Bölüm 2

MAKİNE ÖĞRENMESİ TEKNİKLERİYLE AKCİĞER KANSERİ TAHMİNLEMESİ

Deniz HERAND[1]
Sercan DÜZEL[2]

GİRİŞ

Akciğer, yaşamsal faaliyetlerimizi devam ettirme konusundaki en önemli organlardan biridir. Akciğer organı, milyonlarca hücreden oluşur ve hücreler bölünerek çoğalma sonucu sayısını arttırır. Kanserler hücrelerin belirli dönemlerdeki kontrol sınırları dışında; hızlı bir biçimde artması sonucunda hücrelerin yığılmasıdır. Bu yığılmalar tümör olarak isimlendirilir [1]

Oluşan tümörleri etkileyen faktörler içerisinde dünya sağlık örgütü tarafından da onaylanan kansere neden olan etmenler olarak sınıflandırmalar içerisinde hava kirliliği büyük bir faktör olarak görülmektedir. Kontrolsüz bir biçimde gelişen hava kirlilikleri hücrelerin bölünmelerini olumsuz yönde etkilemekte ve kansere sebep olabilmektedir. [2]

Ülkemizde oranları ölçülebilen bileşik ve hava partikül oranları; O3,PM2.5, PM10, SO2,CO,NO2 olarak tespit edilmiştir. Çalışmam içerisinde 6 adet ana değer üzerinden çalışma yapılması hedeflenmektedir.[3]

Çalışma içerisinde bileşikler ve partikül büyüklüklerinin neden olduğu EQI(Hava kalitesi ölçütü) hesaplanarak hava kalitesi ölçümleri yapılması hedeflenmektedir. Hava kalitesi oranına göre alınan sonuçlar derin öğrenme ve makine öğrenmesi yöntemleriyle kanser sayılarının tahminlemesi hedeflenmektedir. [4]

Ülkemizde 2013 yılında Dünya sağlık örgütü ile paylaşılan il bazlı rakamlar(iller:Antalya ,Edirne ,İzmir, Trabzon) illerindeki total (kadın-erkek) Akciğer kanseri oranlarını paylaşmıştır.2013 yılındaki bilgilere bakılarak oluşturulacak son modelin sonucunun doğruluk oranının belirlenmesi

[1] Doç. Dr., Marmara Üniversitesi, İşletme Fakültesi, İşletme Enformatiği,
ORCID iD: 0000-0001-6813-1427

[2] Marmara Üniversitesi, işletme Fakültesi, İşletme Enformatiği srcndzl17@gmail.com
ORCID iD: xxx

hedeflenmektedir.[5]

Oluşturulan makine öğrenmesi modelimizin,2021-2022 arasında geçen hava partiküllerinin oranları tespit edilerek 2023-2024 arasındaki Akciğer kanser oranları tahminlemesi yapılması hedeflenmektedir. Tahminleme sonuçları Random forest algoritmalarının veriler üzerindeki uyumu üzerine en doğru şekilde kullanılmıştır.[6]

RF kullanılmasının sebebi verileri parçalara bölerek en doğru tahminleme modelini oluşturmaktır. RF kullanım amacı sınıflandırma yaparak verilerin bölünmesi ve Türkiye Cumhuriyeti verileri üzerindeki oranlar ile bağlantı kurarak tahminleme modelinin doğruluğunu arttırmaktır. Hedeflenen çıktımız gelecek yıllardaki hasta oranları hakkında bizlere yol gösterici bir kaynak çalışma olabilir. Çalışma sonuçları Türkiye

Cumhuriyeti'nde bulunamayan her ildeki hava partikül oranları eksik verileri yüzünden sınırlı iller üzerinde tutulmuştur.

1-) MATERYAL VE METOT

1.1-) Verilerin toplanması ve belirlenmesi

Yapılan araştırma içerisinde makine öğrenmesinin eğitilmesinde kullanılan veriler 2602 adet olarak tespit edilip kullanılmıştır. Veriler Amerika Birleşik devletlerindeki "Hava kalitesi ve akciğer kanseri: Lokal Kontrol Yoluyla Analiz" araştırması için toplanan veriler üzerinden eğitilmiş ve modellenmiştir.(Kumer Pial Das ve arkadaşları). Veriler üzerinde Türkiye Cumhuriyeti üzerinde tespit edilebilecek gazlar eşleştirilmiş ve sadece karşılaştırılabilecek veriler kullanılmıştır ve veriler üzerinde çalışılmıştır.[7]

#	Column	Non-Null Count	Dtype
0	Lung Cancer	2602 non-null	float64
1	PM2.5	2602 non-null	float64
2	Land_EQI	2602 non-null	float64
3	Sociod_EQI	2602 non-null	float64
4	Built_EQI	2602 non-null	float64
5	PM10	2602 non-null	float64
6	SO2	2602 non-null	float64
7	NO2	2602 non-null	float64
8	O3	2602 non-null	float64
9	CO	2602 non-null	float64
10	CN	2602 non-null	float64
11	Disel	2602 non-null	float64
12	CS2	2602 non-null	float64
13	Air_EQI	2602 non-null	float64
14	Water_EQI	2602 non-null	float64
15	EQI	2602 non-null	float64
16	AAC	2602 non-null	int64

Şekil 1: Abd üzerinden alınan veri türleri ve tipleri

Alınan verilerin genel özelliklerinden yola çıkarak havadaki gazların oranlarına göre ölüm oranları ve ortalamaları genel olarak gösterilmesi hedeflenmiştir

```
DTR = DecisionTreeRegressor(random_state = 1)
model_result(DTR,"DTR",X,y)
```

```
 mape score of DTR :
19.906199387681983
 r2_ score of DTR :
0.10666636042526567
 mean_squared_error of DTR :
272.569697322468
 mean_absolute_error of DTR :
12.876018626309664
```

Şekil 2: Abd üzerinden alınan veri türleri ve tipleri

```
LR = LinearRegression()
model_result(LR,"LR",X,y)
```

```
 mape score of LR :
17.48205969647939
 r2_ score of LR :
0.32934998129698057
 mean_squared_error of LR :
204.62553351758893
 mean_absolute_error of LR :
11.019386830031708
```

Şekil 3: Abd üzerinden alınan veri türleri ve tipleri

```
model_result(SvR,"SvR",X,y)
```

```
 mape score of SvR :
15.750725056487786
 r2_ score of SvR :
0.45300611962116455
 mean_squared_error of SvR :
166.89616265102998
 mean_absolute_error of SvR :
```

Şekil 4: Abd üzerinden alınan veri türleri ve tipleri

```
df["Lung Cancer"].describe()
```

```
count    2602.000000
mean       69.170907
std        17.418000
min        12.900000
25%        58.000000
50%        68.600000
75%        79.400000
max       169.900000
Name: Lung Cancer, dtype: float64
```

Şekil 5: Abd üzerinden alınan veri türleri ve tipleri

Makine öğrenmesinin eğitilmesi için kullanılan 2602 verinin kanser oranlarına bakılarak 5 adet ortalama Kanser oranı ve kanser oranlarının gerçekleştiği illerdeki havadaki bileşikler gösterilmiştir

1.2-) Kullanılan Algoritmalar ve Makine Öğrenmesi Yöntemleri

Yapılan çalışmada Doğrusal regnesyon ,lineer regnesyon ,DTR , destek vektör makineleri ve Random forest makine öğrenmesi modeleri kullanılmıştır.Çalışma içerisinde temel alınan değerlendirme metriği 'mean absolute, error ve mean absolute percentage error olarak tercih edilmiştir.[8]Çalışma içerisindeki eğitim testleri süresince makine öğrenmesi öğrenmesi modelleri ile denemeler yapılmış ve Random forest algoritması

Lung Cancer	PM2.5	PM10	SO2	NO2	O3	CO
73.9	12.06	15.07	10.661088	123.657648	522.38	4.463225
68.4	11.12	19.99	17.146847	247.742253	540.79	12.875833
76.1	12.36	15.77	23.257118	183.193624	896.42	19.620539
86.4	12.24	14.92	7.630953	127.779935	563.48	2.951976
73.1	12.97	17.90	8.913795	95.198094	561.94	9.362215

Şekil 6: Abd üzerinden alınan veri türleri ve tipleri

tüm metriklere bakılarak en uygun makine öğrenmesi modeli olarak çalışmada kullanılması sonucu kararlaştırılmıştır.

Çalışma sonuçlarımız Amerika şehirleri üzerinden alınan veri kümemiz üzerinden en iyi tahminleri yapabileceğimiz makine öğrenmesi modelini ve modellerini bulmaktır.Birinci veriler Amerika Birleşik Devletleri üzerinde yapılan algoritmalar ve çalışmaları sonucudur.Birinci çalışmamızda Random forest algoritması kullanılmış, en iyi sonucu veren model olduğu sonucuna varılmıştır.

```
Rfr = RandomForestRegressor(max_depth=9, random_state=1)
model_result(Rfr,"Rfr",X,y)
```

```
 mape score of Rfr :
15.39572486045873
 r2_ score of Rfr :
0.4809626213255982
 mean_squared_error of Rfr :
158.3662082530292
 mean_absolute_error of Rfr :
9.697086605997711
```

Şekil 7: Abd üzerinden alınan veri türleri ve tipleri

State	O3	PM2.5	PM10	SO2	CO	NO2
Adana	44.50	19.99	62.88	8.00	596.96	28.08
Ankara	75.74	17.63	35.47	3.73	615.25	61.53
Ağrı	76.38	12.15	92.47	19.74	745.31	18.06
Hatay	66.03	12.00	16.53	10.33	489.78	24.59
Antalya	51.51	18.36	35.71	3.06	455.17	11.53
Gaziantep	27.37	28.17	57.05	12.85	1140.00	98.75
İstanbul	59.24	16.60	32.30	3.36	112.75	16.89
Çankırı	6.23	9.46	43.83	14.65	426.85	27.86
Kars	45.64	6.00	32.02	8.70	717.32	19.78
Konya	65.94	4.80	16.26	7.06	206.69	6.05
Bursa	30.93	32.68	61.59	6.39	543.22	44.25
Edirne	14.26	16.69	76.16	10.83	471.45	9.90
Muğla	52.45	15.17	47.91	18.60	756.78	23.99
Rize	77.57	19.53	64.85	4.38	481.45	7.79
Karabük	17.23	5.56	61.57	9.35	536.92	12.86
Erzurum	98.65	22.29	49.52	3.40	556.65	6.93

Şekil 8: Türkiye'deki hava merkezlerinden toplanan veriler

2.-)ALGORİTMA ÇALIŞMA SONUCU

Kullanılan ve eğitim amacıyla kullanılan veri seti Türkiye Cumhuriyeti'nde bulunan hava partikül verilerine göre belirlenmiştir.Toplanan veriler ilçelerde bulunan hava kontrol merkezi merkezlerinin paylaştığı verilerdir.İller bazlı ortalama alabilmek için iller içindeki ilçelerdeki veriler toplanıp ortalaması alımıştır.Bazı illerimizde her ilçede bulunamayan hava oranları hava verisi bulunan ilçelerden toplanmıştır.[9]İllerden toplanan veri kaynakları yeterli olmadığı için genel bir veri kümesi oluşturulamamıştır. Bundan dolayı; veri kümemiz Amerika veri kümesiyle bağdaşlaşan 6 adet özellik ile sınırlı kalmıştır.

```
DTR = DecisionTreeRegressor(random_state = 1)
DTRcontrol = model_result(DTR,"DTR",X,y)
```

```
 mape score of DTR :
22.25334537763449
 r2_ score of DTR :
-0.1724344403927376
 mean_squared_error of DTR :
357.727601862631
 mean_absolute_error of DTR :
14.263213038416762
```

Şekil 9: Karar ağacı algoritması

Türkiye Cumhuriyeti üzerinden topladığımız veriler Amerika Birleşik Devletleri üzerindeki veriler ve özellikler dikkate alınarak toplanmıştır[10]

Türkiye veri seti baz alınarak daraltılmış, Amerika veri setimizde uyguladığımız eğitim ve test işlemleri sonucunda elde ettiğimiz metrikleri yukarıda görebilirsiniz. Bu elde ettiğimiz sonuçlar üzerinden random forest algoritmasını tahminleme için kullanmaya karar verdik.Buda random forest algoritmasının bu tarz problemler çözümünde ne kadar etkili olduğunu gösteriyor.[11]

```
LR = LinearRegression()
LRcontrol = model_result(LR,"LR",X,y)
```

```
 mape score of LR :
18.308885511995854
 r2_ score of LR :
0.2681829434354954
 mean_squared_error of LR :
223.288528234718
 mean_absolute_error of LR :
11.575696294933845
```

Şekil 10: Lineer Regnesyon algoritması

Çalışmamız için illerimizden toplanan verilerimiz Amerika veri setinde uygulanan eğitimler ve testler sonucunda toplanan metrikler ile çalışması aşağıda sunulmuştur. Elde edilen sonuçlarımız Random Forest algoritmasını

oluşturduğumuz veri kümesi tahminlemesi içinde kullanılması kararlaştırılmıştır. Yaptığımız çalışma sonucunda sınıflandırma ve karar ağacı problemlerinde Random Forest algoritmasının sonuç alabilmek için etkileri resimlerde gösterilmiştir:

```
SvR = SVR(C = 10.0)
SvRcontrol = model_result(SvR,"SvR",X,y)
```

```
 mape score of SvR :
17.334655104915726
 r2_ score of SvR :
0.3367402436906476
 mean_squared_error of SvR :
202.37065192068147
 mean_absolute_error of SvR :
11.039716320019213
```

Şekil 11: Destek vektör algoritması

Çalışmamızın temelinin alındığı Amerika verilerinden eğitilmiş makine öğrenmesi algoritmaları ve yöntemleriyle Türkiye Cumhuriyeti üzerinden toplanan veriler sonucu yanda gösterilmiştir.Çalışmamızın amacı Amerika'daki çalışmanın sonucundaki 100.000 kişideki ortalama Akciğer kanseri hastalarının belirlenmesi ve ölüm oranlarını belirlemektir. ÇalışmamızTürkiye Cumhuriyeti üzerinde toplanan verilerle tahmin yapmak ve bu tahminleri şehirlerin tehlike durumlarını ölçmek amacıyla gerçekleştirilmiştir.Bu süreç içerisinde Türkiye Cumhuriyeti üzerinden daha fazla sayıda veri toplanması hayatı önem taşımaktadır.Sonuçlarımız iller bazlı olarak 100.000 kişi bazlı sonuçlar vermiştir.

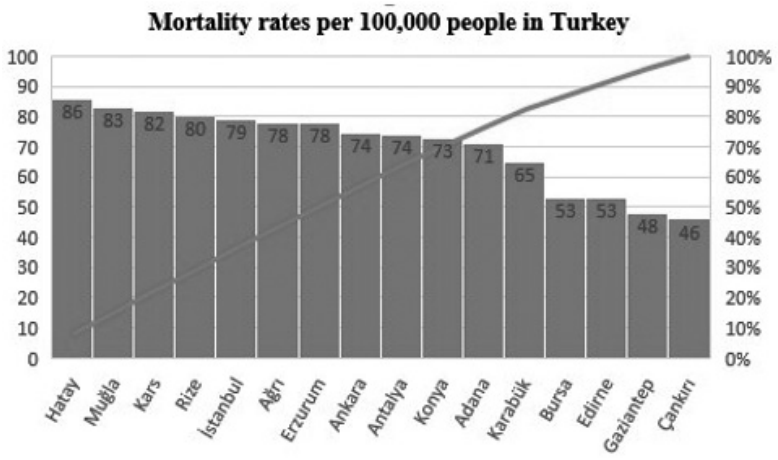
```
Rfr = RandomForestRegressor(max_depth=10, random_state=1)
Rfrcontrol = model_result(Rfr,"Rfr",X,y)
```

```
 mape score of Rfr :
16.72814166494755
 r2_ score of Rfr :
0.3886935100522446
 mean_squared_error of Rfr :
186.51891919758012
 mean_absolute_error of Rfr :
10.560397014218449
```

Şekil 12: Random Forest Algoritması

State	Prediction
Adana	70.933961
Ankara	74.141447
Ağrı	78.675703
Hatay	86.324517
Antalya	74.487641
Gaziantep	48.783261
İstanbul	79.104277
Çankırı	46.246137
Kars	82.034826
Konya	73.407912
Bursa	53.783674
Edirne	53.714318
Muğla	83.821665
Rize	80.460696
Karabük	65.331409
Erzurum	78.162277

Şekil 13: Çalışma sonucunda Türkiye'deki beklenen Akciğer kanser sayıları



Şekil 14: Tahmin edilen Akciğer kanseri sayılarının tablosu

Projemizin amacı hava kirliliğinin sebep olduğu kanser artış ordanlarının tahminlenmesidir. Projemizde amacımız Türkiye'de bulunan 6 adet gaz ölçümü verisinin kanser oranlarını nasıl etkilediğini tahminlemektir.Akiğer kanserinde araştırılan 20 adet gaz içerisinden Türkiye'den (O3,PM2.5,PM10,SO2,CO,NO2) 4 adet gaz verisi ve 2 adet hava partikülü verisi alınabilmiştir.Bunun nedeni Türkiye üzerinde hava kalite oranı amacıyla 6 adet özellik dışında başka bir bileşik araştırılması yapılmamıştır.Çalışmamız 6 adet özelliğin etkilerini ölçmek ve ölçümlerin sonucunu tahminlemek amacıyla yapılmıştır.

SONUÇ

Çalışmamız sonucunda belirtilen iller içerisinde en yüksek kanser oranı beklentisi Hatay ili olduğu görüntülenmiştir. Makine öğrenmesi modelimizde gazları kıyasladığımız zaman değerlerin birbirlerine yakın olduğu görüntülenmektedir fakat ülkemizde rüzgarlarla beraber gelen kum fırtınaları haritalarında güney şehirlerimizin bu kumlardan daha çok etkilendiği söylenebilir. Kum fırtınaları PM2.5 ve PM10 değerlerini doğrudan etkiler[12].

Sonuç olarak projemizde bileşiklerin ve hava partiküllerimizin etkilerini algoritmalarımızla ve derin öğrenme tekniklerimizle makinemiz üzerinde eğitimimizi tamamlanmıştır. Yapılan eğitim sonucunda Oluşturduğumuz makine öğrenmesi ve derin öğrenme modeline uyguladığımız modeller üzerinden 2023-2024 senesinden itibaren hava kirliliğinden dolayı,akciğer kanseri ölüm sayıları tahminleri şekildeki gibi çıkmıştır. Tahminleme modelimiz daha iyi sonuçlar verebilmesi için daha fazla veri elde edilmesi gerekmektedir.Veri kümemizdeki 6 adet özellikle eğitilmiş makine öğrenmesi sonucu şekildeki gibidir

elde ettiğimiz çıktıyla beraber iller bazlı ölüm oranları tahminlenmesi yapılmıştır. Çalışmamız Amerika verilerinin daralttırılarak Türkiye Cumhuriyeti'nde bulunabilen hava değerlerine indirgenmeye çalışılmıştır.Verilerin indirgenmesi sonucunda Amerika verilerinde EQI, AAC gibi kanser oranlarında doğrudan etkili olan sonuçların Türkiye verileri iller bazlı bulunamadığından iller bazlı çalışma sonuçları, tahminlemesi sadece 6 adet veri üzerinden olmuştur.

Yapılan çalışma sadece Amerika çalışmasının Türkiye üzerindeki gazlarla kıyaslanarak değerlendirme amaçlı yapılmıştır.Yapılan çalışma sadece bir tahminleme modelidir ve çalışma sonucunu psikolojik genetik ve toplumsal yaşam alanlarını etkileyebilmektedir, Bundan dolayı çalışmamız gazların kanser oranını arttırabilecek ve akciğerler üzerindeki etkilerinin nasıl sonuçlar verebileceği üzerine yapılmış bir tahminleme çalışmasıdır.

Yapılan çalışma sonucunda EQI, AAC gibi hava kalitesinin ölçümünde doğrudan sonuç veren değerlerin ülkemizde tespiti gerekmektedir.

KAYNAKLAR

1. (Kaynak Yıl:2019 Yayın adı: Akciğer kanseri)(İnternetten erişim yılı 2022)(https://hsgm.saglik.gov.tr/tr/kanser-il-faaliyetleri.html?view=category&id=379&start=25
2. https://dergipark.org.tr/tr/download/article-file/270660 J CONTEMP MED 2016;6 (Case Reports): 131-137 REVIEW DOI: 10.16899/ctd. 80586
3. (Çevre, Şehircilik ve İklim Değişikliği Bakanlığı | Ulusal Hava Kalite İzleme Ağı) (İnternet erişim tarihi 2022) https://www.havaizleme.gov.tr/
4. MEVCUT HAVA KALİTESİ Mevcut Kirletici Maddeler (Erişim tarihi 2022) https://www.accuweather.com/
5. https://ci5.iarc.fr/CI5I-X/Pages/Table10q_sel.aspx (INDICES OF DATA QUALITY(-volX))[5]
6. https://www.researchgate.net/publication/321695891_Air_quality_and_lung_cancer_Analysis_via_Local_Control
7. Air quality and lung cancer: Analysis via Local Control December 2017 Conference: Joint Statistical Meetings(erişim yılı 2022) https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
8. (Yılı 2017Makine öğrenmesi modelleri ve yöntemleri) (Erişi yılı 2022)https://veribilimcisi.com/2017/07/14/mse-rmse-mae-mape-metrikleri-nedir/
9. Adilov, G. , Tınaztepe, G. & Kemali, S. (2011). Genelleştirilmiş Ortalama Fonksiyonu ve Bazı Önemli Eşitsizliklerin Öğretimi Üzerine . Mersin Üniversitesi Eğitim Fakültesi Dergisi , 5 (2) , 294-300 . Retrieved fromhttps://dergipark.org.tr/tr/pub/mersinefd/issue/17374/181436
10. Korkem, Ebru. "Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest Ve Naıve Bayes Sınıflama Yöntemleri Yaklaşım." (2013). MİKROARRAY GEN EKSPRESYON VERİ SETLERİNDE RANDOM FOREST VE NAIVE BAYES SINIFLAMA YÖNTEMLERİ YAKLAŞIMhttp://www.openaccess.hacettepe.edu.tr:8080/xmlui/bitstream/handle/11655/997/5b726bfb-45c9-41f6-a7a0-5c82c0ab86db.pdf?sequence=1
11. https://www.havaizleme.gov.tr/ (Çevre, Şehircilik ve İklim Değişikliği Bakanlığı | Ulusal Hava Kalite İzleme Ağı)(İnternet erişim tarihi 2022)
12. (WMO Sand and Dust Storm – Warning Advisory and Assessment System (SDS-WAS))