Journal of Social Sciences of Mus Alparslan University

anemon

Derginin ana sayfası: http://dergipark.gov.tr/anemon

Araştırma Makalesi ● Research Article

OECD Endüstriyel Üretim Verilerinde Bulunan Kayıp Verilerin kNN Yöntemi İle Tahmini

Estimation Of Missing Data In OECD Industrial Production Data By kNN Method

Serkan Metin*

Öz: Ekonomik İşbirliği ve Kalkınma Örgütü (OECD), daha iyi yaşamlar oluşturmak için çalışan uluslararası bir organizasyondur. Bu amaç doğrultusunda OECD ülkeler hakkında birçok göstergede veri toplamaktadır. Daha doğru analizler yapabilmek için bu verilerin eksiksiz olması gerekmektedir. Fakat ulusal ve uluslararası farklı kaynaklardan toplanan bilgilerde eksiklikler olmaktadır. Bu eksiklikler özellikle istatiksel analiz ve makine öğrenmesi yöntemleri kullanarak çalışmak isteyen araştırmacılara problem çıkartmaktadır. Bu tür analizler için veri setlerinin öncelikle eksik verilerden temizlenmesi gerekmektedir. Genel olarak eksik veriler istatistiksel analizleri üzerinde olumsuz bir etkiye sahiptir. Bu sorunu çözmek için geleneksel ve modern yöntemler vardır. Değişkenler tamamen rastgele eksik (MCAR), rastgele eksik (MAR) ve rastgele eksik değil (MNAR) olabilir. Bu neden ile her değişken ayrı ayrı ele alınmalıdır. Temel Ekonomik Göstergeler veri tabanı içerisindeki endüstriyel üretim başlıklı veriler setinde 34 ülkeye ait 113 eksik veri ve 3933 tam veri olmak üzere 4046 değer bulunmaktadır. Veri setini farklı gruplara ayırmak için çalışmada k-en yakın komşu (kNN) adı verilen makine öğrenimi algoritmasını kullanılmış. kNN algoritması kullanımının basit olduğundan yaygın olarak kullanılmaktadır. Çalışmada kullanılan algoritmaya ait en yakın komşuluk değeri k=15 olarak belirlenmiştir. Eksik verileri tahmin etmede %86,8'lik bir başarı elde edilmiştir.

Anahtar Kelimeler: OECD, Eksik Veri, kNN Algoritması

Abstract: The Organization for Economic Co-operation and Development (OECD) is an international organization that works to create better policies for better lives. For this aim, OECD collects data on countries in many indicators. In order to make more accurate analyses, these data must be complete. But there are deficiencies in the information collected from different national and international sources. These deficiencies are especially problematic for researchers who want to work using statistical analysis and machine learning methods. For such analysis, data sets must first be cleared of missing data. In general, incomplete data has a negative effect on statistical analysis. There are traditional and modern methods to solve this problem. The data can be missing completely at random (MCAR), missing at random (MAR), and not missing at random (MNAR). For this reason, each data must be handled separately. In the data set titled industrial production in the Main Economic Indicators database, there are 4046 values, 113 missing data and 3933 complete data belonging to 34 countries. In order to divide the data set into different groups, the study used a machine learning algorithm called K-Nearest Neighbor(kNN). Because the kNN algorithm is simple to use, it is widely used. The nearest neighborhood value

* Asst. Prof. Dr., Department of Management Information Systems, Faculty of Social Sciences and Humanities, Malatya Turgut Ozal University, Malatya, Turkey
ORCID: 0000-0003-1765-7474, serkan.metin@ozal.edu.tr

Received/Geliş: 01 March/Mart 2021 **Düzeltme/Revised form:** 07 April/Nisan 2021
Accepted/Kabul: 14 April/Nisan 2021 **Published/Yayın:** 31 August/Ağustos 2021

e-ISSN: 2149-4622. http://dx.doi.org/10.18506/anemon.888642

of the algorithm used in the study was determined as k=15. There was an 86.8% success rate in estimating the missing data.

Keywords: OECD, Missing Data, kNN Algorithm

Giriş

Ülkelere ait güvenilir veriler düzenlemek için Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) tarafından veritabanları oluşturulmaktadır. Bu veritabanları içerisinde ülkelere ait 23 kategoriye ayrılmış gösterge bilgileri bulunmaktadır. Temel Ekonomik Göstergeler veri tabanı, tüm OECD üye ülkeleri ve üye olmayan ülkelerin bir kısmı için aylık ve üç aylık istatistikleri ve ilgili istatistiksel metodolojik bilgileri içermektedir. Veritabanı, endüstriyel üretim, bileşik öncü göstergeler, iş eğilimi ve tüketici görüşü anketleri, perakende ticaret, tüketici ve üretici fiyatları, saatlik kazançlar, istihdam, işsizlik, faiz oranları, parasal büyüklükler, döviz kurları, uluslararası ticaret ve ödemeler dengesi verilerini içermektedir (OECD, 2021).

Göstergeler, ulusal istatistik kuruluşları tarafından öncelikle kendi ülkelerindeki kullanıcıların gereksinimlerini karşılayacak şekilde hazırlanmıştır. Çoğu durumda, göstergeler uluslararası istatistiki kılavuzlara ve tavsiyelere göre derlenir. Ancak, ulusal uygulamalar bu kılavuzlardan farklı olabilir ve bu ayrılıklar ülkeler arasındaki karşılaştırılabilirliği etkileyebilir. Veri tabanının kısa vadeli ekonomik analiz için uygunluğunu en üst düzeye çıkarmak için veritabanının içeriğinin sürekli olarak gözden geçirilmesi ve eksik verilerin olmaması gerekmektedir. Fakat ulusal ve uluslararası kurumlar ile hükümetlerin hazırlayıp düzenlediği göstergelere dayalı olarak (Çilingirtürk ve Altaş, 2010, s.74) oluşturulan OECD verilerinde ülkelere ait bazı bilgilerin elde edilememesi nedeniyle eksik veriler oluşmaktadır.

Eksik veri, verilerde boş veya eksik değerlerin bulunmasıdır. İstatiksel analiz yöntemlerinin birçoğu sadece tam veriler üzerinde doğru sonuç vermektedir. Veri seti içerisinde eksik değerler olduğunda (Huang ve Sun, 2016, s.9) yapılan analizler hatalı sonuçlar vermektedir (Zhang, 2012, s.2542). Eksik verilerin oluşumu, gerçek dünya verilerindeki sınıflandırma problemlerini çözen veri bilimcileri içinde en büyük zorluklardan biridir (Choudhury ve Kosorok, 2020, s.1). Çoğu durumda, araştırmacılar eksik değerlere sahip gözlemleri çalışmalarında çıkarırlar (Folch-Fortuny vd., 2016, s.93); ancak eksik verilerin çıkartılması, örneklem boyutunu azaltacağından daha büyük standart hatalarla tahminler yapmaya yol açar (Silva ve Perera, 2017, s.2). Bunun için araştırmacılar eksik verileri silmek yerine farklı yöntemler kullanarak değer atamayı tercih ederler (Ordóñez vd., 2017, s.705). Bu yöntemler, çok değişkenli istatistiksel analizde kullanılan ortalama ve regresyon yöntemini içeren yerine koyma yöntemi, makine öğrenimi ve derin öğrenme temelli yaklaşımlardır. Metasezgisel yöntemlerde eksik veri tamamlama için kullanılmaktadır. Bunlardan Tavlama Benzetimi yöntemi hafıza kullanmayan bir metasezgisel yöntemdir. Tavlama Benzetimi, metalin tavlama işlemini örnek alarak çalışan bir algoritmadır (Fendoğlu, 2020, s.18).

Bu çalışmada, eksik gözlemlerin tahmini için, gözetimli öğrenme algoritmalarından kNN modeli kullanılmıştır. Bu yöntemin geçerliliğini test etmek için, öncelikle veri seti içerisindeki bazı değerler rastgele çıkartılmış ve eğitim için kullanılarak oluşturulan modelin başarısı ölçülmüştür. Elde edilen sonuçların başarılı olduğu belirlendikten sonra yöntem gerçek veri seti üzerine uygulanmıştır.

Eksik Veri Tamamlama Yöntemleri

Eksik veri tamamlama yöntemleri, bir veri seti içerisindeki eksik verileri uygun algoritmalar kullanarak tamamlamayı ve veri setini analizler için uygun hale getirmeyi amaçlar. Eksik veri tamamlama yöntemlerinin başarısı verilerin eksik olma nedenlerine bağlıdır (Pini vd., 2020, s.2). Rubin ve Little eksik verileri, tamamen rastgele eksik (MCAR), rastgele eksik (MAR) ve rastgele eksik değil (MNAR) olarak üç gruba ayırmıştır (Malarvizhi ve Thanamani, 2012, s.5). MCAR, eksikliğin nedeni tamamen rastgeledir, yani bir gözlemin eksik olma olasılığı başka herhangi bir değerle ilişkili değildir (Dondersa vd., 2006, s.1088). MAR, eksikliğin sadece gözlemlenen değişkenlere bağlı olduğu durumdur

(Yoon vd., 2018, s.1). MNAR, bir gözlemin eksik olma olasılığı, gözlemin kendi değeri gibi, gözlemlenmeyen bilgiye de bağlıdır (Dondersa vd., 2006, s.1088). Eksik veriler ile çalışmalarda farklı metotlar kullanılabilir.

Hot-deck Metodu: Anket araştırmalarında oluşan eksik verilerin tamamlanmasında yaygın olarak kullanılan bu yöntemde her bir eksik değer, benzer bir birimdeki yanıtla değiştirilir (Andridge ve Little, 2010, s.1).

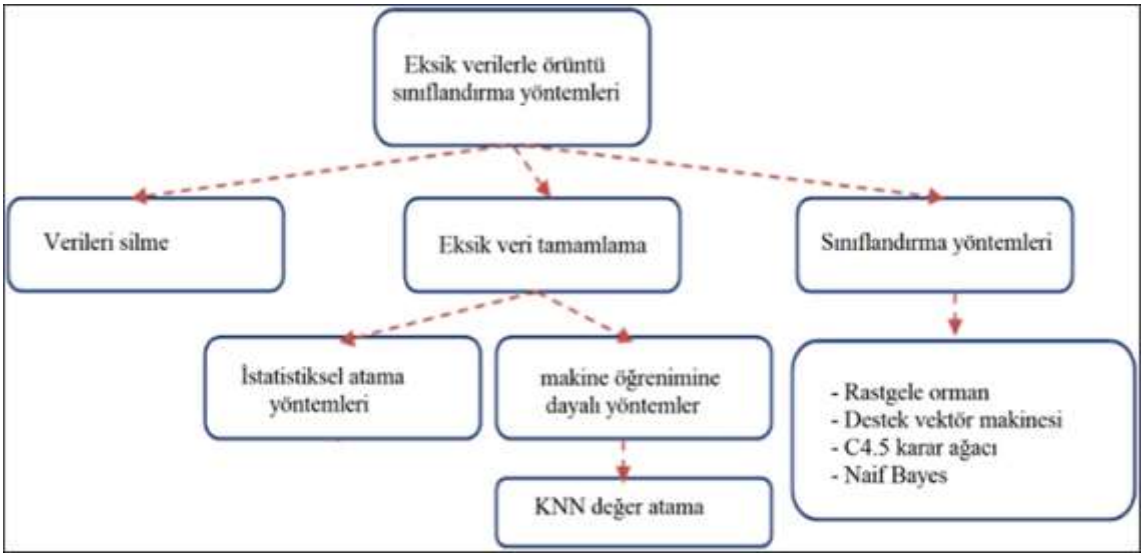
Silme Metodu: Silme yöntemlerinde verimli sonuçlar alabilmek için, eksiklik MCAR türünde olmalı ve eksik veri sayısı mümkün olduğunca az olmalıdır. Silme işlemi iki farklı şekilde yapılır. Birincisi liste halinde silme, ikincisi ise ikili silme yöntemidir (Toka ve Çetin, 2016, s.800)

Ortalama Metodu: Her değişken için gözlemlenen değerlerin ortalaması hesaplanır ve eksik değer atamalarında bu ortalama değerler kullanılır (Jamshidian ve Mata, 2007, s.27).

Regresyon Metodu: Regresyon uygulaması, bir girdi ile çıktı arasındaki veya bir veri noktası ile ilişkili değişken arasındaki ilişkiyi tahmin etmek için kullanılan istatistiksel bir yöntemdir (Osman vd., 2018, s.63282).

Çoklu İsnat Metodu: Bu yöntemde eksik veriler iki aşamalı bir süreç ile tamamlanır. İlk aşamada, eksik değerler, mevcut verilere dayalı bir ispat modeli kullanılarak tahmin edilir. İkinci aşama olan analiz aşamasında, tamamlanmış veri setleri analiz edilir ve elde edilen sonuçlar Rubin kuralları kullanılarak birleştirilir (Sanjar vd., 2020, s.2).

Eksik veri tamamlama için uygun işleme tekniklerinin seçilmesine yardımcı olabilecek diyagram Şekil 1'de verilmiştir.



Şekil 1. Uygun Eksik Veri Tamamlama Tekniğinin Seçilmesi (Idri vd., 2016, s.2865)

Geleneksel yöntemler, az miktarda eksik veri için iyi sonuç verir. Bir veri setindeki eksik veri oranı %5'ten büyük olduğunda, daha gelişmiş teknikler ve modeller kullanılmalıdır (Osman vd., 2018, s.63282). Fuzzy c-Means, Multilayer Perceptrons (MLP) ve k-Nearest Neighbors (kNN) gibi makine öğrenimi teknikleri eksik verilerin tamamlanmasında kullanılan parametrik olmayan gelişmiş tekniklerdir (Kenyhercz ve Passalacqua, 2016, s.3).

kNN ile yapılan eksik veri tamamlama çalışmaları incelendiğinde Minakshi vd.(2014) tarafından yapılan çalışmada, bazı değerlerin eksik olduğu bir öğrenci veri seti kullanılmış ve eksik değerleri belirlemek için Litwise silme, ortalama/mod atama, kNN olmak üzere üç farklı teknik kullanılmıştır. Sonuçlar karşılaştırıldığında kNN yönteminin diğer iki teknikten daha doğru sonuçlar verdiği görülmüştür. Thirumahal ve Patil (2014), eksik değerleri tahmin etmek için kNN ve ARL olmak üzere

iki algoritma kullanmıştır. Sanjar vd. (2020), ulusal ekonomi politikalarını formüle etmek için veri kümelerindeki eksik verileri kNN-MCF yöntemi ile tahmin etmişlerdir. Batista ve Monard (2002) yaptıkları çalışmada, kNN algoritmasının büyük miktarda eksik veri olan veri setlerinde de iyi performans verdiğini göstermişlerdir. Marchang ve Tripathi (2017), çalışmasında kNN'nin üç farklı formatını incelemiştir. kNN-ST (kNN-Spatio-Temporal), kNN-S (kNN-Spatial) ve kNNT (kNN-Temporal). Bu üç yöntemden kNN-ST'nin, kayıp veriler üzerinde çok iyi performans gösterdiğini belirlemişlerdir. Basitliği, kolay anlaşılması ve nispeten yüksek doğruluğu nedeniyle, kNN yaklaşımı, Kanada İstatistikleri, ABD Çalışma İstatistikleri Bürosu ve ABD Nüfus Bürosu'nda yürütülen anketler gibi gerçek veri işleme uygulamalarında da başarıyla kullanılmıştır (Chen ve Shao, 2000, 113). Yapılan diğer bir çalışmada Malarvizhi ve Thanamani (2012), analiz için 5 değişkenli 5000 kayıttan oluşan bir veri seti içerisindeki eksik verileri tahmin etmek için kNN algoritması kullanmışlardır. Zhang vd. (2017), eksik veri atama işlemi için Korelasyon Matrisi kNN (CM-kNN) sınıflandırması olarak adlandırılan, farklı test veri noktalarına farklı k değerleri atayarak veri noktalarını yeniden yapılandırmak için bir korelasyon matrisi oluşturmayı önermişlerdir. Susanti vd. (2018) çalışmalarında, kNN yöntemi kullanılarak eksik değer tahmini yapmışlardır. Çalışmada en iyi k değerini belirleyebilmek için 1,5,10,15,20 değerleri kullanılmıştır. En iyi sonucu k=15 değerinin verdiği belirlenmiştir.

Yöntem ve Uygulama

kNN (k-en yakın komşu) yöntemi, basitliği ve birçok eksik değer tamamlama probleminde kanıtlanmış etkinliği nedeniyle yaygın olarak kullanılmaktadır. Kayıp bir değer için yöntem, en yakın k değişkenlerini arar ve belirlenen komşuların gözlemlenen değerlerinin ağırlıklı ortalamasını hesaplar (Liao vd., 2014, s.3). Makalede kullanılan program Python 3.7 kullanılarak yazılmıştır. Python kütüphanelerinden Scikit-learn makine öğrenimi kitaplığı, en yakın komşu atamasını destekleyen KNNImputer sınıfını içerisinde bulundurmaktadır. KNNImputer, eksik değerleri tahmin etmek için kullanılan yönteme göre veri dönüşümü yapmaktadır. Algoritma adımları:

Adım 1: Eksik bir değer seçin.

Adım 2: O satırdaki diğer değerleri seçin.

Adım 3: Komşuların sayısını seçin.

Adım 4: Diğer ilgili satır elemanlarından Öklid mesafesini hesaplayın.

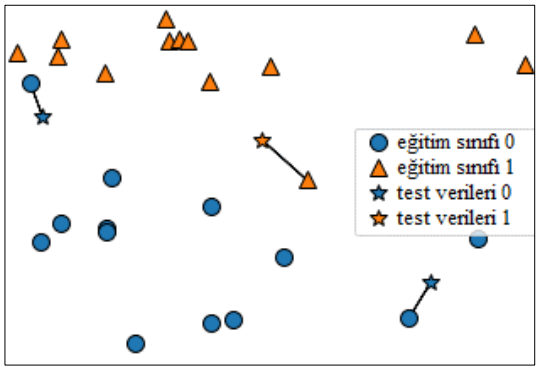
Adım 5: En küçük iki mesafeyi seçin.

kNN Algoritması

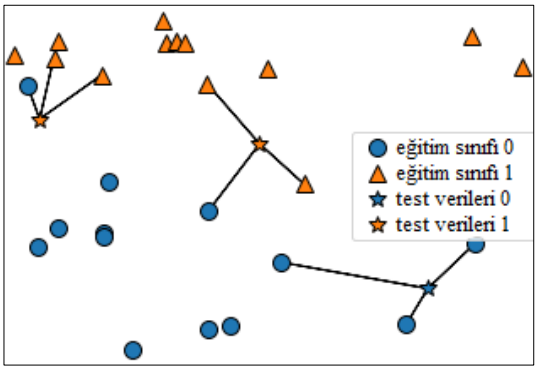
kNN algoritması, makine öğrenimi algoritmalarının en basitlerinden biridir. Bu teknik, eksik verilerin en yakın komşudan gelen benzer verilerle değiştirildiği etkin bir sınıflandırma algoritmasıdır. kNN ile eksik verilerin ele alınması, en yakın komşuların sayısını veya k ile sembolize edilen en yakın gözlemleri belirleyerek başlar, ardından eksik verileri içermeyen her gözlemden en küçük mesafeyi hesaplar. kNN, Öklid mesafesini kullanarak yeni veriler (test verileri) ve önceden bilinen sınıf verileri (eğitim verileri) arasındaki mesafeyi hesaplayarak çalışır (Idri vd., 2016, s.596).

$$d(x,y) = \sqrt{\sum_{j}^{n}(x_j - y_j)^2} \quad (1)$$

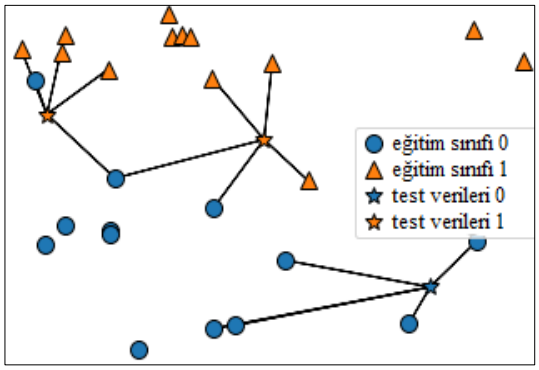
kNN, en yakın komşularının sayısına göre sınıflandırma yaparken belirlenen k değeri için 1 no'lu denklemde verilen formül kullanılarak Öklid uzaklık mesafelerini hesaplar ve bir grup oluşturur. Sınıflandırılması yapılacak veri elemanı, sayısı en fazla olan gruba dahil edilir. k=1,3,5,15 için örnek bir gruplandırma Şekil 2'de verilmiştir.



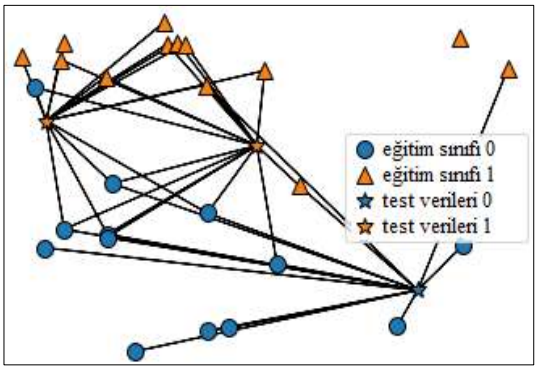
Şekil 2(a). k=1 Değeri İçin Sınıflandırma



Şekil 2(b). k=3 Değeri İçin Sınıflandırma



Şekil 2(c). k=5 Değeri İçin Sınıflandırma



Şekil 2(d). k=15 Değeri İçin Sınıflandırma

Sınıflandırma yapılırken k değerinin doğru seçilmesi algoritmanın performansı üzerinde etkili olmaktadır. Farklı grup elemanlarının aynı sayıda kalmaması için k değerinin tek sayı olarak belirlenmesi doğru bir yaklaşım olarak görülmektedir. Eksik veriler üzerinde çalışılırken Öklid mesafesi için kullanılan denklem değiştirilir. Bunun için aşağıda verilen denklem mesafe hesaplaması için kullanılır.

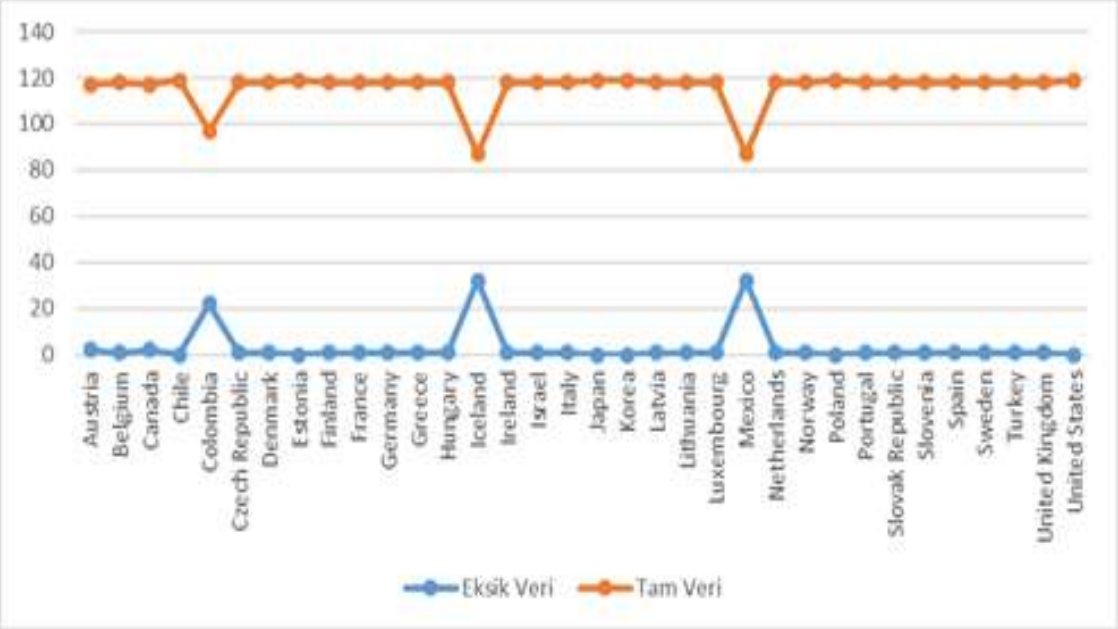
$$d_{xy} = \sqrt{ağırlık * mevcut\ koordinatların\ farklarının\ karesi} \quad (2)$$

$$ağırlık = \frac{toplam\ koordinat\ sayısı}{mevcut\ koordinatların\ sayısı} \quad (3)$$

Bulgular

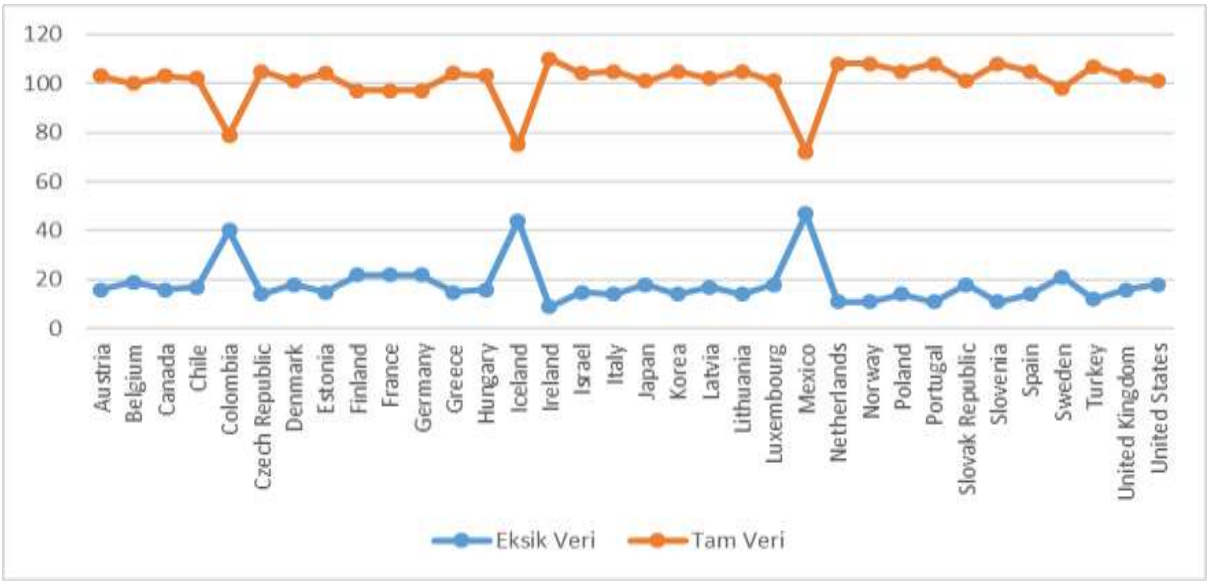
Çalışmada OECD Temel Ekonomik Göstergeler veri tabanı içerisindeki endüstriyel üretim başlıklı veriler kullanılmıştır. Bu başlık altında 34 ülkeye ait büyüme değerleri yüzdesel olarak verilmiştir. Veri setine ait öznitelikler üretim değerleri, fiziksel hacimler, ciro ve çalışma saatleridir.

Gerçek veri seti içerisinde 113 eksik veri ve 3933 tam veri olmak üzere 4046 adet değer bulunmaktadır. Gerçek veri setine ait dağılım Şekil 3'te verilmiştir.



Şekil 3. Gerçek Veri Seti İçerisindeki Eksik Ve Tam Veri Sayısı

kNN algoritmasının yüksek orandaki eksik veriler üzerindeki başarı yüzdesini belirleyebilmek için veri seti içerisinde kolonlardan rastgele veriler çıkartılarak yeni eksik veri seti oluşturulmuştur. Yeni veri setine ait istatiksel değerler Şekil 4'te verilmiştir.



Şekil 4. Veri Seti İçerisindeki Eksik Ve Tam Veri Sayısı

Veri seti Aralık 2010- Ekim 2020 arası toplam 4046 veriden oluşmaktadır. Veri seti içerisinde 485 adet eksik veri bulunmaktadır. Program tarafından endüstriyel üretim veri seti içerisinde belirlenen eksik veri sayı ve oranları Tablo 1'de verilmiştir.

Tablo 1. Endüstriyel Üretim Veri Seti Eksik Veri Bilgileri

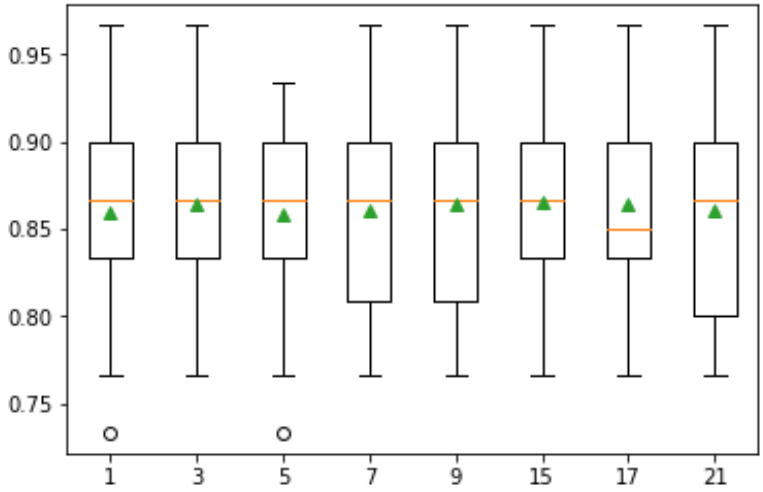
Ülke	n	%	Ülke	n	%
Avusturya	19	16	Japonya	11	9,2
Belçika	11	9,2	Kore	12	10,1
Kanada	15	12,6	Letonya	13	10,9
Şili	11	9,2	Litvanya	8	6,7
Kolombiya	34	28,6	Lüksemburg	5	4,2
Çek Cumhuriyeti	12	10,1	Meksika	38	31,9
Danimarka	15	12,6	Hollanda	9	7,6
Estonya	12	10,1	Norveç	12	10,1
Finlandiya	14	11,8	Polonya	10	8,4
Fransa	12	10,1	Portekiz	13	10,9
Almanya	13	10,9	Slovak Cumhuriyeti	11	9,2
Yunanistan	15	12,6	Slovenya	9	7,6
Macaristan	13	10,9	İspanya	10	8,4
İzlanda	42	35,3	İsveç	13	10,9
İrlanda	15	12,6	Türkiye	11	9,2
İsrail	12	10,1	İngiltere	9	7,6
İtalya	14	11,8	ABD	12	10,1

kNN algoritması ile sınıflandırma yapmak için kullanılan k değerlerine ait başarı yüzdeleri Tablo 2'de verilmiştir.

Tablo 2. k Değerleri İçin Başarı Oranı

k değeri	Başarı Oranı (%)
1	86,3
3	85,4
5	86,6
7	86,3
9	86,1
15	86,8
17	85,6
21	86,3

Çalışmanın sonunda, her sonuç grubu için kutu ve Whisker grafiği oluşturularak sonuçların dağılımı karşılaştırılmalı olarak Şekil 5'te verilmiştir. Şekil 5 incelendiğinden eksik değerleri hesaplarken k değerlerinde çok fazla fark olmadığı, ortalama performansın (yeşil üçgen) birbirine yakın olduğu görülmektedir.



Şekil 5. k Değerleri İçin Oluşturulan Whisker Grafiği

kNN yöntemi için en yüksek doğru tahmin sonucu veren k değeri 15 olarak tespit edilmiştir. Yöntemin eksik verileri doğru tahmin etme oranı kabul edilebilir seviyede olduğu ispatlanmıştır. Bir sonraki adım olarak OECD Temel Ekonomik Göstergeler başlığı altındaki endüstriyel üretim verilerinde bulunan 113 eksik veri tamamlanmıştır. Eksik veri miktarı en fazla olan 3 ülkeye ait tahmin sonuçları Tablo 3, Tablo 4 ve Tablo 5'te verilmiştir.

Tablo 3. Kolombiya Eksik Veri Tahmin Sonuçları

Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler
-2,6	-2,6	1,6	1,6	0,7	,7	1,7	1,7	-3,6	-3,6
5,1	5,1	0,9	0,9	-1,2	-1,2	-1,5	-1,5	NaN	0,8
-1,7	-1,7	-2,4	-2,4	1,6	1,6	-1,6	-1,6	NaN	0,1
2,6	2,6	-4,5	-4,5	-0,6	-,6	-0,0	,0	NaN	0,0
-0,6	-0,6	11,1	11,1	-0,3	-,3	-1,1	-1,1	NaN	-0,4
0,7	0,7	-5,2	-5,2	2,2	2,2	1,7	1,7	NaN	-0,3
-2,1	-2,1	1,9	1,9	-1,1	-1,1	0,2	0,2	NaN	-1,0
0,1	0,1	-0,0	,0	1,4	1,4	3,2	3,2	NaN	-0,2
5,0	5,0	-0,6	-,6	2,7	2,7	-2,9	-2,9	NaN	-1,0
-4,7	-4,7	0,6	,6	-2,2	-2,2	1,2	1,2	NaN	0,5
2,6	2,6	0,4	,4	1,5	1,5	-1,9	-1,9	NaN	0,5
0,8	0,8	0,2	,2	1,1	1,1	1,6	1,6	NaN	-0,3
-1,2	-1,2	-1,1	-1,1	-0,6	-0,6	0,7	0,7	NaN	-0,5
-1,2	-1,2	-0,1	-,1	2,9	2,9	-1,1	-1,1	NaN	0,5
1,9	1,9	1,9	1,9	0,0	0,0	0,3	0,3	NaN	-0,2
-0,6	-0,6	0,4	,4	-5,6	-5,6	1,4	1,4	NaN	-1,4
-4,5	-4,5	0,5	,5	6,1	6,1	2,6	2,6	NaN	-1,2
3,6	3,6	1,0	1,0	-3,0	-3,0	-1,8	-1,8	NaN	1,2
3,0	3,0	-5,8	-5,8	3,5	3,5	1,1	1,1	NaN	1,4
-1,8	-1,8	5,6	5,6	-9,0	-9,0	0,3	0,3	NaN	1,4
-1,1	-1,1	-2,5	-2,5	12,5	12,5	0,6	0,6	NaN	0,2
1,3	1,3	0,7	,7	-4,4	-4,4	0,3	0,3	NaN	-0,2
-1,5	-1,5	-1,0	-1,0	1,4	1,4	-0,4	-0,4	NaN	-0,6
-3,1	-3,1	0,4	,4	-1,2	-1,2	-0,6	-0,6		

Tablo 4. İzlanda Eksik Veri Tahmin Sonuçları

Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler
4,4	4,4	-9,8	-9,8	-1,1	-1,1	0,7	0,7	NaN	0,4
1,0	1,0	12,0	12,0	0,8	0,8	-2,0	-2,0	NaN	0,9
-0,3	-0,3	12,8	12,8	2,2	2,2	-0,8	-0,8	NaN	1,6
-1,0	-1,0	-8,0	-8,0	2,1	2,1	1,6	1,6	NaN	1,7
1,0	1,0	-4,6	-4,6	3,5	3,5	1,4	1,4	NaN	0,7
5,1	5,1	-1,9	-1,9	-1,3	-1,3	-2,4	-2,4	NaN	1,8
7,1	7,1	0,1	0,1	-0,7	-0,7	-2,3	-2,3	NaN	0,8
-2,1	-2,1	0,3	0,3	-4,6	-4,6	8,1	8,1	NaN	2,5
-0,9	-0,9	1,4	1,4	-4,0	-4,0	9,4	9,4	NaN	-1,1
-1,9	-1,9	0,1	0,1	-2,6	-2,6	-4,5	-4,5	NaN	0,5
-0,3	-0,3	1,7	1,7	-0,9	-0,9	-2,8	-2,8	NaN	2,2
-0,2	-0,2	-1,0	-1,0	2,9	2,9	4,3	4,3	NaN	0,0
-0,9	-0,9	-1,1	-1,1	3,7	3,7	5,7	5,7	NaN	0,7
4,2	4,2	-1,0	-1,0	-4,5	-4,5	-2,4	-2,4	NaN	1,8
4,2	4,2	-0,5	-0,5	-4,1	-4,1	-1,3	-1,3	NaN	-0,7
1,8	1,8	-3,4	-3,4	-2,2	-2,2	NaN	1,0	NaN	0,0
4,3	4,3	-0,4	-0,4	-1,3	-1,3	NaN	0,5	NaN	-0,1
-6,4	-6,4	0,7	0,7	2,3	2,3	NaN	-0,5	NaN	0,1
-4,4	-4,4	2,0	2,0	2,2	2,2	NaN	1,2	NaN	0,5
-1,7	-1,7	3,5	3,5	2,5	2,5	NaN	0,5	NaN	0,7
-0,6	-0,6	4,7	4,7	3,9	3,9	NaN	-0,3	NaN	1,3
2,8	2,8	3,3	3,3	-5,4	-5,4	NaN	2,0	NaN	2,6
4,0	4,0	4,6	4,6	-4,0	-4,0	NaN	-0,5	NaN	0,0
-8,6	-8,6	-1,6	-1,6	-0,7	-0,7	NaN	0,7		

Tablo 5. Meksika Eksik Veri Tahmin Sonuçları

Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler	Gerçek Veriler	Tahmini Veriler
0,7	0,7	-1,2	-1,2	0,0	0,0	-0,1	-0,1	NaN	0,3
0,2	0,2	-0,5	-0,5	-0,3	-0,3	0,1	0,1	NaN	-0,1
0,2	0,2	0,5	0,5	0,3	0,3	-0,1	-0,1	NaN	0,1
0,6	0,6	0,4	0,4	0,1	0,1	0,1	0,1	NaN	0,0
0,0	,0	-1,0	-1,0	0,7	0,7	-0,2	-0,2	NaN	0,2
0,8	0,8	0,1	0,1	-1,3	-1,3	-0,1	-0,1	NaN	0,3
-0,4	-0,4	0,3	0,3	0,4	0,4	0,0	0,0	NaN	0,1
0,3	0,3	0,2	0,2	0,4	0,4	-0,9	-0,9	NaN	0,0
0,2	0,2	0,3	0,3	0,4	0,4	0,3	0,3	NaN	0,0
0,6	0,6	-0,3	-0,3	0,5	0,5	-0,5	-0,5	NaN	0,1
0,6	0,6	0,3	0,3	-0,8	-0,8	0,5	0,5	NaN	0,0
0,1	0,1	0,4	0,4	-0,7	-0,7	-0,6	-0,6	NaN	0,1
0,5	0,5	0,2	0,2	0,5	0,5	1,4	1,4	NaN	0,2

0,0	0,0	0,5	0,5	0,3	0,3	-0,5	-0,5	NaN	0,0
0,3	0,3	0,1	0,1	-0,1	-0,1	1,0	1,0	NaN	0,0
0,2	0,2	0,3	0,3	-0,4	-0,4	NaN	0,2	NaN	0,1
0,5	0,5	0,7	0,7	-0,3	-0,3	NaN	0,0	NaN	0,4
0,1	0,1	0,4	0,4	0,1	0,1	NaN	0,0	NaN	0,1
0,6	0,6	-0,2	-0,2	0,2	0,2	NaN	0,2	NaN	0,0
-0,4	-0,4	0,1	0,1	0,3	0,3	NaN	0,1	NaN	-0,1
-0,8	-0,8	0,1	0,1	-0,4	-0,4	NaN	-0,1	NaN	0,1
1,2	1,2	0,2	0,2	-0,0	0,0	NaN	0,4	NaN	0,2
-1,6	-1,6	0,1	0,1	0,3	0,3	NaN	0,0	NaN	0,1
1,5	1,5	0,2	0,2	0,1	0,1	NaN	0,2		

Sonuç

Ekonomik İşbirliği ve Kalkınma Örgütü (OECD), daha iyi yaşam standartları oluşturmak için çalışan uluslararası bir organizasyondur. Amacı, herkes için refah, eşitlik, fırsat ve refahı teşvik eden politikaları şekillendirmektir. Hükümetler, politika yapıcılar ve vatandaşlarla birlikte, kanıta dayalı uluslararası standartlar oluşturmak ve bir dizi sosyal, ekonomik ve çevresel zorluklara çözümler bulmayı amaçlamaktadır. Bu amaç doğrultusunda ülkeler hakkında birçok göstergede veri toplamaktadır. Daha doğru analizler yapabilmek için bu verilerin eksiksiz olması gerekmektedir. Fakat ulusal ve uluslararası farklı kaynaklardan toplanan bilgilerde eksiklikler olmaktadır. Bu eksiklikler özellikle istatiksel analiz ve makine öğrenmesi yöntemleri kullanarak çalışmak isteyen araştırmacılara problem çıkartmaktadır. Bu tür analizler için veri setlerinin öncelikle eksik verilerden temizlenmesi gerekmektedir.

Bu çalışmada kNN algoritması kullanılarak OECD verileri üzerinde belirlenen eksik verilerin tahmin edilmesi amaçlanmıştır. Bu amaç doğrultusunda OECD Temel Ekonomik Göstergeler başlığı altındaki endüstriyel üretim verilerinde bulunan 113 eksik veri tahmin edilmiştir. Algoritmada en iyi sonucu alabilmek için iki aşamalı yöntem denenmiştir. Birinci aşamada en doğru k değeri belirlenmeye çalışılmıştır. Veri seti üzerindeki en başarılı k değeri 15 olarak belirlenmiştir. İkinci aşamada algoritmanın doğruluğunu kıyaslayabilmek için orijinal veri setinden rastgele veriler silinerek test için kullanılmıştır. Test sonuçlar %86,8'lık bir başarı sağladığı belirlenince gerçek veri seti üzerinde tahminler yapılmıştır. En yüksek eksik veri değerine sahip olan Kolombiya 22, İzlanda ve Meksika 32 eksik veri değerleri yüksek doğruluk oranıyla tahmin edilmiştir.

Kaynakça

Andridge, R.R. & Little, R.J.A. (2010). A Review of Hot Deck Imputation for Survey Non-response, Int Stat Rev.,78(1), 40–64.

Batista, G.E.A.P.A. & Monard, M.C. (2002). A study of K-nearest neighbour as an imputation method, Brazilian Research Councils, 1-10.

Chen, J. & Shao, J., 2000. Nearest neighbor imputation for survey data. Journal of Official Statistics, 16(2), 113–131.

Choudhury, A. & Kosorok, M.R. (2020). Missing data imputation for classication problems. National Cancer Institute, 1-27.

Çilingirtürk, A.M. & Altaş, D. (2010). Makro iktisat verilerinde kayıp verilerin regresyona dayalı en yakın komşu "Hot Deck" yöntemi ile tamamlanması. Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 25(2), 73-83.

Dondersa, A.R.T., Heijdenc, G.J.M.G., Stijnend, T., & Moons, K.G.M. (2006). Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 59, 1087-1091.

Fendoğlu, E. (2020). Metasezgisel yöntemlerle rotalama problemlerinin çözümü için çok aşamalı bir yaklaşım. Ankara: Gazi Kitabevi

Folch-Fortuny, A., Arteaga, F., & Ferrer, A. (2016). Missing data imputation toolbox for MATLAB. Chemometrics and Intelligent Laboratory Systems 154, 93–100.

Huang, J. & Sun, H. (2016). Grey relational analysis based k nearest neighbor missing data imputation for software quality datasets. IEEE International Conference on Software Quality, Reliability and Security, 86-91.

Idri, A., Abnane, & I., Abran, A. (2016). Missing data techniques in analogy-based software development effort estimation. The Journal of Systems and Software, 117, 595–611.

Jamshidian, M. & Mata, M. (2007). Advances in Analysis of Mean and Covariance Structure when Data are Incomplete. Handbook of Computing and Statistics with Applications, 1, 21- 44.

Kenyhercz, M.W. & Passalacqua, N.V. (2016). Missing data imputation methods and their performance with biodistance analyses biological distance analysis. Biological Distance Analysis,181-194.

Liao, S.G., Lin, Y., Kang, D.D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F.C., & Tseng, G.C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how?, BMC Bioinformatics, 15(346), 2-12.

Little, R.J.A. & Rubin, D.B. (2020). Statistical analysis with missing data, 3rd Edition, JohnWiley & Sons, Inc, ISBN 9781118596012, 1- 462.

Malarvizhi, R. & Thanamani, A.S. (2012). K-Nearest Neighbor in Missing Data Imputation. International Journal of Engineering Research and Development, 05-07.

Marchang, N. & Tripathi, R. (2017). KNN-ST: Exploiting Spatio-Temporal Correlation for Missing Data Inference in Environmental Crowd Sensing. Ieee Sensors Journal, 1-8.

Minakshi, Vohra, R., & Gimpy. (2014). Missing value imputation in multi attribute data set. International Journal of Computer Science and Information Technologies, 5(4) , 5315-5321.

Ordóñez, G. C., Lasheras, F.S., Juez, F., & Sánchez, A.B. (2017). Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions. Journal of Computational and Applied Mathematics, 311, 704–717.

Osman, M.S., Abu-Mahfouz A.M., & Page, P.R. (2018). A survey on data imputation techniques: Water distribution system as a use case. IEEE 6, 63279- 63291.

Pini, A.S.N., Nelso, M.E., Myer, M.M., Shuffre, L.C., Lucchini, M., Elliott, A.J., Odendaal, H.J., & Fifer, W.P. (2020). The K nearest neighbor algorithm for imputation of missing longitudinal prenatal alcohol data. Researchsquare, 1-19.

Sanjar, K., Bekhzod, O., Kim, J., Paul, A., & Kim, J. (2020). Missing data imputation for geolocation-based price prediction using KNN–MCF method, ISPRS Int. J. Geo-Inf., 9(227), 1-13.

Silva, H., & Perera, A.S. (2017). Evolutionary k-Nearest Neighbor Imputation Algorithm for Gene Expression Data. International Journal on Advances in ICT for Emerging Regions, 10 (1), 1-8.

Susanti, Martha, S., & Sulistianingsih, E. (2018). K nearest neighbor dalam imputasi missing data. Buletin Ilmiah Math. Stat. dan Terapannya, 07(1), 9 -14.

Thirumahal, R. & Patil, D.A. (2014). KNN and ARL based imputation to estimate missing values. Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 2(3),119-124.

Toka, O. & Çetin, M. (2016). Imputation and deletion methods under the presence of missing values and outliers: A comparative study, Gazi University Journal of Science GU J Sci ,29(4), 799-809.

Yoon, J., Jordon, J., & Schaar, M. (2018). GAIN: Missing data imputation using generative Adversarial Nets. Proceedings of the 35 th International Conference on Machine Learning, (80), 1-10.

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for kNN classification. ACM Trans. Intell. Syst. Technol,. 8(3),1-19.

Zhang, S.(2012). Nearest neighbor selection for iteratively kNN imputation. The Journal of Systems and Software, 85, 2541– 2552.

Extended Abstract

The Organization for Economic Cooperation and Development (OECD) was established as part of the system of Western organizations created after World War II. The organization was established after World War II to help distribute financial aid, which was about $ 12 billion, which the United States and Canada did at that time under the Marshall Plan, in order to support and repair the economies of Western Europe. In addition, it was established in place of the Organisation for European Economic Cooperation (OEEC) and within the framework of a broader mandate after the completion of the function of the Organisation for European Economic Cooperation (OEEC), which operated between 1947-1960 with the aim of liberalizing and developing trade payments between European countries. OECD databases are being created to organize reliable data for countries. In these databases, there is indicator information divided into 23 categories for countries. The database of main economic indicators includes monthly and quarterly statistics and related statistical methodological information for all OECD member countries and part of non-member countries. The database includes industrial production, leading indicators, business tendency and consumer opinion surveys, retail trade, consumer and producer prices, hourly earnings, employment, unemployment, interest rates, monetary aggregates, and exchange rates, international trade and balance of payments data. Indicators have been prepared by national statistical organizations to meet the needs of users in their own countries. In most cases, indicators are compiled according to international statistical guidelines and recommendations. In order to maximize the suitability of the database for short-term economic analysis, the content of the database needs to be constantly reviewed and there is no missing data. But OECD data, created based on indicators prepared and organized by national and international institutions and governments, is incomplete due to the inability to obtain some information about countries.

The formation of incomplete data is one of the biggest challenges within data scientists who solve classification problems in real-world data. In most cases, researchers infer observations with incomplete values in their study; however, extracting incomplete data leads to making estimates with larger standard errors as it will reduce the sample size. For this, researchers prefer to assign values using different methods rather than deleting missing data. The success of missing data completion methods depends on the reasons for missing data. It divided incomplete data into three groups: missing completely at random (MCAR), missing at random (MAR), and not missing at random (MNAR). MCAR, the cause of the deficiency is completely random, meaning that the probability of missing an observation is not related to any other value. MAR is the case where the deficiency depends only on the observed variables. The MNAR, the probability of an observation being incomplete, also depends on unobserved information, such as the observation's own value. Incomplete data completion methods are substitution methods, machine learning, and deep learning-based approaches that include the mean and regression method used in multivariate statistical analysis. The properties of the missing dataset are important when determining missing data completion methods. The Hot-deck method is commonly used to complete incomplete data in survey research. The deletion method is used in cases where the lack of data in the data set is of the MCAR type and the number of missing data is small. Multiple elimination method: in this method, the missing data is completed by a two-step process. In the first stage, missing values are estimated using a proof model based on existing data. In the analysis stage, which is the second stage, the completed data sets are analyzed and the results obtained are combined using Rubin's rules.

In this study, the kNN (K-Nearest Neighbors) model from supervised learning algorithms was used to estimate incomplete observations. The kNN algorithm is one of the simplest of machine learning algorithms. This technique is an efficient classification algorithm in which missing data is replaced by similar data from the nearest neighbor. the handling of missing data with kNN begins by determining the number of nearest neighbors or the nearest observations symbolized by k, then calculating the smallest distance from each observation that does not contain the missing data. kNN works by calculating the distance between new data (test data) and previously known class data (training data) using Euclidean distance. In order to test the validity of this method, some values in the dataset were randomly extracted and the success of the model created for training was measured. After it was determined that the results were successful, the method was applied to the actual data set. in order to determine the percentage of success of the kNN algorithm on the high rate of missing data, a new missing data set was created by randomly extracting data from columns in the dataset. December 2010 October 2020 the new data set consists of a total of 4046 data. There are 485 missing data in the data set. The Model was used in two stages. In the first stage, 1,3,5,7,9,15,17,21 values were used to determine the optimal K value for the kNN method. Compared to the results obtained, the optimal value for k was determined to be 15. It has been proven that the rate at which the method correctly estimates incomplete data is at an acceptable level. In the second stage, random data was deleted from the original dataset to compare the accuracy of the algorithm and used for testing. Estimates were made on the actual data set when the Test results were determined to have achieved 86,8% success. Colombia with the highest missing data values of 22, Iceland and Mexico with 32 missing data values were estimated with a high accuracy rate.