Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 22, S. 134-139, Ocak 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 22, pp. 134-139, January 2021

Research Article

COVID-19 Vakalarının Makine Öğrenmesi Algoritmaları ile Tahmini: Amerika Birleşik Devletleri Örneği

Nur Selin Özen*, Selin Saraç[2], Melik Koyuncu[3]

[1*] Toros Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, Mersin, Türkiye (ORCID: 0000-0001-8545-8771), selin.ozen@toros.edu.tr
[2] Toros Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, Mersin, Türkiye (ORCID: 0000-0002-4729-0637), selin.sarac@toros.edu.tr
[3] Çukurova Üniversitesi, Mühendislik Fakültesi, Endüstri Mühendisliği Bölümü, Adana, Türkiye (ORCID: 0000-0003-0513-6276), mkoyuncu@cu.edu.tr

(İlk Geliş Tarihi Aralık 2020 ve Kabul Tarihi Ocak 2021)

(**DOI:** 10.31590/ejosat.855113)

ATIF/REFERENCE: Özen, N. S., Saraç, S. & Koyuncu, M. (2021). COVID-19 Vakalarının Makine Öğrenmesi Algoritmaları ile Tahmini: Amerika Birleşik Devletleri Örneği. *Avrupa Bilim ve Teknoloji Dergisi*, (22), 134-139.

Öz

Koronavirüs, 2019 yılının Aralık ayında ilk olarak Çin'in Wuhan kentinde ortaya çıkmış ve 11 Mart 2020'de Dünya Sağlık Örgütü tarafından pandemi olarak ilan edilmiştir. Vaka sayılarını kontrol altına almak için pek çok ülke karantina, sokağa çıkma yasağı ve sosyal alanların bir süreliğine kapatılması gibi çeşitli önlemler almıştır. Doğrulanmış vaka tahminlemesi pandemide olası planlamalar için büyük önem taşımaktadır. Gelecek verilerinin gerçeğe en yakın bir şekilde tahminlenmesi; pandemi döneminde lojistik, tedarik, hastane personel ve malzeme planlaması için kullanılabileceği gibi aşılama senaryolarında da girdi olarak kullanılabilir. Literatürde doğrulanmış vaka tahmininde makine öğrenmesi, bölmeli model, zaman serisi analizi gibi pek çok yöntem kullanarak tahminleme yapılan çalışmalar vardır. Bu çalışmada, Amerika Birleşik Devletleri'ndeki doğrulanmış vaka sayılarını kullanarak gelecek günlerdeki vaka tahminleri çeşitli makine öğrenmesi modelleri ile üretilmiştir. Python ve R programlama dili kullanılarak yapılan tahminlemeler Prophet, Polinom Regresyon, ARIMA, Doğrusal Regresyon ve Random Forest modelleri ile yapılmıştır. Test verisiyle tahmin edilen verilerin performansları ortalama mutlak yüzde hatası (MAPE), ortalama karekök sapması (RMSE) ve ortalama mutlak hata (MAE) kullanılarak değerlendirilmiştir. Sonuç olarak, MAPE hata metriği baz alınarak en iyi tahminleri veren algoritma Polinom Regresyon olarak bulunmuştur.

Anahtar Kelimeler: COVID-19, Tahminleme, Makine Öğrenmesi.

Prediction of COVID-19 Cases in the United States of America with Machine Learning Algorithms

Abstract

The coronavirus first appeared in Wuhan, China in December 2019 and was declared as a pandemic by the World Health Organization on March 11, 2020. In order to control the number of cases, many countries have taken various measures such as quarantine, curfew and closing social areas for a while. Prediction data can be used in logistics, procurement, hospital personnel and supplies planning and vaccination scenarios. In the confirmed case estimate; in the literature, there are studies that use many methods such as machine learning, compartmental model, and time series analysis in confirmed case prediction. In this study, various machine learning models have been generated to estimate future cases using the number of confirmed cases in the United States. The predictions made using Python and R programming language were made with Prophet, Polynomial Regression, ARIMA, Linear Regression and Random Forest models. The performances of the data estimated by the test data are evaluated using the mean absolute percent error (MAPE), root mean square deviation (RMSE) and mean absolute error (MAE). As a result, the algorithm that gives the best estimates based on the MAPE error metric was found as Polynomial Regression.

Keywords: COVID-19, Prediction, Machine Learning.

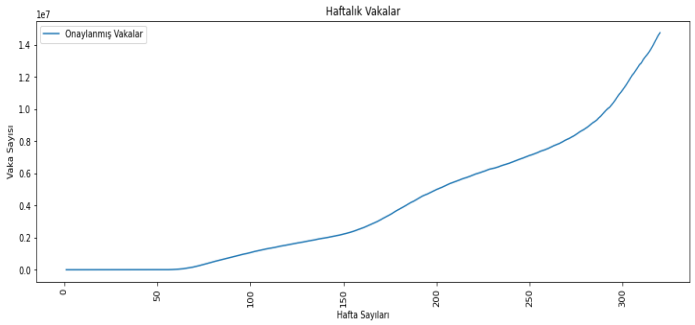
* Sorumlu Yazar: selin.ozen@toros.edu.tr

1. Giriş

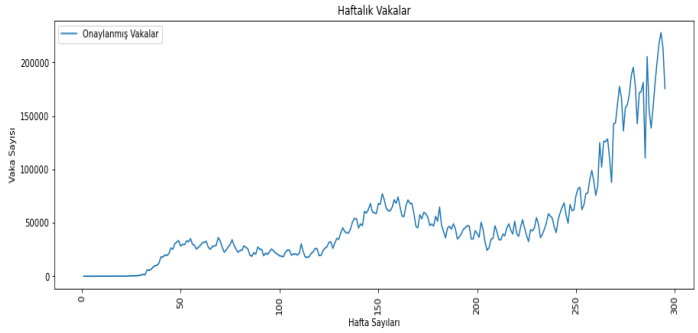
2019 yılının Aralık ayında ilk olarak Çin Halk Cumhuriyeti'nde görülen ve sonra tüm dünyayı etkisi altına alan COVID-19, 11 Mart 2020 tarihinde Dünya Sağlık Örgütü (WHO) tarafından pandemi ilan edilen ve etkileri halen sürmekte olan bir salgın hastalıktır. Tüm dünyada her gün on binlerce insan salgın sebebiyle ölümle karşı karşıyadır ve her gün dünyanın dört bir yanındaki ülkelerden yüz binlerce bireyin test sonucunun pozitif olduğu bildirilmektedir. COVID-19 virüsü, damlacık yoluyla kişilerin kontamine yüzey ya da enfekte olmuş bireyle olan temasıyla yayılmaktadır. Virüsün yayılmasının kontrol altına alınmasındaki en büyük zorluk, bir kişinin virüse yakalanıp günlerce semptom göstermeden taşıyıcı olabilmesidir. Virüsün yayılmasının nedenleri ve tehlikesini göz önünde bulundurarak, neredeyse tüm ülkeler, kısmi veya tam zamanlı karantina ilan etmiştir (Rustam ve ark., 2020). Pandemi süresince yatak kapasitesi, test miktarı ve kişisel koruyucu ekipmanların sayısı gibi hastanelerde talep değişimleri için yapılacak kapasite planlama çalışmaları birçok zor problem içermektedir. Pandemi sürecinde bu tür planlamaların yapılabilmesi ve ilgili modellerin kurulması için vaka sayılarının tahmin edilmesi büyük önem taşımaktadır. Vaka sayısı tahmini geliştirilecek olan birçok senaryo için son derece yardımcı olacak bir veridir.

Makine öğrenmesi algoritmaları deneyim kazandıkça, doğruluğu ve verimliliği gelişen algoritmalardır. Bu da algoritma geliştiricilerinin daha iyi sonuçlar almasını ve geleceğe yönelik hatası düşük tahminler yapmasını sağlamaktadır. Makine öğrenmesi modelleri kullanılarak bağımlı ve bağımsız değişken arasındaki ilişkiyi bulan ve bu ilişkiyi performans metrikleriye somutlaştıran birçok çalışma yapılmıştır. Örneğin finans alanında; borsa fiyat tahminleyen (Nunno, 2014), çevre alanında; ABD'deki bir nehirde tahmini günlük akışı yedi gün öncesine kadar değerlendiren (Papacharalampous & Tyralis, 2018), sağlık alanında; bir hastaneden taburcu olan kişilerin hacmi tahmini yapan (McCoy ve ark., 2018) çalışmalar mevcuttur. Bu da makine öğrenmesi metotlarının kapsama alanının geniş olduğunu ve iyi sonuçlar verebileceğini göstermektedir. Ulaşım, sağlık, ekonomi, tarım, eğitim, seyahat, e-ticaret ve diğer farklı alanlarda makine öğrenmesinin çok boyutlu ve çok çeşitli verileri işlemede iyi olması ve bunu dinamik veya belirsiz ortamlarda yapabilmesi büyük avantajlar sağlamaktadır.

COVID-19 vaka sayısı tahminlemesinde makine öğrenmesi modelleri sıkça kullanılmaktadır. Tahminleme çalışmalarında kullanılan modellerden bazıları: Doğrusal Regresyon, Destek Vektör Regresyonu (Support Vector Regression), Üstel Düzeltme (Exponential Smoothing), LASSO (Rustam ve ark., 2020), ARIMA (Avan & Aslam, 2020; Sahai ve ark., 2020) ve Prophet (Date & Deshmukh, 2020; Sevli & Başer, 2020) gibi modellerdir. Bu çalışmalarla ileriye dönük global ve/veya lokal tahminler yapılmış olup makine öğrenmesi algoritmaları değerlendirilmiştir. Bu değerlendirmeler, kullanılan bölgeye göre değişiklik gösteren veri setleri ve eğitim-test setinin oranlarının farklı olması sebebiyle değişiklik göstermektedir.



Şekil 1. Amerika Birleşik Devletleri Haftalık Kümülatif Vaka Sayıları Grafiği (The New York Times, 2020)



Şekil 2. Amerika Birleşik Devletleri Haftalık Kesikli Vaka Sayıları Grafiği (The New York Times, 2020

Bu çalışmada Amerika Birleşik Devletleri'nin (ABD) onaylanan vakalarının tahmini Prophet, Polinom Regresyon, ARIMA, Doğrusal Regresyon ve Random Forest makine öğrenmesi modelleri kullanılarak, Python ve R paket programlarında gerçekleştirilmiştir. Şekil 1 ve Şekil 2 ABD'nin doğrulanmış vaka sayısı grafiğini göstemektedir. Elde edilen tahminler, test setleriyle karşılaştırılmış ve performans metrikleri üzerinden modellerin değerlendirilmesi yapılmıştır.

2. Materyal ve Metot

ABD'nin gelecek günlerdeki doğrulanmış vaka sayısını tahmin etmek üzere kullanılan makine öğrenmesi algoritmalarını uygulamak için veri seti "https://github.com/nytimes/COVID-19-data " sitesinden elde edilmiştir (The New York Times, 2020). Veri seti, 21.01.2020-06.12.2020 aralığındaki ABD'nin doğrulanmış vaka ve ölüm sayısını içermektedir.

Çalışmamızda; makine öğrenmesi algoritmalarını kodlamak için Python ve R programlama dilleri kullanılmıştır. Python için Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn; R için tidyverse, tidymodels, modeltime, timetk, glmnet ve randomForest gibi kütüphaneler kullanılmıştır. .csv türünde olan veri seti Python ve R yazılım dillerine aktarılmış ve ABD'deki her bir eyaletin verileri günlük olarak, tarih ve doğrulanmış vaka sayıları şeklinde manipüle edilmiştir.

Veri setinin %90'ı her bir algoritmanın eğitilmesi için geri kalan %10'u ise bu eğitime bağlı olarak üretilen tahminlerin doğruluğunu ölçmek için test setine ayrılmıştır. Algoritmaların tahminleri ile üretilen değerlerle test seti karşılaştırılmış ve bu karşılaştırma ortalama mutlak yüzde hatası (MAPE), ortalama karekök sapması (RMSE) ve ortalama mutlak hata (MAE) performans metrikleriyle değerlendirilmiştir. Performans metrikleriyle yapılan değerlendirmeler doğrultusunda makine öğrenmesi modellerinin performansı karşılaştırılmış ve en iyi

performans gösteren model üzerinde yapılabilecek olası senaryolar konusunda önerilerde bulunulmuştur.

2.1. Tahminleme Modelleri

2.1.1. Prophet

Prophet modeli Facebook tarafından geliştirilmiş, Python ve R'da kütüphanesi olan açık kaynak kodlu bir zaman serisi tahminleme algoritmasıdır. Doğrusal olmayan veri için yıllık, günlük, aylık tahminler yapabilen ve bu tahminleri belirtilen tatil günlerini hesaba katarak gerçekleştirebilen bir algoritmadır.

Prophet modelinin temel bileşenleri trend; g(t), mevsimsellik; s(t), tatiller; h(t) ve hata terimi; ε(t)'den oluşup, Denklem (1) olarak ifade edilir.

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t) \quad (1)$$

g(t), zaman serilerisi değerlerindeki periyodik olmayan değişikenleri modelleyen trend fonksiyonudur. s(t) periyodik değişimleri (haftalık, yıllık, aylık) temsil eder ve h(t), tatillerin bir veya daha fazla gün içinde potansiyel olarak düzensiz programlarda meydana gelen etkilerini temsil eder. Hata terimi ε(t), model tarafından barındırılmayan her türlü kendine özgü değişiklikleri temsil eder ve genellikle normal dağılım olarak modellenir (Taylor & Letham, 2018).

2.1.2. Polinom Regresyon

Polinom regresyon, yalnızca bir bağımsız değişken X ile çoklu regresyonun özel bir durumudur. Tek değişkenli polinom regresyon modeli Denklem (2) şeklinde ifade edilir.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3 + \cdots + \beta_k {x_i}^k, \quad (2)$$

$$i = 1, 2, \ldots . n$$

k, polinomun derecesidir ve polinomun derecesi, modelin derecesidir (Ostertagová, 2012).

2.1.3. ARIMA

Box Jenksins modeli olarakta adlandırılan ARIMA modeli, ARMA (p, q), AR (p) ve MA (q) modellerinin bir kombinasyonudur. En iyi kullanımı tek değişkenli zaman serisi modellemesi içindir. AR (p) modelinde, bir değişkenin gelecekteki değerinin, geçmiş gözlemler ve rastgele bir hata teriminin doğrusal bir kombinasyonuna bağlı olduğu varsayılır. AR (p) modelinden farklı olarak, bir MA (q) modeli açıklayıcı değişkenler olarak geçmiş hataları kullanır. AR ve MA yalnızca tek değişkenli durağan zaman serilerine uygulanabilir. Bir zaman serisinin durağanlığını test etmek için birim kökün yapısını test etmemiz gerekir. Seri, seviyede durağan değilse, seriyi d (d=1,2,3,…) kez türev almamız gerekir. Böyle bir zaman serisi modeline ARIMA (p, d, q) modeli denir. ARIMA modelini kullanmak için gereken ilk adım, zaman serisinin durağanlığını test etmek ve bunun için Augmented Dicky Fuller Testi'nin (ADF) kullanılmasıdır. İkinci adım, Otokorelasyon Fonksiyonu (ACF) ve Kısmi Otokorelasyon Fonksiyonu (PACF) grafiklerini oluşturup AR(p) ve MA(q) değerlerini belirlemektir. Son olarak, ACF ve PACF tarafından hesaplanan parametreler içinde en iyisini bulmak için Akaike Bilgi Kriteri'ne (AIC) bakılmalıdır ve bulunan parametrelerle ARIMA modeli oluşturulmuş olur (Sahai ve ark., 2020).

2.1.4. Doğrusal Regresyon

Doğrusal regresyon bir tür regresyon modellemesidir ve makine öğreniminde tahmine dayalı analiz için en kullanışlı istatistiksel tekniktir. Doğrusal regresyondaki her gözlem bağımlı bağımsız değişken olmak üzere iki değere bağlıdır. Bu bağımlı(y) ve bağımsız(x) değişkenler arasında doğrusal ilişkiyi doğrusal regresyon belirler. Denklem (3), y'nin regresyon olarak bilinen x ile nasıl ilişkili olduğunu gösterir.

$$y = \beta_0 + \beta_1 x + \varepsilon \quad (3)$$

ε, doğrusal regresyonun hata terimini temsil eder ve x ve y arasındaki değişlenliği açıklamak için kullanılır. β_0, doğrunun y eksenini kestiği noktayı temsil ederken β_1 eğimi temsil eder (Rustam ve ark., 2020).

2.1.5. Random Forest Regresyonu

Random Forest Regresyonu (RFR), orijinal örneklerden birden fazla örnek çıkarmak için Bootstrap örnekleme yöntemini kullanan ve ardından tahmin için karar ağaçlarını birleştiren istatistiksel bir öğrenme yöntemidir. Girdinin çıktısını yani sonucu almak için karar ağaçlarının verdiği ortalama tahmini alır. Random Forest Regresyon metotu klasik karar ağacı metotunun bir uzantısı olan ve tahmin doğruluğunu iyileştirmek için birden fazla karar ağacından oluşan topluluk öğrenme yöntemidir (Juo ve ark., 2016).

2.2. Performans Metrikleri

2.2.1. Ortalama Mutlak Yüzde Hatası (MAPE)

Ortalama mutlak yüzde hatası (MAPE), gerçek değerlerle tahmin edilmiş değerler arasındaki hatayı yüzdelik olarak temsil eden hata metriğidir. Bu performans metriği Denklem (4)'te gösterilmiştir.

Regresyon ve zaman serileri modellerinde doğruluğu ölçmek için ortalama mutlak yüzde hata sıkça kullanılmaktadır. Gerçek değerler arasında "0" değerleri varsa, 0 ile bölünme olmayacağı için MAPE'nin hesaplanması mümkün değildir. Çok düşük tahmin değerleri için yüzde hatası %100'ü geçemez, fakat gerçek değerlerden uzak tahmin değerleri olduğunda yüzde hatasının üst sınırı yoktur (Chai & Draxler, 2014).

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - y_i'}{y_i} \right| \quad (4)$$

y_i : Gerçek değer

y_i' : Tahmin edilen değer

n : gözlem sayısı

2.2.2. Ortalama Mutlak Hata (MAE)

Ortalama mutlak hata (MAE), kolay yorumlanabilir olduğu sebebiyle regresyon ve zaman serisi problemlerinde sıkça kullanılmaktadır. Denklem (5)'te MAE performans metriğinin denklemi verilmiştir. MAE, iki sürekli değişken arasındaki farkın ölçüsüdür yani gerçekleşen ve tahmin edilen değerler arasındaki farkların mutlak değerlerinin toplamıdır. MAE değeri 0'dan ∞'a kadar değişebilir. (Chai & Draxler, 2014).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - y_i'| \quad (5)$$

y_i : Gerçek değer

y_i' : Tahmin edilen değer

n : gözlem sayısı

2.2.3. Ortalama Karekök Sapması (RMSE)

Bir makine öğrenmesi modelinin, tahminleyicinin tahmin ettiği değerler ile gerçek değerleri arasındaki uzaklığın bulunmasında sıklıkla kullanılan, hatanın büyüklüğünü ölçen ikinci dereceden bir hata metriğidir ve Denklem (6)'da gösterildiği gibi hesaplanmaktadır. RMSE, gerçek değerlerle tahmin edilen değerler arasındaki farkın (tahmin hatalarının) standart sapmasıdır. Yani, tahmin hataları, regresyon hattının veri noktalarından ne kadar uzakta olduğunun bir ölçüsüdür; RMSE ise bu kalıntıların ne kadar yayıldığının bir ölçüsüdür. RMSE değeri 0'dan ∞'a kadar değişebilir. RMSE değerinin sıfır olması modelin hiç hata yapmadığı anlamına gelmektedir. RMSE'nin avantajı, büyük hataları daha fazla cezalandırmasıdır. Bu sebeple bazı durumlara daha uygun olabilir. RMSE, birçok matematiksel hesaplamada istenmeyen mutlak değerin kullanılmasını engeller (Chai & Draxler, 2014).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2} \qquad (6)$$

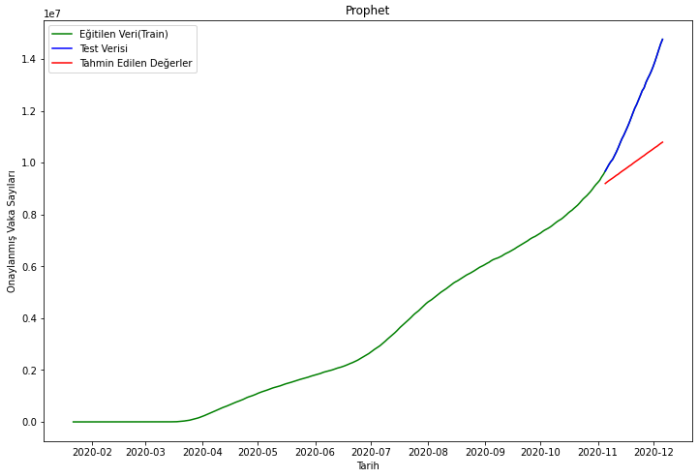
y_i : Gerçek değer

y_i' : Tahmin edilen değer

n : gözlem sayısı

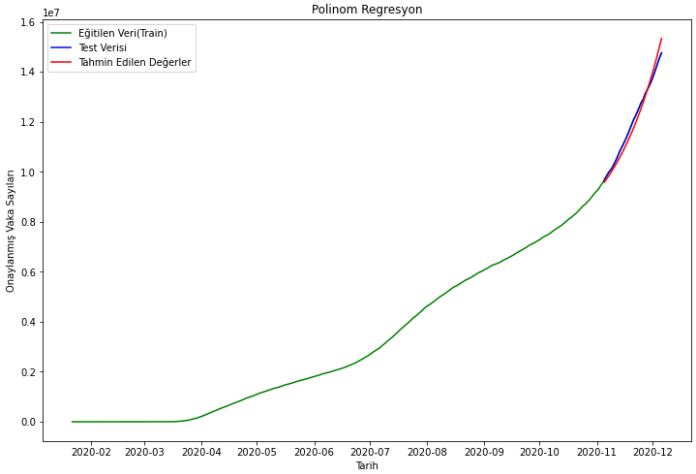
3. Araştırma Sonuçları ve Tartışma

Veri seti, sadece Prophet modelinin çalışması için ayrıca düzenlenmiş, tarih ve vakaların bulunduğu sütun isimleri sırasıyla "ds" ve "y" olarak adlandırılmıştır. Prophet haricinde kullanılan algoritmalar böyle bir düzenleme gerektirmediğinden sütunlar "Date" ve "Confirmed Cases" olarak bırakılmıştır. Şekil 3'te yeşil renkli doğru gerçekleşmiş vakaları temsil ederken, mavi renkli doğru test verisini, kırmızı renkli doğru ise tahmin edilen vaka değerlerini yansıtmaktadır. Hesaplanacak olan hata metrikleri, tahmin değerlerini yansıtan kırmızı doğrunun isabet ettiği vaka sayılarıyla test verisindeki vaka sayıları (mavi doğru) arasında değerlendirilmiştir. Çalışmada ABD verisi için Prophet modeliyle gerçekleştirilen tahminde gerçek veri ile test verisi arasında ortalama %16.1 oranında bir hata elde edilmiştir.



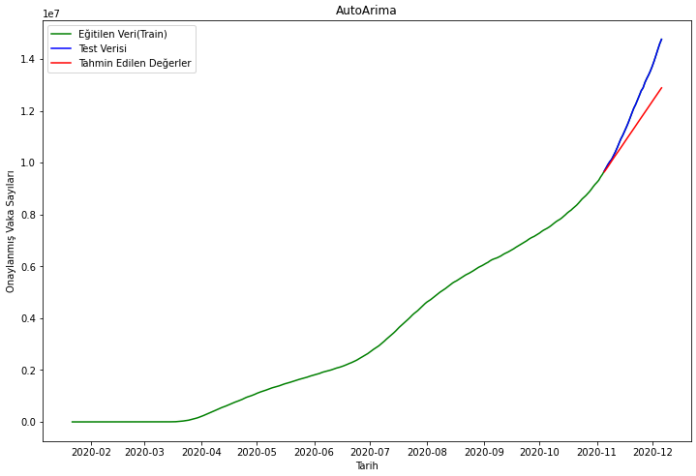
Şekil 3. Prophet ile Tahmin Edilen Vaka Sayısı Grafiği

Şekil 4'te gerçekleşen vaka sayısı, test verisi ve Polinom Regresyon ile tahmin edilen veriler gösterilmiştir. Tahmin edilen değerler ve gerçekleşen veriler arasında çok büyük fark olmadığı açıktır. Polinom Regresyonla oluşturulan ABD'nin doğrulanmış vaka verisine dair tahminler, en iyi MAPE değerini 6. derecede vermiş ve %1.86 olarak bulunmuştur.



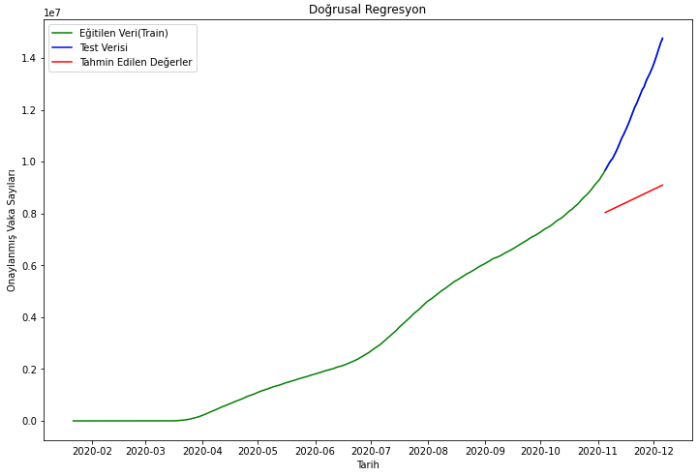
Şekil 4. Polinom Regresyon ile Tahmin Edilen Vaka Sayısı Grafiği

Çalışmada; ARIMA modelinin adımlarını otomatik olarak uygulayan, parametreleri Akaike Bilgi Kriteri (AIC) ve Bayesian Bilgi Kriteri (BIC) kullanarak değerlendiren ve en optimal (p, d, q) değerlerini veren "auto.arima" modülü kullanılmıştır. Python ve R programlama dillerinde bulunan autoarima'nın en büyük avantajlarından biri klasik ARIMA modellerinden daha hızlı uygulanmasıdır. Model çalıştırılığında, otomatik olarak bulunan en optimal (p, d, q) değerlerini AIC ve BIC değerlerine göre inceleyip en optimal (p, d, q) değerlerini sırasıyla (2, 2, 2) olarak bulmuştur. AIC ve BIC değerleri de sırasıyla 5681.68 ve 5699.96'dır. Autoarima ile gerçekleştirilen tahminlerin test verisine göre; MAPE hata değeri %8.63, RMSE değeri 2757813 ve MAE değeri 769664 olarak bulunmuştur. Şekil 5'te görüldüğü gibi, ARIMA'nın gerçeğe yakın sonuçlar verdiği açıktır ve performans metriklerine bakıldığında da ARIMA modelinin tahminleme için başarılı bir model olduğunu yorumu yapılabilir.



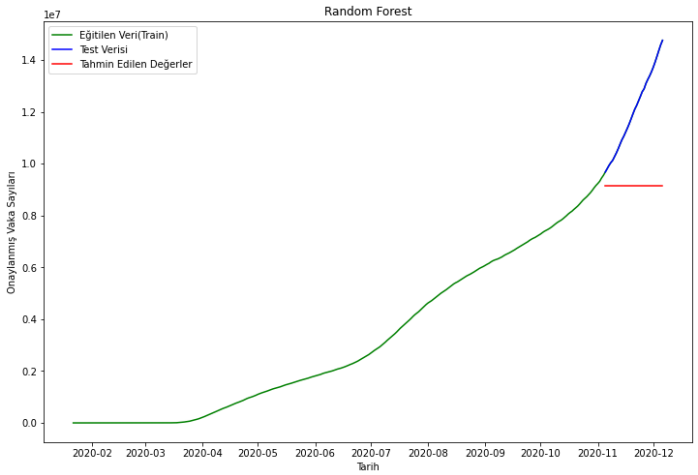
Şekil 5. ARIMA ile Tahmin Edilen Vaka Sayısı Grafiği

Doğrusal Regresyon kullanılarak eğitilen veri ile gerçek değerler Şekil 6'da gösterilmiştir. Test verisi ve tahmin değerleri arasında elde edilen hata oranları MAPE için %27.99, RMSE için 3672745 ve MAE için 3467277 olarak bulunmuştur. ABD'nin doğrulanmış vaka sayıları doğrusal olmadığı gibi dalgalanma da gösterdiği için doğrusal regresyonun hata oranlarının diğer modellere göre yüksek çıkması öngörülen bir sonuçtur. Şekil 6'da görüldüğü gibi, doğrusal regresyonla tahmin edilen vaka değerleri ve test verisi birbiri üstüne tam olarak denk gelmemiştir. Bu da hata metriklerinin yüksek çıkmasını somut olarak desteklemektedir.



Şekil 6. Doğrusal Regresyon ile Tahmin Edilen Vaka Sayısı Grafiği

Random Forest algoritmasının uygulamasında hiperparametre optimizasyonu RandomizedSearchCV modeli kullanılarak gerçekleştirilmiştir ve veri seti için en iyi parametreler tahmin edilmiştir. Random Forest algoritmasından elde edilen sonuçlar incelendiğinde, COVID-19 vaka tahmini için uygun bir algoritma olmadığı açıkça görülmektedir. Test verisi ve tahmin değerleri karşılaştırıldığında elde edilen hata oranları, MAPE için %22.64 olarak hesaplanmıştır. Şekil 7'de Random Forest ile tahmin edilen vaka sayısı grafiği de algortimanın uygunsuzluğunu desteklemektedir.



Şekil 7. Random Forest ile Tahmin Edilen Vaka Sayısı Grafiği

Tablo 1'de algoritmaların hata değerleri verilmiştir. Tabloya bakarak Polinom Regresyon algoritmasıyla ABD'de gelecek günler için üretilen vaka sayılarının doğruluğunun diğer modellere göre daha güvenilir olduğu söylenebilir.

Tablo 1. Algortimaların Performans Metrikleri

Model	%MAPE	RMSE	MAE
Prophet	16.10	2295085	2039241
Polinom Regresyon	1.86	255181	227757
ARIMA	8.63	2757813	769664
Doğrusal Regresyon	27.99	3672745	3467277
Random Forest	22.64	3253735	2874023

4. Sonuç

COVID-19 tüm dünyayı etkisine aldığı gibi her gün bir önceki günden daha fazla kişiyi etkileyen ve hala hızla yayılımını sürdüren bir salgın hastalıktır. Pandemi sürecinde yayılımı modellemek, vaka sayılarını tahminlemek, kaynak planlaması yapmak ve lojistik kısıtlarla ilgili birçok çalışma yapılmıştır ve yapılmaya devam etmektedir. Vaka tahminlemesine dair makine öğrenmesi çalışmalarında çeşitli algoritmalar birbiriyle karşılaştırılırken farklılık gösteren sonuçlar bulmak kaçınılmaz olacaktır. Çünkü üzerinde çalışılan veri seti ve algoritmaların kullandığı parametreler gibi birçok değişken sonuçları etkileyebilir.

Çalışmamızda, makine öğrenmesi algoritmalarıyla vaka tahminlemeleri yaparak bu algoritmalar karşılaştırılmış ve sonuçlar analiz edilmiştir. ABD veri setine uygulanan algoritmalar içinde belirli oranda kullanılmak üzere ayrılan vaka sayısıyla (test verisi) tahmin edilen vaka sayısının en çok uyum gösterdiği algoritma %1.86 MAPE oranıyla Polinom Regresyon olmuştur ve bunu sırasıyla ARIMA, Prophet, Random Forest ve Doğrusal Regresyon algoritmaları takip etmiştir. Pandemi sürecinde problem yaratan durumlardan biri belirsizliklerin fazla ve bu sürecin dinamik olmasıdır. Dolayısıyla mümkün olduğunca gerçeğe yakın senaryolar oluşturmak halk sağlığı açısından büyük önem taşımaktadır. Sonuç olarak, Polinom Regresyon algoritması kullanılarak ABD için yapılacak olan kaynak atama problemlerine, aşı dağıtım modeline, hastaneye gelecek olası hasta sayısına, hastanedeki ekipmanların muhtemel kullanım oranına, yoğun bakımdaki hasta sayısına dair planlamaların düşük hata ile yapılabileceği öngörülmektedir. Gelecek çalışmalarda, farklı ülkelerin vaka sayıları ile bu algoritmalar tekrar denenerek tahmin edilebilir ve performans metriği en başarılı olan algoritmalar kullanılarak planlamalara dair ülke bazlı stratejiler oluşturulabilir.

Kaynakça

Awan, T. M., & Aslam, F. (2020). Prediction of daily COVID-19 cases in European countries using automatic ARIMA model. *Journal of Public Health Research, 9*(3), 227-233. https://doi.org/10.4081/jphr.2020.1765.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?. *Geosci. Model Dev., 7, 1247-1250.* https://doi.org/10.5194/gmd-7-1247-2014.

Date, S., & Deshmukh, S. (2020). Forecasting novel COVID-19 confirmed cases in India using Machine Learning Methods, *International Journal of Computer Sciences and Engineering, 8*(6), 57-62. https://doi.org/10.26438/ijcse/v8i6.5762.

Juo, J., Shi, T., & Chang, J., (2016). Comparison of Random Forest and SVM for Electrical Short-term Load Forecast with Different Data Sources. *7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, 1077-1080, https://doi.org/10.1109/ICSESS.2016.7883252.

Keleş, M. B., Keleş, A., & Keleş, A. (2020). Yapay Zekâ Teknolojisi ile Uçuş Fiyatı Tahmin Modeli Geliştirme. *Turkish Studies - Applied Sciences*, *15*(4). 511-520. https://dx.doi.org/10.29228/TurkishStudies.45993.

McCoy, T. H., Pellegrini, A. M., & Perlis, R. H. (2018). Assessment of Time-Series Machine Learning Methods for Forecasting Hospital Discharge Volume. *JAMA Netw Open, 1*(7). https://doi.org/10.1001/jamanetworkopen.2018.4087.

Nunno, L. (2014). Stock Market Price Prediction Using Linear and Polynomial Regression Models.

Ostertagová, E. (2012). Modelling using polynomial regression. *Procedia Engineering, 48*, 500-506. https://doi.org/10.1016/j.proeng.2012.09.545.

Papacharalampous, G. A., & Tyralis, H., (2018). Evaluation of random forests and Prophet for daily streamflow forecasting. *Advances in Geosciences, 45*, 201-208. https://doi.org/10.5194/adgeo-45-201-2018.

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B., Aslam, W., & Choi, G. S. (2020). COVID-19 Future Forecasting Using Supervised Machine Learning Models, *IEEE Access*, *8*, 101489-101499. https://doi.org/10.1109/ACCESS.2020.2997311.

Sahai, A. K., Rath, N., Sood, V., & Singh, M. P. (2020). ARIMA modelling&forecasting of COVID-19 in top five affected countries. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14*(5), 1419-1427. https://doi.org/10.1016/j.dsx.2020.07.042.

Sevli, O., & Başer, V. G. (2020). COVID-19 Salgınına Yönelik Zaman Serisi Verileri ile Prophet Model Kullanarak Makine Öğrenmesi Temelli Vaka Tahminlemesi. *European Journal of Science and Technology, 19*, 827-835. https://doi.org/10.31590/ejosat.766623.

Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. *The American Statician, 72* (1), 37-45. https://doi.org/10.1080/00031305.2017.1380080.

The New York Times. (2020). Coronavirus (Covid-19) Data in the United States, 14 Aralık 2020 tarihinde github sitesi: https://github.com/nytimes/covid-19-data adresinden alındı.