VBM 655

İstatiksel Veri Analizi

2000-2019 Yılı Dünya Geneli CO_2 Emisyon Değerlerinin Analizi

Tunahan KANBAK

N22130182

5 Ocak 2022

İçindekiler

1. Giriş

Dünya ekosistemi milyarlarca yıldır farklı canlı türlerine ev sahipliği yapmakta olan bir yerdir. İlk canlının yaklaşık 4.5-3.7 milyar yıl önce yaşamaya başladığı yeryüzü ev sahipliği görevini geçtiğimiz milyar yıllarca başarılı bir şekilde yerine getirdi[1]. Bu süre zarfı içerisinde yaşayan her canlı varlık kendi çevresine hem uyum sağladı hem de bir yandan çevresini daha yaşanabilir hale getirdi. Bu canlılara "ekosistem mühendisi" adı verilmekte olup bu canlıların başlattığı değişim genelde yavaş ve kademeli bir şekilde ilerlemiştir. Ancak insanoğlu ekosistem mühendisliğini bambaşka bir boyuta taşıyarak bu değişimlerin çok daha hızlı ve büyük ölçeklerde gerçekleşmesini sağlamıştır.[2]

Günümüzde bu ekosistem mühendisliğinin sonucu olarak en bilinen gösterge ise CO_2 emisyonu verileridir. Bir bölgede gerçekleşen CO_2 emisyonu, bölge içerisinde gerçekleştirilen enerji üretimi, hayvancılık, sanayi üretimleri ve ısınma eylemleri gibi birbiri ile bağlantılı sebepler sonucunda ortaya çıkmaktadır. CO_2 emisyonları dünya üzerindeki doğal hayatı 3 temel yol ile tehlikeye sokmaktadır:

I. Havada bulunan fazla CO_2 sera gazı etkisi göstererek dünya ortalama sıcaklığının artmasına ve mevsim dengesinin bozulmasına sebep olmaktadır.

II. Havada bulunan fazla CO_2 ortalama O_2 oranlarını düşürerek bütün canlıların daha kirli bir hava solumasına sebep olmaktadır.

III. Yayılan fazla CO_2'in bir kısmı doğal su kaynaklarında çözünerek suyun asitlik değerlerine etki etmektedir. Asit değişikliği deniz yaşantısını başlı başına tehlikeye atan bir faktördür.

Bu kapsamda dünya genelinde ülkelerin yaydığı CO_2 emisyon verileri uzun sürelerdir takip edilmekte ve bir takım antlaşmalar (bkz. Kyoto Protokolü) aracılığı ile bu emisyonların kontrol altına alınması üzerinde çalışmalar yapılmaktadır.

Dünya Bankası'na ait veri tabanında her ülkenin yıllar içerisinde gerçekleştirdiği CO_2 emisyon verileri yer almaktadır. Bu raporda, dünya genelinde CO_2 emisyonlarının zaman içerisinde nasıl bir değişim gösterdiği, bu değişimlerde katkısı en fazla olan ülkeler gibi incelemeler yapılmış. Daha sonra ise ülkere ait farklı indikatörler de eklenerek CO_2 emisyonu ile ilişkili olabilecek indikatörler üzerine incelemeler gerçekleştirilmiştir. Son olarak, makine öğrenmesi algoritmaları ile CO_2 emisyonunun farklı indikatörler ile tahminlenip tahminlenemeyeceği hakkında bir inceleme gerçekleştirilmiştir.

2. Metot

2.1. Veri Özellikleri

Gerçekleştirilen veri analizinde Dünya Bankası veri tabanından 8 farklı veri seti alınmıştır:

1. Tarımsal Alan Yüzdesi (TAY): İlgili ülkenin sahip olduğu toprak alanı içerisindeki tarımsal alanın yüzdesel değeridir.

2. Ormanlık Alan Yüzdesi (OAY): İlgili ülkenin sahip olduğu toprak alanı içerisindeki ormanlık alanın yüzdesel değeridir.
3. Elektrik Erişimi Yüzdesi (EEY): İlgili ülkenin sahip olduğu nüfus içerisinde elektrik erişimi bulunan kısmın yüzdesel değeridir.
4. Tarım ve Orman Sektör Kazancı (TOSK): İlgili ülkenin belirtilen senenin sonunda elde ettiği gayrisafi millî hasıla içerisindeki tarım ve orman sektörü katkısının yüzdesel değeridir.
5. Sanayi Sektör Kazancı (SSK): İlgili ülkenin belirtilen senenin sonunda elde ettiği gayrisafi millî hasıla içerisindeki sanayi sektörü katkısının yüzdesel değeridir.
6. Kişi Başı Gayrisafi Milli Hasıla (KBGMH): İlgili ülkenin belirtilen senenin sonunda sahip olduğu gayrisafi milli hasılanın toplam nufüsa olan oranıdır. Amerikan doları cinsinden belirtilmektedir.
7. Toplam Nüfus (TN): İlgili ülkenin belirtilen senenin sonunda sahip olduğu toplam nüfus değeridir.
8. CO_2 Emisyonu (COE): İlgili ülkenin belirtilen senenin sonunda gerçekleştirdiği CO_2 emisyon miktarının kiloton cinsinden değeridir.

2.2. Veri İnceleme Yazılım ve Donanım Bilgisi

Rapor içerisinde yer alan bütün analizler Python yazılım dili ile gerçekleştirilmiştir. Veri analizi sırasında pandas, numpy, scipy, sklearn gibi kütüphanelerden faydalanılmıştır.

Analiz kapsamında hazırlanan kod bütününe ve veri setlerine https://github.com/TunahanKanbak/VBM655 adresinden ulaşılabilir.

2.3. Verinin İncelenmesinde Kullanılan Analitik Metotlar

2.3.1. Permütasyon Testi

Bu raporda yer alan verilerin büyük çoğunluğu log-normal dağılım olarak da tanımlanan çarpık bir dağılıma sahiptir. Bu durum klasik istatistikte yer alan t-testi ve ANOVA gibi analizlerin yapılmasına engel teşkil etmektedir çünkü bu testler verinin normal dağılıma sahip olduğunu varsayar. Bu varsayımların etrafından dolaşabilmek için parametrik olmayan istatistiksel analiz yöntemlerine başvuralabilir. Bu yöntemler arasında en bilinenlerden birisi de permütasyon testidir. Permütasyon testi sıfır hipotezinin doğru olduğunu kabul ederek eldeki veri ile olabilecek bütün permütasyonlar ile grupları tekrar oluşturarak dağılım istatistiklerini (ortalama, t-değeri, F-değeri vs.) tekrar hesaplar. Daha sonra gözlemlenen gerçek istatistiğin permütasyon ile elde edilen dağılımda nereye düştüğünü belirleyerek olasılık değeri olan p-değerini hesaplayabiliriz.[3]

2.3.2. Adımsal Regresyon

İstatistiksel bir modelde bağımlı değişken ile ilişkisi olmayan bağımsız değişkenlerin yer alması model performansını kötü etkileyebilmektedir. Bu durumda önemli bağımsız değişkenlerin önemli bağımsız değişkenlerden ayrıştırılması gerekmektedir.

Adımsal regresyon, bağımlı bir değişkeni açıklamak için kullanılabilecek bağımsız değişkenleri seçmek için kullanılan bir regresyon yöntemidir. Adımsal regresyonda sahip olunan bütün bağımsız değişkenler denkleme eklenir. Daha sonra her bağımsız değişken katsayısı için bir t-değeri aşağıdaki gösterildiği gibi hesaplanır.

$$t_b = \frac{\hat{b}}{SE(\hat{b})}$$

Bu denklemde b_{hat} regresyon yöntemi ile yakınsanan bağımsız değişken katsayısını gösterirken, $SE(b_{hat})$ ise ilgili katsayının standard hatasını tanımlamaktadır. Regresyon katsayıları üzerinden hesaplanan t değerleri ilgili katsayıların istatistiksel olarak ne kadar anlamlı olduğunu gösterir. T-değeri bir katsayı için ne kadar yüksek ise ilgili katsayı bağımlı değişkenin açıklanabilmesi için o kadar anlamlıdır[3].

Adımsal regresyonun gerçekleştirilmesi için bir regresyon yöntemine de ihtiyaç duyulmaktadır. Literatürde en çok karşılaşılan adımsal regresyon yöntemi ise en küçük kareler yöntemidir.

2.3.3. Permutasyon ile Nitelik Önemi

Permütasyon ile nitelik önemi yöntemi de adımsal regresyonda olduğu gibi önemli bağımsız değişkenleri önemsiz bağımlı değişkenlerden ayırmak için kullanılan bir yöntemdir. Bu yöntemde elde edilen model her bağımsız değişkenin rastgele karıştırılması ile test edilir. Böylece modelin performansının hangi parametre tarafından en fazla etkilendiği incelenebilir[4].

3. Analiz ve Değerlendirme

3.1. CO_2 Emisyon Verisinin Dağılımı ve İstatistiksel Özeti

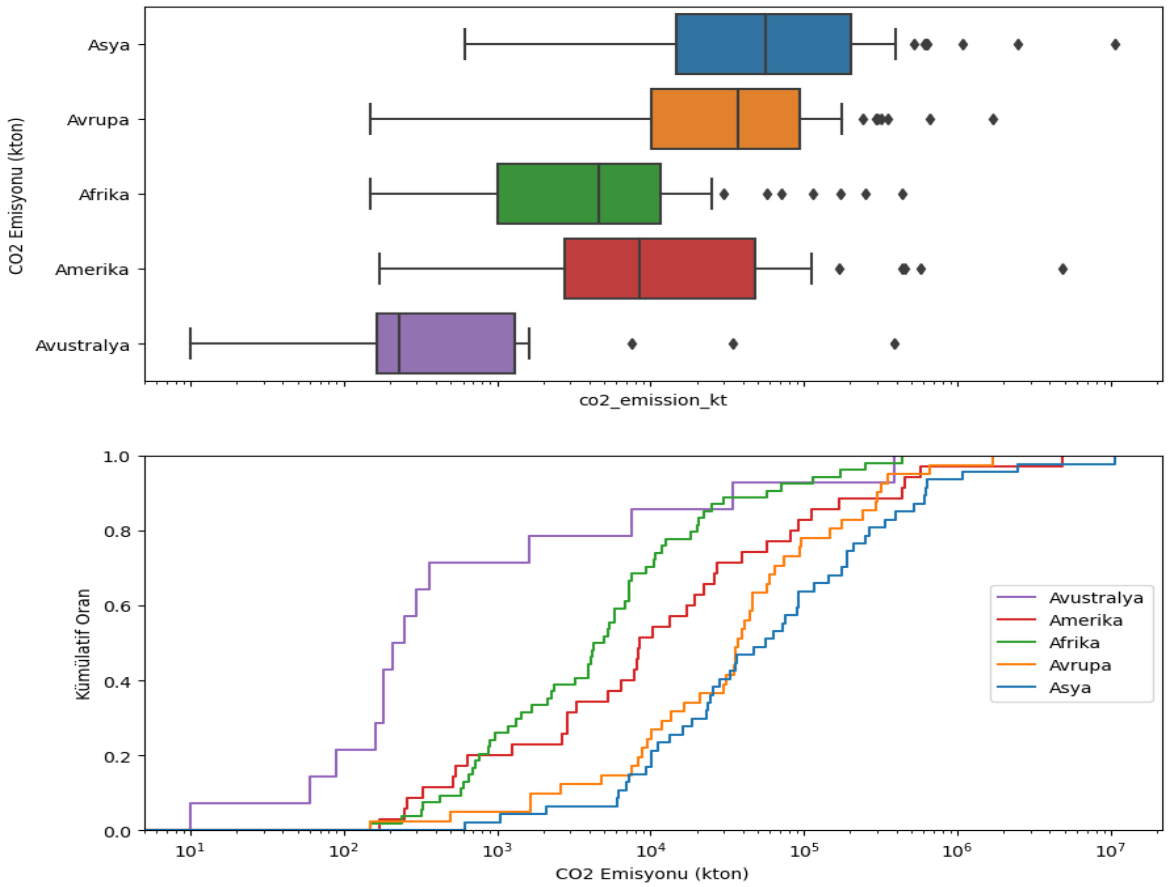
Dünya çağında CO_2 emisyon değerleri en son 2019 yılında toplanabilmiştir. 2019 yılındaki emisyon verisinin özeti aşağıdaki tabloda özetlenmiştir.

Tablo 3.1 CO_2 Emisyon Değeri İstatistiksel Özeti (kton)

Veri Satırı	Ortalama	Standard Sapma	Minimum Değer	Maksimum Değer	Medyan
8	177392.72	877397.77	10	10707219.73	12450

*2.1 numaralı başlıkta yer alan sıralamaya göre

Tablo 3.1'deki istatistiksel veriler incelendiği zaman medyan ile ortalama arasında çok ciddi bir fark olduğu görülmektedir. Bu durum CO_2 emisyonu dağılımında yüksek oranda bir pozitif çarpıklık olduğunun göstergesidir. Çarpık dağılımlarda ortalama değeri aykırı verilerden çok etkilendiği için bu tip dağılımlarda dağılım karakterini ortalama yerine medyan üzerinden tanımlamak daha anlamlı çıkarımlar yapılmasını sağlayabilmektedir.



Şekil 3.1 Kıtalara Göre CO_2 Emisyon Dağılımının Kutu Grafiği Gösterimi

Şekil 3.1'de Asya kıtasının emisyon değerleri diğer kıtalara göre daha fazla görünmektedir. Avustralya ise bütün kıtalar arasında en düşük emisyon dağılımına sahip görünmektedir.

Şekil 3.1'de kıtalara göre CO_2 emisyon verisi görsel olarak incelendiğinde dağılımlar arasında ciddi farklar görülebilmektedir ancak veri grupları içerisinde tek yönlü çok fazla aykırı değer bulunduğu için ayrım logaritmik ölçekte gözlemlenebilmektedir. Bu durum pozitif çarpıklığa sahip verilerin bir özelliği olmakta olup normal dağılımı kabul eden t-testi ve ANOVA gibi istatistiksel analizlerin kullanılamamasına sebep olmaktadır. Dolayısıyla gruplar arası farklılığın (medyan bazında) istatistiksel olarak incelenebilmesi için CO_2 emisyon verisi ile permütasyon testi gerçekleştirilerek medyanlar arası fark test edilmiştir. Yine permütasyon ile ortalamaların varyansı incelenerek kıtaların arasında istatistiksel bir fark olup olmadığı incelenmiştir.

Tablo 3.2 Tek Yönlü ANOVA Analizi

Veri Tanımı	p-değeri
CO_2 Emisyonu (kton)	0.000

Tablo 3.2'de paylaşılan ANOVA analizi ile elde edilen p-değerine baktığımızda gruplar arasındaki fark istatistiksel olarak yakalanabilmiştir.

Permütasyon testi kullanarak en çok emisyona sebep olan ve en az emisyona sahip olan kıtalar hakkındaki hipotezlerimizi test edebiliriz. Ancak daha öncede belirtildiği üzere çarpık verileri

ortalama yerine medyan kullanmak veriyi daha iyi tanımlayacağı için kıtaların medyanları kendi içerilerinde kıyaslanmıştır.:

- En çok emisyona sahip ülke;
  - H_o: X Kıtası Asya ile aynı ya da daha fazla emisyona sahiptir.
  - $H_{alternatif}$: Asya'nın sebep olduğu emisyon daha fazladır. (Tek kuyruk)

Tablo 3.3 Asya Kıtasına ait Emisyon Değerlerinin Diğer Kıtalar ile Karşılaştırılması (p-değeri)

Asya-Avrupa	Asya-Amerika	Asya-Afrika	Asya-Avustralya
0.466	0.005	0.000	0.000

- En az emisyona sahip ülke;
  - H_o: X Kıtası Avustralya ile aynı ya da daha az emisyona sahiptir.
  - $H_{alternatif}$: Avustralya'nın sebep olduğu emisyon daha azdır. (Tek kuyruk)

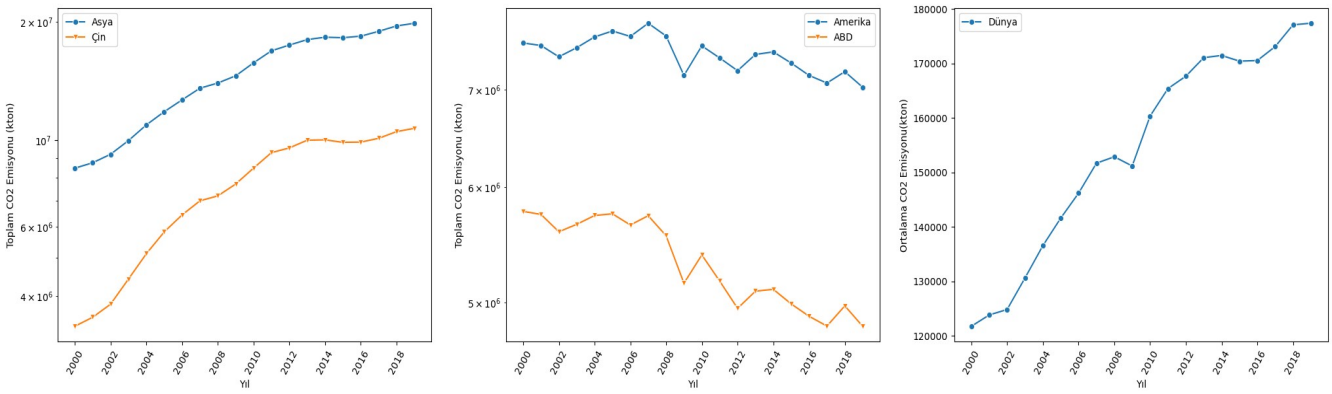
Tablo 3.4 Avustralya Kıtasına ait Emisyon Değerlerinin Diğer Kıtalar ile Karşılaştırılması (p-değeri)

Avustralya-Avrupa	Avustralya-Amerika	Avustralya-Afrika	Avustralya-Asya
0.000	0.002	0.005	0.001

Tablo 3.3'te yer alan p-değerlerini incelediğimiz zaman Avrupa haricindeki kıtalara göre Asya'nın daha çok emisyona sebep olduğu istatistiksel olarak anlamlı bir çıkarımdır.

Tablo 3.4'te yer alan p-değerlerini incelediğimiz zaman ise Avustralya kıtasının en az emisyona sahip kıta olduğu istatistiksel olarak anlamlı bir çıkarımdır.

Benzer bir şekilde 2000-2019 yılları arasında gerçekleşen CO_2 emisyonlarındaki yıllık değişimin manalı bir artış gösterip göstermediği de incelenmiştir. Yapılan incelemelerde kıtaların emisyon değişimlerinden az sayıda ülkenin sorumlu olduğu görülmüştür. Şekil 3.2'de yer alan ilk iki grafiğe bakıldığında Asya ve Amerika kıtasında en çok emisyona sahip ülke ile kıtanın emisyon miktarının neredeyse birebir aynı trende sahip olduğu görülmektedir.



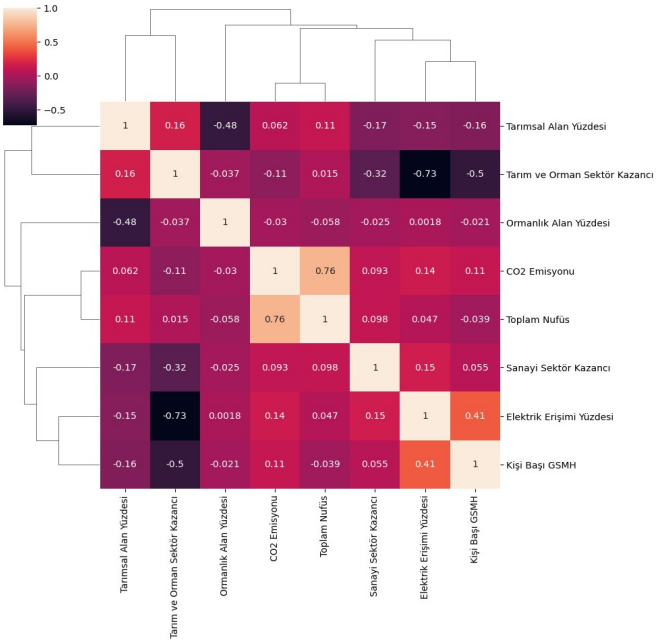
Şekil 3.2 Asya ve Amerika Kıtalarında En Çok Emisyon Miktarına Sahip Ülkeler

Ortalama emisyon değerlerinin artış grafiği incelendiği zaman son yıllara doğru ortalama artışın yavaşladığı görülmektedir. Permütasyon testi ile yapılan incelemede de 2007 yılı öncesi neredeyse her yıl istatistiksel anlama sahip bir artış gözlemlenmiştir. 2007'den itibaren ise sadece 2009-2010 ve 2016-2018 yılları istatistiksel anlama sahip artış gözlemlenebilmiştir.

3.2. İstatistiksel Model Oluşturulması

3.2.1. Bağımsız Verilerin Dağılımı ve İstatistiksel Özeti

Emisyon verisindeki değişimin bir takım ülke indikatörleri ile açıklanabilirliği için istatistiksel model oluşturulmaya çalışılmıştır. Bu kapsamda 2.1 numaralı başlıkta yer alan bağımsız veriler toplanarak veri seti genişletilmişitir. Bu verilere ait istatistiksel özel aşağıdaki tabloda paylaşılmıştır. Şekil 3.3'te ise veri setine ait çapraz korelasyon haritası gösterilmiştir.



Şekil 3.3 Veri Setinde Yer Alan Verilerin İstatistiksel Özeti

Tablo 3.5 Veri Setinde Yer Alan Verilerin İstatistiksel Özeti

Veri Satırı*	Ortalama	Standard Sapma	Minimum Değer	Maksimum Değer	Medyan
1	38.74	21.46	0.45	85.49	39.87
2	33.14	24.39	0.00	98.34	31.78
3	78.03	30.56	0.64	100.00	97.58
4	11.92	11.63	0.03	79.04	8.03
5	26.76	12.50	3.15	86.67	24.67
6	12141.68	19490.83	111.93	178864.85	4139.18
7	35936525.81	135071965.90	9392.00	1407745000.00	7632649.00

*2.1 numaralı başlıkta yer alan sıralamaya göre

Tablo 3.5'te yer alan istatistik özelliklere baktığımız zaman CO_2 emisyon dağılımına benzer şekilde bazı veri satırlarının çarpık bir dağılıma sahip olduğu ortalama ve medyan arasındaki farklardan belli olmaktadır. Verilerin minimum ve maksimum değerleri incelendiği takdirde mantıksal olmayan (örn negatif nüfus) bir veri ile karşılaşılmamıştır.

Şekil 3.3'te yer alan korelasyon haritası birbiri ile ilişkili verilerin kümelenmesi sonucunda elde edilmiştir. Ana gruplar incelendiğinde CO2 Emisyonunun toplam nüfus ile doğrudan ilişkili olduğu görülmüştür. Üst gruplara bakıldığında CO2 Emisyonu dolaylı yoldan Sanayi Sektör Kazancı, Elektrik Erişimi Yüzdesi ve Kişi Başı GSMH grubu ile ilişkili olduğu görülmüştür.

3.2.2. Adımsal Regresyon

Adımsal regresyon için en küçük kareler metotu tercih edilmiştir. Bu rapor kapsamında gerçekleştirilen çalışmalarda bağımsız değişkenlerin sadece tek yönlü etkileri değerlendirilmiştir. Dolayısıyla model belirlenmesi sırasında herhangi bir doğrusal olmayan ilişki kullanılmamıştır. Adımsal regresyon için başlangıçta belirlenen model aşağıda paylaşıldığı şekildedir.

$$COE = \beta_1 * TAY + \beta_2 * OAY + \beta_3 * EEY + \beta_4 * TOSK + \beta_5 * SSK + \beta_6 * KBGMH + \beta_7 * TN + \beta_0$$

Bu model kullanılarak gerçekleştirilen en küçük kareler algoritması ile aşağıdaki sonuçlar elde edilmiştir.

Tablo 3.6 En Küçük Kareler Algoritması ile Oluşturulan İstatistiksel Model Çıktıları

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Katsayı	-5.076e+04	463.62	617.50	404.87	-4122.75	-398.12	4.30	0.0041
SH*	6.2e+04	455.25	390.64	399.06	1173.24	715.50	0.52	5.87e-05
T-istat.	-0.819	1.018	1.581	1.015	-3.514	-0.556	8.171	69.261
p-değeri	0.413	0.309	0.114	0.310	0.000	0.578	0.000	0.000
Model düzeltilmiş R^2 değeri					0.593			

*SH: Standard hata

Adımsal regresyon yöntemine göre Tablo 3.6'teki katsayılar arasından p-değeri en yüksek olan yani istatistiksel olarak en önemsiz katsayısı denklemden çıkartılarak tekrardan modelin oluşturulması gerekmektedir. Bu durumda bir sonraki modelden β_5'e sahip SSK bağımsız verisi çıkartılarak modelin güncel hali aşağıdaki gibi olacaktır.

$$COE = \beta_1 * TAY + \beta_2 * OAY + \beta_3 * EEY + \beta_4 * TOSK + \beta_6 * KBGMH + \beta_7 * TN + \beta_0$$

Yeni denklem ile en küçük kareler algoritması bir daha çalışılarak Tablo 3.6 yeni model için tekrar oluşturulur. Bu işlem bütün p-değerleri daha önceden belirlenen alfa değerinden, 0.05, küçük olana kadar devam eder ve böylece adımsal regresyon tamamlanmış olur. Adımsal regresyon sonucunda elde edilen denklem aşağıda belirtilen şekildedir.

$$COE = -4049.39 * TOSK + 4.6321 * KBGMH + 0.0041 * TN$$

Adımsal regresyon sonucunda elde edilen modelin düzeltilmiş R^2 değeri 0.611 olup ortalama hata kareleri kökü (OHKK) ise 4.78e5'dir. Bu da ortalama hatanın CO_2 emisyon değerinin ortalaması olan 1.5e5'ten daha fazla olduğunu göstermektedir. En küçük karelerin toplamı ile elde edilen bu modelin yüksek hataya ve düşük performansa sahip olmasının en olası sebebi kullanılan indikatörlerin yetersiz olması ya da indikatörler arasında doğrusal olmayan ilişkilerin bulunmasıdır.

Bu problemin üstesinden gelmek için adımsal regresyona doğrusal olmayan (örn. TN * KGBMH) parametreler eklenerek modelin performansı arttırılabilir.

3.3. Yapay Zeka Uygulamaları

3.3.1. Karar Ağaçları, Random Forest ve Boosting Algoritmaları

Bu raporda ele alınan CO_2 emisyonu veri seti gibi tablo halindeki veriler için karar ağaçları, random forest ve boosting algoritmalarının iyi performans verdiği bilinmektedir. Hem veri dağılımından etkilenmeyen hem de veri setindeki doğrusal olmayan ilişkileri yakayalabilen modeller oldukça yüksek performansa sahiptir. Veri özetinin yapıldığı 2.1 numaralı başlıkta yer alan bütün veriler kullanılarak bir 3 farklı model karşılaştırılmıştır.

3 farklı model ile elde edilen performansların değerlendirilebilmesi için veri seti ikiye bölünmüştür. Bunlardan birincisi “öğrenme” veri seti, ikincisi ise “test” veri seti olarak isimlendirilmiştir. Öğrenme setide kendi içerisinde 10 gruba ayrılarak bütün modellerde 10-katlı çapraz doğrulama işlemi gerçekleştirilmiştir. Öğrenme seti ile bulunan en iyi parametreler kullanılarak üç model de test veri setinde denenerek aşağıdaki performans parametreleri elde edilmiştir.

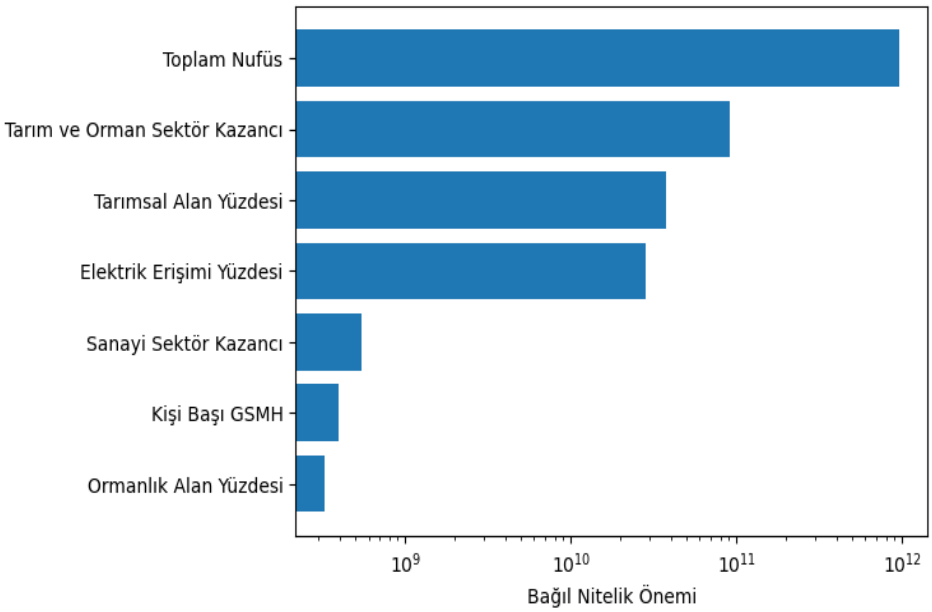
Tablo 3.7 Farklı Makine Öğrenmesi Modellerinin Kıyaslanması

	Karar Ağacı		Random Forest		Adaboost	
	Öğrenme	Test	Öğrenme	Test	Öğrenme	Test
R^2 değeri	7.56e-01	7.10e-01	9.94e-01	9.92e-01	1.00	9.93e-01
OHKK	3.69e+05	4.06e+05	5.76e+04	6.68e+04	1.21e+04	6.13e+04

Tablo 3.7’te yer alan performans sonuçlarını incelediğimiz zaman Random Forest ve Adaboost algortimalarının benzer ve en küçük kareler algoritmasına kıyasla çok daha iyi bir performans gösterdiği değerlendirilmiştir. Random Forest ve Adaboost algoritmalarının 10-katlı çapraz doğrulama sonuçları t-testi aracılığı ile kıyaslanmış olup p-değeri 0.05’in üzerinde bulunmuştur. Dolayısıyla, iki modelin performansı arasında istatistiksel öneme sahip bir farklılık bulunamamıştır.

3.3.2. Boosting Modelinin İyileştirilmesi

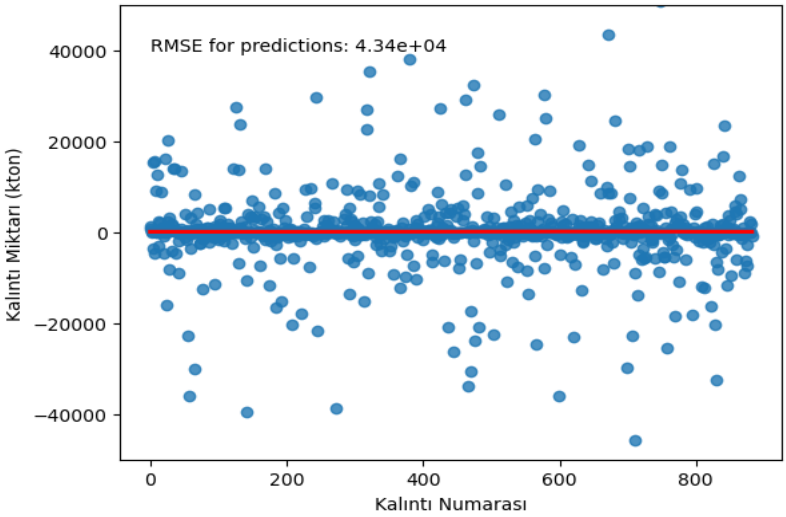
Boosting modelinin iyileştirilmesi amacı ile nitelik seçimi yapılmıştır. Adımsal regresyonda olduğu gibi bağımsız parametrelerin bağımlı parametrenin açıklanabilmesi için ne kadar gerekli olduğunun analizi yapılmıştır. Permütasyon yöntemi ile saptanan nitelik önemleri aşağıdaki şekilde gösterilmiştir.



Şekil 3.4 Permütasyon Analizi ile Bulunan Bağıl Nitelik Önemi

Nitelik önemi analizi sonrasında toplam nüfus parametresinin en önemli ikinci parametreden 10 kat daha önemli olduğu görülebilmektedir. Ek olarak, sanayi sektör kazancı, kişi başı gayrisafi milli hasıla ve ormanlık alan yüzdesinin diğer parametreler yanında önemsiz olabileceği görülmüştür.

Nitelik önemi sonrasında Adaboost algoritması seçilmiş nitelikler ile tekrar modellenerek performansı incelenmiştir.



Şekil 3.5 Boosting Modelinin Test Verisindeki Hata Dağılımı

Seçilen değişkenler kullanılarak çalıştırılan Adaboost algoritmasında öğrenme veri seti için R^2 ve OHKK değerleri sırası ile 1.00 ve 2.03e+03 gelmiş olup test veri seti için bu değerler sırası ile 9.97e-01 ve 4.34e+04 şeklindedir. Adaboost ile elde edilen modelin OHKK değeri en küçük kareler algoritmasının OHKK değerinden 10 kat daha küçüktür. Bu çıktılar CO_2 emisyon değerlerinin belli indikatörler ile tahmin edilebileceğini göstermiştir. Şekil 3.5'te yer alan hata dağılımına bakıldığında da hataların simetrik bir eksende gerçekleştiği görülmektedir. Bu da elimizdeki

modelin pozitif veya negatif yöne meyilli göstermek yerine dengeli bir davranışı olduğunu göstermektedir.

4. Sonuç

Dünyamız için oldukça önemli bir konu olan CO_2 emisyon çıktıları istatistiksel olarak incelendiğinde emisyon miktarının pozitif çarpık bir dağılıma sahip olduğu görülmektedir. Pozitif çarpıklığın sebebi olarak her kıtada diğer ülkelere göre (örn. Amerika, Çin) aşırı gelişmiş ve kalabalık ülkelerin bulunması değerlendirilmişitir. Kıtaların sebep olduğu emisyon miktarlarının dağılımına bakıldığında kıtalar arasında istatistiksel manada farklılık olduğu görülebilmektedir. Ek olarak en çok emisyona sebep olan kıtanın Asya ile Avrupa ve en az emisyona sahip kıtanın Avustralya olduğu istatistiksel olarak manalı çıkarımlar olarak değerlendirilmiştir. Her yıl toplam emisyondaki artış davranışının da istatistiksel olarak manalı olduğu görülmektedir.

Emisyon değerinin açıklanabilmesi için bazı ülke indikatörleri belirlenerek istatistiksel modeller oluşturulmaya çalışılmıştır. İstatistiksel modellerin çıktısı olarak emisyon ve indikatörler arasında doğrusal olmayan ilişkiler bulunabileceği değerlendirilmiştir. En küçük kareler algoritması ile yapılan nitelik seçimi sonucunda TOSK, KBGMH ve TN parametrelerinin emisyon miktarını açıklayan önemli parametreler olduğu görülmüştür. Ancak permütasyon yöntemi ile yapılan nitelik önemi sıralamasına bakıldığı zaman, CO_2 emisyonunu açıklayan önemli parametrelerin TN, TOSK, TAY ve EEY olduğu görülmüştür.

Doğrusal olmayan ilişkiler bulunması ve verinin tablo şeklinde olmasından kaynaklı bazı makine öğrenmesi algoritmaları aracılığı ile CO_2 emisyonu modellenmeye çalışılmıştır. Elde edilen model en küçük kareler algoritmasına göre 10 kat daha az ortalama hataya sahiptir.

Bu çalışmada, CO_2 emisyonunun her geçen sene daha da arttığı ve bu artıştan aykırı olarak değerlendirilebilecek ülkelerin sorumlu olduğu görülmektedir. Öte yandan, CO_2 emisyonunun bazı indikatörler ile yakınsanabileceği de gösterilmiştir. İlerleyen süreçte bu indikatörler arttırılarak ülkelerin CO_2 emisyonunu azaltması için dikkat etmesi gereken kritik indikatörlerin analizi makine öğrenmesi algoritmaları sayesinde gerçekleştirilebilir.

5. Referanslar

[1]Ben K.D. Pearce, Andrew S. Tupper, Ralph E. Pudritz, and Paul G. Higgs. (2018). Constraining the Time Interval for the Origin of Life on Earth. Astrobiology.Mar 2018.343-364.http://doi.org/10.1089/ast.2017.1674

[2]Chu, E. W., & Karr, J. R. (2017). Environmental Impact: Concept, Consequences, Measurement. Reference Module in Life Sciences, B978-0-12-809633-8.02380-3. https://doi.org/10.1016/B978-0-12-809633-8.02380-3

[3]Bruce P. C. & Bruce A. (2017). Practical statistics for data scientists : 50 essential concepts (First). O'Reilly.

[4]Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324