T.C.

FIRAT ÜNİVERSİTESİ

TEKNOLOJİ FAKÜLTESİ

YAZILIM MÜHENDİSLİĞİ BÖLÜMÜ

KORONA VİRÜSÜNÜN RİZE İLİNDEKİ HAVA KALİTESİNE ETKİSİNİN KÜMELEME YÖNTEMLERİ İLE İNCELENMESİ

180541096

Saniye Yurt

DANIŞMAN
Prof. Dr. Resul DAŞ

ELAZIĞ, 2020

KORONA VİRÜSÜNÜN RİZE İLİNDEKİ HAVA KALİTESİNE ETKİSİNİN KÜMELEME YÖNTEMLERİ İLE İNCELENMESİ

Fırat Üniversitesi, Teknoloji Fakültesi, Yazılım Mühendisliği, 23119, Elazığ/TÜRKİYE

SANİYE YURT

180541096@firat.edu.tr

Özet

Çin'den başlayıp bütün dünyayı etkisi altına alan Korona Virüs (COVID-19), solunum yolu ile bulaşan bir virüs türüdür. Virüsün erken teşhisi çok önemlidir. Bu çalışmanın amacı ise virüsün Veri Madenciliği yöntemi kullanılarak hava kalitesindeki değişimin belirlenmesidir. Hava kalite indeksinin belirleyici kriterleri (SO2,PM10,O3,NO2,CO) analiz edilir ve virüsün atmosferdeki etkileri teşhis edilir. Veri seti, Sürekli İzleme Merkezinden indirilmiştir. Rize ilinin mart ayına ait verileri kullanılmıştır. Çalışmada Weka paket programında bulunan Hiyerarşik Kümeleme ve K-Means algoritmaları kullanılmıştır.

Anahtar kelimeler: Kümeleme, K-Means, Hiyerarşik, Korona Virüsü, Veri Madenciliği, COVİD-19

Abstract

Corona Virus (COVID-19), which starts from China and affects the whole world, is a type of virus transmitted through the respiratory tract. Early detection of the virus is very important. The purpose of this study is to determine the change in air quality by using the Data Mining method of the virus. The determining criteria of the air quality index (SO2, PM10, O3, NO2, CO) are analyzed and the effects of the virus in the atmosphere are diagnosed. The data set was downloaded from the Continuous Monitoring Center. March data of Rize province was used. Hierarchical Clustering and K-Means algorithms in Weka package program were used in the study.

Keywords: Clustering, K-Means, Hierarchical, Corona Virus, Data Mining, COVID-19

1)Giriş

Korona virüsü 1960 yılında tespit edilen, birkaç çeşidi olan bir hastalıktır. Daha çok hayvanlarda görülen virüs, ilk kez insanlarda da görülmeye başlandı. Şu anki salgın ilk olarak 2019'un Aralık ayında Çin'in Wuhan şehrinde ortaya çıktı. Dünyayı etkisi altına alan bu virüsten korunmanın en büyük etkenlerinden birisi de insanların kendini evlerinde izole etmesidir.

Koronavirüsle mücadele insan faaliyetlerinden kaynaklanan kirletici emisyonlarının azalmasına sebep olmaktadır. İnsanlar sokağa çıkmayınca, dumanlar gökyüzüne daha az karışınca, çevreyi kirleten kimse olmayınca doğa bundan olumlu anlamdan etkilenmiştir. Bu anlamda, bu çalışmada corona virüs salgınının alınan önlemler neticesinde Rize ilinin HKİ (Hava Kalite İndeksi) 'nin nasıl etkilendiği ele alınacaktır.[2]

2.Veri Madenciliği Nedir?

Veri madenciliği, büyük verilerden anlamlı ve yararlı bilgilerin elde edilmesi sürecidir. Bu süreçte veri kümeleme, sınıflandırma, birliktelik kuralı gibi farklı metotlar kullanılmaktadır.

2.1Çalışmada Kullanılanılan Verilerin Elde Edilmesi

Çalışmada Sürekli İzleme Merkezinden alınan Rize iline ait mart ayına ait hava kalitesi veri seti kullanılmıştır. Veri seti toplam 65 gün kaydından oluşmaktadır ve 6 adet değişkene sahiptir. Bu çalışmada veri madenciliği yöntemlerinden biri olan kümeleme yöntemi kullanılarak korona virüsünün hava kalitesi üzerindeki etkileri tespit edilmeye çalışılmıştır. Hava Kalitesi İndeksi (HKİ) matematiksel hesaplama gerektirmeyen genel bir benzerlik gösterir tek farkı birbirine en uzak olan elamanların arasındaki mesafe iki küme arasındaki uzunluk olarak tayin edilir. HKİ'nin amacı, yaşadığımız bölgedeki hava kalitesi ile sağlığımızı ilişkilendirmemiz için yardımcı olmaktadır. Kolay anlaşılabilmesi için HKİ skalası altı kategoriye bölünmüştür Hava kirleticilerinin halk sağlığı üzerindeki etkisininin ve mevcut hava kalitesinin belirlenmesi amacıyla kullanılır.**(Tablo1)** Hava Kalitesi İndeksinin hesaplanmasında kullanılan

kirleticiler kükürt dioksit (SO2), azot dioksit (NO2), karbon monoksit (CO), ozon (O3) ve partikül madde (PM10)dir. **(Tablo2)**

İyi 0-50	Orta 50-100	Hassas 100-150	Sağlıksız 150-200	Kötü 200-300	Tehlikeli 300-500

***Tablo1**:Hava kalite indeksinin değer aralıkları*

- "**İyi**": HKİ değeri 0-50 aralığındadır.Hava kalitesinin tatmin edici, hava kirliliğinin çok az olduğu veya sağlık riskinin bulunmadığı anlamına gelir.
- "**Orta**": HKİ değeri 51-100 aralığındadır.Hava kalitesi kabul edilebilir, ancak bazı kirleticilerin, toplumun küçük bir kesiminde orta düzeyde sağlık etkisi olabilir.
- "**Hassas**": HKİ değeri 101-150 aralığındadır. Toplumun belli bir kesimi, özellikle belli kirleticilere karşı hassastır. Bu grubun, genel nüfusa göre daha düşük seviyelerde dahi etkilenmeleri muhtemeldir.
- "**Sağlıksız**": HKİ değeri 151-200 aralığındadır. Toplumun tüm kesimleri sağlık etkileri ile karşılaşmaya başlayabilir
- "**Çok sağlıksız**": HKİ değeri 2001-300 aralığındadır.Sağlık alarmı için bir tetikleme noktasıdır. Toplumun tüm kesimleri, çok ciddi düzeyde etkilenebilir.
- "**Tehlikeli**": HKİ değeri 300'ün üzerindedir. Acil durum alarmı için bir tetikleme noktasıdır. Toplumun tüm kesimleri, büyük bir ihtimalle etkilenecektir.

PM10	Solunabilir partiküllerdir. Ana kaynakları çimento fabrikaları,termik santraller,egzoz gazları vb.
SO2	Hava kirliliğine ve asit yağmurlarına sebep olur (SO_2) Ana kaynakları, termik santraller ve endüstriyel kazanlardır.
CO	Solunabilir partiküllerdir. Ana kaynakları yanmamış odun, kömür ve doğal gaz gibi karbonlu yakıtların dumanlarında bulunur.
NO2	NO_2 fosil yakıtlarının, yani Gaz, Kömür ve Yağ'ların yanması sonucunda ortaya çıkmaktadır.
NOX	NOX fosil yakıtlarının yanması ile ortaya çıkmaktadır.
O3	Atmosferdeki oluşumu günün saatleri boyunca gelişir.Güneş ışınları ve yıldırımlardan meydana gelir.

***Tablo2**:Hava kalite indeksi parametrelerinin açıklamaları*

3) Yöntem Ve Metodolojiler

Veri madenciliği konusunda çok sayıda yöntem ve algoritma geliştirilmiştir. Veri madenciliğininasıl amacı, veri yığınlarını kullanarak temel modellerden birini oluşturmaktır. Bu model üzerinden, veriler arasındaki ilişkiyi ortaya çıkarmak ve veri yığınında olmayan farklı bir verinin yorumlanmasını sağlamaktır. Söz konusu veri madenciliği modellerini temel olarak şu şekilde gruplandırılabilir.

1. Sınıflandırma
2. Kümeleme
3. Birliktelik Kuralları

3.1.Kümeleme

Veri madenciliğinin temel konuları arasında yer alan kümeleme yöntemleri, verileri birbiri ile benzer alt kümelere ayırma işlemi olarak bilinmektedir. Uygulamada çok sayıda kümeleme yöntemi kullanılmaktadır. Bu yöntemler, değişkenler arasındaki benzerlik lerden yada farklılıklardan yararlanarak bir kümeyi alt kümeye ayırmakta kullanılmaktadır.

3.2.1 Uzaklık ölçütleri

Kümeleme yöntemlerinin bir çoğu uzaklık ölçütlerine dayanmaktadır bu yüzden iki nokta arasındaki uzaklığın hesaplanmasına gerek vardır.

a)Öklit Uzaklığı: En çok kullanılan uzaklık ölçütü olarak bilinir. Pisagor bağıntısı ile hesaplanır.

$$\mathbf{d(A,B) = \sqrt{(X1 - X2)^2 + (Y1 - Y2)^2}}$$

b)Manhattan Uzaklığı: Bu uzaklık, gözlemler arasındaki mutlak uzaklığın toplamı alınarak hesaplanır**.**

$$\boldsymbol{d(i,j) = \sum_{k=1}^{p} (|xik - xjk|)}$$

c)Minkowski Uzaklığı: P sayıda değişken göz önüne alınarak gözlem değerleri arasındaki uzaklığın hesaplanması söz konusu ise Minkowski uzaklık bağıntısı hesaplanır.

$$d(i,j) = \sum_{i=1}^{p} \left(|x_{ik} - x_{jk}|^{m} \right)^{1/m}$$

3.1.1.Hiyerarşik Kümeleme

Hiyerarşik kümeleme yöntemleri, kümelerin bir ana küme olarak ele alınması ve sonra aşamalı olarak bir küme biçiminde biçerdiği alt kümelere ayrılması veye ayrı ayrı olan kümelerin aşamalı olarak birleşrierilmesidir.

a)En Yakın Algoritması

En yakın komşu algoritmasına "tek bağlantı kümeleme yöntemi" adı da verilmektedir. Başlangıçta tüm gözlem değerleri birer küme olarak değerlendirilir. Adım adım bu kümeler birleştirilerek yeni kümeler elde edilir. Bu yöntemde i ve j arasındaki uzaklık öklik uzaklık ölçütü ile belirlenir.

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

b)En Uzak Komşu Algoritması

Bu yönteme "tam bağlantı kümesi" de denmektedir. En yakın algoritma ilearasındaki en temel fark birbirinr uzak olan elemanları arasındaki mesafe iki küme arasındaki uzunluk olarak tayin edilir.

3.1.2.Hiyerarşik Olmayan Kümelemeler

a) K-Ortalamalar Yöntemi

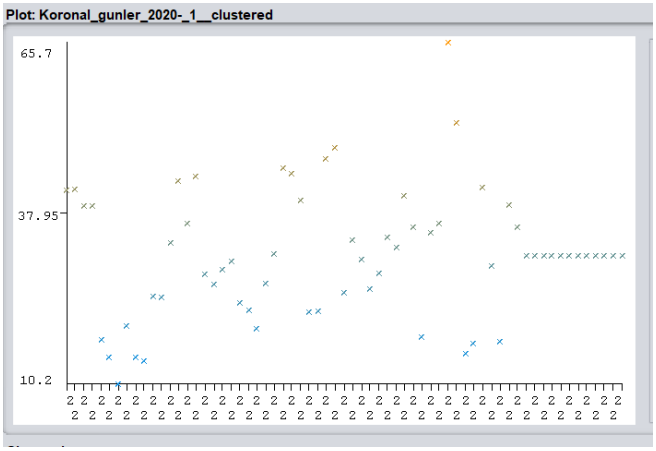
K-Means kümeleme algoritması verileri K giriş parametre sayısı kadar kümeye bölmektedir. K-means kümeleme algoritması değerlendirilmesinde genelolarak karesel hata kriteri SSE kullanılır. En düşük SSE değerine sahip kümeleme sonucu en iyi sonucu verir. verilerin bulundukları kümenin merkez noktalarına olan uzaklıklarının karelerinin toplamı (1) nolu eşitlik ile hesaplanmaktadı

$$\mathbf{SSE} = \sum_{\mathbf{l=1}}^{\mathbf{K}} \sum \mathbf{dist^2(m_i, x)}$$

4.Deneysel Çıktılar

Bu veri setinde 15.05.2020-12.03.2020 tarihleri arasında 66 güne ait veriler vardır. Toplamda 6 paremetreye sahiptir. Makalede kullanılan K-Means yöntemi ve hiyerarşik yöntem için 65 satır 6 sütun içeren bir giriş matrisi oluşturulmuştur.

WEKA (Waikato Environment for Knowledge Analysis) K-Mans yönteminin uygulanması sonucunda aşağıdaki Şekil 1'deki karar kümeleme grafiği edilmiştir.

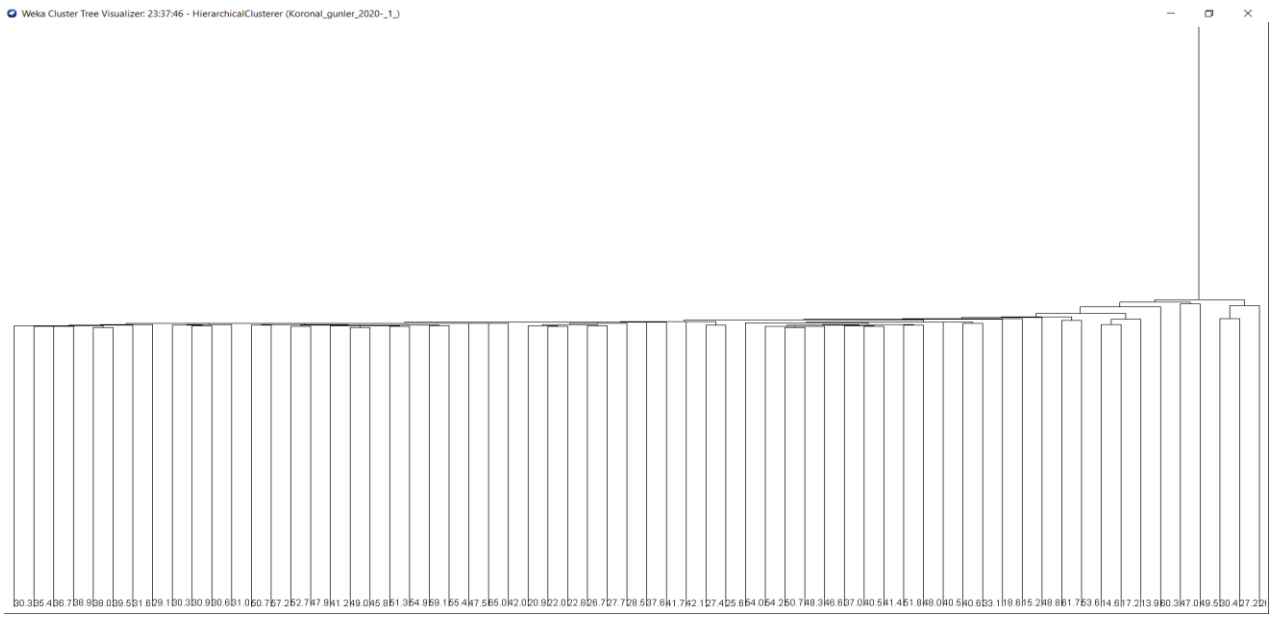


Şekil1:K-Means algoritması sonucu çıkan grafik

X ekseni: Tarih **Y ekseni**: PM10

Yukardaki şekil'e göre 2 kümeleme yapılmıştır.65 verinin 50 tanesi (%77) bir küme kalan 15 tanesi (%23) bir küme olmak üzere 2 kümeye ayrılmıştır.PM10 değeri yasakların başlaması ve bilinçli halkın dışarıya daha az çıkması ile toplu tasıma aracı ve dogaya verilen zarar azaldıgı için bir azalma görülmektedir.

WEKA programında hiyerarşik kümeleme yönteminin uygulanması sonucunda aşağıdaki Tablo3'deki dendogram grafiği elde edilen sonuçlar tablo şeklinde aşağıda verilmiştir.



Tablo3:hiyerarşik kümeleme yöntemi ile elde edilen dendogram grafiği

5.Sonuç

Yöntemin Adı	**Birinci Küme**	**İkinci Küme**
HİYERARŞİK YÖNTEM	**64 (%98)**	**1(%2)**
K-MEANS YÖNTEMİ	**50(%77)**	**15(%23)**

Tablo3*:K-Means ve Hiyerarşik kümeleme yöntemlerinin kümelenme oranları*

Bu makale çalışmasında Hava İzleme Sisteminden rize iline ait koronalı günlere ait alınan verilerin hiyerarşik ve k-means kümeleme yöntemleri uygulanarak bir kümeleme yapılmıstır. Bu kümeleme koronalı günlerdeki hava kalitesinin artısındaki değişimi gözlemlememizi sağlamıştır.Buna ilişkin değişimleri tablo4 ‘den inceleyebilirsiniz. Uygulanan kümeleme yöntemlerinin sonucunda k-means yönteminin hiyerarşik yönteminden daha iyi bir kümeleme gerçekleştirdiği görülmüştür.

İstasyon	**Parametre**	**Birim**	**Min.değer**	**Min.değer**	**Max.değer**	**Max.tarih**
Rize	PM10	µg/m³	14,88	30.03.2020	65,73	01.04.2020
Rize	SO2	µg/m³	4,05	10.04.2020	7,12	06.03.2020
Rize	CO	µg/m³				
Rize	NO2	µg/m³	13,14	04.04.2020	61,87	06.03.2020
Rize	NOX	µg/m³	15,54	04.04.2020	128,29	06.03.2020
Rize	O3	µg/m³	7,36	27.02.2020	50,73	12.04.2020

***Tablo4**:parametrelerin tarihlere göre min ve max değişkenliğini gösteren tablo*

Kaynakça

[1] https://www.iskteknik.com/ipucu/hava-kalitesi-indeksi
[2] https://www.havaizleme.gov.tr/
[3] Papatya yayınları Dr.Yalçın ÖZKAN VERİ MADENCİLİĞİ YÖNTEMLERİ KİTABI
[4] https://www.mynet.com, 18.04.2019.
[5] https://www.msn.com/tr, 18.04.2019.
[6] https://www.milliyet.com.tr/gundem/beklenen-aciklama-corona-virus-etkisi-istanbulda-hava-kirliligi-yuzde-30-azaldi-6172692
[7] https://www.academia.edu/9860553/Veri_Madencili%C4%9Fi_-_K%C3%BCmeleme_ve_K%C3%BCmeleme_Y%C3%B6ntemleri
[8] http://bilgisayarkavramlari.sadievrenseker.com/category/datamining/
[9] https://yzm5550.files.wordpress.com/2012/02/hafta-10-veri-madencilic49fi_kumeleme_model_degerlendirme.pdf
[10] https://www.bbc.com/news/coronavirus