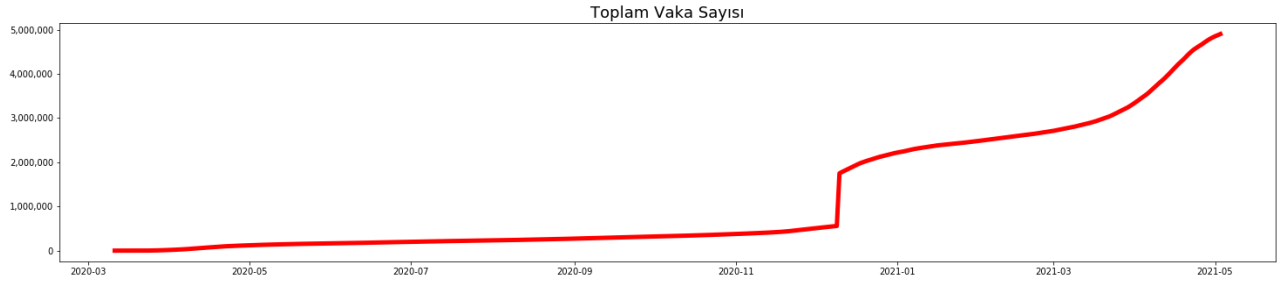
Veri setindeki geri kalan "NaN" değerler "0(sıfır)" nümerik sayısı ile doldurulmuştur.

3) *Veri görselleştirme*

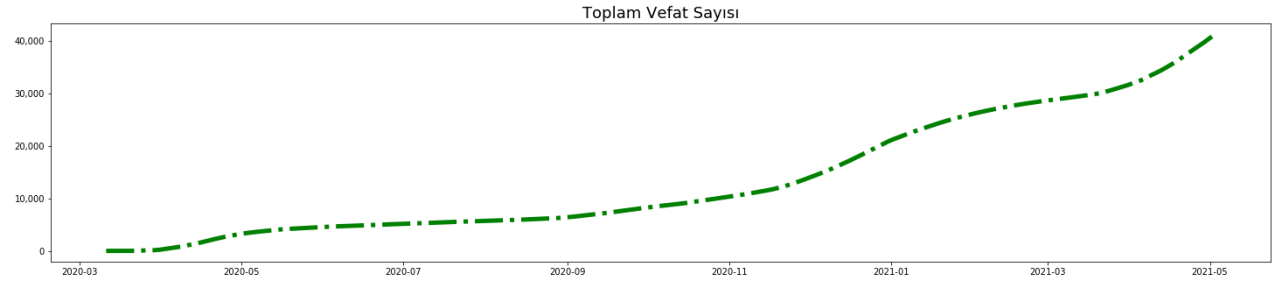
Şekil 2, 3 ve 4'te Türkiye Cumhuriyeti Sağlık Bakanlığının sağladığı veriler ile Toplam Vaka Sayısı, Toplam İyileşen Hasta Sayısı ve Toplam Vefat Sayısının tarihsel olarak seyrinin görselleştirilmesi yapılmıştır.

Tablo 1. Veri Setinin özellikleri

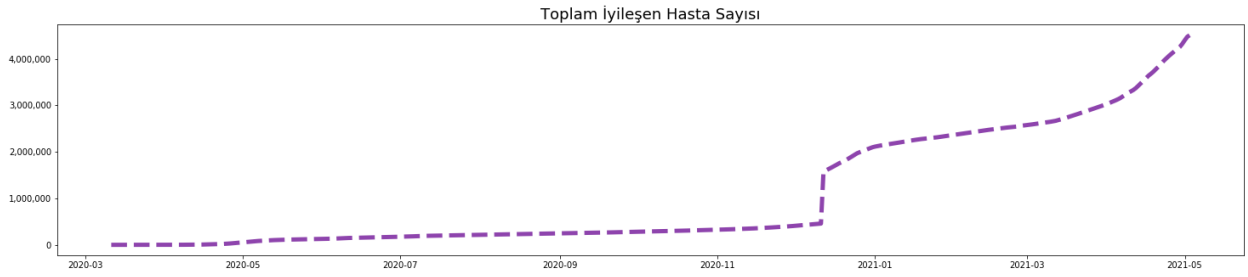
Değişken adı	Türü	Açıklama
Tarih	Nominal	Vakanın görüldüğü tarih
Toplam Test Sayısı	Nümerik	Doğrulanmış toplam test sayısı
Toplam Vaka Sayısı	Nümerik	Doğrulanmış toplam vaka sayısı
Toplam Vefat Sayısı	Nümerik	Hastalık kaynaklı toplam vefat sayısı
Hastalarda Zatürre Oranı (%)	Nümerik	Günlük hastalığa yakalananlardaki zatürre oranı
Ağır Hasta Sayısı	Nümerik	Hastalığı yoğun bakımda geçirenlerin sayısı
Toplam İyileşen Hasta Sayısı	Nümerik	Hastalıktan kurtulanların toplam sayısı
Bugünkü Vaka Sayısı	Nümerik	Günlük doğrulanmış vaka sayısı
Bugünkü Hasta Sayısı	Nümerik	Günlük doğrulanmış hasta sayısı
Bugünkü Test Sayısı	Nümerik	Günlük doğrulanmış test sayısı
Bugünkü Vefat Sayısı	Nümerik	Günlük hastalık kaynaklı vefat sayısı
Bugünkü İyileşen Sayısı	Nümerik	Günlük hastalıktan kurtulanları sayısı



Şekil 2. Türkiye'deki Covid-19 Toplam Vaka Sayısının seyri



Şekil 3. Türkiye'deki Covid-19 Toplam İyileşen Hasta Sayısının seyri



Şekil 4. Türkiye'deki Covid-19 Toplam Vefat Sayısının seyri

B. Zaman Serisi, Makine Öğrenmesi ve Yapay Sinir Ağı Modelleri

Zaman serileri; zamana göre (kronolojik) sıralanan veri dizileridir, veri satırları periyodik (saat, gün, ay, yıl vb.) bir döngü ile sıralanırlar. Zaman serileri sayısal olarak ifade edilebilecek olayların ve işlemlerin bir zaman damgası ile ilişkilendirildiği basit veri setleri olarak da tanımlanabilir [9].

Makine Öğrenmesi, kökleri yapay zekâ ve istatistiğe dayanan veri kümelerinden bilgi üretmemize imkân veren yöntemlere sahip bir araştırma dalıdır [10]. Makine Öğrenmesi, matematiksel ve istatistiksel işlemler ile veriler üzerinden çıkarımlar yaparak tahminlerde bulunan sistemlerin bilgisayarlar ile modellenmesidir. Günümüzde Makine Öğrenmesi için birçok metodoloji ve algoritma mevcuttur. Makine Öğrenmesi temelde öğrenme yöntemine göre üç gruba ayrılır; Supervised (Gözetimli), Unsupervised (Gözetimsiz) ve Reinforcement (Takviyeli).

Yapay Sinir Ağları nöron adı verilen bir çok işlem biriminden oluşur. Genelde nöronlar katman denilen mantıksal gruplar içinde yer alır. Ağ, 3 ya da daha çok katmandan oluşan hiyerarşik bir yapıya sahiptir. Bu ağda 1 girdi, 1 veya daha çok gizli ve 1 çıktı katmanı bulunmaktadır [11].

1) SARIMAX

Otoregresif ve hareketli ortalama modeller, sabit ve doğrusal verilerle çalışır. Bununla birlikte, çoğu durumda veriler sabit değildir. Durağan olmayan verilerle başa çıkmak için Otoregresif Entegre Hareketli Ortalama (ARIMA) modelleri kullanılır. Bir ARIMA modeli, otoregresif (AR) terimler, hareketli ortalama (MA) terimler ve farklılaştırma işlemleri (I) olmak üzere üç bölümden oluşur. Farklılaştırma işlemi, modelleme için sabit bir seri oluşturmak için kullanılır. Bu işlemde bir değer, değerin farkı ile önceki değeriyle değiştirilir. Mevsimsel ARIMA (SARIMA) olarak bilinen ARIMA modelinin genelleştirilmiş bir formu, verilerdeki mevsimselliği işlemek için kullanılır. Bu ARIMA modelleri sınıfı, modeldeki mevsimsel AR, MA ve farklılaşan terimleri kullanarak verilerdeki mevsimsellikle açıkça ilgilenir. Dış değişkenler ayrıca bir dışsal regresör terimi aracılığıyla modele eklenebilir. Dışsal regresörlü (SARIMAX (seasonal autoregressive integrated moving average with external or exogenous regressors)) mevsimsel ARIMA, kullanıcının dış değişkenlerin etkilerini modele eklemesini sağlar. Dışsal değişkenler, bir modeli etkileyen ancak ondan etkilenmeyen değişkenler olarak tanımlanır. Hava durumu, bir binanın enerji tüketim modeli bağlamında dışsal bir değişken olarak kabul edilir [12].

2) Ekstrem Gradyan Arttırma

Ekstrem Gradyan Artırma algoritması, karar ağaçları ve makine öğrenmesinde sıklıkla tercih edilen bir araç haline gelen bir uygulamadır. Geliştirilen bu yaklaşım sınıflandırma, reg-resyon ve sıralama görevlerinde yüksek performans sağlayan denetimli öğrenme alanında önemli bir araç olarak kabul edilmektedir [13].

Ekstrem Gradyan Artırma algoritmasındaki artırılmış ağaçlar, regresyon ve sınıflandırma ağaçlarına bölün-müştür. Bu algoritmanın özü, amaç fonksiyonu değerinin optimize edilmesine dayanmaktadır [14]. Ekstrem Gradyan Artırma algoritmasının en önemli özelliği, tüm senaryolarda ölçekle-nebilirliğidir. Sistem, tek bir makinedeki mevcut popüler çözümlerden 10 kattan daha hızlı çalışır ve dağıtılmış veya hafıza sınırlı ayarlarda milyarlarca örneğe ölçeklenir [15].

3) Lineer Regresyon

Lineer Regresyon, bağımsız değişkenlere dayanan bir hedef tahmin değerini modeller. Çoğunlukla değişkenler ve tahmin arasındaki ilişkiyi bulmak için kullanılır. Farklı regresyon modelleri, bağımlı ve bağımsız değişkenler, kullanılan bağımsız değişkenlerin sayısı arasındaki ilişkiye göre farklılık gösterir [16].

4) Karar Ağacı

Karar ağacı yöntemi pek çok test gerçekleştirerek, hedefi tahmin etmede en iyi sırayı bulmaya çalışır. Her bir test karar ağacındaki dalları oluşturur ve bu dallar da diğer testlerin gerçekleşmesine neden olur. Bu durum, test işleminin bir son düğümünde (terminal node) sonlanmasına kadar devam eder. Karar ağacının iki temel işlemi olan bölme (splitting) ve budama (pruning) işlemlerinin sonlanması uygulanan durdurma kriterine göre olmaktadır. Bu işlemler ve durdurma kriteri kısaca şöyle açıklanabilir

• Bölme: Bu işlem, verilerin daha küçük alt kümelere ayrılmasını sağlayan tekrarlı bir süreçtir. İlk tekrar tüm veriyi içeren kök düğüm ile başlar. Bundan sonraki tekrarlar, verinin alt kümelerini içeren üretilmiş düğümler üzerinde işlem yapmaktadır. Her bölme işleminde, değişkenler analiz edilir ve en iyi bölme seçilir.

• Budama: Bir ağaç oluşturulduktan sonra, istenmeyen alt ağaçlar veya düğümler bulunabilir. Budama işlemi ile bunlar çıkarılarak karar ağacı daha genel bir biçimde ifade edilebilmektedir.

• Durdurma kriteri: Ağaç oluşturma algoritmaları çeşitli durdurma kuralları içermektedir. Bu kurallar genellikle, maksimum ağaç derinliği, bir düğümde bölme için ele alınan minimum eleman sayısı ve yeni bir düğümde olması gereken minimum eleman sayısı gibi çeşitli faktörlere dayanır [16].

5) Gradyan Artırma

Gradyan artırma yöntemi, kayıp fonksiyonunu en aza indiren ek bir model bulmayı amaçlayan sayısal bir optimizasyon algoritması olarak görülebilir. Böylelikle, gradyan artırma algoritması yinelemeli olarak her adımda kayıp işlevini en iyi azaltan yeni bir karar ağacı yani “zayıf öğrenen" ekler. Daha kesin olarak, regresyonda algoritma modeli bir tahminle başlar, bu genellikle kayıp fonksiyonunu en üst düzeyde azaltan bir karar ağacıdır (regresyon için ortalama hata karesidir), ardından her adımda yeni bir karar ağacı mevcut artığa takılır ve artığı güncellemek için önceki modele eklenir. Algoritma, kullanıcı tarafından sağlanan maksimum yineleme sayısına ulaşılana kadar yinelemeye devam eder. Bu süreç, sözde aşamalı olarak adlandırılır. Her yeni adımda, önceki adımlarda modele eklenen karar ağaçları değiştirilmez. Karar ağaçlarının kalıntılara uydurulmasıyla model, iyi performans göstermediği bölgelerde iyileştirilir [18].

6) Yapay Sinir Ağı

TensorFlow, makine öğrenimi algoritmasındaki tüm hesaplama işlemlerini ve verileri temsil etmek için veri akış grafiklerini kullanır. TensorFlow'da, grafikteki düğümler matematiksel işlemleri ve bilgi vermenin başlangıcını veya çıktı vermenin sonunu temsil eder. Kenarlar, düğümler arasında iletişim kuran çok boyutlu veri dizilerini (tensörler) temsil eder [19]. Bu tensörler tüm düğümlere akar ve nihayetinde makine öğrenimi sürecini tamamlar. TensorFlow ayrıca hesaplama grafiklerinin görüntülerini kolayca görüntülemek için TensorBoard adlı kullanışlı bir görselleştirme aracı sağlar.

TensorFlow'daki çoğu algoritma, sinir ağlarına dayanır. Yapay sinir ağları veya bağlantısal sistemler olarak da adlandırılan sinir ağları, başlangıçta hayvan beyinlerini oluşturan biyolojik sinir ağlarından esinlenmiştir. Bir sinir ağı, deneysel bilgiyi depolamak ve onu kullanıma hazır hale getirmek için doğal bir eğilime sahip olan basit işlem birimlerinden oluşan büyük ölçüde paralel dağıtılmış bir işlemcidir [20]. Bir sinir ağının temel öğeleri arasında bir nöron, bir dizi sinaps, bir toplayıcı ve bir aktivasyon işlevi bulunur. Bir nöron, bir sinir ağının çalışması için temel olan bir bilgi işlem birimidir. Nöronlar arasındaki her bağlantı (sinaps) kendine ait bir ağırlık veya güçle karakterize edilir ve başka bir nörona bir sinyal iletebilir. Nöronların ilgili sinaptik güçleri ile ağırlıklandırılan giriş sinyallerini toplamak için bir toplayıcı kullanılır. Bir nöronun çıktısının genliğini sınırlamak için bir aktivasyon işlevi kullanılır.

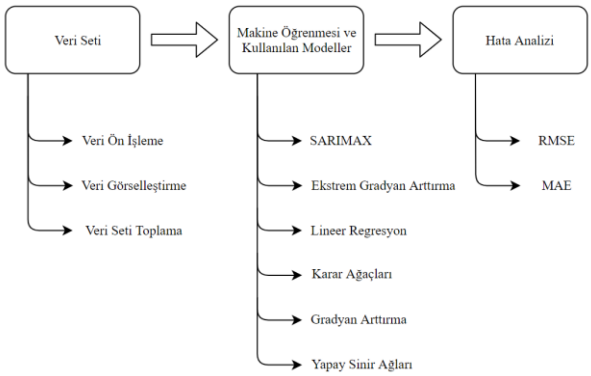
C. *Hata Analizi*

Modelin performansı değerlendirilirken gerçek değerler ile modelin ürettiği tahminlerin uyumuna bakılmıştır. Bu metriklerden Kök Ortalama Kare Hata (Root Mean Square Error (RMSE)) ve Ortalama Mutlak Hata (Mean Absolute Error (MAE)), regresyon modellerinde tahminlerin doğruluğunu ölçmek kullanılmaktadır.

RMSE ve MAE ise birer hata ölçüsü olması nedeniyle düşük sonuçlar, performans ile ters orantılı olarak yüksek performansı gösteren ölçülerdir [21]. Örneğin RMSE sıfıra eşit olması durumunda iyi bir performans göstermektedir [22].

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \quad (1)$$

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} |e_t| \quad (2)$$



Şekil 5. Materyal ve Metot bölümünün akış şeması

III. Araştırma Sonuçları ve Tartışma

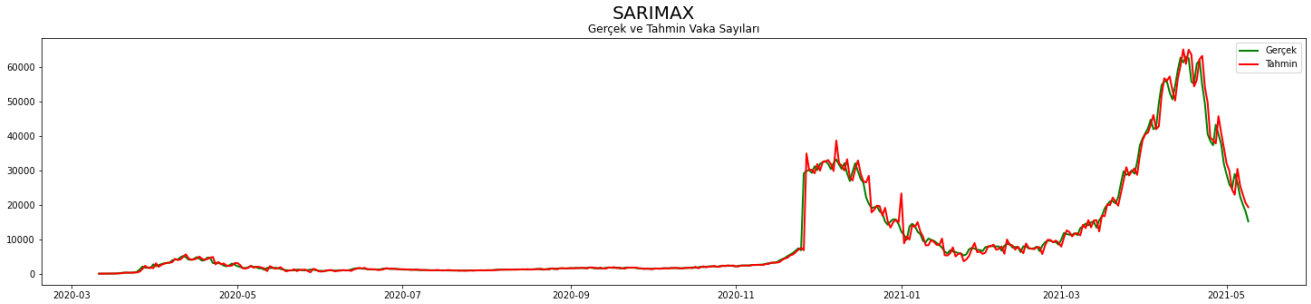
Bu çalışmada SARIMAX, Ekstrem Gradyan Arttırma, Lineer Regresyon, Karar Ağacı, Gradyan Arttırma, Yapay Sinir Ağı modelleri kullanılarak Türkiye genelindeki doğrulanmış vakalara yönelik tahminlemeler yapılmıştır. Yapılan tahminler süreçteki gerçek sayılarla karşılaştırılarak tahmin başarısı yorumlanmıştır.

Modelin performansı değerlendirilirken gerçek değerler ile modelin ürettiği tahminlerin uyumuna bakılmıştır. Tahmin edilen değerlerin gerçek değerlerden uzaklıkları temel alınarak RMSE ve MAE değerlerine bakılarak modellerin başarılarının karşılaştırılması yapılmıştır.

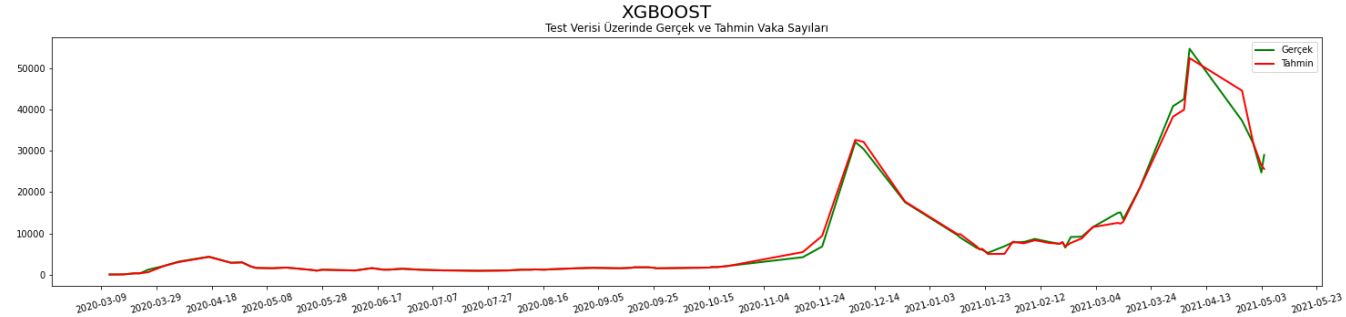
Türkiye genelindeki doğrulanmış gerçek vaka değerlerinin, SARIMAX zaman serisi modelinin tahmin ettiği vaka değerleri ile karşılaştırılmasına ait grafik Şekil 6'da görülmektedir.

Test veri seti üzerinde Türkiye genelindeki doğrulanmış gerçek vaka değerleri ile Ekstrem Gradyan Arttırma, Lineer Regresyon, Karar Ağacı, Gradyan Arttırma, Yapay Sinir Ağı modellerinin tahmin ettiği değerlerin karşılaştırılmasına ait grafikler sırasıyla Şekil 7, 8, 9, 10 ve 11'de görülmektedir. Grafiklerde gerçek değerler yeşil, modellerin tahmin değerleri ise kırmızı renkte gösterilmiştir.

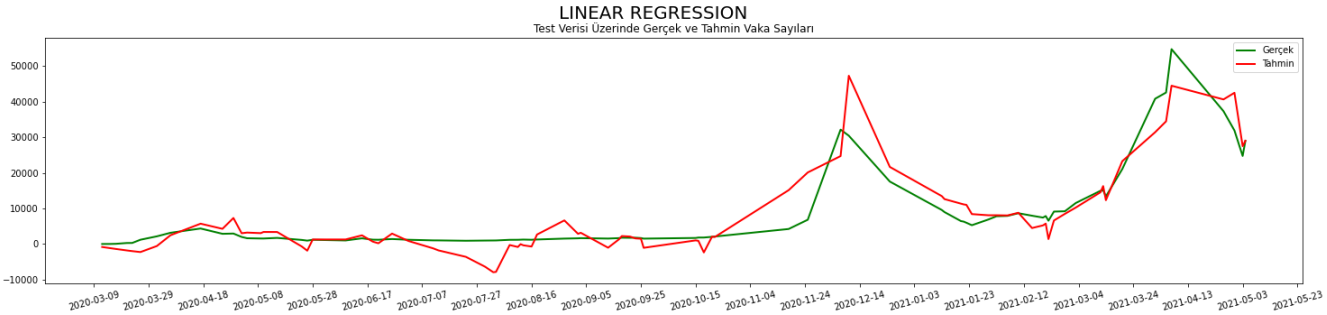
Zaman Serisi, Makine Öğrenmesi ve Yapay Sinir Ağı Modelleri ile yapılan testlerde, modellerin performanslarının değerlendirilmesinde kullanılan RMSE ve MAE değerleri Tablo 3 ve Şekil 12'de görülmektedir.



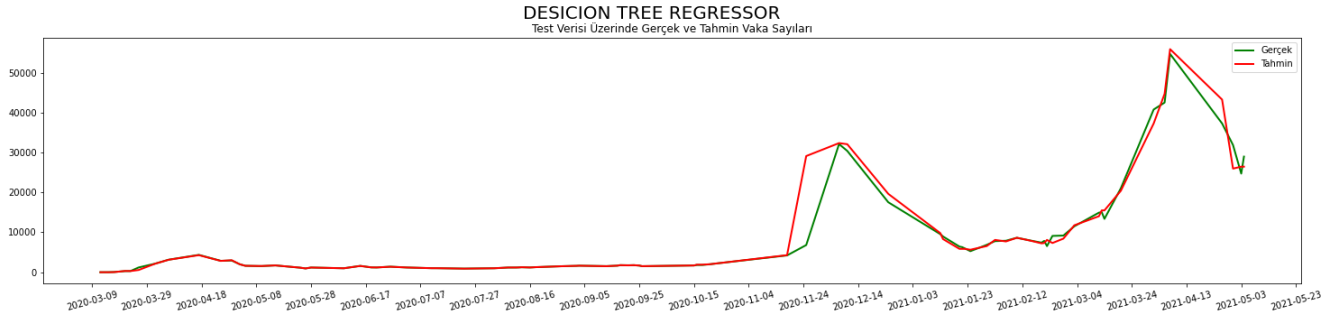
Şekil 6. Gerçek vaka değerlerinin, SARIMAX modelinin tahmin ettiği vaka değerleri ile karşılaştırılması



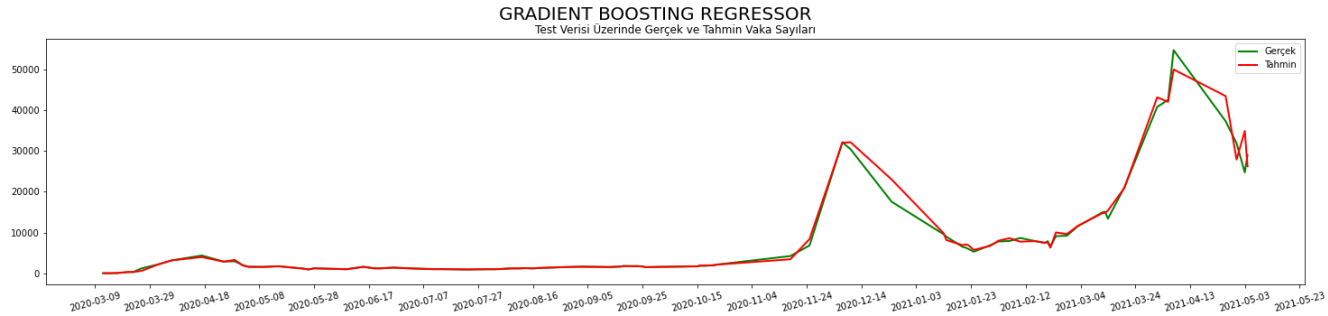
Şekil 7. Test verisi üzerinde, gerçek vaka değerleri ile Ekstrem Gradyan Arttırma modelinin tahmin ettiği vaka değerlerinin karşılaştırılması



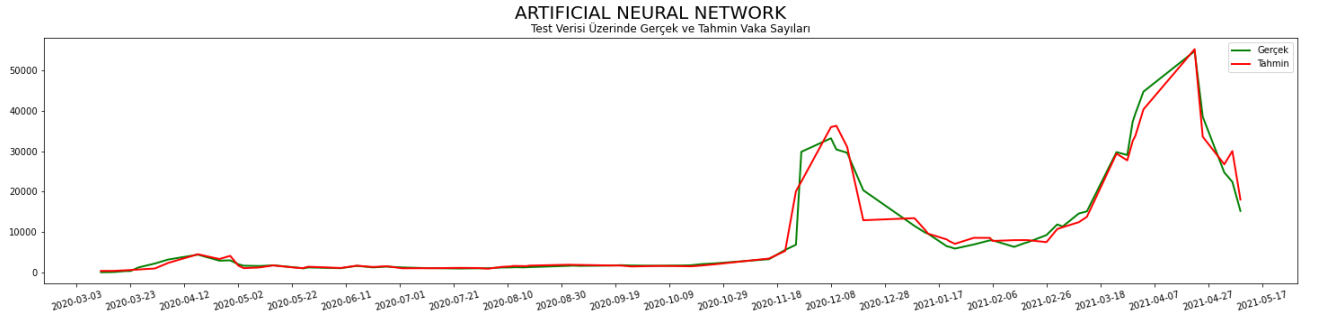
Şekil 8. Test verisi üzerinde, gerçek vaka değerleri ile Lineer Regresyon modelinin tahmin ettiği vaka değerlerinin karşılaştırılması



Şekil 9. Test verisi üzerinde, gerçek vaka değerleri ile Karar Ağacı modelinin tahmin ettiği vaka değerlerinin karşılaştırılması



Şekil 10. Test verisi üzerinde, gerçek vaka değerleri ile Gradyan Arttırma tahmin ettiği vaka değerlerinin karşılaştırılması



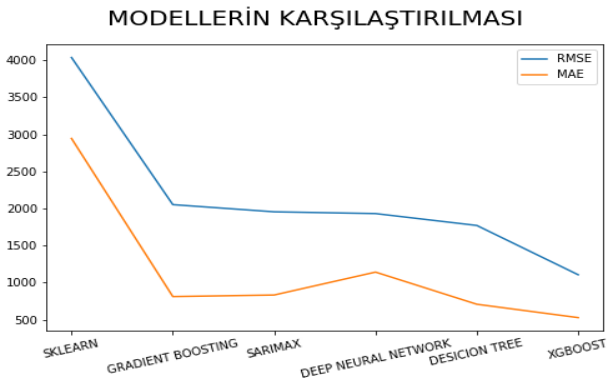
Şekil 11. Test verisi üzerinde, gerçek vaka değerleri ile Yapay Sinir Ağının tahmin ettiği vaka değerlerinin karşılaştırılması

Deneysel sonuçlar, SARIMAX modeli için RMSE değerini 1954 ve MAE değerini 833 vermiştir. Ekstrem Gradyan Arttırma modeli için RMSE 1106 değerini ve MAE değerini 528 vermiştir. Lineer Regresyon modeli için RMSE 4036 değerini ve MAE değerini 2944 vermiştir. Karar Ağacı modeli için RMSE 1770 değerini ve MAE değerini 708 vermiştir. Gradyan Arttırma modeli için RMSE 2051 değerini ve MAE 811 değerini vermiştir. Yapay Sinir Ağı modeli için RMSE 1931 değerini ve MAE 1142 değerini vermiştir.

Lineer Regresyon yöntemi bu araştırma için kötü sonuçlar vermektedir. Ekstrem Gradyan Arttırma ve Karar Ağacı en iyi sonuçları vermektedir.

Tablo 2. Modellerin RMSE ve MAE değerleri

Model	RMSE	MAE
Lineer Regresyon	4036	2944
Gradyan Arttırma	2051	811
SARIMAX	1954	833
Yapay Sinir Ağı	1931	1142
Karar Ağacı	1770	708
Ekstrem Gradyan Arttırma	1106	528



Şekil 12. Modellerin RMSE ve MAE ile karşılaştırılması

IV. SONUÇ

2019 yılının sonunda Çin'in Wuhan eyaletinde ortaya çıkarak hızla yayılan Covid-19, birçok insanın hastalığına sebep olup dünya genelinde pandemiye neden olmuştur. Araştırmanın gerçekleştiği sırada, virüs yayılımı devam etmekte olup farklı bilim dallarınca araştırmalara devam edilmektedir. Bilgisayar bilimleri alanında, Zaman Serisi, Makine Öğrenmesi ve Yapay Sinir Ağı teknikleri ile virüsün dünya üzerindeki yayılımı, önlenmesi ve tedavisi konusunda çalışmalar yapılabilmektedir. Bu çalışmada, Türkiye Cumhuriyeti Sağlık Bakanlığının sağladığı güncel Covid-19 salgını verileri ile kullanılan modeller yardımıyla vaka tahminlemesi yapılmıştır.

Doğrulanmış vaka, ölüm ve iyileşen hasta sayısı her geçen gün artış göstermektedir. Türkiye genelindeki doğrulanmış gerçek vaka değerleri ile SARIMAX, Ekstrem Gradyan Arttırma, Lineer Regresyon, Karar Ağacı, Gradyan Arttırma, Yapay Sinir Ağı modelleri kullanılarak 11 Mart 2020 – 9 Mayıs 2021 tarihleri arasında tahminlemeler yapılmıştır.

Ekstrem Gradyan Arttırma ve Karar Ağacı modelleri üzerinde tahmin edilen değerler büyük oranda doğruluk göstermiştir. Gerçek ve tahmin değerleri arasındaki model performanslarını ölçmek için RMSE ve MAE hata metrikleri kullanılmıştır. Şekil 2'deki grafikte doğrulanmış vaka sayılarında sürekli bir artış olmakla beraber artışın tam olarak düzenli olmadığı görülmektedir. Son dönemdeki ani yükseliş ve düşüşler, modellerin performansını etkilemiştir. Bununla birlikte Türkiye Cumhuriyetinin alacağı tedbirler, uygulayacağı sokağa çıkma yasakları ve karantina süreçleri Covid-19 vakalarının seyrini büyük oranda etkileyebilir.

Bu sonuçlar ışığında elimizdeki gerçek veriler ile başarılı sonuçlar elde edilebileceği görülmüştür. Ancak hem elimizdeki verinin az olması hem de gelecekteki vaka sayılarını etkileyebilecek birçok farklı etken bulunduğundan ötürü geleceğe yönelik tahminlemelerde bulunmanın zor olacağı öngörülmektedir.

V. KAYNAKÇA

[1] K. McIntosh and S. Perlman, "Coronavir uses, including severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS)," Mand. Douglas Bennetts Princ. Pract. Infect. Dis. Updat. Ed. 8th Ed Phila. PA Elsevier Saunders, 2015.

[2] P. K. Chan and M. C. Chan, "Tracing the SARS coronavirus," J. Thorac. Dis., vol. 5, no. Suppl 2, p. S118, 2013.

[3] R. J. de Groot et al., "Commentary: Middle East respiratory syndrome coronavirus (MERS CoV): announcement of the Coronavirus Study Group," J. Virol., vol. 87, no. 14, pp. 7790 7792, 2013

[4] E. R. Ahmet Görkem and S. ÜNAL, "2019 Koronavirüs Salgını Anlık Durum ve İlk İzlenimler," FLORA, vol. 25, p. 8, 2020.

[5] https://covid19.who.int

[6] https://covid19.saglik.gov.tr/

[7] https://covid19.who.int

[8] https://covid19.saglik.gov.tr/TR-66935/genel-koronavirus-tablosu.html

[9] George E. P. Box; Gwilym M. Jenkins; Gregory C. Reinsel; Greta M. Ljung, Time Series Analysis: Forecasting and Control (2015), 1.

[10] T. M. Mitchell, Machine Learning, New York, 1997.

[11] Doç. Dr. Birgül Kutlu; Yrd. Doç. Dr. Bertan Badur (2009), Yapay Sinir Ağları İle Borsa Endeksi Tahmini, Yönetim Dergisi, 63, 28.

[12] C. Chatfield, Time-series Forecasting. Chapman & Hall/CRC, 2001.

[13] Mitchell, Rory; Frank, Eibe (2017), "Accelerating the XGBoost Algorithm Using GPU Computing", PeerJ Computer Science, Vol.3; 127-164.

[14] Zheng, Huiting; Yuan, Jiabin; Chen, Long (2017), "Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation", Energies, Vol. 10, No. 8; 1168-1188.

[15] Chen, Tianqi; Guestrin, Carlos (2016), "XGBoost: A Scalable Tree Boosting System", in Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, ACM, 785-794.

[16] Dey, A. (2016). Machine Learning Algorithms: A Review, International Journal of Computer Science and Information Technologies, 7(3): 1174-1179.

[17] C. Bounsaythip, R. R. Esa, 2001, Overview of data mining for customer behavior modeling, VTT Information Technology Research Report, Version:1, 1-53.

[18] Friedman, J. 2002. Stochastic Gradient Boosting. Computational statistics & data analysis, 38(4), 367-378.

[19] Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. Tensorflow: A system for large-scale machine learning. arXiv 2016, arXiv:1605.08695

[20] Haykin, S.S. Neural Networks and Learning Machines; Pearson: Upper Saddle River, NJ, USA, 2009; Volume 3.

[21] Wang, W. & Xu, Z. (2004). A Heuristic Training for Support Vector Regression, Neurocomputing, 61: 259 275

[22] Çınaroğlu, S. (2017). Sağlık Harcamasının Tahmininde Makine Öğrenmesi Regresyon Yöntemlerinin Karşılaştırılması, Uludağ Üniversitesi Mühendislik Fakültesi Dergisi, 22(2): 179-200.