

## **COVID-19 SALGINI İLE MÜCADELEDE KÜMELEME ANALİZİ ILE ÜLKELERİN SINIFLANDIRILMASI**

**Mert DEMİRCİOĞLU<sup>1</sup>**

ORCID ID: 0000-0002-2287-2067

**Sevgi EŞİYOK<sup>2</sup>**

ORCID ID: 0000-0003-0208-6242

### **ÖZ**

Ülkelerin sağlık harcamaları, sağlık ekipman kapasiteleri ve sağlık kaynakları gibi göstergeler sağlık sonuçlarını etkileyen en önemli kriterlerdir. Bu nedenle tüm dünya ülkeleri için bu kriterlere ayrılan payların benzerliklerinin incelenmesi, ülkelerin sağlık kriterlerine göre sınıflandırılması sonuçların analiz edilmesi açısından yol gösterici olmaktadır. Araştırmanın amacı, 2019 yılının son aylarından beri dünya gündeminin en önemli konularından biri haline gelen COVID-19 salgının sonuçlarının ülkeler bazında incelenmesidir. Araştırmada OECD ve AB üyesi olan 36 ülkenin son dönemde sağlık verileri değerlendirilmiştir, benzerlik gösteren ülkeler tespit edilmiş ve Türkiye'nin bu ülkeler içerisindeki yeri belirlenmeye çalışılmıştır. Verilerin analizinde kümeleme algoritmalarından olan K-ortalamalar yöntemi kullanılmış ve değerler WEKA yazılımı ile çözümlenmiştir. K-ortalamalar yöntemi ile analiz sonucunda ülkeler, ikili, üçlü ve dörtlü kümelere ayrılmıştır.

**Anahtar Kelimeler:** *Kümeleme Analizi, Hiyerarşik Olmayan Kümeleme, Sağlık Göstergeleri, COVID-19*

**CLASSIFYING COUNTRIES WITH CLUSTERING ANALYSIS IN STRUGGLING COVID-19 PANDEMIC**

### **ABSTRACT**

Indicators such as health expenditures, health equipment capacities and health resources of the countries are the most important criteria affecting health results. For this reason, examining the similarities of the shares allocated to these criteria for all countries of the world, classifying the countries according to the health criteria are guiding in terms of analyzing the results. The aim of the research is to examine the results of the COVID-19 pandemic by country which has become one of the most important topics of the world agenda since the last months of 2019. In research, recent health data of 36 countries that are OECD and EU members were evaluated, countries showing similarities have been identified and Turkey's place in these countries has tried to determine. K-means method, one of the clustering algorithms, was used in the analysis of the data and values were analyzed with WEKA software. As a result of the analysis with the K-means method, countries were divided into binary, triple and quadruple clusters.

**Keywords:** *Clustering Analysis, Non-hierarchical Cluster, Health Indicators, COVID-19*

---

<sup>1</sup>Dr. Öğr. Üyesi, Çukurova Üniversitesi, İ.I.B.F., İşletme Bölümü, [ndemirciooglu@cu.edu.tr](mailto:ndemirciooglu@cu.edu.tr)

<sup>2</sup>Yüksek Lisans Öğrencisi, Çukurova Üniversitesi, Sosyal Bilimler Enstitüsü, [sevgiesiyok@gmail.com](mailto:sevgiesiyok@gmail.com)

## **1.GİRİŞ**

Sağlık endüstrisi hem Türkiye'de hem de dünyada hayatı önem taşıyan önemli bir sektördür. Bu alanda yapılacak her bir iyileştirme toplum sağlığı açısından büyük önem teşkil etmektedir. Bu sebeple sektörde gerçekleşen her bir veri analizinin büyük titizlikle gerçekleştirilmesinin maddi yararının yanında manevi yararının da önemi büyütür.

Her sektörde olduğu gibi sağlık verileri de her geçen gün kapasite bakımından artmaktadır. Günümüzde teknolojinin de hızla gelişmesi ve yaygınlaşması ile depolanan ve işlenmemiş halde bulunan çok fazla miktarda veri bulunmaktadır. Büyük ve karmaşık koleksiyonundan ötürü sağlık veri kümelerini geleneksel veri işleme yöntemleri ile analiz etmek oldukça güç hale gelmiştir.

1990'lı yılların ortalarında ortaya çıkan "Veri Madenciliği" kavramı araştırmacılara biyomedikal ve sağlık alanındaki verilerin işlenmesinde büyük yararlar sağlamaktadır. Veri madenciliği, klinik ve idari kararlar için yeni biyomedikal ve sağlık bilgisi ortaya çıkarabilir, büyük deneysel verilerden bilimsel hipotezler üretebilir (Marinov vd., 2012: 2431).

Veri madenciliği yazılımı, büyük verilerden toplanan anlamlı verilerin analiz edilmesi için birçok analiz aracına sahiptir (Ramkumar vd., 2017). Veri madenciliği algoritmaları olan Naive Bayes, KNN, Apriori, karar ağaçları, K-ortalamalar gibi makine öğrenimi algoritmaları bu verileri analiz etmede büyük kolaylık sağlamaktadır (Chen vd., 2017: 1). Sağlık hizmetleri sektörü de dahil olmak üzere bu algoritmalar çok büyük miktardaki anlamsız verilerden anlamlı ve analiz edilebilecek verilerin elde edilmesinde sıkılıkla başvurulan yöntemlerdir.

Kümeleme analizi özellikle büyük hacimli verilerin gruplandırılması için önemli veri madenciliği konularından biridir (Kurasova vd., 2014: 740). Verilerin birbirlerine göre benzer ve ayrı yönlerini inceleyen, bu özelliklere göre verileri kümeleyip analiz eden bu yöntemde en yaygın kullanılan algoritma K-means (K-ortalamalar) yöntemidir. Hiyerarşik olmayan kümeleme yöntemleri arasında sayılan bu yöntemde temel amaç küme içerisindeki ortalama hata payını minimize etmektir. Yani küme içerisindeki verilerin tüm veriler içerisinde birbirine en benzer olanlarından seçilmesidir.

Bu çalışmada 2019 yılında Çin'in Hubei eyaletinin yönetim merkezi ve en büyük şehri olan Wuhan'da başlayan ve aylar içerisinde küresel bir salgın haline gelen COVID-19 salgınınının Türkiye'de ve tüm dünya ülkelerinde gerçekleşen değerleri ve bu ülkelerdeki sağlık göstergeleri baz alınarak bir kümeleme analizi gerçekleştirilmiştir. Verilerin analizi için bir veri madenciliği yazılımı olan WEKA, algoritma olarak ise kümeleme algoritmalarından K- algoritması kullanılmıştır.

## **2. KAVRAMSAL ÇERÇEVE**

Veri madenciliği, en genel tanımı ile bilgisayarlarda depolanan büyük miktarlardaki anlamlı ve anlamsız verilerden anlamlı verilere ulaşabilmek için kullanılan analiz yöntemidir. Veri madenciliğinin amacı veriler arasındaki yararlı, geçerli ve anlaşılabilir ilişkiler kurmaktır. Teknolojik gelişmelere paralel olarak verilerin dijital ortamlarda saklanması nedeniyle veri tabanlarının hacimlerinde olağanüstü bir artış meydana gelmiştir. Bu durum geleneksel sorgulama ve raporlama araçlarının dev veri yiğinları karşısında etkisiz kalmasına yol açmıştır. Bunun sonucunda veri tabanlarında öz bilgi keşif sürecinin temelini oluşturan veri madenciliği önem kazanmıştır. Veri madenciliği modellerinde kullanılan algoritma türüne göre; verilerin içerisinde gömülü olan belirli türde öz bilgiler keşfedilmekte, farklı kullanıcı gereksinimleri karşılanmakta ve verilerin farklı yorumlamaları yapılmaktadır (Karacan ve Yeşilbudak, 2010: 7). Özette veri madenciliği, kuruluşların en önemli bilgilere odaklanmalarına yardımcı olur (Jackson, 2002: 267-270). Tüketimin ve üretimin gün geçiktikçe arttığı günümüzde veri madenciliği çalışmaları tıbbi maddelerin analizi, finansal tahminler, hedef pazarlama, ürün tasarımları gibi birçok alanda gelişmeye devam etmektedir.

Veri madenciliği, birçok alanda olduğu gibi tıbbi teşhis ve sağlık hizmetlerinin etkinlik ölçümlerinde de hayatı bir teknik olarak varsayılar (Ahamad, vd., 2016). Sağlık göstergelerinin analizi için veri madenciliği modelleri sıkılıkla başvurulan yöntemlerdir. Kümeleme analizi de bunlardan en yaygın kullanılanlar arasındadır.

Kümeleme analizi, bir araştırmada incelenen birimleri, aralarındaki benzerliklerine göre belirli gruplar içinde toplayarak sınıflandırma yapmayı, birimlerin ortak özelliklerini ortaya koymayı ve bu sınıflar ile ilgili genel tanımlar yapmayı sağlayan bir yöntemdir. Kümeleme analizinin genel amacı, gruplanmamış verileri benzerliklerine göre sınıflandırmak ve araştırmacıya uygun, işe yarar ve özetleyici bilgiler elde etmede yardımcı olmaktadır (Çelik, 2013: 179). Yöntemde birimler, değişkenler arası benzerlik ya da uzaklıklara dayalı olarak hesaplanan bazı ölçütlerden yararlanarak homojen grplara ayrılır. Analiz sonucunda kümeleri oluşturan elemanlar birbirine benzerlik, başka kümelerin elemanlarından ise farklılık gösterirler. Dolayısıyla kümeleme analizi bir veri setini oluşturan birimleri grup içi değişimin minimum, gruplar arası değişimin maksimum olacağı alt grplara ayırmaktadır (Altıntaş, 2012: 62).

Bu bölümde literatürde sağlık alanında veri madenciliği algoritmaları ile gerçekleştirilen çalışmalar incelenmiş, kümeleme analizi çalışmaları üzerine yoğunlaşılmıştır.

Sığırlı vd. (2006) Avrupa Birliği'ne üye ülkelerin sağlık düzeyi ölçütlerini baz alarak birbirlerine göre konumlarını incelemiştir. Çalışmanın sonucunda Türkiye, Slovakia, Macaristan ve Çek Cumhuriyeti dışındaki diğer ülkelerin sağlık harcamaları ve sağlık harcamalarının gayri safi yurt içi hasıla içerisindeki payı bakımından farklılık gösterdikleri tespit edilmiştir.

Ersöz (2009) çalışmasında OECD ülkelerini sağlık düzeyleri ve sağlık harcamaları açısından kümeleme analizi ile test etmiştir Çalışmanın sonucunda Türkiye'nin, Kore Cumhuriyeti, Meksika, Polonya ve Slovakya ile on dört sağlık değişkeni açısından benzer olduklarını tespit etmiştir. ABD ise sağlık ölçütleri açısından diğer ülkelerden önemli ölçüde ayrılmıştır.

Barlin (2010) çalışmasında Türkiye'nin sağlık hizmet ölçütleri bakımından OECD ülkeleri arasındaki konumunu incelemek amacıyla kümeleme analizi yapmıştır. Türkiye'nin Çek Cumhuriyeti, Macaristan, Kore, Meksika, Polonya ve Slovakya ile benzer olduğu tespit edilmiştir

Altıntaş (2012) çalışmasında Türkiye ve AB ülkelerinde 2008 yılında gerçekleşen sağlık göstergelerini çok değişkenli istatistik yöntemleri ile sınıflandırılmıştır. Çalışmanın sonucunda Türkiye'nin AB ülkelerinin gerisinde olduğunu tespit etmiştir.

Balasubramanian ve Umarani (2012) Hindistan'ın Krishnagiri Bölgesi'ndeki yüksek floraklı içme suyu kullanan insanlardan oluşan bir örneklem üzerinde çalışmışlardır. Analiz yöntemi olarak kümeleme analizi kullanılmıştır. Çalışmada sudaki florur seviyesinin hastanın yaşı, bu suya ne kadar süre maruz kaldığı gibi değişkenler de göz önünde bulundurularak hangi tür dış rahatsızlıklarına neden olduğu tespit edilmeye çalışılmıştır.

Hemant ve Pushpavathi (2012) çalışmalarında şehir hastanesinden edindikleri 768 adet diyabet verisini analiz etmişlerdir. Veriler öncelikle K-ortalamalar yöntemi ile test edilmiştir. Daha sonra verilerin analizi için farklı farklı sınıflandırma algoritması kullanılmıştır. En iyi sonuca bagging algoritması ile ulaşmışlardır.

İşler ve Narin (2012) çalışmalarında konjestif kalp yetmezliği rahatsızlığı bulunan 29 hastadan ve kontrol grubunda yer alan 54 kişiden elde edilen kalp hızı değişkenliğini verilerini kullanarak K-ortalamalar kümeleme yöntemi ile analiz etmişlerdir. Analiz WEKA yazılımı ile gerçekleştirilmiştir. Sonuç olarak, sadece dört kümeyi kullanıldığı durum için en yüksek %98,79 başarıya ulaşıldığı tespit edilmiştir.

Çelik (2013) çalışmasında 81 ile 10 sağlık değişkeni içeren TÜİK'in 2010 yılına ait verilerini kullanmıştır. Araştırmada aynı yapıyı gösteren il gruplarının belirlenmesi amaçlanmıştır. Küme sayılarının belirlenmesinde 81 ilin kümeleme analiziyle 7, 10 ve 15 kümeye ayırdığı sonuçlar incelenmiştir. Analiz sonucunda, sağlık verilerine göre en kötü durumdaki iller de belirlenmiştir.

Girginer (2013) çalışmasında Türkiye'nin sağlık göstergeleri bakımından AB ülkeleri arasındaki yerini tespit etmeye çalışmıştır. Bunu yaparken çok boyutlu ölçekte analizi yönteminden faydalanyanmıştır. Araştırma sonucunda, Türkiye'nin sağlık göstergeleri bakımından Estonya, Macaristan, Litvanya, Slovakya, Polonya, Romanya, Letonya, Bulgaristan ile benzerlik gösterdiği sonucuna ulaşmıştır.

Alptekin (2014) çalışmasında 27 Avrupa Birliği ülkesi ve Türkiye'nin sağlık hizmetleri ile sınıflandırılmasını amaçlamıştır Yöntem olarak bulanık kümeleme analizi kullanılmıştır. Çalışmada Türkiye'nin sağlık istatistikleri açısından Avrupa Birliği ülkelerine kıyasla konumu araştırılmıştır. Veriler 2012 Dünya Sağlık Raporu'ndan elde edilmiştir. Ülkeler iki farklı gruba ayrılmıştır. Türkiye, Bulgaristan, Kıbrıs, Estonya, Macaristan, Letonya, Litvanya, Polonya, Romanya ve Slovakya ile aynı kümeye yer almıştır.

Çınaroğlu ve Avcı (2014) çalışmalarında Sağlık Bakanlığı 2011 yılı istatistik verilerinden elde edilen 26 farklı sağlık göstergesi kullanılmışlardır. Seçilen sağlık göstergeleri bakımından Kalkınma Bakanlığı ve Türkiye İstatistik Kurumu tarafından belirlenen 12 istatistikti bölge birimi hiyerarşik kümeleme yöntemi kullanılarak kümelenmiştir. Hiyerarşik kümeleme sonucunda 12 istatistikti bölgenin 5 farklı kümeye toplandığı ve bu kümelerin oluşumunda belirleyici faktörün sağlık hizmetlerine erişim ve coğrafi yakınlık olduğu sonucuna ulaşılmıştır

Alptekin ve Yeşilaydın (2015) çalışmalarını 34 OECD ülkesinin sağlık göstergelerini baz alarak gerçekleştirmiştir. Yöntem olarak bulanık kümeleme analizini kullanmışlar, araştırma sonucunda ülkeleri 5 farklı kümeye ayırmışlardır. Türkiye, Estonya, Macaristan, Meksika, Polonya ve Şili ile aynı kümeye yer almıştır.

Ahamad vd. (2016) çalışmalarında Virginia Tıp Fakültesi'nin veri tabanından elde ettikleri diyabet, obezite ve hipertansiyon verilerini kümeleme analizi ile incelemiştir. Kümeleme algoritmalarından K-ortalamalar yöntemini kullanmışlardır. Diyabet, obezite ve hipertansiyon gibi risk faktörleri arasındaki farkları açıkça ayırt edebilen on kümeler oluşturulmuştur. Verilerin analizi için WEKA yazılımı kullanılmıştır. Hastane verileri on ayrı kümeye ayrılmıştır. Bu kümelerin obezite, hipertansiyon ve diyabet gibi risk faktörleri arasında belirgin bir şekilde farklılaşlığı tespit edilmiştir. Kümelerin risk faktörlerinin göstergesi olduğu sonucuna varılmıştır.

Kolay ve Erdoğmuş (2016) çalışmalarında UCI (University of California- Irvine) Machine Learning laboratuvarından alınan göğüs kanseri verilerini kullanmışlardır. Veriler, K-ortalamalar kümeleme yöntemi kullanılarak Matlab programında analiz edilmiştir. Daha sonra, hiçbir ön işleme yapmadan, Weka Veri Madenciliği yazılımı kullanılarak veriler, çeşitli makine öğrenme teknikleri kullanılarak sınıflandırılmıştır. Sonuç olarak kümeleme uzaklık ölçütlerinin başarıyı yaklaşık olarak %12 değiştirdiği ve sınıflandırma başarısının yöntemlere göre %45-%79 arasında değiştiği görülmüştür.

Rahman ve Sarma (2016) çalışmalarında Amerikan Kanser Topluluğu'ndan aldıkları prostat kanseri verilerini analiz etmişlerdir. Verilerin testi için K-ortalamalar ve Apriori algoritmasının kombinasyonundan oluşan yeni bir yöntem denenmiş ve çıkan sonuçlar çeşitli doğruluk ölçülerine göre Naive Bayes, rassal orman, SVM ve C4.5 algoritmalarının verdiği sonuçlar ile kıyaslanmıştır. Prostat kanseri verileri için önerilen metodun diğer algoritmala göre daha iyi değerler verdiği sonucuna ulaşılmıştır.

Songur (2016) çalışmasında sağlık göstergelerine göre ekonomik kalkınma ve işbirliği örgütü ülkelerini kümeleme algoritmaları ile analiz etmiştir. Çalışma sonucunda, Türkiye'nin İsrail, Meksika ve Şili gibi ülkelerle aynı kümede yer aldığı tespit etmiştir. Ayrıca bu ülkelerin Bismarck finansman modeli kullanımının ortak yanları olduğu sonucuna varılmıştır.

Kumar ve Khatri (2017) çalışmalarında UCI (University of California- Irvine) Machine Learning laboratuvarından alınan kronik böbrek hastalığı verilerini analiz etmişlerdir. Veriler, J48, Naive Bayes, SVM, rassal orman ve K-NN gibi algoritmalarla analiz edilmiştir. Algoritmalar ROC, kappa istatistikleri, RMSE ve MAE gibi performans ölçütleri ve TP oranı, FP oranı gibi doğruluk ölçütlerine göre karşılaştırılmış, rassal orman sınıflandırıcısının diğerlerine göre daha doğru sonuç verdiği tespit edilmiştir.

Mut ve Akyürek (2017) çalışmalarında OECD ülkelerinin sağlık göstergelerini ele almışlardır Yöntem olarak kümeleme analizini kullanmışlardır. Veriler hem Ward yöntemi hem de K ortalamalar yöntemi ile test edilmiştir. Her iki yöntemin sonucuna göre Türkiye, Meksika ve Şili ile aynı kümede yer almıştır.

Ramkumar vd. (2017) günümüzde yaygın bir kanser türü haline gelen karaciğer kanserinin tahmini için çalışmışlardır. Bunun için BUPA araştırma laboratuvarındaki 20 hastadan örnek veri seti toplamışlardır. Veri setinde 7 tane özellik bulunmaktadır. Analizde Bayes teorimi kullanılmış ve karaciğer kanserinin alkol tüketen insanlarda daha sık görüldüğü sonucuna varılmıştır. 20 örnek içeren veri setindeki alkol tüketen 3 kişinin karaciğer kanseri olma riski diğer 17 kişiden daha fazla çıkmıştır.

Silitonga (2017) çalışmasında Hindistan'da bulunan Hajji Adam Malik Hastanesi'nin 2014 yılındaki hasta profili verilerini kullanmıştır. Hastaneyeye gelen hastaların giriş çıkış tarihleri, hastalık nedenleri gibi değişkenler ele alınmıştır. Araştırmanın amacı Adam Hajji Malik Hastanesi'ne başvuran hastaların hastalık eğilimi örtüsünü bulmaktır. Analiz için K-ortalamalar yöntemi kullanılmış, veriler WEKA yazılımında test edilmiştir. Çalışma sonucunda en sık rastlanılan hastalık tespit edilmiştir.

Sneha vd. (2018) çalışmalarında University of California in Irvine (UCI)'nın makine öğrenmesi veri setlerindeki karaciğer hastalıkları verilerini kullanmışlardır. Veriler Hindistan'ın kuzeydoğu kesimindeki popülasyona aittir. Veriler 416 karaciğer hastalığı olan ve 167 sağlıklı insandan oluşan örneklemi içermektedir. Araştırmada altı tane algoritma kullanılmış ve elde edilen sonuçlar karşılaştırılmıştır. Karaciğer hastalıklarının erken evrelerde tespit edilerek hastalara daha kaliteli bir yaşam şansı verilmesi amaçlanmıştır.

Aggarwal ve Jain (2019) çalışmalarında meme kanseri verilerini Naive Bayes, J48, SMO, REP algoritmaları ile analiz etmişlerdir. Tüm algoritmaların verdiği sonuçlar karşılaştırılmıştır. En uygun sonuçlara J48 ve REP algoritmaları ile ulaşmışlardır.

Akbuğday (2019) çalışmasında 699 örnek ve 10 özellikten oluşan meme kanseri verilerini makine öğrenmesi algoritmalarını kullanarak WEKA yazılımı ile analiz etmiştir. Meme Kanseri veri seti Naive Bayes, K-NN ve destek vektör makinesi (SVM) algoritmaları ile ayrı ayrı test edilmiştir. Veriler üç ayrı algoritma kullanılarak sınıflandırılmış ve elde edilen sonuçlar karşılaştırılmıştır.

Uzun vd. (2019) çalışmalarında son zamanlarda yaygın olarak kullanılan kriyoterapi ve immünoterapi tedavi yöntemlerinin sigil tedavisinde ne kadar başarılı olacağını araştırmışlardır. Bunun için lojistik regresyon ve karar ağacı algoritmaları kullanılarak tahminler yapılmıştır. Çalışmada açık erişime sahip olan UCI veri tabanındaki veriler kullanılarak WEKA yazılımı üzerinde algoritmalar test edilmiştir. Sonuç olarak, seçilen sigil tedavi yönteminin başarısı karar ağacı sınıflandırıcısı ile %85,56 oranında doğru olarak tespit edilmiştir.

Watomakin ve Emanuel (2019) çalışmalarında Diyabet, Sindirim ve Böbrek Hastalıkları Ulusal Enstitüsü'nün 768 adet veriden oluşan veri setini kullanmışlardır. Verileri Naive Bayes ve SVM algoritmaları ile analiz etmişler ve iki algoritmanın sonuçlarını kıyaslamışlardır. Bu veriler için SVM algoritmasının Naive Bayes algoritmasından daha üstün olduğu sonucuna varılmıştır.

Çetintürk ve Gençtürk (2020) çalışmalarında OECD ülkelerinin 2003-2017 yılları arasında gerçekleştirdikleri sağlık harcamalarını miktar ve oransal bakımdan değerlendirek harcama türlerine göre benzer OECD ülkelerini tespit etmeyi ve Türkiye'nin OECD ülkeleri içindeki yerini belirlemeyi amaçlamışlardır. Araştırmada, ülkelerin benzer yanları ve farklılıklarını göz önüne alarak gruplandırmaya yarayan kümeleme analizini kullanılmıştır. 36 OECD ülkesinin 2003-2017 yılları arasında sağlık hizmetlerinde kullandıkları 14 farklı harcama değişkeni Ward yöntemi kullanılarak analize tabi tutulmuştur. Analiz sonucunda ülkeler sağlık harcama türlerine göre kümelenmiş ve Türkiye'nin çeşitli sağlık harcama türlerinde en fazla benzerlik gösterdiği ülkeler; Estonya, Letonya, Meksika, Çek Cumhuriyeti, Lüksemburg, Belçika ve Avustralya olarak belirlenmiştir. OECD ülkelerinin sınıflandırılmasında Türkiye'nin konumuna bakıldığından; sağlık harcamalarında gelişmiş ülkelerin gerisinde kaldığı ve nispeten düşük sağlık harcamalarına sahip ülkelerle aynı kümeye yer aldığı görülmektedir.

Prabadevi vd. (2020) çalışmalarını meme kanseri verileri üzerinde gerçekleştirmiştir. Uygulamada Naive Bayes, karar ağacı, lojistik regresyon, rassal orman, SVM ve ANN algoritmaları kullanılmıştır. Veriler altı ayrı algoritma ile incelenmiş ve elde edilen sonuçlar karşılaştırılmıştır. Analiz için WEKA yazılımı kullanılmıştır.

### **3. YÖNTEM**

Basit tanımlıyla veri madenciliği, büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme iśidir. Bu sayede veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik kestirimlerde bulunmak da mümkündür. Bu anlamlıyla veri madenciliği, bir kurumda üretilen tüm verilerin belirli yöntemler kullanılarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi su yüzüne çırpmaya süreci olarak değerlendirilebilir (Özkan, 2008: 38).

Teknolojinin hızla geliştiği günümüzde her alandaki veri akışı da büyük bir hız kazanmıştır. Üretim ve tüketimin olduğu her alanda veriler işlem yapmayı zorlaştıracak derecede büyümüş ve çoğu alanda anlamsız ve gereksiz bilgi yiğinlarına dönüştülerdir. İşte tam da bu alanda anlamlı veri ile anlamsız veriyi ayırt etmek adına yepeneli bir kavram olan Veri Madenciliği” ortaya çıkmıştır. Günümüzde pazarlama, bankacılık, sigortacılık, elektronik ticaret, sağlık ve endüstri gibi birçok alanda veri madenciliğinin geniş bir kullanım alanı bulunmaktadır (Gemici, 2012: 11).

Veri madenciliğinde kullanılan modeller tahminleyici ve tanımlayıcı olmak üzere iki ana başlık altında toplanır. Tahminleyici modeller ileriye dönük tahminler geliştirmeyi hedeflerken tanımlayıcı modeller mevcut durumu değerlendirmeyi ve bundan sonuçlar çıkarmayı hedeflemektedir (Gemici, 2012: 28).

Veri madenciliğinde verilerin işlenmesi için kullanılan birçok yöntem ve algoritma vardır. Çalışma kapsamında sıkılıkla kullanılan yöntemlere deðinilecektir. Söz konusu veri madenciliği modellerini temel olarak şu şekilde grupperlendirilebilir:

- 1)Sınıflandırma
- 2)Kümeleme
- 3)Birliktelik Kuralları (Özkan, 2008: 44)

Bu modeller arasında sınıflandırma tahminleyici, kümeleme ve birliktelik kuralları tanımlayıcı modellerdir (Gemici, 2012: 28).

Bu çalışma kapsamında tanımlayıcı modeller kategorisinde yer alan kümeleme analizi üzerinde durulacaktır.

Veriyi birbirine benzeyen alt kümelere ayırmaya işlemeye adı verilir (Özkan, 2008: 155). En basit tanımlıyla kümeleme analizi, bir dizi veri nesnesini alt kümeler halinde böölümleme işlemidir. Her alt kümeye bir kümekdir. Her kümekdeki nesneler birbirine benzer ancak diğer kümelerdeki nesnelere benzemez (Han vd., 2001: 481).

Literatürde birçok kümeleme analizi yöntemleri mevcuttur. Bu çalışmada kapsamında kümeleme analiz yöntemleri, en yaygın kullanılan iki ayrı grup altında incelenmiştir.

Hiyerarşik kümeleme yöntemleri arasında en yakın komşu algoritması ve en uzak komşu algoritması ele alınacaktır. Hiyerarşik olmayan kümeleme metodu olarak ise aynı zamanda uygulamada da kullanılacak olan K-ortalamalar yöntemi incelenecaktır.

Kümeleme analizinde en önemli adım bir mesafe ölçüsü belirlemektir (Madhulatha, 2012: 719). Bunlardan en yaygın kullanılanları Öklid uzaklık bağıntısı, Manhattan uzaklık bağıntısı, Minkowski uzaklık bağıntısı ve Mahalanobis uzaklık bağıntısıdır.

**-Öklid Uzaklık Bağıntısı:** Kümeleme analizi algoritmalarında en fazla kullanılan uzaklık ölçüsü Öklid uzaklık bağıntısıdır. Genel tanımı ile denklem 1'deki gibi formülize edilebilir:

i ve j gözlem noktaları, k değişken sayısı olmak üzere,

$$d(i,j) = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2} \quad (1)$$

**-Manhattan Uzaklık Bağıntısı:** Bu uzaklık gözlemler arasındaki mutlak uzaklıklarının toplamı alınarak hesaplanır. Söz konusu bağıntı denklem 2'deki gibidir:

$$d(i,j) = \sum_{k=1}^p (|x_{ik} - x_{jk}|) \quad i,j = 1,2, \dots, n; \quad k = 1,2, \dots, p \quad (2)$$

**-Minkowski Uzaklık Bağıntısı:** p sayıda değişken göz önüne alınarak gözlem değerleri arasındaki uzaklığın hesaplanması söz konusu ise kullanılır ve denklem 3'deki gibi hesaplanır (Özkan, 2008: 133).

$$d(i,j) = \left[ \sum_{k=1}^p (|x_{ik} - x_{jk}|^m) \right]^{\frac{1}{m}} \quad i,j = 1,2, \dots, n; \quad k = 1,2, \dots, p \quad (3)$$

Minkowski uzaklık ölçüsü m = 1 için Manhattan uzaklık ölçüsüne, m = 2 için ise Öklid uzaklık ölçüsüne eşit olacaktır. Minkowski uzaklık ölçüsü genel bir uzaklık ölçüsü, Öklid

ve Manhattan uzaklık ölçüleri ise Minkowski uzaklık ölçüsünün özel bir durumudur (Çelik, 2013: 180).

**-Mahalonobis Uzaklık Bağıntısı:** Mahalonobis uzaklık bağıntısı, iki değişken arasında bir ilişki olduğu durumlarda kullanılır. Değişkenler arasındaki kovaryans veya korelasyonu göz önüne alır.  $p$  değişkenli bir analizde  $i$  ve  $k$  gözlemleri arasındaki Mahalonobis uzaklık ölçüsü denklem 4'teki gibi hesaplanır. Buradaki  $S$ ,  $p \times p$  tipinde örneklem kovaryans matrisini göstermektedir (Çelik, 2013: 180).

$$Md_{ij} = (X_{ik} - X_{jk})' S^{-1} (X_{ik} - X_{jk}) \quad (4)$$

### 3.1. Hiyerarşik Kümeleme

Hiyerarşik kümeleme yöntemleri kümelerin bir ana küme olarak ele alınması ve sonra aşamalı olarak içерdiği alt kümelere ayrılması veya ayrı ayrı ele alınan kümelerin aşamalı olarak bir küme biçiminde birleştirilmesi esasına dayanır (Özkan, 2008: 136). Hiyerarşik bir kümeleme yöntemi veri nesnelerini bir küme hiyerarşisinde gruplayarak çalışır. Veri nesnelerini bir hiyerarşî biçiminde temsil etmek veri özetlemesi ve görselleştirme için yararlıdır (Han vd., 2001: 494).

#### ➤ En Yakın Komşu Algoritması

Bir algoritma kümeler arasında minimum mesafeyi kullandığında bu, “en yakın komşu algoritması” olarak adlandırılır (Han vd., 2001: 498). “Tek bağlantı kümeleme yöntemi” olarak da bilinen bu yöntemde başlangıçta tüm yöntem değerleri birer küme olarak değerlendirilir. Adım adım gözlemler arasındaki mesafeler hesaplanarak kümeler birleştirilir ve yeni kümeler elde edilir. İki kümenin içeriği gözlemler arasında birbirine en yakın olanların uzaklığı iki kümenin birbirine olan uzaklığı olarak kabul edilir (Özkan, 2008: 136-137).

#### ➤ En Uzak Komşu Algoritması

Bir algoritma kümeler arasında maximum mesafeyi kullandığında bu, “en uzak komşu algoritması” olarak adlandırılır (Han vd., 2001: 499). Bu yönteme “tam bağlantı kümeleme yöntemi” adı da verilmektedir. Bu yöntem en yakın komşu algoritmasına çok benzer ancak bu kez kümeler arasındaki uzaklık belirlenirken iki kümenin birbirine en uzak elemanları arasındaki mesafe iki küme arasındaki uzunluk olarak belirlenir (Özkan, 2008: 143).

### 3.2. Hiyerarşik Olmayan Kümeleme

Kümeleme analizinin temel sürümü hiperarşik olmayan kümeleme yöntemidir. Bu yöntem, başlangıç grup sayılarını belirtme ve yakınsanan nesnelerinin yeniden gruplanması tâhsis edilmesi esasına dayanır (Madhulatha, 2012: 721). Bu metoda en yaygın kullanılan

yöntem K-ortalamalar (K-means) yöntemidir. Bu yöntem küme içerisindeki ortalama hatayı en aza indirmeyi amaçlar (Özkan, 2008: 149-155).

### ➤ K-ortalamalar Yöntemi

K-ortalamalar yöntemi kümenin merkezine en yakın olan veriyi kümeye atar. Kümenin merkezi kümeye tüm verilerin aritmetik ortalamasıdır. K-ortalamalar yönteminin işlem adımları şu şekildedir (Madhulatha, 2012: 721).

N boyutlu uzayda N örnekli kümelerin olduğu varsayılsın. Bu uzayı  $\{C_1, C_2, \dots, C_n\}$  biçiminde kümeye ayırdığımızı düşünelim.  $\sum n_k = N$  ( $k=1, 2, \dots, n$ ) olmak üzere  $C_k$  kümelerinin ortalama vektörü  $M_k$  denklem 5'teki gibi hesaplanır:

$$M_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ik} \quad (5)$$

$X_k$  değeri  $C_k$  kümelerine ait olan i. veridir. Kare hata adı verilen Her iterasyonda minimize edilmesi amaçlanan aşağıda gösterildiği gibi hesaplanan kare hata değeri her kümeye için hesaplanır.  $C_k$  kümeleri için kare hata her bir  $C_k$  örneği ile onun merkezi arasındaki Öklid uzaklıklarını toplamıdır. Bu hataya “küme içi değişme” adı da verilir. Kare hata değeri denklem 6'daki gibi hesaplanabilir:

$$e_i^2 = \sum_{i=1}^{n_k} (X_{ik} - M_k)^2 \quad (6)$$

K kümelerini içeren bütün kümeler uzayı için kare hata, küme içindeki değişimelerin toplamıdır. Söz konusu değer denklem 7'deki gibi hesaplanır:

$$E_k^2 = \sum_{k=1}^K e_k^2 \quad (7)$$

K-ortalamalar yöntemi şu adımları izler:

- 1) K küme sayısı belirlenir.

- 2) Her bir kümenin merkezi belirlenir. ( $M_1, M_2, \dots, M_n$ )
- 3) Kümelerin merkez değerleri ( $M_k$ ) ile gözlem değerleri arasındaki uzaklıklar hesaplanır. Gözlem değeri hangi merkeze yakın ise o merkezin kümesine dahil edilir.
- 4) Kümelerde herhangi bir değişiklik olmayıncaya kadar ikinci ve üçüncü adımlara devam edilir (Özkan, 2008: 149-150).

#### **4. UYGULAMA**

Bu çalışmada tüm dünyayı etkileyen COVID-19 salgını verileri ve ülkelerin sağlık göstergeleri kullanılarak bir kümeleme analizi gerçekleştirilmiştir. Araştırma verileri, OECD, World Health Organization, Johns Hopkins University kurumlarının resmi sitelerinden ve data.worldbank, worldometers, wikipedia açık erişimli internet sitelerinden 1 Haziran 2020 tarihinde elde edilmiştir.

Analiz için 10 farklı değişken kullanılmıştır ve şu şekildedir:

**Vaka/Nüfus Oranı (D1):** Her ülke için ilk vakanın gerçekleştiği günden 1 Haziran 2020 tarihine kadar açıklanan vaka sayısının ülke nüfusuna oranıdır. Her bin kişide gerçekleşen değerler baz alınmıştır.

**Ölüm/Nüfus Oranı (D2):** Her ülke için ilk vakanın gerçekleştiği günden 1 Haziran 2020 tarihine kadar açıklanan kayıtlı ölüm sayısının ülke nüfusuna oranıdır. Her bin kişide gerçekleşen değerler baz alınmıştır.

**İyileşen/Nüfus Oranı (D3):** Her ülke için ilk vakanın gerçekleştiği günden 1 Haziran 2020 tarihine kadar açıklanan kayıtlı iyileşen sayısının ülke nüfusuna oranıdır. Her bin kişide gerçekleşen değerler baz alınmıştır.

**Test Sayısı/Nüfus Oranı (D4):** Her ülke için ilk vakanın gerçekleştiği günden 1 Haziran 2020 tarihine kadar yapılan test sayısının ülke nüfusuna oranıdır. Her bin kişide gerçekleşen değerler baz alınmıştır.

**Doktor Sayısı (D5):** Her ülke için bin kişiye düşen doktor sayısı alınmıştır.

**Hemşire Sayısı (D6):** Her ülke için bin kişiye düşen hemşire sayısı alınmıştır.

**Sağlık Harcaması (D7):** Her ülkenin gayri safi yurt içi hasılası içerisindeki sağlık için harcanan orandır.

**Yatak Sayısı (D8):** Her bin kişiye düşen yatak sayısı alınmıştır.

**Yoğun Bakım Yatak Sayısı (D9):** Her 100 000 kişiye düşen yoğun bakım yatak sayısı alınmıştır.

**Yaşlı Nüfus/Nüfus Oranı (D10):** Her ülkenin yaşlı nüfusunun toplam nüfusa oranı alınmıştır.

Çalışma AB ve OECD üyesi olan 36 ülkenin verileri ile gerçekleştirilmiştir. Bu 36 ülkenin sağlık hizmet verileri ile salgın sürecinde gerçekleşen vaka, ölüm ve iyileşen kişi sayıları karşılaştırılırken Türkiye'nin de bu sıralamadaki yeri incelenmiş, hangi grupta yer aldığı değerlendirilmiştir. Salgının başladığı yer kabul edilen Çin'in yapılan test sayısı açıklanmadığı için ülkelere dahil edilememiştir. Veriler Weka yazılımı kullanılarak analiz edilmiştir. Analizde K-ortalamalar yöntemi kullanılmış, 36 ülke ikili, üçlü ve dörtlü kümelere ayrılmıştır.

Ülkelerin K-ortalamalar yöntemi ile analizi sonucu ülkeler iki kümeye ayrılmış ve Tablo 1'de görülen şekilde ülkeler kümelenmiştir. 36 ülkenin %64'ü (23 ülke) ilk kümede, %36'sı (13 ülke) ise ikinci kümede yer almıştır.

**Tablo 1. Ülkelerin İki Gruba Ayrılması**

| Küme 1      |            | Küme 2     |
|-------------|------------|------------|
| Almanya     | Litvanya   | ABD        |
| Avustralya  | Macaristan | Belçika    |
| Avusturya   | Meksika    | Fransa     |
| Bulgaristan | Norveç     | Hollanda   |
| Çekya       | Polonya    | İngiltere  |
| Danimarka   | Portekiz   | İrlanda    |
| Estonya     | Romania    | İspanya    |
| Finlandiya  | Rusya      | İsveç      |
| Hırvatistan | Slovakya   | İsviçre    |
| Japonya     | Slovenya   | İtalya     |
| Letonya     | Türkiye    | İzlanda    |
| Yunanistan  |            | Kanada     |
|             |            | Lüksemburg |

Tablo 2'de iki kümenin incelenen 10 değişkenin küme ortalamaları ve genel ortalamaları verilmiştir. İki kümenin en büyük farkları, vaka/nüfus oranı (D1), ölüm/nüfus oranı (D2) ortalamalarında olmuştur. İkinci kümenin vaka/nüfus oranı birinci kümenin vaka/nüfus oranının 3,83 (4,2911/1,1184) katıdır. Ayrıca ikinci kümenin ölüm/nüfus oranı birinci kümenin ölüm/nüfus oranının 8,2574 (0,3881/0,047) katıdır. Vaka sayılarına paralel olarak ikinci kümenin test sayısı/nüfus oranı birinci kümenin üstünde olmasına rağmen bu oran 1,55 (66,2157/42,6457) olarak vaka/nüfus oranı ve ölüm/nüfus oranı kadar yüksek değildir.

**Tablo 2. İkili Kümelemede Ortalama Değerler**

| Değişken                             | Tüm Veriler | Küme 1  | Küme 2  |
|--------------------------------------|-------------|---------|---------|
| <b>D1: Vaka Sayısı/Nüfus</b>         | 2,2641      | 1,1184  | 4,2911  |
| <b>D2: Ölüm Sayısı/Nüfus</b>         | 0,1702      | 0,047   | 0,3881  |
| <b>D3: İyileşen Sayısı/Nüfus</b>     | 1,3818      | 0,8333  | 2,3521  |
| <b>D4: Test Sayısı/Nüfus</b>         | 51,1571     | 42,6457 | 66,2157 |
| <b>D5: Doktor Sayısı</b>             | 3,5969      | 3,6817  | 3,4469  |
| <b>D6: Hemşire Sayısı</b>            | 9,2344      | 8,0604  | 11,3115 |
| <b>D7: Sağlık Harcaması</b>          | 8,4372      | 7,5265  | 10,0485 |
| <b>D8: Yatak Sayısı</b>              | 4,8211      | 5,5448  | 3,5408  |
| <b>D9: Yaşlı Nüfus/Nüfus</b>         | 18,4722     | 18,6087 | 18,2308 |
| <b>D10: Yoğun Bakım Yatak Sayısı</b> | 12,7361     | 12,8565 | 12,5231 |

İkinci durumda ülkeler 3 kümeye ayrılmıştır. 36 ülkenin %58'i (21 ülke) ilk kümede, %31'i (11 ülke) ikinci kümede, %11'i (4 ülke) ise üçüncü kümede yer almıştır. Üçlü kümelere ayrılan ülkeler Tablo 3'te görüldüğü gibidir.

**Tablo 3. Ülkelerin Üç Gruba Ayrılması**

| Küme 1      | Küme 2     | Küme 3    |
|-------------|------------|-----------|
| Almanya     | Macaristan | ABD       |
| Australya   | Meksika    | Belçika   |
| Avusturya   | Norveç     | Fransa    |
| Bulgaristan | Polonya    | Hollanda  |
| Çekya       | Romanya    | İngiltere |
| Danimarka   | Rusya      | İspanya   |
| Estonya     | Slovakya   | İsveç     |
| Finlandiya  | Slovenya   | İsviçre   |
| Hırvatistan | Yunanistan | İtalya    |
| Japonya     |            | Kanada    |
| Letonya     |            | Portekiz  |
| Litvanya    |            |           |

Tablo 4'de üç kümenin incelenen 10 değişkenin küme ortalamaları ve genel ortalamaları verilmiştir. Üç kümenin en büyük farkları, vaka/nüfus oranı (D1), ölüm/nüfus oranı (D2) ortalamalarında olmuştur. Birinci kümede yer alan ülkelerin vaka/nüfus oranı genel ortalamanın altında kalırken ikinci ve üçüncü kümede yer alan ülkelerin vaka/nüfus oranları genel ortalamanın üstünde yer almaktadır. Ölüm sayısı/nüfus oranında ise birinci kümenin ortalama oranının 0,25 (0,0423/0,1702) katı olarak ortalamanın çok altında çıkmıştır. İkinci kümenin ölüm sayısı/nüfus oranının ortalama oranının 2,48 (0,4213/0,1702) katı olmuştur. Üçüncü küme ise ölüm sayısı/nüfus oranın bakımından genel ortalamanın altında kalmıştır ve birinci ve ikinci kümelerin ortasında bir ölüm sayısı/nüfus oranının ortalama oranına 0,89 (0,1511/0,1702) sahiptir. İyileşen sayısı/nüfus oranları incelendiğinde birinci küme ortalamanın altında kalırken ikinci küme ortalamanın çok az üstüne çıkmıştır. Üçüncü küme ise vaka sayısı oranının yüksek ve ölüm oranının da düşük çıkışmasına paralel olarak ortalamanın çok üstündedir. Test sayısı/nüfus açısından bakıldığından ise birinci ve ikinci kümelerin değerleri birbirine yakınen üçüncü kümenin test sayısı/nüfus oranı 1,88 (96,0275/51,1571) katı olarak genel ortalamanın çok üstündedir. Yaşlı nüfusun nüfusa oranına bakıldığından ise birinci ve ikinci küme ortalamanın çok az üstünde olmasına karşın üçüncü küme ortalamanın çok altında kalmıştır. Bunun en önemli nedeni ise genç nüfusu fazla olan ülkemizdir.

**Tablo 4. Üçlü Kümelemede Ortalama Değerler**

| Değişken                             | Tüm Veriler | Küme 1  | Küme 2  | Küme 3  |
|--------------------------------------|-------------|---------|---------|---------|
| <b>D1: Vaka Sayısı/Nüfus</b>         | 2,2641      | 0,9785  | 3,8389  | 4,6825  |
| <b>D2: Ölüm Sayısı/Nüfus</b>         | 0,1702      | 0,0423  | 0,4213  | 0,1511  |
| <b>D3: İyileşen Sayısı/Nüfus</b>     | 1,3818      | 0,748   | 1,5166  | 4,3383  |
| <b>D4: Test Sayısı/Nüfus</b>         | 51,1571     | 41,7526 | 52,7946 | 96,0275 |
| <b>D5: Doktor Sayısı</b>             | 3,5969      | 3,6995  | 3,6091  | 3,025   |
| <b>D6: Hemşire Sayısı</b>            | 9,2344      | 8,3971  | 10      | 11,525  |
| <b>D7: Sağlık Harcaması</b>          | 8,4372      | 7,6114  | 10,8082 | 6,2525  |
| <b>D8: Yatak Sayısı</b>              | 4,8211      | 5,7776  | 3,5491  | 3,2975  |
| <b>D9: Yaşlı Nüfus/Nüfus</b>         | 18,4722     | 18,9524 | 19,6364 | 12,75   |
| <b>D10: Yoğun Bakım Yatak Sayısı</b> | 12,7361     | 11,6381 | 11,5091 | 21,875  |

Üçüncü durumda ülkeler 4 kümeye ayrılmıştır. 36 ülkenin %22'si (8 ülke) ilk kümede, %25'i (9 ülke) ikinci kümede, %8'i (3 ülke) üçüncü kümede, %44'ü (16 ülke) ise dördüncü kümede yer almıştır. Dörtlü kümeye ayrılan ülkelerin kümelenmesi Tablo 5'teki gibi gerçekleşmiştir.

**Tablo 5. Ülkelerin Dört Gruba Ayrılması**

| Küme 1     | Küme 2    | Küme 3     | Küme 4      |                |
|------------|-----------|------------|-------------|----------------|
| Almanya    | ABD       | Irlanda    | Bulgaristan | Meksika        |
| Avustralya | Belçika   | İzlanda    | Çekya       | Polonya        |
| Avusturya  | Fransa    | Lüksemburg | Estonya     | Romanya        |
| Danimarka  | Hollanda  |            | Hırvatistan | Rusya          |
| Finlandiya | İngiltere |            | Japonya     | Slovakya       |
| İsviçre    | İspanya   |            | Letonya     | Slovenya       |
| Norveç     | İsveç     |            | Litvanya    | <b>Türkiye</b> |
| Portekiz   | İtalya    |            | Macaristan  | Yunanistan     |
|            | Kanada    |            |             |                |

Tablo 6'da dört kümenin incelenen 10 değişkenin küme ortalamaları ve genel ortalamaları verilmiştir. Dört kümenin en büyük farkları, vaka/nüfus oranı (D1), ölüm/nüfus oranı (D2), iyileşen sayısı/nüfus oranı (D3), test sayısı/ nüfus oranı (D4) ortalamalarında olmuştur. Birinci kümede ve dördüncü kümede yer alan ülkelerin vaka/nüfus oranı genel ortalamanın altında kalırken ikinci ve üçüncü kümede yer alan ülkelerin vaka/nüfus oranları genel ortalamanın üstünde yer almaktadır. Ölüm sayısı/nüfus oranında ise dördüncü kümenin ortalama oranının 0,21 (0,035/0,1702) katı ve birinci kümenin ortalama oranının 0,55 (0,0934/0,1702) katı olarak ortalamanın çok altında çıkmıştır. İkinci kümenin ölüm sayısı/nüfus oranı ortalama oranın 2,46 (0,4746/0,1702) katı olmuştur. Üçüncü küme ise ölüm sayısı/nüfus oranı bakımından genel ortalamanın çok az üstündedir. İyileşen sayısı/nüfus oranları incelendiğinde birinci küme ve ikinci küme genel ortalamaya yakınen üçüncü küme genel ortalamanın çok üstünde ve dördüncü küme ise genel ortalamanın çok altında kalmıştır. Test sayısının nüfusa oranına bakıldığından birinci, ikinci ve dördüncü küme değerleri birbirine yakın iken üçüncü küme genel ortalamanın çok üstünde bir değere sahiptir.

**Tablo 6. Dörtlü Kümelemede Ortalama Değerler**

| Değişken                        | Tüm Veriler | Küme 1  | Küme 2  | Küme 3  | Küme 4  |
|---------------------------------|-------------|---------|---------|---------|---------|
| <b>D1:Vaka Sayısı/Nüfus</b>     | 2,2641      | 2,0076  | 3,9381  | 5,572   | 0,8304  |
| <b>D2:Ölüm Sayısı/Nüfus</b>     | 0,1702      | 0,0934  | 0,4746  | 0,1828  | 0,035   |
| <b>D3:İyileşen Sayısı/Nüfus</b> | 1,3818      | 1,6921  | 1,272   | 5,2603  | 0,5611  |
| <b>D4:Test Sayısı/Nüfus</b>     | 51,1571     | 58,9106 | 50,5586 | 119,687 | 34,7676 |
| <b>D5:Doktor Sayısı</b>         | 3,5969      | 4,3975  | 3,3622  | 3,41    | 3,3638  |

|                               |         |        |         |         |         |
|-------------------------------|---------|--------|---------|---------|---------|
| D6: Hemşire Sayısı            | 9,2344  | 12,38  | 9,5322  | 14,6767 | 6,4738  |
| D7: Sağlık Harcaması          | 8,4372  | 10,225 | 10,8456 | 6,9467  | 6,4681  |
| D8: Yatak Sayısı              | 4,8211  | 4,5638 | 3,4578  | 3,46    | 5,9719  |
| D9: Yaşlı Nüfus/Nüfus         | 18,4722 | 19,5   | 19,4444 | 14,3333 | 18,1875 |
| D10: Yoğun Bakım Yatak Sayısı | 12,7361 | 12,6   | 12,3778 | 13,4667 | 12,8688 |

## 5. SONUÇ

Ülkelerin gelişmişlik düzeyi göstergelerinden biri sağlıklı bireyler ve bu sağlıklı bireylerin oluşturduğu sağlıklı toplumlardır. Dolayısıyla sağlık hizmetlerinin üretimi ve sunulması sadece bireysel düzeyde değil toplumsal düzeyde de büyük öneme sahiptir (Mut ve Akyürek, 2017: 411). Bu sebeple bireylere sunulan sağlık hizmetleri yeterli miktarda ve kalitede olmalıdır.

Günümüzde yaşanan teknolojik gelişmeler, nüfusla birlikte artan talep miktarları, sunulan hizmetten beklenen yüksek kalite değerleri her sektörü olduğu gibi sağlık sektörünü de yakından etkilemiştir. Bu nedenle sunulan sağlık hizmetlerinin gerçekleşen değerlerinin kayıt altına alınması sunulan hizmetin verimlilik performansının analizi açısından son derece önem teşkil etmektedir. Bu amaçlar doğrultusunda Avrupa Birliği (AB), Dünya Sağlık Örgütü (DSÖ) ve Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) gibi uluslararası kuruluşlar sağlık göstergeleri belirlemiştir. Sağlık göstergeleri, hem sağlık hizmetlerinin planlanması, yönetilmesi ve hizmete ilişkin politikaların üretimi ve denetlenmesine hem de hizmet sunulan toplumun sağlık statüsünün izlenmesine ve değerlendirilmesine yarayan göstergelerdir. Sağlık göstergeleri toplumun sağlık statüsünde meydana gelen değişimleri takip etmeye olanak sağlamaktır (Mut ve Akyürek, 2017: 412).

Bu çalışmada 2019 yılının son aylarında Çin'de başlayan ve aylar içerisinde tüm dünyaya yayılan COVID-19 salgını verilerinin ülkelerin sağlık göstergeleri doğrultusunda analizi amaçlanmıştır. Türkiye'nin de içerisinde bulunduğu, OECD ve AB üyesi olan 36 ülkenin verileri ele alınmıştır. Çin'e ait test sayıları açıklanmadığı için Çin, analize dahil edilememiştir. Ülkelerin salgının başladığı tarihten 1 Haziran 2020 tarihine kadar gerçekleşen kayıtlı vaka, ölüm, iyileşen kişi sayıları ve gerçekleştirilen test sayıları dikkate alınmıştır. Değerler her ülkenin nüfus sayısına oranlanmıştır. Sağlık göstergesi olarak ise doktor sayısı, hemşire sayısı, yaşlı nüfus sayısı, yatak sayısı, yoğun bakım yatak sayısı ve sağlık harcama oranlarından oluşan toplam 6 değişken ele alınmıştır. Analiz salgın sürecinde gerçekleşen değerler ve sağlık göstergelerinden oluşan toplam 10 değişken ile gerçekleştirılmıştır. Verilerin analizi için kümleme algoritmalarından olan K-ortalamalar yöntemi kullanılmış ve WEKA programı aracılığıyla değerler çözümlenmiştir.

36 ülke iki, üç ve dörderli kümelere ayrılmış ve her üç durum için de çıkan sonuçlar yorumlanmıştır.

İlk durumda ülkeler, K-ortalamalar yöntemi ile iki kümeye ayrılmıştır. 36 ülkenin %64'ü (23 ülke) ilk kümede, %36'sı (13 ülke) ise ikinci kümede yer almıştır. Türkiye; Almanya, Japonya, Danimarka gibi ülkelerin yer aldığı Küme 1'de yer almıştır. ABD, Fransa, UK, İtalya, İspanya, İsveç gibi salgının yoğun olarak yaşadığı ülkeler ise Küme 2'de yer almıştır. İkinci kümede yer alan ülkelerin COVID – 19 ile mücadelede birinci ülkelerdeki kadar başarılı olamamış, bu ülkelerde vaka ve ölüm sayıları birinci kümedeki ülkelerde göre daha yüksek olmuştur.

İkinci durumda ülkeler algoritmaya göre 3 kümeye ayrılmıştır. 36 ülkenin %58'i (21 ülke) ilk kümede, %31'i (11 ülke) ikinci kümede, %11'i (4 ülke) ise üçüncü kümede yer almıştır. Almanya, Polonya, Rusya gibi ölüm/nüfus oranının ortalamaya göre düşük olduğu ülkeler birinci kümede yer alırken ABD, Fransa, İtalya gibi ölüm/nüfus oranının yoğun olarak yaşadığı ülkeler ikinci kümede yer almışlardır. Türkiye ise ölüm/nüfus oranının ortalama değerlere göre düşük yaşadığı İrlanda, İzlanda ve Lüksemburg ile aynı kümede yer almıştır. Ayrıca test sayısının nüfusa oranı bakımından üçüncü kümede yer alan ülkemiz ve diğer ülkeler nispeten daha az nüfusa sahip olan İrlanda, İzlanda ve Lüksemburg diğer iki kümeden daha yüksek bir değere sahiptir. Ayrıca yine ülkemizin içinde bulunduğu üçüncü küme yaşılı nüfusun nüfusa oranında ortalamanın çok altında kalmıştır.

Üçüncü durumda ise ülkeler algoritmaya göre 4 kümeye ayrılmıştır. 36 ülkenin % 22'si (8 ülke) ilk kümede, % 25'si (9 ülke) ikinci kümede, %8'i (3 ülke), üçüncü kümede, %44'ü (16 ülke) ise dördüncü kümede yer almıştır. Birinci kümede salgında başarılı olmuş gelişmiş ülkeler yer almıştır. Dördüncü kümede de COVID-19 ile mücadelede başarılı olan Japonya, Meksika, Polonya, Türkiye gibi ülkeler yer almaktadır. İkinci kümede COVID – 19'un yaygın olarak görüldüğü ABD, İtalya, İspanya, İngiltere gibi ülkeler yer almaktadır. Üçüncü kümede ise nüfusları az olan ve bu yüzden nüfusa göre vaka sayısı yüksek çıkan İrlanda, İzlanda ve Lüksemburg ülkeleri yer almaktadır.

Bu araştırmada salgın ile mücadelede yola çıkılarak sağlık göstergeleri açısından dünya ülkeleri, benzerliklerine göre kümelenmiştir. Türkiye'nin de içinde bulunduğu 36 ülkenin birbirlerine göre konumları değerlendirilmiştir. Çalışmanın sağlık sektörüne verilen değerin ne derece önemli olduğunu anlaşıldığı şu günlerde sağlık hizmet değerlerinin performansının ölçümünde kullanılması için yol gösterici olacağı düşünülmektedir.

## KAYNAKÇA

- Aggarwall, G., Jain S. (2019). "Analysis of Genes Responsible for The Development of Cancer Using Machine Learning." International Conference on Inventive Systems and Control (ICISC 2019)
- Ahamad, M. , Ahmed, M. , Uddin, M. (2016) "Clustering As Data Mining Technique in Risk Factors Analysis of Diabetes, Hypertension and Obesity." EJERS, European Journal of Engineering Research and Science, 1- 6.
- Akbuğday, B. (2019) "Classification of Breast Cancer Data Using Machine Learning Algorithms", Medical Technologies National Conference (TIPTEKNO).
- Alptekin, N. (2014). "Comparison of Turkey and European Union Countries' Health Indicators by Using Fuzzy Clustering Analysis." International Journal of Business and Social Research, 4-10
- Alptekin, N. ve Yeşilaydin, G. (2015). OECD Ülkelerinin Sağlık Göstergelerine Göre Bulanık Kümeleme Analizi ile Sınıflandırılması. İşletme Araştırmaları Dergisi, 7(4), 137-155.
- Altıntaş, T. (2012). Türkiye ve Avrupa Birliği'ne Üye Ülkelerin Sağlık Göstergeleri Açılarından Çok Değişkenli İstatistik Yöntemlerle Karşılaştırılması. Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.
- Barlin, S. F. (2010). Türkiye'de Sağlık Harcamalarının Ekonomik Kalkınma Üzerine Etkileri. Yüksek Lisans Tezi, Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Maliye Anabilim Dalı, Balıkesir.
- Çelik, Ş. (2013). "Kümeleme Analizi ile Sağlık Göstergelerine Göre Türkiye'deki İllerin Sınıflandırılması", Doğu Üniversitesi Dergisi, 14(2), 175-194.
- Cetintürk, İ., Gençtürk, M. (2020). "OECD Ülkelerinin Sağlık Harcama Göstergelerinin Kümeleme Analizi ile Sınıflandırılması", Süleyman Demirel Üniversitesi Vizyoner Dergisi, 11-26.
- Çınaroğlu, S., Avcı, K. (2014). "İstatistik Bölgelerinin Seçilen Sağlık Göstergeleri Bakımından Kümelenmesi", Hacettepe Sağlık İdaresi Dergisi, 17-2.
- Ersöz, F. (2009). "OECD'ye Üye Ülkelerin Seçilmiş Sağlık Göstergelerinin Kümeleme ve Ayırma Analizi ile Karşılaştırılması.", Türkiye Klinikleri Journal of Medical Sciences, 29(6), 1650-1659.
- Gemicı, B. (2012). Veri Madenciliği ve Bir Uygulaması. Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü Ekonometri Anabilim Dalı Ekonometri Programı Yüksek Lisans Tezi, İzmir.

- Girginer, N. (2013).” Çok Boyutlu Ölçekleme ve Kümeleme Analizi ile Sağlık Göstergeleri Bakımından Türkiye'nin AB Üyesi Ülkelerle Karşılaştırılması”, İktisat İşletme ve Finans Dergisi, 28(32), 55-72.
- Han, J., Kamber, M. ve Pei J. Data Mining: Concepts and Techniques, 2001. Academic Press, San Diego, California, USA.
- Hemant, P., Pushpavathi, T. (2012). “A Novel Approach to Predict Diabetes By Cascading Clustering and Classification”. Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)
- İşler, Y., Narin, A. (2012). “WEKA Yazılımında K-ortalama Algoritması Kullanılarak Konjestif Kalp Yetmezliği Hastalarının Teşhisi”. SDU Teknik Bilimler Dergisi, 2-4.
- Jackson, J. (2002). “Data Mining: A Conceptual Overview”. Communications of the Association for Information Systems, 8, 267:296.
- Karacan, H., Yeşilbudak, M. (2010). “Kullanıcı Merkezli İnteraktif Veri Madenciliği: Bir Literatür Taraması”. Bilişim Teknolojileri Dergisi, 3,1, 17-21.
- Kolay, N., Erdoğmuş, P. (2016). “Makine Öğrenmesi Teknikleri ile Göğüs Kanserinin Sınıflandırılması”. Tıp Teknolojileri Kongresi.
- Kumar, N., Khatri, S. (2017). “Implementing WEKA for Medical Data Classification and Early Disease Prediction”. IEEE International Conference on Computational Intelligence and Communication Technology (IEEE-CICT 2017).
- Kurasova, O., Marcinkevicius, V., Medvedev, V., Rapecka, A. ve Stefanovic, P. (2014) “Strategies for Big Data Clustering”. IEEE 26th International Conference on Tools with Artificial Intelligence.
- Madhulatha, T. (2012). “An Overview on Clustering Methods”. IOSR Journal of Engineering, 2(4).
- Marinov, M., Yoo, I., Alafaireet, P., Pena, K., Gopidi, R., Chang, J., Hua, L. (2012). “Data Mining in Healthcare and Biomedicine a Survey of The Literature.” J Med Syst, 36
- Mitrpanont, J., Sawangphol,W., Vithantirawat, T., Paengkaew, P., Suwannasing, P. Daramas, A., Chen, Y.(2017). “Study on Using Python vs Weka on Dialysis Data Analysis”. 2nd International Conference on Information Technology (INCIT)
- Mut, S., Akyürek, Ç. E. (2017). “OECD Ülkelerinin Sağlık Göstergelerine Göre Kümeleme Analizi ile Sınıflandırılması”. International Journal of Academic Value Studies, 3(12), 411-422.
- Özkan, Dr. Y. (2008). *Veri Madenciliği Yöntemleri*, Papatya Yayıncılık Eğitim, 1. Baskı, İstanbul.

Prabadevi, B., Deepa, N., Krithika L.B, Vinod, V. (2020). “Analysis of Machine Learning Algorithms on Cancer Dataset.”

Ramkumar, N., Prakash, S., Ashok kumar, S., Sangeetha, K. (2017).” Prediction of Liver Cancer Using Conditional Probability Bayes Theorem”. International Conference on Emerging Trends in Information Technology and Engineering (IC-ETITE).

Siğırlı, D., Ediz, B., Cangür, Ş., Ercan, İ. ve Kan, İ. (2006). “Türkiye ve Avrupa Birliği’ne Üye Ülkelerin Sağlık Düzeyi Ölçütlerinin Çok Boyutlu Ölçekleme Analizi ile İncelenmesi”. İnönü Üniversitesi Tıp Fakültesi Dergisi, 13(2), 81-85.

Silitonga, P. (2017) “Clustering of Patient Disease Data by Using K-Means Clustering”. International Journal of Computer Science and Information Security (IJCSIS), 15- 7.

Sneha, V., Basavaraju, D., Tanvi, T. (2018). “Effective Analysis and Diagnosis of Liver Disorder”. International Conference on Communication and Signal Processing.

Songur, C. (2016). “Sağlık Göstergelerine Göre Ekonomik Kalkınma ve İşbirliği Örgütü Ülkelerinin Kümeleme Analizi”. Sosyal Güvenlik Dergisi, 6(1)

Uzun, R., İŞLER, Y., TOKSAN, M. (2019). “WEKA Yazılım Paketinin Siğil Tedavi Yöntemlerinin Başarısının Tahmininde Kullanımı”. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 7

Watomakin, D., Emanuel, A. (2019.). “Comparison of Performance Support Vector Machine Algorithm and Naive Bayes for Diabetes Diagnosis.” 2019 5th International Conference on Science in Information Technology (ICSITech).

[worldometers.info/coronavirus/](http://worldometers.info/coronavirus/) (Erişim Tarihi: 01.06.2020)

<https://www.oecd.org/> (Erişim Tarihi: 01.06.2020)

[https://en.wikipedia.org/wiki/List\\_of\\_countries\\_by\\_hospital\\_beds](https://en.wikipedia.org/wiki/List_of_countries_by_hospital_beds) (Erişim Tarihi: 01.06.2020)

<https://data.worldbank.org/> (Erişim Tarihi: 01.06.2020)

<https://www.who.int/> (Erişim Tarihi: 01.06.2020)

<https://coronavirus.jhu.edu/map.html> (Erişim Tarihi: 01.06.2020)