

MANİSA
CELAL BAYAR
ÜNİVERSİTESİ

Hava Kalitesi Kontrol Parametreleri
İlişkileri

Air Quality Control Parameters
Relationships

Öznur PAYLAN

Hava Kalitesi Kontrol Parametreleri İlişkileri

Air Quality Control Parameters Relationships

Öznur PAYLAN Yazılım Mühendisliği, Manisa Celal Bayar Üniversitesi

Anahtar Kelimeler

Birliktelik Analizi

Hava

Hava Bileşenleri

Weka

Parametreler

Veri Madenciliği

Özet İncelenen konuda hava kalitesi parametreleri sınıflandırılmaktadır. 5 temel hava kalitesini belirleyen parametreler kullanılmıştır. Bunlar partikül maddeler (PM_{10}), karbon monoksit (CO), kükürt dioksit (SO_2), azot dioksit (NO_2), ozon (O_3). Hava değerini oluşturan parametreler arası ilişkiler ortaya konulmaktadır. Kullanılan verilerde eksik veri, tutarsız veri, analiz için gerekli olmayan verilerin incelenmesi yapılmıştır. Bu bağlamda veri seti oluşturulmuştur. Çalışma hava parametrelerine sınıflandırma yapmayı amaçlar. Parametreler arasında ki ilişkiler ile sınıflandırma yapılmıştır. Python ile veriler arasındaki istatistiksel oranlar hesaplanmıştır. Weka programına göre en iyi 10 sınıflandırma değeri gösterilmiştir. En iyi sonuçlara göre değerlendirme yapılmıştır.

1. Giriş

İnsan yaşamının getirdiği yoğun şehirleşme birçok açıdan kirliliğe sebep olmaktadır. Bunlardan biri de hava kirliliğidir. Hava kirliliği küresel ölçekte bir etki alanına sahiptir. Hava kirliliğini çözmek için atmosferik kirleticiler konsantrasyonlarını izlemek ve analiz etmek konusuna odaklanılması gerekmektedir. Bu konu halk sağlığını doğrudan etkiler. Bu nedenle bu makalede hava kalitesi parametreleri arasındaki ilişkilerin ortaya konulması amaçlanmıştır. Hava kirliliğini/ kalitesini daha iyi anlayabilmemiz için bir tür sınıflandırma kullanılmaktadır. Hava Kalitesi İndeksi (HKİ) sınıflandırma sistemi ile havadaki kirleticilerin konsantrasyonuna göre hava kalitesi iyi, orta, kötü vb. şekilde derecelendirilmiştir. 5 temel kirleticiler için hava kalitesi indeksi hesaplanmaktadır. Bunlar:

- Partikül maddeler (PM_{10})
- Karbon monoksit(CO)
- Kükürt dioksit (SO_2)
- Azot dioksit(NO_2)
- Ozon (O_3)

Partikül maddeler katı ve sıvı damlacıkların karışımından oluşur. Partikül maddeler sanayi, yakıt yanması, tarım ve ikincil kimyasal reaksiyonlar sebebi ile ortaya çıkar. Kanser, solunum hastalıkları, kalp problemleri gibi problemler ortaya çıkarır. Karbon monoksit gaz, yağ, tahta, kömür içeren karbon bazlı maddelerin eksik yanması sonucu oluşur. Kükürt dioksit fosil yakıt yanması sonucu oluşur. Solunum yolu hastalıklarına neden olur. Azot

dioksit fosil yakıtların, yeni gaz, kömür ve yağların yüksek yanması sonucu meydana çıkar. Asit yağmurlarına neden olur. Trafikte oluşan azot oksitler ve uçucu organik bileşenlerin güneş ışığıyla değişimi ile ozon oluşur. Hava kalite parametrelerin insan üzerinde çeşitli zararlı etkileri vardır. [1]

Literatüde yapılmış örneklerine bakılarak asıl yapmak istenen partiküller arasındaki ilişkileri ortaya koymaktır.

Tablo 1: EPA Hava Kalitesi İndeksi. [1]

Hava Kalitesi İndeksi (IAQ) Değerler	Sağlık Endişe Seviyeleri	Renkler	Anlamı
<i>Hava Kalitesi İndeksi bu aralıkta olduğunda..</i>	<i>..hava kalitesi koşulları..</i>	<i>..bu renkler ile sembolize edilir..</i>	<i>..ve renkler bu anlama gelir.</i>
0 – 50	İyi	Yeşil	Hava kalitesi memnun edici ve hava kirliliği az riskli veya hiç risk teşkil etmiyor.
51 – 100	Orta	Sarı	Hava kalitesi uygun fakat alışılmadık şekilde hava kirliliğine hassas olan çok az sayıda insan için bazı kirlleticiler açısından orta düzeyde sağlık endişesi oluşabilir.
101- 150	Hassas	Turuncu	Hassas gruplar için sağlık etkileri oluşabilir. Genel olarak kamunun etkilenmesi olası değildir.
151 – 200	Sağlıksız	Kırmızı	Herkes sağlık etkileri yaşamaya başlayabilir, hassas gruplar için ciddi sağlık etkileri söz konusu olabilir.
201 – 300	Kötü	Mor	Sağlık açısından acil durum oluşturabilir. Nüfusun tamamının etkilenme olasılığı yüksektir.
301 – 500	Tehlikeli	Kahverengi	Sağlık alarmı: Herkes daha ciddi sağlık etkileri ile karşılaşabilir.

2. Verilerin Oluşturulması

Hava değer parametrelerini T.C Çevre, Şehircilik ve İklim Değişikliği Bakanlığının sunmuş olduğu Sürekli İzleme Merkezi'nden (SİM) alınmıştır [2]. Veriler illere göre her il için 10 günlük veri çekilmiştir. Her il için 5 tane hava değer parametresi dikkate alınarak yapılmıştır. Veriler her il için ayrı olarak indirilmiştir. Daha sonra veri seti oluşturma aşamasında veriler belirli kriterlere göre düzenlenmiştir. İlk olarak veriler birleştirilmiştir. Daha sonra Excel formatındaki veriler için makine öğrenmesi (ML) algoritmalarının kullanabileceği düzenlemeler yapılmıştır. Eldeki veriler için veri ön işleme yöntemleri uygulanmıştır.

2.1. Kullanılan Veri Ön İşleme Yöntemleri [3] 1-Eksik veri: Veri seti içerisindeki eksik olan verilerin doldurulması gerekmektedir. Bu işlem yapılmazsa hatalı sonuçlar elde etmemize neden olur. Eksik veri doldurmak için farklı yöntemler kullanılmakta. Global bir değişken kullanmak, ortalama kullanmak, özniteliği en fazla olan veriyi kullanmak, regresyon ile tahmin ederek boş verileri doldurmak gibi farklı yöntemleri vardır. Bu veri setinde boş değerler ortalama bir değer ile doldurulmuştur.

D15									
	A	B	C	D	E	F	G	H	I
1	SEHIRLER	TARİH	SO2	NO2	CO	O3	PM10	HAVA KALİTESİ	
2	ADANA	01.11.2022	3.04	11.14	438.95	64.05	47.99	iyi	
3	ADANA	02.11.2022	1.93	9.10	383.02	58.09	33.61	iyi	
4	ADANA	03.11.2022	2.36	7.92	452.69	66.01	29.79	iyi	
5	ADANA	04.11.2022	4.41	8.27	488.94	68.00	27.41	iyi	
6	ADANA	05.11.2022	3.87	8.31	525.01	62.61	30.88	iyi	
7	ADANA	06.11.2022	3.14	7.32	488.14	66.41	26.60	iyi	
8	ADANA	07.11.2022	2.52	4.70	391.79	54.40	19.84	iyi	
9	ADANA	08.11.2022	2.44	5.98	367.83	47.89	12.58	iyi	
10	ADANA	09.11.2022	2.69	5.74	387.41	43.80	17.26	iyi	
11	ADANA	10.11.2022	7.96	9.53	567.37	39.46	20.03	iyi	
12	AFYON	01.11.2022	8.93	62.69	601.27	29.30	33.81	iyi	
13	AFYON	02.11.2022	8.98	115.77	1032.12	23.40	37.93	iyi	
14	AFYON	03.11.2022	12.58	110.03	1282.73	18.67	66.78	iyi	
15	AFYON	04.11.2022	77.47	826.47	14.62	49.32	iyi		
16	AFYON	05.11.2022	18.85	124.63	1289.80	16.30	73.89	kotu	

Şekil 1: Excel formülü ile ortalama değeri atılması.

2-Ayrık/Uç verileri tespit etme: Veri setini bozan normal değerin çok üstündeki verilerin tespiti ve veri setinde olması istenmeyen verilerin kaldırılması.

3-Yeni öznitelik oluşturma: Mevcut öznitelikler kullanarak yeni öznitelik oluşturmaktır. Hava parametreleri kullanarak havanın iyi veya kötü olması belirlenmiştir. Değerler ulusal hava kalitesi indeksine göre baz alınmıştır.

Tablo 2: Ulusal Hava Kalitesi İndeksi Kesme Noktaları. [1]

İndeks	HKİ	SO ₂ [µg/m ³]	NO ₂ [µg/m ³]	CO [µg/m ³]	O ₃ [µg/m ³]	PM10 [µg/m ³]
		1 Sa. Ort.	1 Sa. Ort.	8 Sa. Ort.	8 Sa. Ort.	24 Sa. Ort.
İyi	0 - 50	0-100	0-100	0-5500	0-120	0-50
Orta	51 - 100	101-250	101-200	5501-10000	121-160	51-100
Hassas	101 - 150	251-500	201-500	10001-16000	161-180	101-260
Sağlıksız	151 - 200	501-850	501-1000	16001-24000	181-240	261-400
Kötü	201 - 300	851-1100	1001-2000	24001-32000	241-700	401-520
Tehlikeli	301 - 500	>1101	>2001	>32001	>701	>521

Tablo 3: İndeks Hesaplama Parametrelerinin Sınır Değerleri. [1]

Parametre	SO ₂ [µg/m ³]	NO ₂ [µg/m ³]	CO [µg/m ³]	O ₃ [µg/m ³]	PM10 [µg/m ³]
	1 Sa. Ort.	1 Sa. Ort.	8 Sa. Ort.	8 Sa. Ort.	24 Sa. Ort.
Ulusal Sınır Değer	410	270	10.000	120	70
AB Üye Ülkeleri Sınır Değeri	350	200	10.000	120	50

H2	=EĞER(VE(C2>0;C2<410;D2>0;D2<270;E2>0;E2<10000;F2>0;F2<120;G2>0;G2<70);"iyi";"kotu")								
	A	B	C	D	E	F	G	H	I
1	SEHIRLER	TARİH	SO2	NO2	CO	O3	PM10	HAVA KALİTESİ	
2	ADANA	01.11.2022	3.04	11.14	438.95	64.05	47.99	iyi	
3	ADANA	02.11.2022	1.93	9.10	383.02	58.09	33.61	iyi	
4	ADANA	03.11.2022	2.36	7.92	452.69	66.01	29.79	iyi	
5	ADANA	04.11.2022	4.41	8.27	488.94	68.00	27.41	iyi	
6	ADANA	05.11.2022	3.87	8.31	525.01	62.61	30.88	iyi	
7	ADANA	06.11.2022	3.14	7.32	488.14	66.41	26.60	iyi	
8	ADANA	07.11.2022	2.52	4.70	391.79	54.40	19.84	iyi	
9	ADANA	08.11.2022	2.44	5.98	367.83	47.89	12.58	iyi	
10	ADANA	09.11.2022	2.69	5.74	387.41	43.80	17.26	iyi	
11	ADANA	10.11.2022	7.96	9.53	567.37	39.46	20.03	iyi	

Şekil 2: Excel formülü ile hava değeri oluşturma.

Verilerin alınması ve düzenlenmesinden sonra kullanılabilmesi için veriler birleştirilerek csv uzantılı dosya haline getirilmiştir.

	A	B	C	D	E	F	G	H	I	
1	SEHIRLER,TARİH,SO2,NO2,CO,O3,PM10,HAVA KALİTESİ									
2	ADANA,44866.0006481481,3.04,11.14,438.95,64.05,47.99,iyi									
3	ADANA,44867.0006481481,1.93,9.1,383.02,58.09,33.61,iyi									
4	ADANA,44868.0006481481,2.36,7.92,452.69,66.01,29.79,iyi									
5	ADANA,44869.0006481481,4.41,8.27,488.94,68,27.41,iyi									
6	ADANA,44870.0006481481,3.87,8.31,525.01,62.61,30.88,iyi									
7	ADANA,44871.0006481481,3.14,7.32,488.14,66.41,26.6,iyi									
8	ADANA,44872.0006481481,2.52,4.7,391.79,54.4,19.84,iyi									
9	ADANA,44873.0006481481,2.44,5.98,367.83,47.89,12.58,iyi									
10	ADANA,44874.0006481481,2.69,5.74,387.41,43.8,17.26,iyi									
11	ADANA,44875.0006481481,7.96,9.53,567.37,39.46,20.03,iyi									
12	AFYON,44866.0006481481,8.93,62.69,601.27,29.3,33.81,iyi									
13	AFYON,44867.0006481481,8.98,115.77,1032.12,23.4,37.93,iyi									
14	AFYON,44868.0006481481,12.58,110.03,1282.73,18.67,66.78,iyi									
15	AFYON,44869.0006481481,10.5566666666667,77.47,826.465555555556,14.62,49.3244444444445,iyi									
16	AFYON,44870.0006481481,18.85,124.63,1289.8,16.3,73.89,kotu									
17	AFYON,44871.0006481481,18.71,67.26,817.01,21.18,69.19,iyi									
18	AFYON,44872.0006481481,9.35,54.21,627.24,21.81,43.88,iyi									
19	AFYON,44873.0006481481,5.03,23.27,317.46,34.58,22.99,iyi									
20	AFYON,44874.0006481481,5.32,59.59,620.62,18.82,46.48,iyi									
21	AFYON,44875.0006481481,7.26,79.78,849.94,18.35,48.97,iyi									
22	AGRI,44866.0006481481,34.86,43.27,1354.19,39.08,119.4,kotu									
23	AGRI,44867.0006481481,39.4,40.2,1227.23,31.01,128.77,kotu									
24	AGRI,44868.0006481481,39.52,40.21,1207.09,31.76,131.66,kotu									
25	AGRI,44869.0006481481,47.92,42.36,1503.85,28.35,149.15,kotu									
26	AGRI,44870.0006481481,28.39,33.89,1026.75,41.68,131.66,kotu									
27	AGRI,44871.0006481481,48.28,39.92,1568.41,33.45,131.66,kotu									
28	AGRI,44872.0006481481,52.49,41.66,1443.5,33.54,140.03,kotu									

Şekil 3: Uygun hale gelen verileri bir kısmı.

Oluşturulan veri setinin parametreler arasındaki ortalama değeri, mod, medyan vb. değerlerin hesaplamaları yapılmıştır. Python kullanarak JupiterNotebook’da oluşturulmuştur. Öncelikle gerekli kütüphaneler yüklenmiştir. Daha sonra veri seti yüklendi. isnull() kullanılarak boş değer kontrolü yapıldı. describe() komutu ile istatistiksel veriler hesaplandı.

	SO2	NO2	CO	O3	PM10
count	499.000000	499.000000	499.000000	499.000000	499.000000
mean	14.846473	47.818589	1074.389644	25.761892	58.068628
std	12.269851	33.426670	659.961862	13.929503	39.265260
min	0.360000	2.080000	165.890000	5.060000	1.170000
25%	6.860000	27.635000	588.635000	16.120000	32.050000
50%	10.370000	40.000000	913.790000	22.930000	48.940000
75%	19.705000	58.455833	1419.460000	32.890000	70.660000
max	98.560000	200.500000	4018.140000	109.340000	223.540000

Sekil 4: İstatiksel veriler.

3. Önerilen Yöntemler

Makalede kullanılan yöntem ve tekniklere bakacak olursak adım adım şu şekildedir;

3.1 Veri Bulma ve Veri Ön İşleme

Veriler SİM'den alındıktan sonra veriler için veri ön işleme yapıldı. Excel ile veriler düzenlendi ve veri ön işleme adımları uygulandı. Veri setindeki boş değerler için Excel ile ortalama değer atıldı.

$$= ORTALAMA(D1; D13; D14; D16; D17; D18; D19; D20; D21) \quad (1)$$

Buradaki formülde belirtilen yer için değerleri oluşturan sütunun ortalama değeri atanmıştır. Boş değerler Python isnull() komutu ile kontrol edilmiştir. Buradan çıkan sonuçlarla JupiterNotebook kullanılarak Python kodu ile veriler arasında istatiksel hesaplamalar yapılmıştır.

```
df = pd.read_csv('veri.csv', encoding = 'unicode_escape')
```

Burada veri adındaki csv dosyası kullanmak için df adındaki değişkene atanmıştır.

```
df.describe()
```

Veri seti ile ilgili birkaç işlem yapıldıktan sonra df'deki veri setine describe komutu ile istatiksel veriler hesaplandırılıyor.

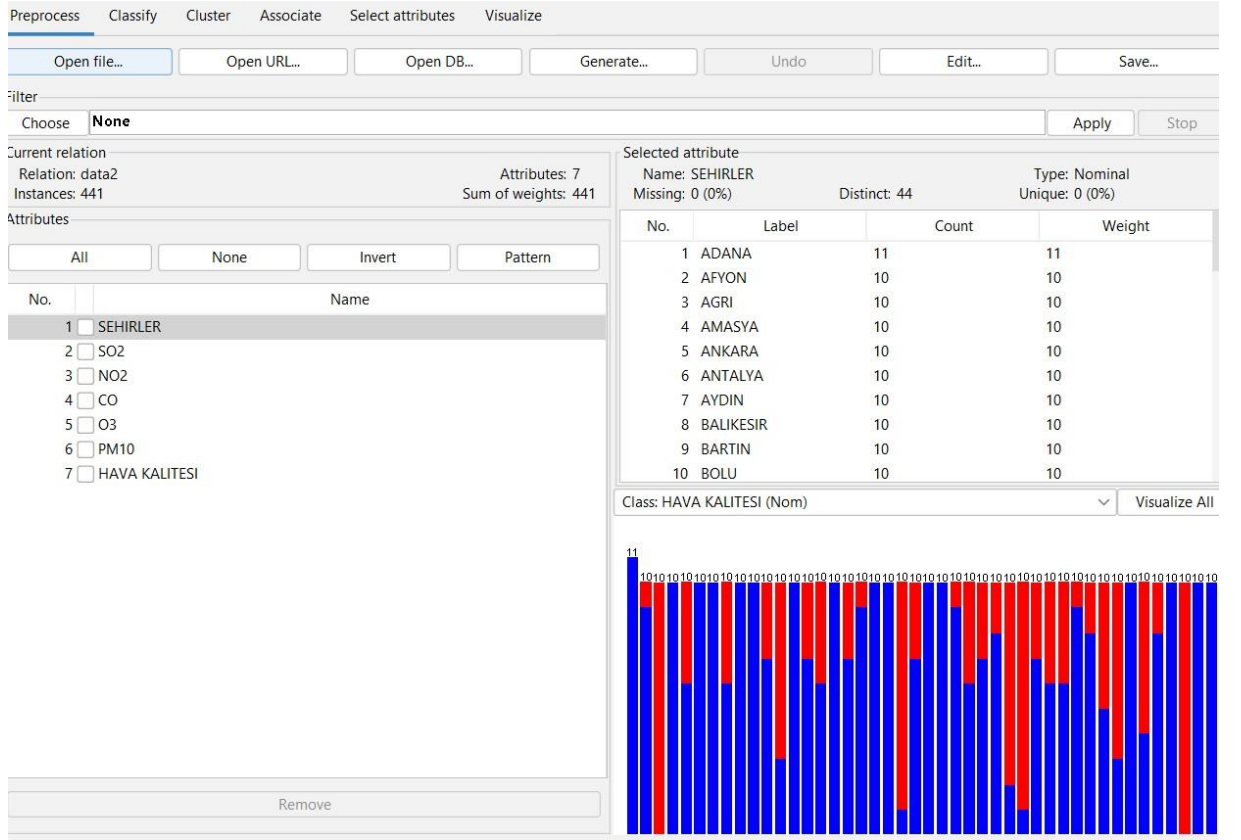
3.2 Weka ile Verilerin Sınıflandırılması

Weka temelde makine öğrenmesi algoritmalarını ve veri ön işleme gibi gereksinimleri bir arada sunan bir veri madenciliği programıdır. Wakiato Üniversitesi tarafından geliştirilmiştir. Java dili üzerinde geliştirilmiştir. Weka yazılımı dosya uzantısı olarak arff kullanır [4]. Verileri basit bir dosyadan okur. Veritabanı üzerinden de veri çekebilir. Weka ile pek çok kütüphane hazır olarak gelmektedir.

Weka ile veri ön işleme (data preprocessing), sayısal analiz (regression), sınıflandırma (classification), gruplandırma (clustering) gibi veri ile ilgili pek çok işlem yapılabilmektedir. Bunun yanında bu işlemler doğrultusunda çıkan sonuçların görsel olarak gösterilmesini sağlayan görüntü araçları bulunmaktadır.

Veri ön işleme aşamasından sonra Weka ile ilgili işlemlerimizde Weka'nın Explorer uygulaması altında ki sınıflandırma sekmesini kullanıyor olacağız.

Veri setimizle ilgili öncelikle yapmak istediğimiz adımlar csv uzantılı veri setimizi Weka'nın daha kolay uygulayabileceği arff uzantılı dosya haline getirmektir. Weka bu işlemi kendi içerisinde yapılmasına olanak sağlar. Kullanılacak veri seti programa yüklendikten sonra düzenleme yapılarak istenilen formata getirilir.



Şekil 5: Veri seti içe aktarıldı.

Generate...	Undo	Edit...	Save...
-------------	------	---------	---------

Viewer							
Relation: data2							
No.	1: SEHIRLER Nominal	2: SO2 Numeric	3: NO2 Numeric	4: CO Numeric	5: O3 Numeric	6: PM10 Numeric	7: HAVA KALITESI Nominal
1	ADANA	3.04	11.14	438.95	64.05	47.99	iyi
2	ADANA	1.93	9.1	383.02	58.09	33.61	iyi
3	ADANA	2.36	7.92	452.69	66.01	29.79	iyi
4	ADANA	4.41	8.27	488.94	68.0	27.41	iyi
5	ADANA	3.87	8.31	525.01	62.61	30.88	iyi
6	ADANA	3.14	7.32	488.14	66.41	26.6	iyi
7	ADANA	2.52	4.7	391.79	54.4	19.84	iyi
8	ADANA	2.44	5.98	367.83	47.89	12.58	iyi
9	ADANA	2.69	5.74	387.41	43.8	17.26	iyi
10	ADANA	7.96	9.53	567.37	39.46	20.03	iyi
11	AFYON	8.93	62.69	601.27	29.3	33.81	iyi
12	AFYON	8.98	115.77	1032.12	23.4	37.93	iyi
13	AFYON	12.58	110.03	1282.73	18.67	66.78	iyi
14	AFYON	10.556...	77.47	826.46...	14.62	49.3244...	iyi
15	AFYON	18.85	124.63	1289.8	16.3	73.89	kotu
16	AFYON	18.71	67.26	817.01	21.18	69.19	iyi
17	AFYON	9.35	54.21	627.24	21.81	43.88	iyi
18	AFYON	5.03	23.27	317.46	34.58	22.99	iyi
19	AFYON	5.32	59.59	620.62	18.82	46.48	iyi
20	AFYON	7.26	79.78	849.94	18.35	48.97	iyi
21	AGRI	34.86	43.27	1354.19	39.08	119.4	kotu
22	AGRI	39.4	40.2	1227.23	31.01	128.77	kotu
23	AGRI	39.52	40.21	1207.09	31.76	131.66	kotu
24	AGRI	47.92	42.36	1503.85	28.35	149.15	kotu

Şekil 6: arff' ye çevrilen dosya.

Veri ile ilgili tüm işlemler bittikten sonra artık verimizi sınıflandırmamız gerekmektedir. Ayrıca Weka parametreler arasındaki ilişkileri de görsel aracı ile bize sunmaktadır.



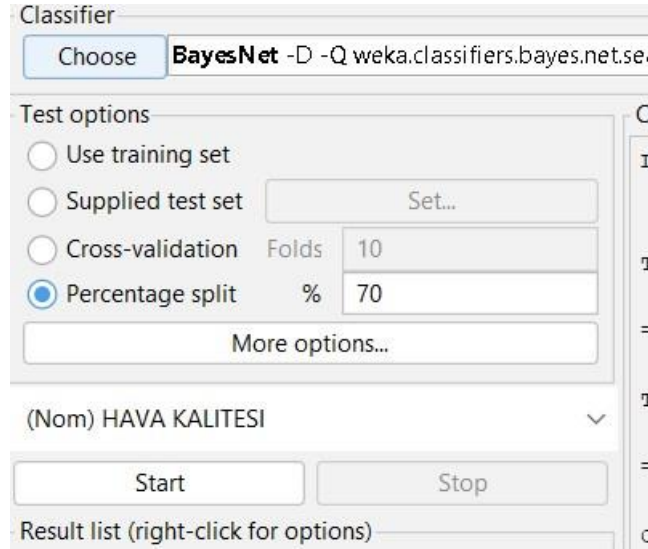
Şekil 7: Weka ile parametreler arası ilişkiler.

Burada Şekil 7 'ya bakacak olursak bazı verilerin bir doğru etrafında yayılım gösterdiğini, bazı verilerin ise doğrudan çok uzak olduğunu görürüz. Bu verilerin dağılımının doğru etrafında yayılan verilere göre normal olmadığını göstermektedir. Bunun en büyük nedeni ise her şehirde oluşacak hava değerlerinin birbirinden farklı oluşacağı içindir.

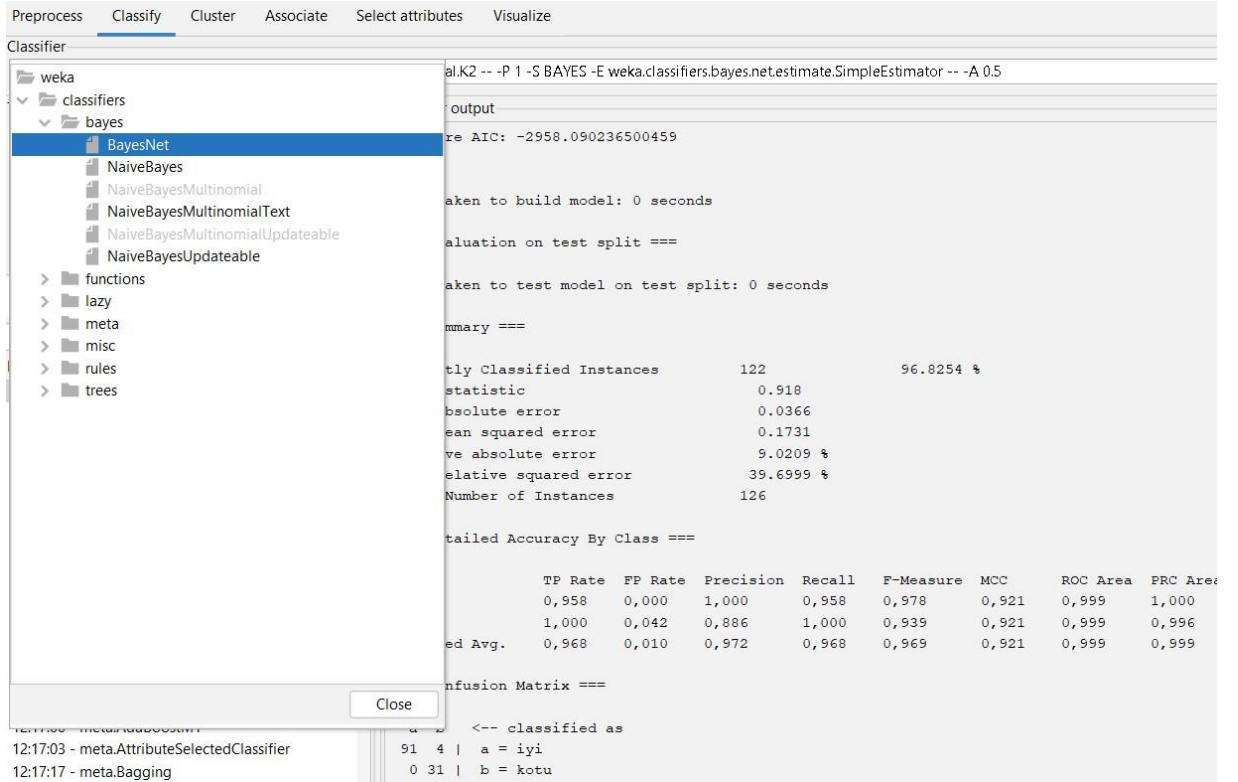
4. Deney ve Tartışma

Bu bağlamda hazırlanan veri seti ile ilgili sınıflandırma işlemi yapılacaktır ve en iyi 10 sınıflandırma sonucu gösterilecektir.

Verilerimizi %70 eğitim verisi (train) %30 test olarak ayırdık. Bu oranlar ML algoritmalarında genelde bu şekilde alındığı için %70'e %30 olarak ayrılmıştır. Burada program sayesinde train verisi ile bağımlı-bağımsız veriler arasındaki ilişki öğretilir, test ile ne kadar öğrendiği sorgulanır. Bu oran kullanılan her sınıflandırma algoritması için aynı olarak seçilmiştir. Şekil 8-9 da gösterilmiştir.



Şekil 8: Verilerin ayrılması.



Şekil 9: Uygulanacak sınıflandırma seçimi.

Sınıflandırma işlemi ile birlikte en iyi sonuçları öğrenmek için karmaşıklık matrislerine de bakılır. Karmaşıklık matrisi ML de kullanılan sınıflandırma modellerinin performansını değerlendirmek gerçek değer ve tahmin verilerini karşılaştırmak için kullanılır. Sınıflandırma tahminleri 4 temel değerlendirmeden birine sahip olacaktır. Bunlar:

- Doğruya doğru (True Pozitive – TP)
- Yanlış yanlış (True Negative – TN)
- Doğruya yanlış (False Pozitive – FP)

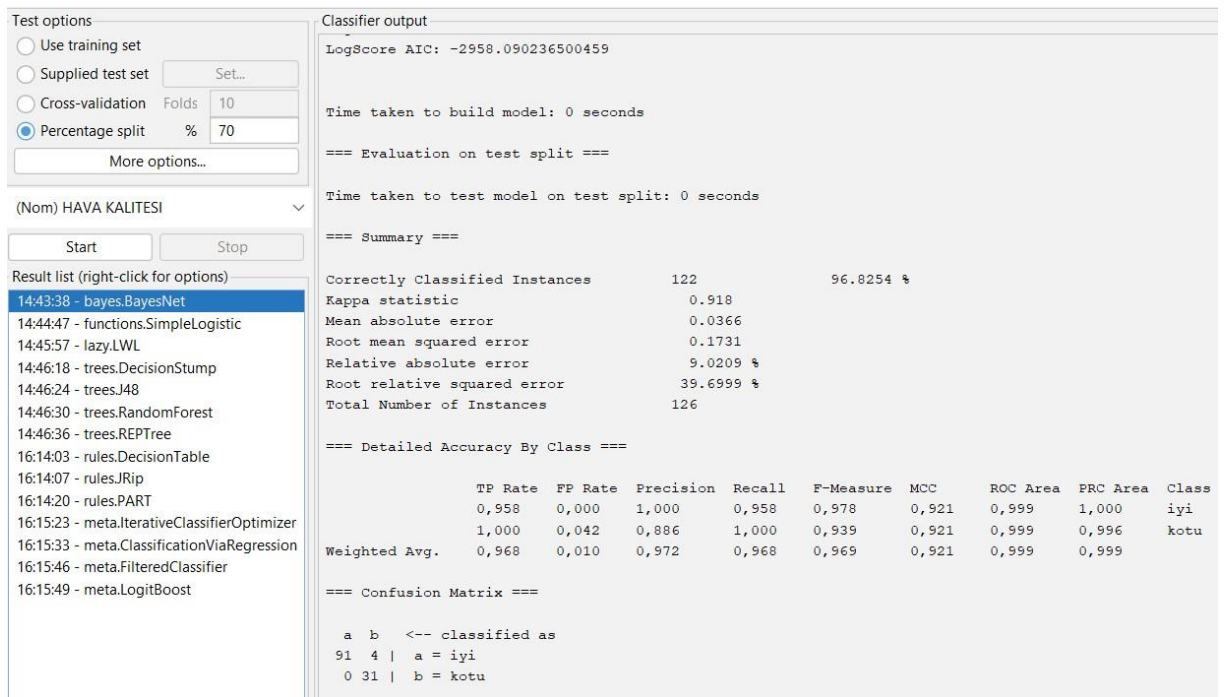
- Yanlış doğru (False Negative – FN)

4.1 Kullanılan Sınıflandırma Algoritmaları

Sınıflandırma aşamasında veriler ile ilgili en iyi sonucu veren 10 sınıflandırma algoritması verilecektir. En iyi sonucu veren algoritmalar seçilirken karmaşıklık matrislerine ve doğru sınıflandırma oranına bakılmıştır.

4.1.1 BayesNet Sınıflandırıcı

Bayes ağları, meydana gelen bir olayı anlamak ve bilinen birkaç olası nedenden herhangi birinin katkıda bulunan faktör olma olasılığını tahmin etmek için kullanılır.



Şekil 10: BayesNet sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı % 96,8254 olarak bulunmuştur.

4.1.2 SimpleLogistic Sınıflandırıcı

Doğrusal lojistik regresyon modelleri oluşturmak için kullanılan sınıflandırıcıdır. Başka bir değişken kullanarak tek bir ikili değişkeni tahmin etmek için kullanılan istatistiksel bir testtir. İkili değişken arasındaki sayısal ilişkiyi belirlemek için kullanılır.

Test options

☐ Use training set

☐ Supplied test set

☐ Cross-validation Folds 10

☒ Percentage split % 70

(Nom) HAVA KALITESI

Result list (right-click for options)

- 14:43:38 - bayes.BayesNet
- 14:45:57 - lazy.LWL
- 14:46:24 - trees.J48
- 14:46:30 - trees.RandomForest
- 14:46:36 - trees.REPTree
- 16:14:03 - rules.DecisionTable
- 16:14:07 - rules.JRip
- 16:15:23 - meta.IterativeClassifierOptimizer
- 16:15:33 - meta.ClassificationViaRegression
- 16:15:49 - meta.LogitBoost
- 17:47:32 - functions.SimpleLogistic

Classifier output

-101.78 +
[PM10] * 1.46

Time taken to build model: 0.03 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances	125	99.2063 %
Kappa statistic	0.9788	
Mean absolute error	0.0185	
Root mean squared error	0.0905	
Relative absolute error	4.5723 %	
Root relative squared error	20.7504 %	
Total Number of Instances	126	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,989	0,000	1,000	0,989	0,995	0,979	1,000	1,000	iyi
	1,000	0,011	0,969	1,000	0,984	0,979	1,000	1,000	kotu
Weighted Avg.	0,992	0,003	0,992	0,992	0,992	0,979	1,000	1,000	

=== Confusion Matrix ===

a b <-- classified as

94 1 | a = iyi

0 31 | b = kotu

Şekil 11: SimpleLogistic sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı % 99,2063 olarak bulunmuştur.

4.1.3 Locally Weighted Learning (LWL) Sınıflandırıcı

LWL, tüm işlev alanı için küresel bir model oluşturmak yerine, komşu verilere dayalı her ilgi noktası için yerel bir model oluşturan bir fonksiyon yaklaşımıdır. LWL yöntemleri parametrik değildir.

Choose **SimpleLogistic** -I 0 -M 500 -H 50 -W 0.0

Test options

☐ Use training set

☐ Supplied test set

☐ Cross-validation Folds 10

☒ Percentage split % 70

(Nom) HAVA KALITESI

Result list (right-click for options)

- 14:43:38 - bayes.BayesNet
- 14:45:57 - lazy.LWL
- 14:46:24 - trees.J48
- 14:46:30 - trees.RandomForest
- 14:46:36 - trees.REPTree
- 16:14:03 - rules.DecisionTable
- 16:14:07 - rules.JRip
- 16:15:23 - meta.IterativeClassifierOptimizer
- 16:15:49 - meta.LogitBoost

Classifier output

Using linear weighting kernels

Using all neighbours

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.04 seconds

=== Summary ===

Correctly Classified Instances	125	99.2063 %
Kappa statistic	0.9788	
Mean absolute error	0.0079	
Root mean squared error	0.0891	
Relative absolute error	1.9566 %	
Root relative squared error	20.4302 %	
Total Number of Instances	126	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,989	0,000	1,000	0,989	0,995	0,979	0,996	0,999	iyi
	1,000	0,011	0,969	1,000	0,984	0,979	0,996	0,982	kotu
Weighted Avg.	0,992	0,003	0,992	0,992	0,992	0,979	0,996	0,995	

=== Confusion Matrix ===

```

a b  <-- classified as
94  1 | a = iyi
  0 31 | b = kotu

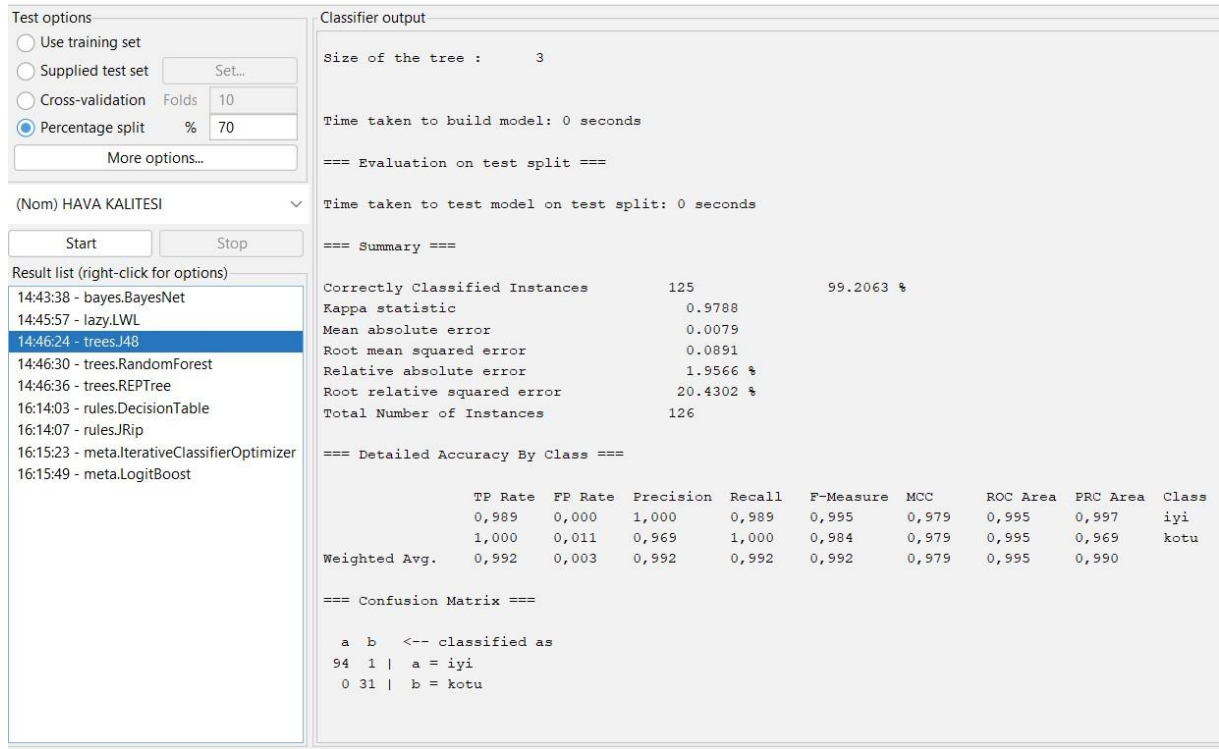
```

Şekil 12: LWL sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı % 99,2063 olarak bulunmuştur.

4.1.4 J48 Sınıflandırıcı

Ross Quinlan'ın Weka'da Java'yı temsil eden J48, J olarak da bilinen önceki ID3 algoritmasının bir uzantısıdır. C4.5 tarafından oluşturulan karar ağaçları, sınıflandırma için kullanılır. Bu nedenle C4.5 genellikle istatistiksel bir sınıflandırıcı olarak alınır. J48 verilerin kategorik ve sürekli olarak incelenmesinde önemli rol oynar.

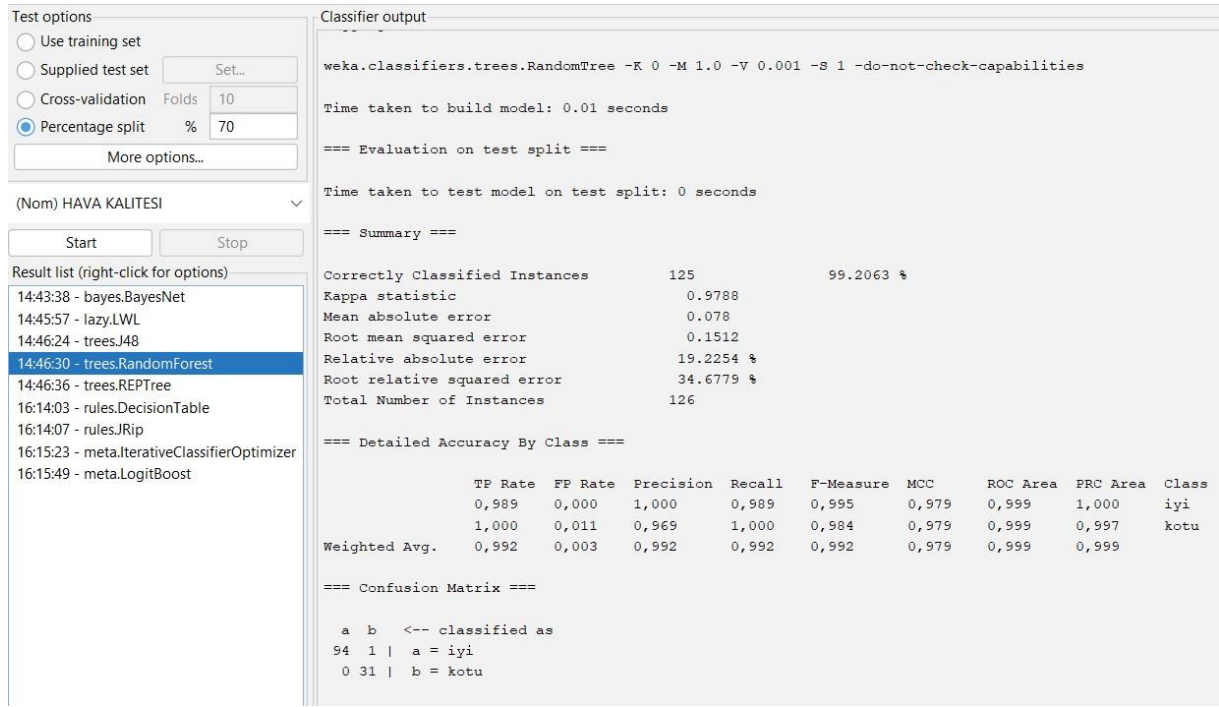


Şekil 13: J48 sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı % 99,2063 olarak bulunmuştur.

4.1.5 RandomForest Sınıflandırıcı

Regresyon veya sınıflandırma verilerini çözmek için kullanılır. Bu karar ağacı tabanlı öngörücüler, iyi hesaplama performansları ve ölçeklenebilirlikleri ile tanınırlar.

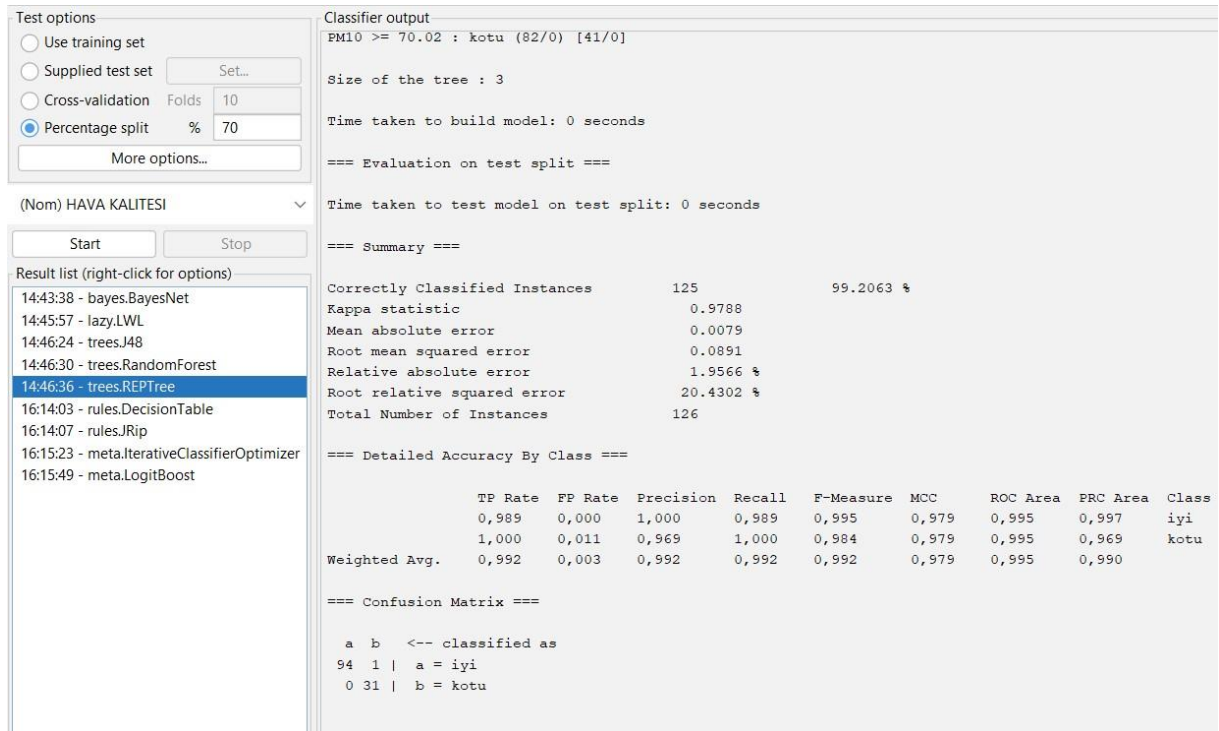


Şekil 14: RandomForest sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı % 99,2063 olarak bulunmuştur.

4.1.6 Reduced Error Pruning (REPTree) Sınıflandırıcı

Bölünme kriteri olarak bilgi kazancını kullanır. Bunnunla bir karar/regresyon ağacı oluşturur ve azaltılmış hata budama algoritması kullanarak daha hızlı bir karar ağacı öğreticisidir.



Şekil 15: REPTree sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı % 99,2063 olarak bulunmuştur.

4.1.7 DecisionTable Sınıflandırıcı

Makine öğreniminden elde edilen çıktıyı temsil etmenin en basit yolu, onu girdi ile aynı forma koymaktır. Veri kümesini, orijinal veri kümesiyle aynı sayıda öznelik içeren bir karar tablosu ile özetler.

The screenshot shows the WEKA software interface. On the left, the 'Test options' panel is visible with 'Percentage split' selected at 70%. Below it, the 'Result list' shows various classifiers, with 'rules.DecisionTable' selected. The main panel on the right displays the 'Classifier output' for the DecisionTable classifier. It includes evaluation statistics, a summary table, a detailed accuracy table by class, and a confusion matrix.

Test options

- ☐ Use training set
- ☐ Supplied test set (Set...)
- ☐ Cross-validation (Folds: 10)
- ☒ Percentage split (%: 70)
- More options...

(Nom) HAVA KALITESİ

Start Stop

Result list (right-click for options)

- 14:43:38 - bayes.BayesNet
- 14:45:57 - lazy.LWL
- 14:46:24 - trees.J48
- 14:46:30 - trees.RandomForest
- 14:46:36 - trees.REPTree
- 16:14:03 - rules.DecisionTable**
- 16:14:07 - rules.JRip
- 16:15:23 - meta.IterativeClassifierOptimizer
- 16:15:49 - meta.LogitBoost

Classifier output

Evaluation (for feature selection): CV (leave one out)
Feature set: 6,7

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances	125	99.2063 %
Kappa statistic	0.9788	
Mean absolute error	0.0142	
Root mean squared error	0.0884	
Relative absolute error	3.4938 %	
Root relative squared error	20.2709 %	
Total Number of Instances	126	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,989	0,000	1,000	0,989	0,995	0,979	0,995	0,997	iyi
	1,000	0,011	0,969	1,000	0,984	0,979	0,995	0,969	kotu
Weighted Avg.	0,992	0,003	0,992	0,992	0,992	0,979	0,995	0,990	

=== Confusion Matrix ===

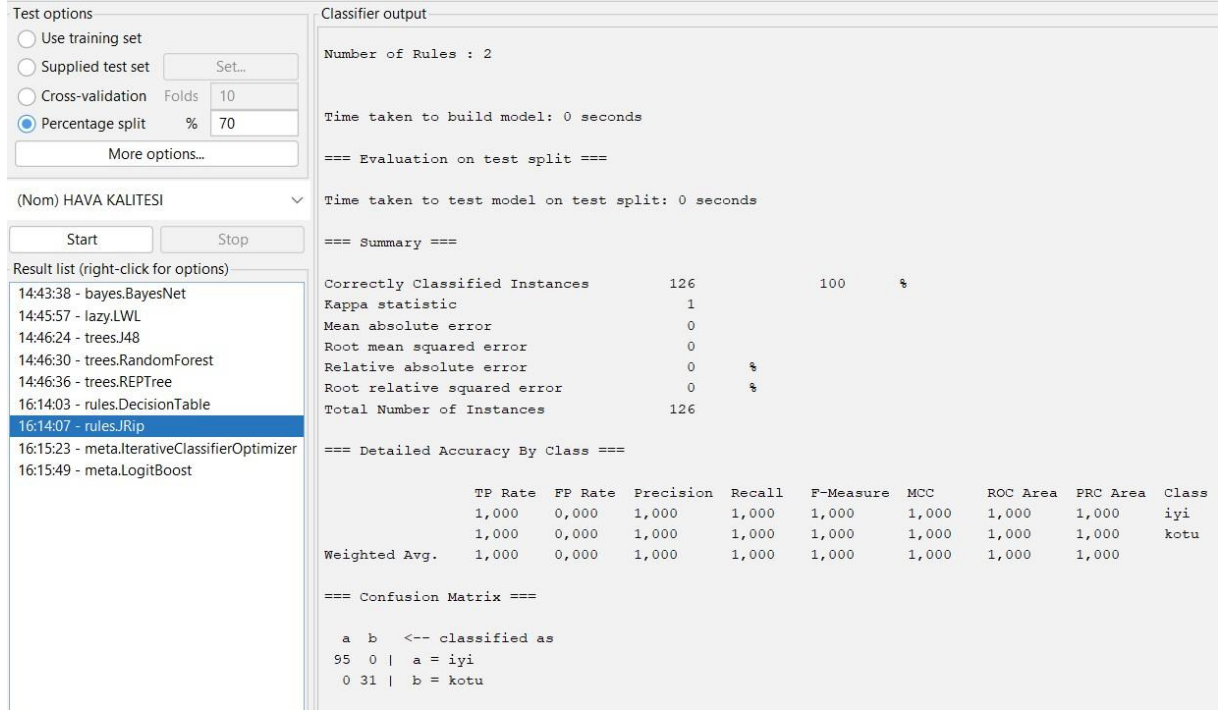
```
a b <-- classified as
94 1 | a = iyi
0 31 | b = kotu
```

Şekil 16: DecisionTable sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı % 99,2063 olarak bulunmuştur.

4.1.8 JRip sınıflandırıcı

Temel bir artımlı azaltılmış budama algoritmasıdır. Hata azaltma için kullanılan sınıflandırıcıdır.



Şekil 17: JRip sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı %100 olarak bulunmuştur. Veri setimize göre en iyi sonucu veren sınıflandırıcı JRip sınıflandırıcıdır. Oranın %100 olması veri setimize en uygun sınıflandırıcı olduğunu ve doğru öğrendiğinin göstergesidir. Karmaşıklık matrisi doğru bir eğri sağlamıştır. Bu da sınıflandırma aşamasında overfitting olmadığının modelin doğru biçimde öğrenildiğini göstermektedir.

4.1.9 IterativeClassifierOptimizer Sınıflandırıcı

Yinelemeli Sınıflandırıcı Optimizer (ICO) çapraz doğrulama kullanır ve verilen sınıflandırıcı için yineleme sayısını optimize eder.

Test options

☐ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 10
☒ Percentage split % 70

More options...

(Nom) HAVA KALITESI

Start

Stop

Result list (right-click for options)

14:43:38 - bayes.BayesNet
14:45:57 - lazy.LWL
14:46:24 - trees.J48
14:46:30 - trees.RandomForest
14:46:36 - trees.REPTree
16:14:03 - rules.DecisionTable
16:14:07 - rules.JRip
16:15:23 - meta.IterativeClassifierOptimizer
16:15:49 - meta.LogitBoost

Classifier output

Number of performed iterations: 5

Time taken to build model: 0.04 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances	125	99.2063 %
Kappa statistic	0.9788	
Mean absolute error	0.0079	
Root mean squared error	0.0891	
Relative absolute error	1.9599 %	
Root relative squared error	20.4299 %	
Total Number of Instances	126	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,989	0,000	1,000	0,989	0,995	0,979	0,995	0,997	iyi
	1,000	0,011	0,969	1,000	0,984	0,979	0,995	0,969	kotu
Weighted Avg.	0,992	0,003	0,992	0,992	0,992	0,979	0,995	0,990	

=== Confusion Matrix ===


```

a b  <-- classified as
94  1 | a = iyi
  31 | b = kotu

```

Şekil 18: IterativeClassifierOptimizer sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı % 99,2063 olarak bulunmuştur.

4.1.10 LogitBoost Sınıflandırıcı

Hızlandırıcı bir sınıflandırma algoritmasıdır. LogitBoost ve AdaBoost, her ikisinin de ek bir lojistik regresyon gerçekleştirmesi açısından birbirine yakındır. Aradaki fark, AdaBoost'un üstel kaybı en aza indirmesi, LogitBoost'un ise lojistik kaybı en aza indirmesidir.

Test options

☐ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 10
☒ Percentage split % 70

More options...

(Nom) HAVA KALITESI

Start

Stop

Result list (right-click for options)

14:43:38 - bayes.BayesNet
14:45:57 - lazy.LWL
14:46:24 - trees.J48
14:46:30 - trees.RandomForest
14:46:36 - trees.REPTree
16:14:03 - rules.DecisionTable
16:14:07 - rules.JRip
16:15:23 - meta.IterativeClassifierOptimizer
16:15:49 - meta.LogitBoost

Classifier output

Number of performed iterations: 10

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances	125	99.2063 %
Kappa statistic	0.9788	
Mean absolute error	0.0079	
Root mean squared error	0.0891	
Relative absolute error	1.9599 %	
Root relative squared error	20.4299 %	
Total Number of Instances	126	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,989	0,000	1,000	0,989	0,995	0,979	0,995	0,997	iyi
	1,000	0,011	0,969	1,000	0,984	0,979	0,995	0,969	kotu
Weighted Avg.	0,992	0,003	0,992	0,992	0,992	0,979	0,995	0,990	

=== Confusion Matrix ===


```

a b  <-- classified as
94  1 | a = iyi
  31 | b = kotu

```

Şekil 19: LogitBoost sınıflandırıcı modeli.

Burada doğru sınıflandırma oranı % 99,2063 olarak bulunmuştur.

Burada en iyi sonuç veren sınıflandırma algoritmaları ve sonuçları verilmiştir. Tüm algoritmalar için aynı veriler ve aynı train-test bölümlendirmesi yapılmıştır. Buna göre en iyi sonuç veren algoritmalar;

- BayesNet
- SimpleLogistic
- Locally Weighted Learning (LWL)
- J48
- RandomForest
- Reduced Error Pruning (REPTree)
- DecisionTable
- JRip
- IterativeClassifierOptimizer • LogitBoost

olarak bulunmuştur. Bu algoritmaların diğer Weka programındaki sınıflandırma algoritmalarına kıyasla daha iyi sonuç verdiği gözlemlenmiştir. Bu 10 algoritma içerisinde verileri en iyi sınıflandıran algoritma ise JRip sınıflandırma algoritmasıdır.

5. Sonuç

Hava kalitesi kontrol parametreleri ilişkilerinin kontrolü için öncelikle veri seti hazırlanmış ve veri ön işleme aşaması gerçekleştirilmiştir. Hava değerlerine göre parametreler arası değerler ortaya konulmuş ve Weka ile verileri sınıflandırma işlemi yapılmıştır. Yapılan sınıflandırma işlemlerinde veriler üzerinde en iyi sınıflandırma yapan 10 sınıflandırma algoritması sunulmuş ve başarıları gözlemlenmiştir. Yapılan sınıflandırma esnasında veriler aynı tutulmuştur. Üzerinde durulan 10 algoritmanın diğer sınıflandırma algoritmalarına kıyasla daha iyi sonuç verdiği söylenebilir. Bu 10 algoritma arasında ise JRip sınıflandırma algoritmasının daha iyi sonuç verdiği gözlemlenmiştir. JRip algoritmasında oranın %100 olması veri setimize en uygun sınıflandırıcı olduğunun ve doğru öğrendiğinin göstergesidir.

YAZAR

Öznur PAYLAN

Manisa Celal Bayar Üniversitesi Yazılım Mühendisliği Bölümü. Verilerin bulunması, araştırılması yapılmıştır. Veri ön işleme adımları uygulanmıştır. Python ile verilerin değerleri hesaplanmıştır. Weka ile veriler parametreler arası ilişkiler oluşturulmuştur.

REFERANSLAR

- [1] Hava Kalitesi İndeksi. 2018, Çevre, Şehircilik ve İklim Değişikliği Bakanlığı, <https://sim.csb.gov.tr/#pagetop>
- [2] Veri. 2019, Çevre, Şehircilik ve İklim Değişikliği Bakanlığı | Ulusal Hava Kalite İzleme Ağı, <http://sim.csb.gov.tr/Services/AirQuality/>
- [3] Naz İlkbahar, 2019, Data Preprocessing, <https://medium.com/@ilkbaharnaz/data-preprocessing-veri-%C3%B6n-i-CC%87%C5%9Fleme-85236484f913>
- [4] Ceyhun Enki Aksan, 2019, Weka nedir?, <https://ceaksan.com/tr/weka-nedir>

KAYNAKLAR

- Material for MkDocs, 2009 : <https://waikato.github.io/weka-wiki/>
- Jason Brownlee, 2016: <https://machinelearningmastery.com/use-classification-machine-learning-algorithmsweka/>
- Devin Soni, 2018 : <https://towardsdatascience.com/introduction-to-bayesian-networks-81031eed94e> Uğur Akel, 2020 : <https://bilisimkitabi.com/confusion-matrix>
- weka.sourceforge. 2008 : <https://weka.sourceforge.io/doc.dev/overview-summary.html> Executable
- Book Project, 2014-2022 : <https://oneapi-src.github.io/oneDAL/index.html> scikit-learn
- developers, 2007 – 2022: <https://scikit-learn.org/stable/>