T.C.
İNÖNÜ ÜNİVERSİTESİ
SOSYAL BİLİMLERİ ENSTİTÜSÜ

ÇOKLU DOĞRUSAL BAĞLANTI VE AYKIRI DEĞER SORUNU İÇİN RIDGE-ROBUST-BOOSTING TOPLULUK REGRESYON YAKLAŞIMI

DOKTORA TEZİ

DANIŞMAN
Prof. Dr. Mehmet GÜNGÖR

HAZIRLAYAN
Ayşegül HAN

MALATYA-2023

T.C
İNÖNÜ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI

DOKTORA TEZİ

ÇOKLU DOĞRUSAL BAĞLANTI VE AYKIRI DEĞER SORUNU İÇİN RIDGE-ROBUST-BOOSTING TOPLULUK REGRESYON YAKLAŞIMI

Ayşegül HAN

Danışman
Prof. Dr. Mehmet GÜNGÖR

MALATYA-2023

ONUR SÖZÜ

Prof. Dr. Mehmet GÜNGÖR'ün danışmanlığında doktora tezi olarak hazırladığım **"ÇOKLU DOĞRUSAL BAĞLANTI VE AYKIRI DEĞER SORUNU İÇİN RIDGE-ROBUST-BOOSTING TOPLULUK REGRESYON YAKLAŞIMI"** başlıklı bu çalışmanın, bilimsel ahlak ve geleneklere aykırı düşecek bir yardıma başvurmaksızın tarafımdan yazıldığını ve yararlandığım bütün yapıtların hem metin içinde hem de kaynakçada yöntemine uygun biçimde gösterildiğini belirtir, bunu onurumla doğrularım.

TEŞEKKÜR

Tez çalışmam süresince bana her zaman destek olan ve bu tezin oluşmasında büyük rol oynayan değerli hocam, danışmanım Prof. Dr. Mehmet GÜNGÖR'e,

Bilgi ve öneri anlamında hiçbir zaman desteğini esirgemeyen çok değerli hocalarım Doç. Dr. Yunus BULUT ve Dr. Öğr. Üyesi Gökhan KONAT'a,

Katkı ve yardımlarından dolayı Prof. Dr. Latif ÖZTÜRK ve Prof. Dr. Ali KARCI'ya,

Çok zorlu yollardan geçmiş olsalar bile hayatım boyunca bunu bana yansıtmadan öğrenim hayatımı tamamlamamı sağlayan, hayatımın her alanında hep arkamda durup bana destek olan canım annem ve canım babama,

Canım kardeşim Ömer'e

Sonsuz teşekkür ederim.

ÖZET

Bu çalışmada, aykırı değer ve çoklu doğrusal bağlantı sorunları birlikte ele alınmış ve bu sorunlara karşı güçlü sonuçlar elde edebilmek amacıyla Ridge-Robust-Boosting Topluluk Regresyon modeli önerilmiştir. Yapılan simülasyon çalışması sonucunda, Ridge-Robust-Boosting Topluluk Regresyon modelinin farklı çoklu doğrusallık düzeylerinde ve aykırı değer oranlarında diğer regresyon modellerine göre daha üstün bir performans sergilediği görülmüştür. Ayrıca gerçek verilerle yapılan ikinci uygulama sonucunda da Ridge-Robust-Boosting Topluluk Regresyon modelinin diğer regresyon modellerine kıyasla çok daha iyi performans gösterdiği gözlemlenmiştir. Bu nedenle hem çoklu doğrusal bağlantı hem de aykırı değerlerin bulunduğu veri setlerinde daha doğru ve güvenilir tahminler elde etmek için Ridge-Robust-Boosting Topluluk Regresyon modelinin kullanılması önerilmektedir.

Anahtar Kelimeler: Aykırı Değer, Çoklu Doğrusal Bağlantı, Ridge Regresyon, Robust Regresyon, Gradient Boosting Regresyon, Ridge-Robust-Boosting Topluluk Regresyon.

ABSTRACT

In this study, outlier and multicollinearity problems are considered together, and the Ridge-Robust-Boosting Ensemble Regression model is proposed to obtain robust results against these problems. As a result of the simulation study, it was observed that the Ridge-Robust-Boosting Ensemble Regression model outperformed other regression models at different multicollinearity levels and outlier rates. In addition, as a result of the second application with real data, it was observed that the Ridge-Robust-Boosting Ensemble Regression model performed much better than the other regression models. Therefore, it is recommended to use the Ridge-Robust-Boosting Ensemble Regression model to obtain more accurate and reliable predictions in data sets with both multicollinearity and outliers.

Keywords: Outlier, Multicollinearity, Ridge Regression, Robust Regression, Gradient Boosting Regression, Ridge-Robust-Boosting Ensemble Regression.

İÇİNDEKİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

1. GİRİŞ

Regresyon analizi, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi modelleyerek gelecekteki değerleri tahmin etmek için sıkça kullanılan önemli bir istatistiksel yöntemdir. Ancak, bu analiz sırasında çeşitli sorunlarla karşılaşılabilir. Örneğin, çoklu doğrusallık, bağımsız değişkenler arasında yüksek düzeyde korelasyonun olduğu durumlarda ortaya çıkan bir sorundur ve bu durum, modelin istikrarını ve doğruluğunu olumsuz etkileyebilir. Ayrıca, aykırı değerler, diğer verilere göre belirgin şekilde farklı değerlere sahip olan anormal gözlemlerdir ve bu değerler, modelin parametrelerini yanıltarak sonuçların güvenilirliğini azaltabilir. Bu nedenle, regresyon analizi yaparken bu tür sorunları dikkate almak ve uygun çözüm yollarını aramak önemlidir.

Topluluk modelleri, farklı temel modellerin bir araya getirilerek tek bir optimal tahmin modeli üretmeyi amaçlayan bir makine öğrenmesi tekniğidir. Bu yaklaşım, her temel modelin kendi başına sağlayabileceğinden daha doğru ve güvenilir tahminler elde etmeyi hedefler. Temel olarak, topluluk modelleri, çeşitli modellerin birleştirilmesiyle oluşturulur ve her bir modelin güçlü yanlarını kullanarak zayıf yönlerini telafi etmeye çalışır. Bu şekilde, tahminlerin hassasiyeti ve güvenilirliği artırılabilir.

Bu çalışmada, topluluk öğrenme algoritmalarının regresyon analizi alanındaki potansiyeli vurgulanmaktadır. Ridge Regresyon, Robust Regresyon ve Gradient Boosting Regresyon modelleri, Ridge-Robust-Boosting Topluluk Regresyon modelinin temel taşları olarak birleştirilmiştir. Bu model, Ridge Regresyon'un çoklu doğrusallıkla başa çıkma yeteneği, Robust Regresyon'un aykırı değerlerin etkisini azaltma gücü ve Gradient Boosting Regresyon'un karmaşık yapıları modelleme yeteneğini bir araya getirerek çoklu doğrusallık ve aykırı değerlerin aynı anda ele alınmasını sağlamaktadır. Önerilen Ridge-Robust-Boosting Topluluk Regresyon modeli, Ridge Regresyon, Robust Regresyon ve Gradient Boosting Regresyon modellerinin avantajları bir araya getirerek daha sağlam bir analiz yaklaşımını sunmaktadır.

2. REGRESYON MODELLERİNİN İNCELENMESİ

Regresyon modelleri, verilerdeki ilişkileri anlamak için güçlü bir araçtır ve doğru bir şekilde kullanıldığında, gelecekteki olası gelişmeleri tahmin etmek, belirli bir değişkenin diğer değişkenler üzerindeki etkisini ölçmek veya en uygun çözümü sağlamak gibi birçok uygulama alanına sahiptir. Bu nedenle, regresyon modellerinin incelenmesi, modelin doğruluğunu ve veriye uygunluğunu değerlendirmek için önemlidir.

Bu doğrultuda, çalışmanın bu bölümünde doğrusal regresyon analizi, varsayımların bozulması durumunda kullanılan regresyon modelleri, genelleştirilmiş regresyon modelleri ve gelişmiş regresyon modelleri başlıkları altında detaylı olarak incelenmiştir.

2.1. Doğrusal Regresyon Analizi

Regresyon analizi, bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini matematiksel olarak modellemek ve bu etkiyi kullanarak bağımlı değişkenin değerini tahmin etmek için kullanılır. Aynı zamanda, regresyon analizi bağımsız değişkenlerin bağımlı değişken üzerindeki etkisini analiz ederek, bu ilişkinin ne kadar güçlü veya zayıf olduğunu ve hangi bağımsız değişkenlerin daha önemli olduğunu anlamak için de kullanılır. Bu nedenle, regresyon analizi, verilerdeki ilişkileri anlamak, tahminlerde bulunmak ve kararlar vermek için önemli bir istatistiksel araçtır.

Basit doğru denklemi aşağıdaki gibi ifade edilmektedir:

$$y = ax + b$$

Basit doğrusal regresyon modeli aşağıdaki gibi gösterilmektedir (Arkes, 2019: 17);

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$(i = 1,2,3, \dots, N)$$

Burada; Y bağımlı (sonuç, açıklanan, yanıt, bağlanan) değişken, X bağımsız (açıklayan, açıklayıcı, girdi, tahmin, regresör) değişken, β_0 sabit terim (kesme terimi), β_1 eğim katsayısı ve ε rassal hata terimini ifade etmektedir.

Çoklu doğrusal regresyon modelinin iki veya daha fazla bağımsız değişkeni bulunmaktadır ve aşağıdaki gibi gösterilmektedir (Arkes, 2019: 33):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \varepsilon_i$$

Burada; K, bağımsız değişkenlerin sayısını ifade etmektedir. $\beta_1, \beta_2, \dots, \beta_K$ eğim katsayılarıdır ve regresyon katsayıları (parametreleri) olarak adlandırılmaktadır.

Basit ve çoklu doğrusal regresyon denklemleri anakütle regresyon denklemi olarak kabul edilmektedir. Ancak anakütleden örneklemeyle ilgili tesadüfilik ve sonucu

etkileyen rassal olaylar nedeniyle gerçek regresyon denklemini tahmin etmek mümkün değildir. Gerçek katsayı tahminleri olmadan da gerçek hata terimi elde edilememektedir. Bu nedenle örneklem regresyon modeli aşağıdaki gibi elde edilmektedir (Panik, 2005: 673):

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$$

Burada; $\hat{\beta}_0$, $\hat{\beta}_1$ ve $\hat{\varepsilon}_i$ üzerindeki "şapkalar" (ˆ), bunların tahmin edilen değerler olduğunu göstermektedir. Ayrıca, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ şeklindedir.

Regresyon katsayılarını tahmin etmek için yaygın olarak kullanılan yöntem, en küçük kareler (EKK) yöntemidir. EKK yöntemi, karesi alınmış hata teriminin toplamını aşağıdaki gibi en aza indirmektedir (Panik, 2005: 674):

$$min\left\{\sum\left(Y_i - \hat{Y}_i\right)^2 = \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2\right\}$$

Burada toplam, tüm gözlemler üzerinden alınmaktadır.

KKT'nın en küçüklenmesi için her $\hat{\beta}$ katsayısına göre KKT'nın kısmi türevinin alınması gerekmektedir. Kısmi türev sonrasında elde edilen denklemler sıfıra eşitlenmektedir. Bu da aşağıdaki gibi gösterilmektedir (Panik, 2005: 674):

$$\frac{\partial \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{\partial \hat{\beta}_0} = 0$$

$$\frac{\partial \sum\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2}{\partial \hat{\beta}_1} = 0$$

$\hat{\beta}_0$ ve $\hat{\beta}_1$ parametrelerine göre kısmi türev alınıp gerekli sadeleştirmeler yapıldığında aşağıda belirtilen eş zamanlı doğrusal denklem sistemi (normal denklemler) elde edilmektedir.

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i = \sum Y_i \\ \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 = \sum X_i Y_i \end{cases}$$

Normal denklemler, $\hat{\beta}_0$ ve $\hat{\beta}_1$ parametreleri için Cramer yöntemi ile çözüldüğünde aşağıda belirtilen en küçük kareler tahmincileri elde edilmektedir (Panik, 2005: 674):

$$\hat{\beta}_0 = \frac{(\sum Y_i)(\sum X_i^2) - (\sum X_i Y_i)(\sum X_i)}{n(\sum X_i^2) - (\sum X_i)^2}$$

$$\hat{\beta}_1 = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n(\sum X_i^2) - (\sum X_i)^2}$$

Burada; $x_i = X_i - \bar{X}$ ve $y_i = Y_i - \bar{Y}$ (i=1, 2, …, n) olarak yani x_i, X_i'nin ortalamadan i. sapmasını ve y_i ise Y_i'nin ortalamadan i. sapmasını belirttiğinde $\hat{\beta}_0$ ve $\hat{\beta}_1$ parametreleri aşağıdaki gibi gösterilmektedir:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ denklemi, en uygun $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ en küçük kareler doğrusunun önemli bir özelliğini, yani $(\bar{X}, \bar{Y})$ ortalama noktasından geçtiğini göstermektedir.

2.1.1.1. En Küçük Kareler Varsayımları

EKK, doğrusal modeller için en yaygın tahmin yöntemidir. En iyi tahminleri üretmek için dikkate alınması gereken EKK varsayımları aşağıdaki gibidir (Frost, 2019: 195):

Varsayım 1: Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki doğrusaldır.

Varsayım 2: Bağımsız değişkenlerle hata terimi arasında ilişki yoktur.

Varsayım 3: Hata terimi, sıfır ortalama ve sabit varyans ile normal dağılım göstermektedir.

Varsayım 4: Bağımsız değişkenler arasında çoklu doğrusal bağlantı olmamalıdır.

Varsayım 5: Hata terimi, sabit bir varyansa sahiptir.

Varsayım 6: Hata terimlerinin birbirini izleyen değerleri arasında ilişki yoktur.

2.1.1.1.1. Aykırı Değer

Aykırı değerler, bir veri kümesindeki diğer veri noktalarının çoğundan önemli ölçüde sapan ve genel örüntüye uymayan veri noktalarıdır. Aykırı değerlerin varlığı istatistiksel analizleri önemli ölçüde etkilemekte, ortalama ve standart sapma gibi ölçümleri çarpıtmakta ve doğru bir şekilde tespit edilip ele alınmadığı takdirde hatalı sonuçlara yol açmaktadır. Aykırı değerlerin etkili bir şekilde yönetilmesi, veri analizlerinin ve yorumlarının geçerliliğini ve güvenilirliğini sağlamak için çok önemlidir.

Bir veri kümesinde aykırı değer olmasının nedenlerini aşağıdaki gibi ifade etmek mümkündür (Gujarati, 2004: 541):

- **Ölçüm hataları:** Verileri ölçmek için kullanılan yöntemler yeterince hassas olmadığında veya veri toplama sırasında hata meydana geldiğinde ortaya çıkabilir.

- **Örnekleme hataları:** Örneklem boyutu çok küçükse veya örneklem rastgele seçilmemişse bu durum ortaya çıkabilir. Bu gibi durumlarda, örneklem, popülasyonu tam olarak yansıtmayabilir ve bu da aykırı değerlere neden olur.
- **Veri giriş hataları:** Aykırı değerler veri giriş hatalarından da kaynaklanabilir. Örneğin, bir kişinin yaşı 50 yerine 500 olarak girilirse, bu bir aykırı değerle sonuçlanacaktır.
- **Aşırı olaylar:** Aykırı değerler, nadir görülen ancak veriler üzerinde önemli bir etkiye sahip olan aşırı olaylardan da kaynaklanabilir. Örneğin, hisse senedi fiyatları üzerine yapılan bir çalışmada, ani ve beklenmedik bir piyasa çöküşü aykırı veri noktalarına neden olabilir.
- **Sistematik yanlılık:** Veri toplama süreci belirli bir gruba karşı önyargı olduğunda veya belirli veri noktaları sistematik olarak hariç tutulduğunda görülebilir.
- **Ölçüm birimleri:** Aykırı değerler farklı ölçüm birimlerinin kullanılmasından da kaynaklanabilir. Örneğin, bir veri kümesi hem inç hem de santimetre cinsinden ölçümler içeriyorsa, bu durum aykırı değerlere neden olabilir.

Aykırı değerleri tespit etmek için kullanılan en yaygın yöntemler aşağıdaki gibidir:

- **Dağılım grafiği:** Bağımlı ve bağımsız değişkenler arasındaki ilişkiyi görsel olarak incelemek için bir dağılım grafiği kullanılabilir.
- **Artık grafiği:** Aykırı değerler, diğer veri noktalarına kıyasla büyük artıklara sahip veri noktaları olarak tanımlanabilir.
- **Cook mesafesi:** Belirli bir gözlem noktasının regresyon katsayıları üzerindeki etkisini ölçerek, bu gözlem noktasının aykırı bir değer olup olmadığını değerlendirmemize yardımcı olur.
- **Mahalanobis mesafesi:** Bir gözlem noktasının diğer noktalardan ne kadar uzak olduğunu değerlendirmek için değişkenler arasındaki ilişkileri dikkate alıp hesaplayarak değişkenler arasındaki korelasyonların ve dağılımların etkisi altında aykırı değerleri tespit etmeye yardımcı olur.
- **Z-skoru:** Z-skoru, bir gözlemin ortalamadan kaç standart sapma uzakta olduğunun bir ölçüsüdür. Aykırı değerler, belirli bir eşikten (örneğin 3 veya 4) daha büyük Z-skorlarına sahip veri noktaları olarak tanımlanabilir. Z değeri aşağıdaki gibi hesaplanmaktadır:

$$Z = \frac{x_i - \bar{x}}{S}$$

$$x_i \sim N(\mu, \sigma^2)$$

$$S = \sqrt{\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$X \sim N(\mu, \sigma^2)$ iken $Z = \frac{x-\mu}{\sigma} \sim N(0,1)$'dir.

- **Kutu grafiği:** Veri kümesindeki aykırı değerleri hızlıca tespit etmeye ve genel veri dağılımı hakkında fikir sahibi olmaya yardımcı olmaktadır. Kutu grafiği, verilerin beş sayı özeti ile temsil edilmektedir: en küçük değer (minimum), alt çeyrek (Q_1), medyan (Q_2 veya ortanca), üst çeyrek (Q_3) ve en büyük değer (maksimum). Aynı zamanda, aykırı değerleri tespit etmek için "Tukey outlier" yöntemini de kullanmaktadır. Tukey outlier yöntemine göre, bir değer Q_1'den $1.5(Q_3 - Q_1)$ daha küçük veya Q_3'ten $1.5(Q_3 - Q_1)$ daha büyükse, o değer bir aykırı değer olarak kabul edilmektedir.

Aykırı değerler tespit edildikten sonra uygulanabilecek yaygın yaklaşımlar şunlardır:

- **Aykırı değerleri kaldırmak:** Aykırı değer olarak tanımlanan veri noktalarının belirlenmesi ve veri setinden çıkarılmasıdır. Ancak, bu yaklaşım bilgi kaybına yol açabilmekte ve regresyon modelinin doğruluğunu etkileyebilmektedir
- **Verileri dönüştürmek:** Bir başka yaklaşım da verileri aykırı değerlerden daha az etkilenecek şekilde dönüştürmektir. Örneğin, verilerin logaritması veya karekökü alınarak ya da veriler çeşitli şekilde standartlaştırılarak dönüştürülebilir.
- **Sağlam regresyon teknikleri kullanmak:** Sağlam regresyon teknikleri aykırı değerlere karşı daha az duyarlıdır ve çok sayıda aykırı değer içeren veri kümelerini işlemek için kullanılabilir. Özellikle aykırı değerlerin çıkarılması zor olduğunda veya örneklem boyutu küçük olduğunda kullanışlıdır.
- **Aykırı değerleri saklamak:** Bazı durumlarda, aykırı değerleri veri kümesinde tutmak ve bunları ayrı olarak analiz etmek uygun olabilir. Bu yaklaşım, aykırı değerler verilerin önemli bir alt kümesini temsil ettiğinde veya değişkenler arasındaki ilişki hakkında değerli bilgiler sağladığında faydalı olmaktadır.

2.1.1.1.2. Çoklu Doğrusal Bağlantı

Doğrusal regresyon modelleri, bağımsız değişkenler arasında karmaşık ilişkilerin olmadığını varsayar. Ancak çoklu regresyon analizinde, bağımsız değişkenler arasında yakın veya mükemmel doğrusal ilişkiler varsa bu problem yaratabilir. Bu durumda, en düşük varyansa sahip tahminci olan En Küçük Kareler (EKK) tahmini güvenilirliğini yitirebilir (Albayrak, 2005: 109).

Frisch'e (1934) göre, gerçek değişkenler arasındaki farklı doğrusal ilişkileri tahmin etmek zor olabilir. Bu zorluk, bağımlı ve bağımsız değişkenlerin ayrı ayrı ele alınamaması ve tüm değişkenlerin hata içerdiği durumlarda daha da artar. Çoklu doğrusallık sorunu, doğrusal regresyon modellerinde bağımsız değişkenler arasındaki güçlü ilişkilerin tahminleri yanıltıcı hale getirebileceği bir durumu ifade eder.

Bir veri kümesinde çoklu doğrusal bağlantı bulunmasının nedenlerini aşağıdaki gibi ifade etmek mümkündür:

- **Yetersiz veri:** Regresyon analizinde kullanılan veri seti yeterince büyük değilse, değişkenler arasında daha fazla korelasyon olabilir.
- **Ölçüm Hatası:** Bağımsız değişkenlerin ölçümünde hata olması, bu değişkenler arasında yanlış bir ilişki meydana getirebilir.
- **Değişken Dönüşümleri:** Değişken dönüşümleri (örneğin, logaritmik veya karekök dönüşümleri) kullanıldığında, değişkenler arasındaki ilişkiler değişebilir ve doğrusal bağlantı oluşabilir.
- **Yüksek çok boyutlu veri:** Yüksek boyutlu veri setlerinde, değişkenler arasındaki korelasyon artabilir.

Çoklu doğrusal bağlantıyı tespit etmek için kullanılan en yaygın yöntemler aşağıda verilmiştir;

- **Korelasyon Matrisi:** Bağımsız değişkenler arasındaki korelasyonları ölçmek için korelasyon matrisi oluşturulabilir. Yüksek korelasyonlar, çoklu doğrusal bağlantının bir göstergesi olabilir.
- **Varyans Enflasyon Faktörü (Variance Inflation Factor (VIF)):** VIF, her bağımsız değişkenin varyansının diğer bağımsız değişkenlerle nasıl arttığını gösteren bir istatistiksel ölçüdür. VIF değeri 10'dan büyükse, çoklu doğrusal bağlantı sorunu olduğu söylenebilir.

$$VIF = \frac{1}{1 - R^2}$$

- **Koşul İndeksi (Condition Index (CI)):** Koşul indeksi, çoklu doğrusal bağlantının etkisini daha detaylı olarak anlamak için kullanılan bir ölçümdür. Bu indeks, regresyon katsayılarındaki varyansın ne kadar arttığını göstermektedir. Yüksek koşul indeksi değerleri, Çoklu doğrusal bağlantının regresyon katsayılarına etkisinin daha fazla olduğunu ve regresyon tahminlerinin daha az güvenilir olabileceğini göstermektedir.

$$CI = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{\frac{en\ büyük\ özdeğer}{en\ küçük\ özdeğer}}$$

$$(\lambda_{max} = \lambda_1 > \lambda_2 > \cdots > \lambda_k = \lambda_{min})$$

- **Tolerans:** Tolerans, VIF'in tersidir ve bir bağımsız değişkenin diğer bağımsız değişkenlerle olan ilişkisini göstermektedir. Tolerans değeri, 0.1'den küçükse çoklu doğrusal bağlantı sorunu olabilir.

Çoklu doğrusal bağlantının tespit edilmesinden sonra ele alınabilecek yaygın yaklaşımlar şunlardır:

- **Değişken Seçimi:** Aralarındaki korelasyon yüksek olan bazı bağımsız değişkenleri modelden çıkarmak, çoklu doğrusal bağlantıyı azaltabilir.
- **Değişken Dönüşümleri:** Değişkenleri farklı dönüşümlere tabi tutarak, ilişkileri düzeltmek ve çoklu doğrusal bağlantıyı azaltmak mümkün olabilir.
- **Yeniden Ölçeklendirme:** Bağımsız değişkenlerin ölçeğini standartlaştırmak veya normalleştirmek, çoklu doğrusal bağlantı etkisini azaltmaya yardımcı olabilir.
- **Yanlı Tahmin Yöntemlerinin Kullanılması:** Ridge ve LASSO gibi düzenlileştirme teknikleri, çoklu doğrusal bağlantının etkisini azaltabilir ve regresyon modelinin performansını iyileştirebilir.

2.1.1.1.3. Değişen Varyans

Değişen varyans (heteroskedastisite), hata terimlerinin farklı gözlemler veya bağımsız değişkenler seviyeleri için farklı varyanslara sahip olduğu bir durumu ifade etmektedir (Basu, 2005: 501-502).

Değişen varyans nedenlerini aşağıdaki gibi ifade etmek mümkündür (Albayrak, 2008: 113):

- Bağımsız değişken değerleri değiştikçe, varyansın da değişmesi sonucu değişen varyans ortaya çıkabilir.
- Bağımlı değişkenin kesikli olması durumunda ortalama ve varyansın ilişkili olması değişen varyansa yol açabilir.
- Değişen varyansın bir diğer nedeni, modelleme hatasıdır. Farklı bir model tahmini veya bazı önemli değişkenlerin model dışı bırakılması değişen varyansın kaynağı olabilir.
- Değişen varyans, aykırı değerlerin örneklemdeki diğer gözlemlerden, çok küçük veya çok büyük olması sonucunda da görülebilir. Bu gözlemlerin atılması özellikle örneklem büyüklüğü küçükse regresyon çözümlerinin sonuçlarını önemli ölçüde değiştirebilir.
- Veri derleme teknikleri veya ölçme araçları vasıtasıyla varyans küçülebilir.

Değişen varyansı tespit etmek için kullanılan en yaygın yöntemler aşağıda verilmiştir;

- **Artık Grafiği:** Artıklar, gözlem değerleri ile regresyon modelinden tahmin edilen değerler arasındaki farklardır. Eğer artık grafiğinde düzensiz bir dağılım varsa ve artıkların bağımsız değişkenlerin değerlerine bağlı olarak genişleyen bir deseni varsa, değişen varyansın varlığından şüphelenilebilir.
- **Breusch-Pagan Testi:** Breusch-Pagan testi, artıkların bağımsız değişkenlerle olan ilişkisini analiz etmektedir. Eğer değişen varyans varsa, regresyon modelinde kullanılan bağımsız değişkenlerle artıklar arasında anlamlı bir ilişki olduğu tespit edilmektedir.
- **White Testi:** Bağımsız değişkenler ile hataların kareleri arasındaki ilişkiyi inceleyen bir istatistiksel testtir.
- **Goldfeld-Quandt Testi:** Goldfeld-Quandt testi, veri setini iki alt gruba bölmekte ve bu alt gruplardaki artıkların varyanslarını karşılaştırmaktadır. Eğer alt gruplar arasında anlamlı bir fark varsa, değişen varyansın varlığından söz edilebilir.

Değişen varyansın tespit edilmesinden sonra ele alınabilecek yaygın yaklaşımlar şunlardır:

- **Ağırlıklı Regresyon:** Değişen varyansın varlığı durumunda, gözlemlerin ağırlıklı regresyon ile analiz edilmesi bir seçenek olabilir. Ağırlıklı regresyon, değişen

varyansın olduğu bölgelerde daha düşük ağırlıklar kullanarak modelin güvenilirliğini artırabilir.

- **Düzeltme:** Bazı regresyon modelleri, değişen varyansı düzelten düzenlemelerle kullanılabilir. Örneğin, düzeltilmiş standart hatalar kullanarak, değişen varyansın etkisini azaltmak mümkündür.
- **Değişken Dönüşümleri:** Değişkenleri uygun dönüşümlere tabi tutmak, değişen varyansı ele almanın başka bir yolu olabilir. Örneğin, logaritmik ya da karekök dönüşümleri, değişkenlerin dağılımlarını düzelterek değişen varyansın etkilerini azaltabilir.

2.1.1.1.4. Otokorelasyon

Otokorelasyon, hata terimlerinin birbirini izleyen değerleri arasında ilişki olması durumunu ifade eder (Ünver ve Gamgam, 1996: 345). Otokorelasyonun neden olduğu durum, hata terimlerinin zaman içinde veya ardışık örnekler arasında benzer desenler sergilemesidir. Otokorelasyonun varlığı, regresyon modelinin güvenilirliğini etkileyebilir ve istatistiksel sonuçların yanıltıcı olmasına yol açabilir.

Otokorelasyon nedenlerini aşağıdaki gibi ifade etmek mümkündür:

- **Örnekleme Hatası:** Verilerin yetersiz veya rastgele seçilmemesi nedeniyle örnekleme hataları ortaya çıkabilir ve bu da otokorelasyona yol açmaktadır.
- **Uygun Model Seçimi:** Yanlış veya eksik regresyon modeli seçimi, otokorelasyonun ortaya çıkmasına neden olmaktadır.
- **Zaman Serisi Verileri:** Zaman serisi verilerinde, bir gözlem bir önceki gözlemle korelasyon içerebilir ve bu durum otokorelasyona yol açmaktadır.

Otokorelasyonu tespit etmek için kullanılan en yaygın yöntemler aşağıda verilmiştir:

- **Artık Grafiği:** Artık grafiğinde, düzensiz bir dağılım varsa ve artıkların birbiriyle ilişkili bir deseni varsa otokorelasyon olabilir.
- **Durbin-Watson İstatistiği:** Durbin-Watson istatistiği, artıklar arasındaki ilişkinin derecesini ölçmektedir. Durbin-Watson istatistiği değeri, 0 ile 4 arasında değişmektedir. Bu değer 2'ye yaklaştıkça otokorelasyonun olmadığını, 0'a yaklaştıkça pozitif otokorelasyonun olduğunu ve 4'e yaklaştıkça ise negatif otokorelasyonun olduğunu ifade etmektedir.

Otokorelasyonun tespit edilmesinden sonra ele alınabilecek yaygın yaklaşımlar şunlardır:

- **Model Düzeltmeleri:** Otokorelasyonu azaltmak için daha uygun bir regresyon modeli seçilebilir veya modelde ek düzenlemeler yapılabilir.
- **Otoregresif Hata Modelleri:** Zaman serisi verilerinde otokorelasyonu ele almak için otoregresif hata modelleri kullanılabilir.
- **Robust Standart Hata:** Otokorelasyonun etkilerini azaltmak için robust standart hatalar kullanılabilir.
- **Örneklem Büyüklüğünü Artırma:** Örneklem büyüklüğünü artırarak otokorelasyonun etkisini azaltmak mümkün olabilir.

2.1.1.2. Regresyon Katsayıları İçin Hipotez Testleri

Katsayı tahmininin istatistiksel olarak anlamlı olup olmadığını test etmek için kullanılan t istatistiği, aşağıdaki gibi ifade edilmektedir (Arkes, 2019: 91):

$$t = \frac{\hat{\beta}}{sh(\hat{\beta})}$$

Burada; $\hat{\beta}$ katsayı tahminini ve $sh(\hat{\beta})$ katsayı tahmininin standart hatasını göstermektedir.

Bir regresyon modelindeki tüm bağımsız değişkenlerin birlikte istatistiksel bir önemi olup olmadığını anlamak için ise F testi kullanılmaktadır.

F testinin nasıl hesaplandığını görmek için aşağıdaki terimlerin tanımlanması gerekmektedir (Gujarati, 2011: 13):

$$Toplam\ Kareler\ Toplamı\ (TKT) = \sum y_i^2 = \sum (Y_i - \bar{Y})^2$$

$$Açıklanan\ Kareler\ Toplamı\ (AKT) = \sum \hat{y}_i^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

$$Kalıntı\ Kareler\ Toplamı\ (KKT) = \sum e_i^2$$

Bu terimler aşağıdaki gibi göstermek mümkündür (Gujarati, 2011: 13):

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$$

$$TKT = AKT + KKT$$

Regresyon modelinin genel anlamlılık sınaması için kullanılan F test istatistiği aşağıdaki gibi ifade edilmektedir (Arkes, 2019: 99);

$$F_{n-k}^{k-1} = \frac{AKT/(k-1)}{KKT/(n-k)}$$

Burada; n örneklem büyüklüğünü, k tahmin edilen regresyon katsayı sayısını ve $(n-k)$, serbestlik derecesini ifade etmektedir.

2.1.1.3. Belirlilik Katsayısı R^2

R^2 ile gösterilen belirlilik katsayısı, oluşturulan regresyon modeli etrafındaki veri noktalarının dağılımını değerlendirmektedir. Başka bir ifadeyle R^2, tahmin edilen regresyon çizgisinin uyum iyiliğinin genel bir ölçüsüdür ve tüm bağımsız değişkenler tarafından açıklanan bağımlı değişkendeki toplam varyasyonun (değişimin) oranını ya da yüzdesini vermektedir.

R^2 aşağıdaki gibi tanımlanmaktadır (Panik, 2005: 692):

$$R^2 = \frac{AKT}{TKT} = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum(Y_i - \bar{Y})^2}$$

$$R^2 = 1 - \frac{KKT}{TKT} = 1 - \frac{\sum e_i^2}{\sum(Y_i - \bar{Y})^2}$$

R^2, 0 ile 1 arasında yer almaktadır. R^2, bağımsız değişken sayısının artan bir fonksiyonudur. Yani, modele bir bağımsız değişken eklendiğinde R^2 değeri artmaktadır (Gujarati, 2011: 44).

Modele eklenen bağımsız değişkenler R^2 değerini arttırdığı için modele dahil edilen bağımsız değişken sayısını açıkça hesaba katan bir R^2 ölçüsünün kullanılması önerilmektedir. Buna da düzeltilmiş R^2 ($\bar{R}^2$) adı verilmektedir. $\bar{R}^2$ aşağıdaki gibi hesaplanmaktadır (Gujarati, 2011: 14):

$$\bar{R}^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k}\right)$$

2.2. Varsayımlardan Sapmalar Durumunda Kullanılan Regresyon Modelleri

Regresyon analizinde varsayımlardan sapmaların tespit edilmesi ve bu sapmaların giderilmesi önemlidir. Bu amaçla, farklı regresyon modelleri geliştirilmiştir. Varsayımlardan sapmaların olduğu durumlarda kullanılan bu alternatif modeller, regresyon analizinin güvenilirliğini artırmaya ve daha doğru sonuçlar elde etmeye yardımcı olmaktadır. Bu alternatif modeller, regresyon analizinin daha güvenilir ve doğru sonuçlar vermesini sağlamaktadır.

2.2.1. Doğrusal Olmayan Regresyon

Doğrusal olmayan regresyon, bir bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki ilişkinin doğrusal olmadığı durumlarda bu ilişkiyi modellemek için kullanılan istatistiksel bir tekniktir (Bates ve Watts, 1988: 32). Doğrusal olmayan regresyon denklemi, doğrusal regresyon denkleminden daha genel bir formdadır ve değişkenler arasındaki özel ilişkiye bağlı olarak birçok farklı fonksiyonel form alabilir. Doğrusal olmayan regresyon denkleminin genel bir biçimini aşağıdaki gibi göstermek mümkündür:

$$y = f(x, \beta) + \varepsilon$$

Burada; y bağımlı değişken, x bağımsız değişken, β tahmin edilecek parametrelerin vektörü, $f(.)$ doğrusal olmayan fonksiyon ve ε hata terimidir.

Doğrusal olmayan regresyon modelleri genellikle matematiksel olarak karmaşık olabilir ve bu nedenle iteratif parametre tahmin yöntemleri kullanılır. Örneğin, Gauss-Newton veya Levenberg-Marquardt gibi yöntemler, modelin parametrelerini tahmin etmek için sıkça kullanılır. Bu yöntemler, modelin karmaşıklığına uygun olarak parametreleri iteratif olarak günceller ve en iyi uyum sağlayan parametre değerlerini bulmaya çalışır (Montgomery vd., 2012: 391).

2.2.2. Ridge Regresyon

Ridge regresyon analizi, Hoerl (1962) tarafından önerilen istatistiksel bir yöntem olup özellikle çoklu doğrusal bağlantı problemlerinde kullanılan bir L2 düzenlileştirme tekniğidir. Ridge regresyon analizindeki temel amaç, regresyon katsayılarını sınırlamaktır. Bu amacı gerçekleştirmek için KKT terimine regresyon katsayılarının karelerinin toplamının bir çarpanı eklenir. Bu çarpan, "ceza" terimi olarak adlandırılır. Bu ekstra terim, regresyon katsayılarının büyüklüğünü kontrol ederek modelin aşırı öğrenmeye eğilimini azaltmaya yardımcı olur (Zou, 2020: 457).

Ridge regresyon denklemi aşağıdaki gibidir (Aktaş ve Yılmaz, 2003: 189):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \lambda\Sigma(\beta_i^2) + \varepsilon$$

Burada; y bağımlı değişken, $x_1, x_2, \ldots, x_p$ bağımsız değişkenler, $\beta_0, \beta_1, \ldots, \beta_p$ regresyon katsayıları ve ε hata terimidir. $i = 1, 2, \ldots, p$'dir. $\lambda \geq 0$ bir ayar parametresidir ve ayrı olarak belirlenir. Ayar parametresi λ, bu iki terimin regresyon katsayısı tahminleri üzerindeki göreli etkisini kontrol eder. $\lambda = 0$ olduğunda, model EKK modeli ile aynı olur, dolayısıyla ceza teriminin bir etkisi olmaz. Ancak, λ sonsuza doğru gittikçe, ceza büyür

ve ridge regresyon katsayı tahminleri sıfıra yaklaşır, ancak $\lambda = \infty$ olmadıkça hiçbir katsayıyı sıfıra ayarlamaz.

Genel olarak, L2 düzenleme tekniği, ilgisiz tahmin edicilerin seçiminden ve çoklu doğrusal bağlantıdan kaynaklanan model karmaşıklığını azaltmaya yardımcı olur. Ridge regresyonu modeldeki tüm p tahmincilerini içerir. λ'daki artış katsayıların büyüklüğünü azaltır ancak hiçbir öngörücüyü nihai modelin dışında bırakmaz. Küçültme cezasının sadece eğim terimine uygulandığına, kesişme noktasına uygulanmadığına dikkat etmek önemlidir.

2.2.3. LASSO Regresyon

LASSO (Least Absolute Shrinkage and Selection Operator) regresyon, çoklu doğrusal bağlantı sorununu ele almak ve değişken seçimi yapmak amacıyla kullanılan L1 düzenlileştirme tekniğidir. Doğrusal regresyonun bir türüdür ve Ridge regresyon gibi, EKK regresyon denkleminde bir ceza terimi eklemek suretiyle çalışır. Ancak, LASSO'nun özelliği katsayıların karelerinin toplamı yerine katsayıların mutlak değerlerinin toplamını kullanmasıdır (James vd., 2013: 219).

LASSO regresyon denklemi aşağıdaki gibidir (Hastie vd., 2009: 28):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \lambda \sum |\beta_j| + \varepsilon$$

Burada; y bağımlı değişken, $x_1, x_2, \ldots, x_p$ bağımsız değişkenler, $\beta_0, \beta_1, \ldots, \beta_p$ regresyon katsayıları, λ ceza terimi ve ε hata terimidir. $j = 1, 2, \ldots, p$'dir.

LASSO regresyonunun amacı, Ridge regresyonunda olduğu gibi KKT ve ceza terimini en aza indirmektir. Ancak, Ridge regresyonunun aksine, LASSO regresyon katsayılarının bazılarını sıfıra indirme eğiliminde olabilir. Bu özellik, etkin bir değişken seçimi yapabilme yeteneği sunar, yani gereksiz veya etkisiz bağımsız değişkenleri modelden çıkararak daha sade ve genelleştirilebilir modeller oluşturmayı amaçlar.

2.2.4. ElasticNet Regresyon

ElasticNet regresyon, EKK regresyon denkleminde hem Ridge regresyonun hem de LASSO regresyonun ceza terimlerini birleştirerek bu iki yöntemin bireysel sınırlamalarını aşmayı hedefler. Ridge regresyonun düzenlemesi, regresyon katsayılarının büyüklüklerini sınırlayarak aşırı öğrenmeyi önlemeye yardımcı olurken, LASSO regresyonun düzenlemesi ise bazı regresyon katsayılarını sıfıra indirerek etkisiz değişkenleri modelden çıkarır. ElasticNet, bu iki yaklaşımın avantajlarını bir araya getirerek daha esnek ve güçlü bir çözüm sunmayı amaçlar.

ElasticNet regresyon denklemi aşağıdaki gibidir (Zou ve Hastie, 2005: 303):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \lambda_1 \Sigma \beta_i^2 + \lambda_2 \sum |\beta_j| + \varepsilon$$

Burada; y bağımlı değişken, $x_1, x_2, \ldots, x_p$ bağımsız değişkenler, $\beta_0, \beta_1, \ldots, \beta_p$ regresyon katsayıları, λ_1 ve λ_2 ceza terimi ve ε hata terimidir. $j = 1, 2, \ldots, p$'dir.

ElasticNet regresyon, Ridge ve LASSO'nun avantajlarını birleştirerek regresyon katsayılarını hem sınırlar hem de seçim yaparak daha geniş bir değişken uzayını ele alır. Bu sayede model hem genelleştirilebilir hem de daha iyi bağımsız değişken seçimi yapabilir.

2.2.5. Temel Bileşenler Regresyon

Temel Bileşenler Regresyon, bağımsız değişkenlerin sayısını azaltmak ve verinin varyansını daha iyi açıklamak için kullanılan bir istatistiksel yöntemdir. Bu yaklaşım, yüksek boyutlu veri setlerindeki fazla değişken sayısının karmaşıklığını ele alarak analizi daha etkili hale getirmeyi amaçlar. Temel bileşenler regresyonu, bağımsız değişkenleri daha az ve birbirinden bağımsız temel bileşenlere dönüştürerek veri boyutunu azaltır.

Temel bileşenler regresyon denklemi aşağıdaki gibidir (James vd., 2013: 379):

$$y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_p Z_p + \varepsilon$$

Burada; y bağımlı değişken, $Z_1, Z_2, \ldots, Z_p$ seçilen temel bileşenler, $\beta_0, \beta_1, \ldots, \beta_p$ regresyon katsayıları ve ε hata terimidir.

Temel bileşenler regresyonda, verinin içerdiği varyansı maksimize eden temel bileşenler oluşturulur. Bu bileşenler, önem sırasına göre sıralanır. İlk bileşen, verinin toplam varyansının en büyük bölümünü açıklar. İkinci bileşen, kalan varyansın en büyük bölümünü açıklar ve bu şekilde devam eder. Bu yöntem, veriyi daha yoğun ve öz şekilde temsil ederek gereksiz gürültüyü azaltır ve daha uygun bir analiz sağlar. Aynı zamanda, çoklu doğrusallık gibi sorunları çözmeye yardımcı olur. Çünkü temel bileşenler, birbirleriyle ilişkisiz olarak oluşturulurlar (Jolliffe, 2002: 167).

2.2.6. Kısmi En Küçük Kareler Regresyonu

Kısmi En Küçük Kareler (KEKK) regresyonu, bağımsız değişkenler arasında yüksek korelasyon ve bağımsız değişken sayısının gözlem sayısını aştığı durumlarda kullanılan bir yöntemdir. Kısmi EKK regresyon denklemi aşağıdaki gibidir (Hastie vd., 2009: 40-41):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Burada; Y bağımlı değişkeni, $X_1, X_2, \dots, X_p$ bağımsız değişkenleri, $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ regresyon katsayılarını ve ε hata terimini ifade etmektedir.

KEKK regresyonu, yüksek boyutlu veri setlerindeki çoklu doğrusallık sorununu azaltmaya ve gereksiz gürültüyü engellemeye yardımcı olur. "Küçük n, büyük p" olarak adlandırılan durumlarda da kullanılır. Bu tür durumlarda, bağımsız değişken sayısı (p), gözlem sayısından (n) daha fazla olabilir. Bu durum, geleneksel regresyon modellerinde aşırı uyum sorununa neden olabilir. Ayrıca yüksek boyutlu verilerdeki karmaşıklığı ele alırken aşırı uyum sorununu azaltmayı ve ilişkili değişkenler arasındaki ilişkiyi dikkate alarak daha güçlü ve anlamlı tahminler elde etmeyi amaçlayan önemli bir istatistiksel yöntemdir.

2.2.7. Genelleştirilmiş En Küçük Kareler Regresyonu

Genelleştirilmiş En Küçük Kareler (GEKK) regresyonu, sıradan EKK yönteminin bir genellemesi olarak, verilerde korelasyonlu hataların veya değişen varyansın varlığı durumunda kullanılan bir yöntemdir. Bu yaklaşım, regresyon katsayılarını ve hataların varyans-kovaryans matrisini aynı anda tahmin ederek, verilerdeki korelasyonları veya değişen varyansları düzelten bir yöntem olarak kullanılır. GEKK tahmincisi, ağırlıklı kalıntı kareler toplamını minimize ederek elde edilir (Kariya ve Kurata, 2004: 35):

$$min \sum (\varepsilon_i^T \Sigma^{-1} \varepsilon_i)$$

Burada; ε_i gözlenen değerlerin tahmin edilen değerlerden sapmalarının bir vektörü, ε_i^T, ε_i'nin transpozu, Σ hataların varyans-kovaryans matrisi ve Σ^{-1} bunun tersidir.

Genelleştirilmiş EKK regresyonu, sıradan EKK yönteminin sınırlamalarını aşarak korelasyonlu hatalar veya değişen varyans gibi gerçek veri özelliklerini ele alarak daha hassas ve güvenilir tahminler elde etmeyi amaçlayan bir istatistiksel yöntemdir.

2.2.8. Ağırlıklandırılmış En Küçük Kareler Regresyonu

Ağırlıklandırılmış En Küçük Kareler (AEKK) regresyonu, temel olarak EKK yöntemiyle benzer bir şekilde bağımsız değişkenlerin bağımlı değişkeni açıklamak için kullanılan bir regresyon analizi türüdür. Ancak, AEKK regresyonu, farklı örneklem birimlerinin verilerindeki önem farklılıklarını dikkate almak üzere bir ağırlıklandırma faktörü kullanır. AEKK tahmincisi, ağırlıklı kalıntı kareler toplamını minimize ederek elde edilir (Greene, 2003: 225):

$$min \sum (w_i (y_i - \beta_0 - \beta_1 x_i)^2)$$

Burada; w_i i. gözleme atanan ağırlık, y_i bağımlı değişken, x_i bağımsız değişken ve β_0 ve β_1 sırasıyla kesişim ve eğim katsayılarıdır.

Ağırlıklar genellikle her bir gözlem için hata teriminin tahmini varyansının tersi şeklinde seçilir (Greene, 2003: 226):

$$w_i = 1/\sigma_i^2$$

Burada; σ_i^2, i gözlem için hata teriminin tahmini varyansıdır. Tahmini varyans, bağımsız değişkenler üzerindeki kalıntı karelerin ön regresyonundan elde edilmektedir.

2.2.9. Zaman Serisi Regresyonu

Zaman Serisi Regresyonu, bir bağımlı değişken ile zaman içinde gözlenen bir veya daha fazla bağımsız değişken arasındaki ilişkiyi anlamak ve modellemek için kullanılan bir yöntemdir. Bağımsız değişkenlerin gözlenen değerlerini kullanarak bağımlı değişkenin gelecekteki değerlerini tahmin etmek ve değişkenler arasındaki ilişkinin büyüklüğünü ve yönünü anlamak için kullanılır. Bir zaman serisi regresyon modelinin genel formu aşağıdaki gibidir (Chatfield, 2019: 19):

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \varepsilon_t$$

Burada; Y_t, t zamanındaki bağımlı değişken, β_0 kesişim terimi, $\beta_1, \beta_2, \ldots, \beta_k$ sırasıyla $X_{1t}, X_{2t}, \ldots, X_{kt}$ bağımsız değişkenleri için regresyon katsayıları ve ε_t, t zamanındaki hata terimidir.

2.2.10. Robust Regresyon

Robust regresyon, aykırı değerler veya etkili gözlemler gibi verilerin sıradan regresyon analizinin sonuçlarını çarpıtabileceği durumlarda değişkenler arasındaki ilişkiyi modellemek için kullanılan bir istatistiksel yöntemdir (Wilcox, 2017: 517-518).

Robust regresyon teknikleri, regresyon modelinin uyumuna olan etkisini azaltarak aykırı değerlerden daha az etkilenmesini sağlamak amacıyla geliştirilmiştir. Bu teknikler arasında En Küçük Mutlak Sapma, Theil-Sen, En Küçük Medyan Kareler, En Az Kırpılmış Mutlak Değer ve M Regresyon gibi farklı yöntemler bulunmaktadır.

2.2.10.1. En Küçük Mutlak Sapma Regresyonu

En Küçük Mutlak Sapma (Least Absolute Deviation (LAD)) regresyonu, bir regresyon modelinin gözlenen bağımlı değişken ile tahmin edilen bağımlı değişken arasındaki mutlak artıkların toplamını en aza indiren bir tür robust regresyon yöntemidir. Bu yöntem, verilerdeki aykırı değerlere karşı hassasiyeti yüksek olan EKK regresyonuna sağlam bir alternatif sunar. LAD regresyonunun temel amacı, mutlak artıkların toplamını

minimize ederek regresyon katsayılarını elde etmektir. Bu amacı gerçekleştirmek için aşağıdaki denklem kullanılır (Powell, 1984: 305):

$$\min_{\beta} \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right|$$

$$(i = 1, \dots, n \; ve \; j = 0, \dots, p-1)$$

Burada; y_i, bağımlı değişken vektörü, x_{ij} bağımsız değişken vektörü, β_j regresyon katsayısıdır.

LAD regresyonu, EKK'dan farklı olarak, mutlak artıkların toplamını minimize ederek aykırı değerlere karşı daha dayanıklı bir yaklaşım sunar. Bu özelliği sayesinde verilerdeki anormallikleri dikkate alarak daha sağlam ve güvenilir tahminler elde etmek mümkün hale gelir.

2.2.10.2. Theil-Sen Regresyon

Theil-Sen tahmincisi, ilk olarak Theil (1950) tarafından önerilmiş ve daha sonra Sen (1968) tarafından genişletilmiş bir regresyon yöntemidir. Geleneksel parametrik regresyon yöntemlerine bir alternatif olarak geliştirilmiş olan bu yöntem, iki değişken arasındaki doğrusal ilişkinin eğimini tahmin etmek için kullanılan parametrik olmayan bir yaklaşımdır (Sen, 1968: 1379).

Theil-Sen tahmincisi, veri noktaları arasındaki eğilimi yakalamak için tüm veri noktalarının çiftlerinin eğimlerini hesaplayıp bu eğimlerin medyanını alır. Bunu aşağıdaki gibi ifade etmek mümkündür (Sen, 1968: 1380):

$$medyan\left(\frac{y_i - y_j}{x_i - x_j}\right)$$

Burada, i ve j olası tüm veri noktası çiftlerini kapsayan indekslerdir.

Theil-Sen regresyonu, değişkenler arasındaki ilişkinin normallik ve doğrusallık varsayımlarının geçerli olmadığı durumlarda özellikle kullanışlıdır. Verilerin aykırı değerler içerdiği veya değişkenler arasında doğrusal olmayan ilişkilerin olduğu senaryolarda özellikle etkili sonuçlar sağlar. Ayrıca, bu yöntem, etkili noktaların varlığına karşı EKK yöntemine göre daha dayanıklıdır.

2.2.10.3. En Küçük Medyan Kareler Regresyon

En Küçük Medyan Kareler (Least Median Squares - LMS) regresyon tahmincisi, aykırı değerlere ve etkili gözlemlere karşı hassas olan geleneksel yöntemlere alternatif

olarak Rousseeuw (1984) tarafından geliştirilmiştir. Bu yöntem, doğrusal regresyon modelinin katsayılarını tahmin etmek için kullanılan sağlam bir yöntemdir.

LMS tahminleri, aşağıdaki gibi elde edilmektedir (Öztürk, 2003: 37):

$$\min_{\beta} \underset{i}{\text{med}} \left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j \right)^2$$

Burada; y_i, bağımlı değişken vektörü, x_{ij} bağımsız değişken vektörü, β_j regresyon katsayısıdır.

LMS regresyonunda amaç, EKK yöntemlerinde olduğu gibi kalıntı kareler toplamını minimize etmek yerine, kalıntı kareler medyanını minimize etmektir. Bu yaklaşım, özellikle verilerin analiz sonuçlarını önemli düzeyde etkileyebilecek aykırı değerler içerdiği durumlarda son derece kullanışlıdır.

2.2.10.4. En Az Kırpılmış Mutlak Değer Regresyonu

En Az Kırpılmış Mutlak Değer (Least Trimmed Absolute (LTA)) regresyon tahmincisi, doğrusal regresyon modelinin katsayılarını tahmin etmek için kullanılan bir başka sağlam yöntemdir. LTA regresyonunda amaç, en büyük mutlak kalıntılara sahip gözlemlerin belirli bir yüzdesini kırptıktan sonra, geriye kalan mutlak kalıntıların toplamını en aza indirmektir. Bu yöntem, verilerin analiz sonuçlarını önemli düzeyde etkileyebilecek az sayıda etkili gözlem içerdiğinde özellikle kullanışlıdır.

LTA tahminleri, aşağıdaki gibi elde edilmektedir (Türkay, 2004: 94):

$$\min_{\beta} \sum_{i=1}^{k} |r|_{i:n}$$

Burada, $|r|_{1:n} \leq |r|_{2:n} \leq \cdots \leq |r|_{n:n}$ küçükten büyüğe doğru sıralanmış kalıntıların mutlak değerleri ve k toplama dahil edilen mutlak kalıntı değerlerinin sayısıdır. Kırpılacak gözlem sayısı k, genellikle toplam gözlem sayısının %10 veya %20 gibi küçük bir yüzdesi olarak ayarlanır. Bu yaklaşım, verilerdeki aykırı değerlerin ve etkili gözlemlerin etkilerini sınırlayarak daha güvenilir regresyon tahmincileri elde etmek amacıyla kullanılır.

2.2.10.5. M-Regresyon

M-regresyon, regresyon katsayılarının tahmininde aykırı gözlemlerin etkisini azaltmak için bir ağırlıklandırma fonksiyonu kullanan robust regresyon türlerinden biridir. İlk olarak Huber (1973) tarafından, verilerdeki aykırı değerlere karşı hassas olabilen EKK yöntemine sağlam bir alternatif olarak tanıtılmıştır.

M-regresyonunun amacı, gözlenen bağımlı değişken ile regresyon modelinden tahmin edilen bağımlı değişken arasındaki ağırlıklı mutlak artıkların toplamını en aza indirmektir. M-regresyon tahminleri, aşağıdaki gibi elde edilmektedir (Wilcox, 2017: 537):

$$\min_{\beta}\left\{\sum_{i=1}^{n} w_i(|y_i - \beta_0 - \beta_1 x_i|)\right\}$$

Burada; y_i, i. gözlem için bağımlı değişken, x_i bağımsız değişken, β_0 ve β_1 regresyon modelinin kesişim ve eğim katsayıları ve w_i ise i. gözleme atanan ağırlıktır. Her bir gözleme atanmış ağırlık, aykırı gözlemlerin etkisini azaltmak için seçilen bir ağırlıklandırma fonksiyonuna bağlı olarak belirlenir. Yaygın bir ağırlıklandırma fonksiyonu, aşağıdaki gibidir (Wilcox, 2017: 538):

$$w_i = \begin{cases} (1-(r_i/c)^2)^2 & r_i \leq c\ ise \\ 0 & r_i > c\ ise \end{cases}$$

Burada; r_i, i. gözlem için standartlaştırılmış artıktır ve c tahmin edicinin sağlamlık derecesini belirleyen bir ayarlama parametresidir. Standartlaştırılmış artık $r_i = (y_i - \hat{y_i})/s$ olarak tanımlanır; burada, $\hat{y_i}$ bağımlı değişkeninin i. gözlem için tahmin edilen değeridir ve s standart sapmayı göstermektedir.

2.2.11. Kantil Regresyon

Kantil regresyon, bir veya daha fazla bağımsız değişken olduğu durumda, bağımlı değişkenin kantillerini tahmin etmek için kullanılan bir istatistiksel yöntemdir. Geleneksel regresyon yöntemlerinin aksine, bu yöntem bağımlı değişkenin farklı kantil seviyelerindeki dağılımını tahmin eder. Bu, bağımlı değişkenin tüm dağılımı yerine belirli yüzdelik dilimlerine odaklanmayı sağlar (Koenker, 2005: 112).

Kantil regresyonunun denklemi aşağıdaki gibidir (Huang vd., 2017: 1):

$$Q(\tau|X) = \beta_0(\tau) + \beta_1(\tau)X_1 + \beta_2(\tau)X_2 + \cdots \beta_p(\tau)X_p$$

Burada; $Q(\tau|X)$, τ yüzdeliği için tahmin edilen regresyon kantil fonksiyonunu, $\beta_0(\tau), \beta_1(\tau), \ldots, \beta_p(\tau)$, τ yüzdeliği için tahmin edilen regresyon katsayılarını, $X_1, X_2, \ldots, X_p$ ise bağımsız değişkenlerdir.

Kantil regresyonu, farklı yüzdelikler için farklı regresyon katsayılarını tahmin etmek için kullanılır ve bu sayede veri dağılımının farklı kısımlarına odaklanır. Bu yöntem, veri setinde aykırı değerlerin veya değişen varyansın etkilerini azaltmaya yardımcı olabilir ve tahminlerin daha sağlam ve esnek olmasına olanak tanır.

2.3. Genelleştirilmiş Regresyon Modelleri

Genelleştirilmiş regresyon modelleri, normal olmayan hata dağılımlarını veya bağımlı ve bağımsız değişkenler arasındaki doğrusal olmayan ilişkileri analiz etmek amacıyla kullanılan geleneksel doğrusal regresyon modelini genişleten istatistiksel bir model sınıfını ifade eder. Bu modeller, farklı türlerdeki bağımlı değişkeni (nominal, ordinal vb.) modellemek için kullanılabilir.

Genelleştirilmiş regresyon modelleri, daha esnek bir analiz sunarak normal dağılım ve doğrusallık varsayımlarını ihlal eden verileri ele almak için kullanışlıdır. Bu tür modeller, farklı veri türleri ve dağılımlarıyla çalışırken daha gerçekçi sonuçlar elde etmek amacıyla geliştirilmiştir.

2.3.1. Lojistik Regresyon

Lojistik regresyon analizi, genellikle kategorik sınıflar arasındaki ilişkiyi anlamak amacıyla kullanılan istatistiksel bir yöntemdir. Bu yöntem, bağımlı değişkenin kategorik olduğu durumlarda tercih edilir. Çünkü geleneksel regresyon analizi bu tür durumlarda uygun değildir. Özellikle sınıflandırma problemlerinde kullanılan lojistik regresyon, örnekler arasındaki ilişkiyi anlamak ve yeni gözlemleri sınıflandırmak için oldukça kullanışlıdır. Lojistik regresyon analizinin amacı, kategorik sınıflar arasındaki ilişkiyi anlamak ve açıklamaktır. Bu analiz, bağımsız değişkenlerin kategorik sonuçlar üzerindeki etkilerini belirlemek için kullanılır (Çokluk, 2010: 1360-1361).

Lojistik regresyon analizi, bağımsız değişkenlere bağlı olarak log-odds'un lineer bir fonksiyonu olarak ifade edilir. Lojistik regresyon denklemi aşağıdaki gibi gösterilir (Oğuzlar, 2005: 22):

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Burada; p, yani ilgilenilen olayın olasılığı, $x_1, x_2, \ldots, x_k$ bağımsız değişkenler ve $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ regresyon katsayılarıdır.

2.3.2. Probit Regresyon

Probit regresyon analizi, iki kategorik sınıftan oluşan bir bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi modellemek için kullanılan bir regresyon analizi türüdür. Bu yöntem, bağımsız değişkenlerin değerleri göz önüne alındığında, bağımlı değişkenin belirli bir kategoride (örneğin, ikili bir sonuç) olma olasılığını modellemek amacıyla kullanılır. Bu özelliği ile lojistik regresyona benzerlik gösterir. Probit regresyon denklemi aşağıdaki gibidir (Agresti, 2015: 183):

$$\Phi^{-1}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Burada; $\Phi^{-1}(p)$, bağımlı değişkeninin başarı olasılığında değerlendirilen standart normal dağılımın kümülatif dağılım fonksiyonunun tersidir. $x_1, x_2, \ldots, x_k$ bağımsız değişkenler ve $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ değerleri ise bağımsız değişkenler ile p başarı olasılığı arasındaki ilişkinin gücünü ve yönünü temsil eden katsayılardır.

Probit regresyon ve lojistik regresyon birçok açıdan benzerdir, ancak yorumlamaları ve uygulamalarında bazı farklılıklar bulunmaktadır. Örneğin, probit regresyon katsayıları doğrudan odds oranları olarak yorumlanmaz, bu konuda lojistik regresyondan ayrılır. Ayrıca, probit regresyon genellikle verilerin değişen varyans sergilediği durumlarda tercih edilirken, lojistik regresyon hataların sabit varyanslı olduğunu varsayar.

2.3.3. Poisson Regresyon

Poisson regresyonu, sayım değişkenleri ile bağımsız değişkenler arasındaki ilişkiyi modellemek için kullanılan istatistiksel bir yöntemdir. Bu analiz türü, özellikle bir belirli zaman aralığında, mekânda veya hacimde meydana gelen olayların sayısını temsil eden sayım verilerini (bir belirli süre içinde gerçekleşen kazalar veya hastane ziyaretlerinin sayısı gibi) incelemek için kullanılır.

Poisson regresyon modeli, bağımlı değişkenin bir Poisson dağılımını takip ettiği varsayımına dayanır. Bu dağılımın temel özelliği, sabit zaman veya mekânda meydana gelen olayların olasılık dağılımını açıklamasıdır. Poisson dağılımı, ayrık ve pozitif değerler alan bir rassal değişkenin frekans dağılımını modellemek için kullanılır. Poisson regresyon denklemi aşağıdaki gibidir (Deniz, 2005: 61-62):

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Burada; μ sayım değişkeninin beklenen değeri, $x_1, x_2, \ldots, x_k$ bağımsız değişkenler, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ değerleri ise bağımsız değişkenler ile sayım değişkeninin beklenen değeri arasındaki ilişkinin gücünü ve yönünü temsil eden katsayılardır.

Poisson regresyonunun temel varsayımı, bağımlı değişkenin ortalama ve varyansının eşit olmasıdır. Bu durum, “eşit varyanslılık” veya “eşit dağılım” olarak adlandırılır. Bu varsayımın ihlal edildiği durumlarda, negatif binom regresyon gibi alternatif yöntemler tercih edilir.

2.3.4. Negatif Binom Regresyon

Negatif binom regresyonu, aşırı dağılım gösteren yani varyansın ortalamadan büyük olduğu sayım verilerini modellemek için kullanılan bir regresyon analizi türüdür. Bu yöntem, Poisson regresyon modelinin genelleştirilmiş bir versiyonudur ve sayım verilerinin istatistiksel analizinde kullanılır. Negatif binom regresyon denklemi aşağıdaki gibidir (Agresti, 2015: 249):

$$\ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Burada; μ sayım değişkeninin beklenen değeri, $x_1, x_2, \ldots, x_k$ bağımsız değişkenler, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ değerleri ise bağımsız değişkenler ile sayım değişkeninin beklenen değeri arasındaki ilişkinin gücünü ve yönünü temsil eden katsayılardır.

Negatif binom regresyonunun temel özelliği, Poisson regresyonunun varsayımlarını genişletmesidir. Negatif binom regresyonu, sayım verilerinin varyansının ortalamasından farklı olduğu ve aşırı dağıldığı durumları ele alabilir.

2.4. Gelişmiş Regresyon Modelleri

Gelişmiş regresyon modelleri, geleneksel regresyon modellerine kıyasla daha karmaşık ilişkileri modelleme yeteneği sunar. Bu modeller, bağımsız değişkenler arasındaki etkileşimleri, doğrusal olmayan ilişkileri ve çok boyutlu veri yapılarını doğru bir şekilde ele alabilir. Aynı zamanda eksik veriler, aykırı değerler ve diğer veri kalitesi sorunlarına da etkili çözümler sunarlar.

Gelişmiş regresyon modellerinin getirdiği avantajlar aşağıdaki gibi sıralanabilir:

- Daha doğru tahminler ve kestirimler sunarlar, çünkü daha karmaşık ilişkileri ve etkileşimleri modelleyerek gerçek verilerdeki durumları daha iyi yansıtabilirler.
- Geleneksel modellere göre daha karmaşık bağımlılık yapılarını yakalayabilirler. Böylece, gerçek ilişkileri daha uygun bir şekilde açıklayabilirler.
- Eksik verileri işleme kapasiteleri vardır. Veri eksikliğinin analiz sonuçlarına etkisini minimize edebilirler.
- Aykırı değerleri etkili bir şekilde ele alabilirler, böylece anormal değerlerin analiz sonuçlarına olumsuz etkisini azaltabilirler.

Gelişmiş regresyon modelleri, özellikle karmaşık ilişkilerin ve veri yapılarının bulunduğu durumlarda daha iyi sonuçlar elde etmeyi sağlarlar. Bu tür modeller, araştırmacıların gerçek verileri daha etkili bir şekilde anlamalarına ve açıklamalarına yardımcı olabilir.

2.4.1. Makine Öğrenimi Regresyon Modelleri

Makine öğrenimi regresyonu, sürekli değerleri tahmin etmek için kullanılan bir analiz türüdür. Bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi belirlemek ve tahminlerde bulunmak amaçlanır. Geleneksel regresyon analizinden farklı olarak, makine öğrenimi regresyonu, doğrusal olmayan ilişkileri ve karmaşık veri yapılarını ele alabilir. Makine öğrenimi regresyonunda odaklanılan konu, en iyi uyum sağlayan bir modeli oluşturmaktır. Model, veri örüntülerini yakalayarak tahminler yapabilir. Performansı değerlendirmek için ortalama hata kareleri, ortalama mutlak hata veya R^2 gibi ölçütler kullanılabilir.

Makine öğrenimi regresyonunda önemli adımlardan biri bağımsız değişken seçimidir. Bu aşamada, en anlamlı ve etkili bağımsız değişkenleri seçerek model oluşturulur. Aşırı uyumu engellemek için ise hiperparametreler ayarlanır.

2.4.1.1. Doğrusal Regresyon

Makine öğrenimi kapsamında doğrusal regresyon, bir veya daha fazla bağımsız değişkenine dayalı olarak sürekli bir bağımlı değişkeni tahmin etmek için yaygın olarak kullanılan bir denetimli öğrenme algoritmasıdır. Bu algoritma, verilere dayalı olarak en iyi uyum sağlayan doğrusal bir denklemi bulmaya çalışır ve bu denklemi kullanarak tahminler yapar.

Doğrusal regresyon denklemi aşağıdaki gibidir:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Burada; y bağımlı değişkeni, $x_1, x_2, \ldots, x_p$ bağımsız değişkenlerini, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ katsayıları temsil eder ve bu katsayılar, bağımsız değişkenlerin bağımlı değişken üzerindeki etkilerini gösterir. ε hata terimi, tahmin edilen değerler ile gerçek değerler arasındaki farkı ifade eder.

2.4.1.2. Polinom Regresyon

Makine öğreniminde polinom regresyon, bağımsız değişken ile bağımlı değişken arasındaki ilişkiyi n. dereceden bir polinom fonksiyonu olarak modelleyen bir regresyon türüdür. Bu yöntem, doğrusal regresyonun sınırlamalarını aşarak daha karmaşık ilişkileri yakalamak için kullanılır. n. dereceden polinom regresyonu için denklem aşağıdaki gibidir (Özen vd., 2021: 136):

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n$$

Burada; y bağımlı değişkeni, x bağımsız değişkeni ve $b_0, b_1, b_2, \dots, b_n$ katsayıları temsil eder.

Polinom regresyonunun temel amacı, verilere en iyi uyan $b_0, b_1, b_2, \dots, b_n$ katsayılarının değerlerini bulmaktır. Bu genellikle, tahmin edilen değerler ile gerçek değerler arasındaki farkların karelerinin toplamını en aza indirme prensibine dayalı bir en küçük kareler yaklaşımı kullanılarak gerçekleştirilir. Bu sayede veriler arasındaki karmaşık ve doğrusal olmayan ilişkiler daha esnek şekilde tahmin edilir.

2.4.1.3. Destek Vektör Regresyon

Vapnik (1982) tarafından önerilen Destek Vektör Regresyonu (DVR), bir veya daha fazla bağımsız değişkenin bağımlı bir değişkenle olan doğrusal olmayan ilişkilerini modellemek için kullanılan bir makine öğrenimi regresyon algoritmasıdır. DVR, sınıflandırma için kullanılan Destek Vektör Makineleri (DVM) ile aynı temel ilkeleri benimser. Ancak, regresyon analizi yapabilmesi için kayıp fonksiyonunda değişiklikler yapar (Smola ve Schölkopf, 2004: 200):

DVR'nin temel amacı, tahmin edilen ve gerçek değerler arasındaki hatayı en aza indirirken, verileri farklı sınıflara maksimum düzeyde ayıran bir hiper düzlemi daha yüksek boyutlu bir uzayda bulmaktır. Bu hiper düzlem, regresyon analizi için uygun hale getirilmiştir ve bağımsız değişkenler ile bağımlı değişken arasındaki karmaşık ilişkileri modellemek üzere optimize edilmiştir.

2.4.1.4. Karar Ağacı Regresyon

Karar ağacı regresyonu, olası kararların ve bu kararların sonuçlarının bir ağaç benzeri modelini oluşturmayı amaçlayan bir makine öğrenimi regresyon türüdür. Belirli bir denklemi takip etmek yerine, parametrik olmayan bir yaklaşım kullanarak bir ağaç benzeri model oluşturur. Algoritma, verileri seçilen bir bağımsız değişken değerine göre böler, daha küçük alt kümeler oluşturur ve her bir alt kümeye bir regresyon modeli uygular. Yeni veri noktaları için tahminler, ağacın kökten yaprak düğümlere yolculuğu ile gerçekleştirilir. Yaprak düğümlerindeki regresyon modelleri temel alınarak tahminler sağlanır (Nagalla vd., 2017: 476).

Karar ağacı regresyonu, sürekli değişkenlerin tahmin edilmesi için kullanışlıdır ve hem sayısal hem de kategorik verileri işleyebilir. Bağımsız değişkenler ile bağımlı değişken arasında doğrusal olmayan ilişkiler olduğunda özellikle etkilidir. Ancak, ağaç derinleştikçe aşırı uyum riski artar ve küçük veri değişikliklerine duyarlı hale gelir. Bu

nedenle, aşırı uymayı önlemek ve tahmin doğruluğunu artırmak için ağacın derinliği, yaprak başına minimum örnek sayısı ve bölme kriteri gibi algoritma hiperparametreleri ayarlanmalıdır (Uyanık vd., 2020: 7).

2.4.1.5. Gradient Boosting Regresyon

Gradient Boosting Regresyonu, ilk olarak Breiman (1996) tarafından tanıtmış ve uygun bir kayıp fonksiyonu üzerinde bir optimizasyon tekniği olarak temsil edilebileceğini belirtmiştir. Daha sonra, Gradient boosting algoritmasının genişletilmiş bir versiyonu Friedman (2002) tarafından geliştirilmiştir. Bu algoritmanın öğrenme süreci, sağlam bir sınıflandırıcı elde etmek için yeni modellerin sırayla eğitilmesi şeklindedir (Natekin ve Knoll, 2013: 21).

Bir eğitim seti $S = \{x_i, y_i\}_1^N$ verildiğinde, gradient boosting, kayıp fonksiyonu $L(y, F(x))$'i en aza indirerek, x tahmin değişkenlerini kullanarak y bağımlı değişkenlerini bulmayı amaçlar. Gradient boosting karar ağacı, ağırlıklı bir fonksiyon toplamı aracılığıyla $F(x)$'in eklemeli bir yaklaşımını oluşturur:

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x)$$

Burada; ρ_m, yeni adımın (m. adım) öğrenme oranını (learning rate) temsil eder. Öğrenme oranı, yeni adımın önemini ayarlamak için kullanılır. Bu değer, $[0, 1]$ aralığında bir sayıdır ve yeni adımın katkısının ne kadar olacağını belirler. $h_m(x)$, m. adımda eklenen yeni tahmin fonksiyonunu temsil eder. Bu fonksiyonlar topluluktaki karar ağacı modelleridir. Algoritma yaklaşımı yinelemeli olarak gerçekleştirir.

$$F_0(x) = argmin_\alpha \sum_{i=1}^{N} L(y_i, \alpha)$$

Ardışık temel öğrenicileri en aza indirmeyi amaçlar:

$$(\rho_m h_m(x)) = argmin_{\rho,h} \sum_{i=1}^{N} L\big(y_i, F_{m-1}(x_i) + \rho h(x_i)\big)$$

Her h_m yeni bir eğitim seti $D = \{x_i, r_{mi}\}_{1=1}^N$ ile eğitilir; burada, ρ her bir iterasyonda eklenen tahmini fonksiyonun (base learner) ağırlığını, r_{mi} kalıntıları temsil eder (Caruana vd., 2004: 18). Kalıntılar şu şekilde hesaplanır:

$$r_{mi} = \left[\frac{\delta\big(y_i, F(x)\big)}{\delta F(x)}\right]_{F_m(x)=F_{m-1}(x)}$$

Daha sonra, ρ_m değeri bir çizgi arama optimizasyonu gerçekleştirilerek hesaplanır. Bu arada, yinelemeli görev uygun şekilde düzenlenmezse bu algoritma aşırı uyum sağlayabilir (Friedman, 2001: 1190). İkinci dereceden kayıp fonksiyonu gibi belirli kayıp fonksiyonları için, h_m yanlış kalıntılara mükemmel bir şekilde uyarsa, sonraki iterasyonda yanlış kalıntılar sıfır olur ve iterasyon erken sona erer.

Gradient boosting prosedürü aşağıdaki algoritma ile özetlenmiştir:

Tablo 2.1. Gradient Boosting Algoritması

Girdi: eğitim verileri $S = ((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m))$
Bir diferansiyel kayıp fonksiyonu $L(y;\ F(x))$.
İterasyon sayısı T.
Prosedür:
1) Modeli aşağıdakileri kullanarak sabit bir değerle başlatın
$F_m = argmin_\alpha \sum_{i=1}^{N} L(y_i, \alpha)$
2) ***for*** $m = 1, \dots, T$:
(i) Kalıntıları hesaplayın
$r_{mi} = \left[\frac{\delta(y_i, F(x))}{\delta F(x)}\right]_{F_m(x)=F_{m-1}(x)}$, $i = 1, \dots, n$ için
(ii) Eğitim setini kullanarak bir temel öğreniciyi eğitin
$D = \{x_i, r_{mi}\}_{1=1}^{N}$
(iii) Çizgi arama optimizasyonunu gerçekleştirerek ρ_m'yi elde edin.
$(\rho_m h_m(x)) = argmin_{\rho,h} \sum_{i=1}^{N} L\big(y_i, F_{m-1}(x_i) + \rho h(x_i)\big)$
(iv) Modeli güncelleyin:
$F_m(x) = F_{m-1}(x) + \rho_m h_m(x)$
end for
Çıktı: Nihai model $F_m(x)$'i döndürür.

Gradient Boosting Regresyon'un amacı, veri setindeki karmaşık ilişkileri yakalayarak ve tahmin etmek istediğiniz bağımlı değişkeni en iyi şekilde tahmin etmek için bir model oluşturmaktır. Özellikle doğrusal olmayan, etkileşimli ve karmaşık ilişkileri içeren veri setlerinde, basit regresyon modelleri yetersiz kalabilir. Bu tür durumlarda Gradient Boosting Regresyon, daha karmaşık ve esnek bir model oluşturarak veriyi daha iyi anlayabilir ve tahminlerde bulunabilir.

Gradient boostingin ana avantajı, diğer boosting algoritmaları gibi, önceki modelin hatalarını düzeltmek için eğitildiğinden, giriş verilerinden karmaşık kalıpları öğrenebilmesidir. Bununla birlikte, bu algoritma kullanılarak oluşturulan bir model, giriş verileri gürültülü ise aşırı uyum sağlayabilir ve gürültüyü modelleyebilir (Natekin ve Knoll, 2013: 21; Zhang vd., 2019: 3).

2.4.1.6. XGBoost Regresyon

XGBoost algoritması, gradyan artırma çerçevesini kullanan karar ağacı tabanlı bir topluluktur. Sınıflandırma ve regresyon uygulamaları için kullanılan ölçeklenebilir ve son

derece uygun bir algoritmadır. Chen ve Guestrin tarafından 2016 yılında geliştirilmiştir ve gradyan artırma algoritmasına kıyasla çeşitli ilerlemelere sahiptir. Gradyan artırmanın aksine, XGBoost kayıp fonksiyonu aşırı uyumu önleyen bir düzenleme terimi içerir (Li ve Chen, 2020: 1756):

$$L_M\big(F(x_i)\big) = \sum_{i=1}^{n} L\big(y_i, F(x_i)\big) + \sum_{m=1}^{M} \Omega(h_m)$$

Burada; $F(x_i)$, M. iterasyonda i. örneğe ilişkin tahmini, $L(*)$ ise tahmin edilen sınıf ile bağımlı değişkenin gerçek sınıfı arasındaki farkları hesaplayan bir kayıp fonksiyonunu temsil etmektedir. $\Omega(h_m)$, düzenlileştirme terimini ifade eder ve şu şekilde formüle edilir:

$$\Omega(h_m) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$$

Burada; γ karmaşıklık parametresini temsil eder ve bir iç düğümü bölmek için gereken minimum kayıp azaltma kazancını kontrol eder. γ'ye yüksek bir değer atamak daha basit ağaçlara yol açar. Bu arada, T ağaçtaki yaprak sayısını, λ bir ceza parametresini ve ω yaprak düğümlerinin çıkışını gösterir. Bu arada, Gradient boosting karar ağacındaki birinci dereceden türevin aksine, XGBoost'ta amaç fonksiyonunun ikinci dereceden Taylor yaklaşımı kullanılarak aşağıdaki gibi ifade edilir:

$$L_M \approx \sum_{i=1}^{n}\left[g_i f_m(x_i) + \frac{1}{2}h_i f_m^2(x_i)\right] + \Omega(h_m)$$

Burada, g_i ve h_i kayıp fonksiyonunun birinci ve ikinci türevlerini göstermektedir.

XGBoost'a düzenlilik teriminin eklenmesi nedeniyle, aşırı uyuma duyarlı değildir (Liang vd., 2020: 765). Gradyan artırma algoritması gibi, XGBoost da modelin aşırı uyum sağlamasını önlemek için maksimum ağaç derinliği, öğrenme oranı ve alt örnekleme kullanır.

XGBooost algoritmasını kullanmanın bazı avantajları, algoritma bu tür durumların üstesinden gelebildiği için veri normalleştirme ve özellik ölçeklendirme gibi minimum özellik mühendisliği gerektirmesidir. Ayrıca, eksik değerleri işleme kapasitesine sahiptir. Bu algoritma, girdi özelliklerini daha iyi anlamak ve ayrıca bağımsız değişken seçimi yapmak için kullanılabilecek özellik önemi çıktısı verebilir. XGBoost, çoğu makine öğrenimi algoritmasından daha hızlıdır, büyük veri kümelerini işleyebilir ve aşırı uyuma eğilimli değildir. Ayrıca diğer makine öğrenimi algoritmalarından daha iyi performans

gösterir. Bununla birlikte, çok sayıda hiperparametre içermesi ve ayarlamayı zorlaştırması gibi bazı sınırlamaları vardır (Nobre ve Neves, 2019: 3-4).

2.4.2. Bulanık Regresyon

Bulanık regresyon, belirsiz veya kesin olmayan verilere sahip sistemleri modellemek amacıyla bulanık mantığı kullanarak gerçekleştirilen bir regresyon analizi türüdür. Bu yöntem, 1967'de Zadeh tarafından geliştirilen bulanık mantığa dayanır (Zadeh, 1996: 103).

Genel bulanık doğrusal regresyon yöntemi aşağıdaki gibidir (Tanaka vd., 1982: 904):

$$\tilde{Y}_i = \tilde{A}_0 + \tilde{A}_1 X_1 + \cdots + \tilde{A}_j X_{ij}$$

Burada; X_i bağımsız değişkenleri, $\tilde{A}_j$ ise bulanık katsayıları ifade etmektedir. $\tilde{A}_j$ aşağıdaki gibi ifade edilmektedir:

$$\tilde{A}_j = \left(a_j, c_j\right)_L$$

$$\tilde{A}_j = \left(a_j, c_j\right)_R$$

Burada; a_j, $\tilde{A}_j$'nin merkezini ve c_j, merkezden sağa ve sola yayılmayı belirtmektedir. L ve R ise bulanık katsayı $\tilde{A}_j$'nin referans fonksiyonları olarak ifade edilmektedir. $\tilde{A}_j$ katsayısı simetrik olduğu zaman $L(X) = R(X)$ olmaktadır.

$\tilde{A}_j = \left(a_j, c_j\right)$ şeklinde ifade edildiğinde, model aşağıdaki gibi gösterilmektedir:

$$\tilde{Y}_j = (a_0, c_0) + (a_1, c_1)X_1 + \cdots + \left(a_j, c_j\right)X_{ij}$$

$$(i = 1,2, \ldots, n \; ve \; j = 0,1, \ldots, k)$$

$\tilde{Y}_j$'nin beklenen değerlerinin genel üyelik fonksiyonu, Zadeh'in genişleme ilkesi yardımıyla aşağıdaki gibi tanımlanır (Wang ve Tsaur, 2000: 356):

$$\mu_{(Y_j)}(y) = 1 - \frac{|y_i - y|}{e_j}$$

Bulanık regresyon analizi, geleneksel regresyon analizinden daha fazlasını sunar. Bu yöntemde, bağımlı değişken ve bağımsız değişkenler arasındaki ilişkiyi saptamak ve tahmin modelleri oluşturmak için sağ-sol veya alt-üst limitler belirlenir. Bu yaklaşım, Tanaka vd. tarafından 1982'de geliştirilip bulanık küme teorisini temel alarak oluşturulmuştur. Bu yöntem, bulanık modelleri kullanarak doğrusal regresyon analizi uygular. Bulanık doğrusal regresyonun önemli bir avantajı, değişkenin alabileceği en

büyük veya en küçük değerleri tahmin edebilme yeteneğidir. Bu sayede, karar verici değişkenin sınır değerlerini de gözlemleme şansı sağlar.

2.4.3. Panel Veri Regresyon

Panel veri regresyon hem yatay kesit hem de zaman serisi boyutlarını içeren verilerin analiz edilmesi için kullanılan istatistiksel bir yöntemdir. Panel veri regresyonu, aynı bireyler veya varlıklar kümesinin zaman içinde gözlemlenmesiyle gerçekleştirilir ve bu, değişikliklerin ve eğilimlerin analiz edilmesine olanak sağlar. Panel veri regresyonu için denklem aşağıdaki gibidir (Polat ve Kızılkan, 2022: 1986):

$$y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$$

Burada; y_{it} bağımlı değişkeni, x_{it} bağımsız değişkeni, β_0 kesişim terimini, β_1, x_{it} için katsayıyı ve ε_{it} hata terimini ifade etmektedir.

Bu denklem, her bir birey veya varlık için zaman serisi boyunca gözlemlenen bağımlı ve bağımsız değişkenleri kullanarak regresyon analizi yapmayı sağlar. Panel veri regresyonu, bireyler arasındaki farklılıkları ve zaman içindeki değişimleri ele alarak daha kapsamlı ve dinamik sonuçlar elde etmeyi mümkün kılar.

2.4.4. Mekansal Regresyon

Mekânsal regresyon, coğrafi konumla ilişkili değişkenler arasındaki ilişkileri modellemek amacıyla kullanılan bir istatistiksel tekniktir. Bu yöntem, gözlemlerin coğrafi konumunu dikkate alarak analiz yapar ve mekânsal otokorelasyonu, yani yakın coğrafi konumlarda bulunan gözlemlerin birbirine benzerlik gösterme eğilimini regresyon modeline entegre eder. Mekânsal regresyon denkleminin genel formu aşağıdaki gibidir (LeSage ve Pace, 2009: 2);

$$Y = X\beta + \rho W_Y + \varepsilon$$

Burada; Y bağımlı değişken matrisi, X bağımsız değişkenler matrisini, β katsayılar vektörünü, ρ mekânsal otoregresif parametreyi, W_Y mekânsal ağırlıklar matrisini ve ε hata terimini temsil etmektedir. Mekânsal otoregresif parametre ρ verilerdeki mekânsal bağımlılığı yakalamakta ve mekânsal ağırlıklar matrisi W_Y gözlemler arasındaki mekânsal ilişkiyi tanımlamaktadır.

Mekânsal regresyon, coğrafi faktörlerin etkilerini ve mekânsal yapıları içeren veri setlerinde kullanılır. Bu yöntem sayesinde, coğrafi konumun analiz sonuçlarına etkisi göz önünde bulundurularak daha gerçekçi sonuçlar elde edilir. Aynı zamanda, mekânsal

otokorelasyonun varlığı veya mekânsal bağımlılığın etkisi de regresyon modeline dahil edilerek daha doğru sonuçlar elde edilmesi sağlanır.

2.4.5. Dalgacık Regresyon

Dalgacık regresyon, verileri farklı frekans bantlarına ayırmak ve her bant için ayrı bir regresyon modeli oluşturmak amacıyla dalgacık dönüşümlerini kullanan bir regresyon analizi yöntemidir. Bu yaklaşım, özellikle farklı ölçeklerde karmaşık örüntüler ve korelasyonlar sergileyen veri setlerini analiz etmek için etkili bir tekniktir. Dalgacık regresyon için denklem aşağıdaki gibidir (Donoho ve Johnstone, 1995: 1200);

$$y(t) = b_0 + \sum b_j \phi_j(t)$$

Burada; $y(t)$ bağımlı değişken, b_0 sabit terim ve b_j dalgacık katsayılarıdır. $\phi_j(t)$ ise j. dalgacık fonksiyonudur ve t zamanındaki sinyalin j. frekanstaki bileşeni temsil etmektedir.

Dalgacık regresyon yöntemi, verilerin ölçek ve frekans bileşenlerini ayrıştırarak farklı düzeylerdeki yapıları yakalama yeteneği sunar. Her bir dalgacık fonksiyonu, belirli bir frekans bandını temsil eder ve verilerin bu bantlardaki değişiklikleri yakalaması amaçlanır. Bu sayede, verilerin farklı ölçeklerdeki örüntülerini analiz etmek ve karmaşık ilişkileri anlamak daha kolay hale gelir.

3. SİMÜLASYON ÇALIŞMASI

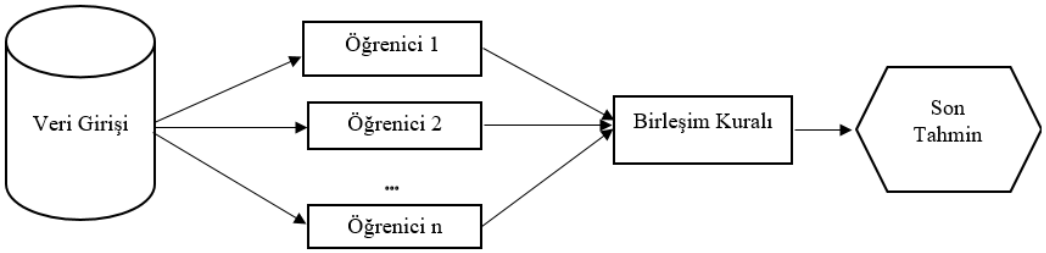
3.1. Topluluk Modeli Yönteminin Oluşturulması

Topluluk öğrenmenin temel çalışmalarından biri olarak ele alınan Dasarathy ve Sheela (1979) çalışmalarında, çoklu bileşen sınıflandırıcıları kullanarak özellik uzayını bölmek için bir yöntem sunmuşlardır. Daha sonra 1990 yılında Hansen ve Salamon, benzer yapay sinir ağı sınıflandırıcılarından oluşan bir topluluğun uygulanmasının tek bir sınıflandırıcıdan daha üstün tahmin performansı sağladığını göstermiştir. Aynı yıl Schapire (1990) zayıf bir sınıflandırıcıyı güçlü bir sınıflandırıcıya dönüştürmek için geliştirilen bir yöntem olan boosting tekniğini önermiş ve bu teknik AdaBoost, Gradient Boosting ve XGBoost gibi günümüzdeki güçlü algoritmaların temelini oluşturmuştur (Polikar, 2012: 4).

Topluluk öğrenmesi, iki veya daha fazla makine öğrenmesi algoritmasını birleştirerek, bunları oluşturan algoritmaların ayrı ayrı kullanılmasına kıyasla daha üstün performans elde etmek için kullanılan bir tekniktir. Tek bir modele güvenmek yerine, bireysel öğrenicilerin tahminlerini birleştirerek daha doğru tahminler elde etmeyi amaçlar.

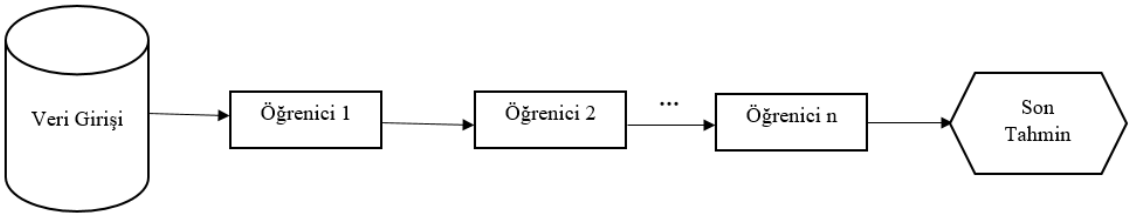
Topluluk modelleri genellikle paralel ve sıralı topluluklar olarak sınıflandırılabilir. Paralel yöntemler farklı temel sınıflandırıcıları bağımsız olarak eğitir ve tahminlerini bir birleştirici kullanarak birleştirir (Liu vd., 2016: 63). Paralel topluluk algoritmaları, topluluk üyelerindeki çeşitliliği teşvik etmek için temel öğrenicilerin paralel üretimini kullanır. Sıralı topluluklar ise temel modellere bağımsız olarak uymaz. Bunun yerine, yinelemeli olarak eğitilirler ve her yinelemedeki modeller, bir önceki modelin yapılan hatalarını düzeltmeyi öğrenir. Şekil 1 ve 2'de paralel ve sıralı topluluk öğrenimini gösteren diyagramlar yer almaktadır:

Şekil 3.1. Paralel Topluluk Öğrenimi Diyagramı



Kaynak: Mienye ve Sun, 2022: 99131.

Şekil 3.2. Sıralı Topluluk Öğrenimi Diyagramı



Kaynak: Mienye ve Sun, 2022: 99131.

Çalışmanın ana amacı, aykırı değerler ve çoklu doğrusal bağlantı sorunlarına karşı daha dirençli sonuçlar elde etmektir. Bu amaç doğrultusunda, Ridge Regresyon, Robust Regresyon ve Gradient Boosting Regresyon modellerini bir araya getiren ve üçlü bir birleşim olan "Ridge-Robust-Boosting Topluluk Regresyon" modeli önerilmiştir. Bu model, bu üç farklı regresyon modelinin avantajlarını birleştirerek daha güçlü ve dirençli bir regresyon analizi yaklaşımı sunmaktadır.

Ridge-Robust-Boosting Topluluk Regresyon modelinin avantajları aşağıdaki şekildedir:

- ***Daha İyi Performans:*** Ridge-Robust-Boosting Topluluk Regresyon modeli, farklı temel modellerin güçlü yönlerini birleştirerek daha yüksek tahmin performansı sergiler. Ridge regresyon, çoklu doğrusal bağlantıyı etkili bir şekilde ele alarak modelin istikrarını artırırken, Robust Regresyon aykırı değerlere karşı dayanıklılık sağlar. Gradient boosting regresyon ise veri setindeki karmaşık ilişkileri yakalamak için uygundur. Bu üç farklı yaklaşımın kombinasyonu, veri setindeki çeşitli zorlukları daha iyi ele almayı mümkün kılar.
- ***Daha İyi Genelleme:*** Ridge-Robust-Boosting Topluluk Regresyon modeli, farklı modellerin tahminlerini birleştirerek daha iyi bir genelleme yeteneği sağlar. Bu, modelin eğitim veri seti dışındaki yeni verilere daha iyi adapte olmasını ve gerçek veriler üzerinde daha güvenilir tahminler yapmasını sağlar.
- ***Daha İyi Esneklik:*** Topluluk modeli, farklı modellerin güçlerini birleştirerek daha esnek bir yapı sunar. Bu, modelin farklı veri setlerine ve değişen koşullara daha iyi uyum sağlamasına olanak tanır. Model, özel gereksinimlere uyacak şekilde özelleştirilebilir ve optimize edilebilir.
- ***Daha İyi Güvenilirlik:*** Ridge-Robust-Boosting modeli, farklı modellerin tahminlerini birleştirerek daha güvenilir tahminler sunar. Çünkü farklı modellerin

tahminlerinin birleşimi, her bir modelin tek başına yapabileceğinden daha düşük hata payına sahip olma eğilimindedir. Bu da modelin güvenilirlik düzeyini artırır.

Ridge-Robust-Boosting Topluluk Regresyon modelinde kullanılan paketler ve hiperparametreler aşağıdaki gibidir:

3.2. Kodlama Aşamaları

Verilerin Oluşturulması: Monte Carlo Simülasyonun Gibbs Algoritması ile veri üretildi.

- Veri setine %20, %30 ve %40 oranlarında aykırı değerler ve düşük (0.3), orta (0.6) ve yüksek (0.9) düzeyde çoklu doğrusal bağlantı eklendi.

Gibbs Örnekleme Algoritması, çok boyutlu problemlerde yaygın olarak kullanılan bir Markov Zinciri Monte Carlo (MCMC) algoritmasıdır ve Geman ve Geman (1984) tarafından geliştirilmiştir. Bu algoritma, özellikle çok boyutlu veri analizinde ve istatistiksel çıkarımlarda yaygın olarak kullanılır. Temel fikri, birçok parametrenin olduğu karmaşık bir ortak olasılık dağılımından örnek çekmek yerine, daha düşük boyutlu koşullu dağılımlardan örnek çekmektir. Bu nedenle, Gibbs Örnekleme Algoritması, diğer değişkenler sabitken bir değişkenin koşullu dağılımından örnek çekmeyi içerir.

Eğitim ve Test Serisi: Veri seti %80 eğitim ve %20 test serisi olarak bölündü.

- Eğitim serisi, bir modelin eğitildiği ve öğrendiği veri parçasını ifade eder. Bu veri parçası, modelin içsel yapısını ve özelliklerini anlaması için kullanılır. Eğitim serisindeki veriler, modelin parametrelerini belirlemek ve ilişkileri öğrenmek için kullanılır. Model, eğitim serisindeki verilere göre ayarlanır ve optimize edilir.
- Test serisi, eğitilen modelin performansını değerlendirmek ve genelleme yeteneğini test etmek için kullanılan veri parçasını ifade eder. Bu veri parçası, modelin daha önce görmediği verileri içerir ve modelin bu verilere nasıl tepki verdiğini ölçmek için kullanılır. Test serisindeki veriler, modelin gerçek dünya verileriyle ne kadar iyi çalıştığını değerlendirmek için kullanılır.

Hiperparametreleri Ayarlama: Grid Search (Izgara Arama) yöntemi ile en iyi hiperparametreler belirlenmeye çalışıldı.

Grid search yöntemi, belirli bir aralıkta bulunan hiperparametrelerin tüm olası kombinasyonlarını deneyerek, en iyi performansı sağlayan hiperparametreleri belirlemeye çalışır. Bu, manuel deneme yanılma sürecini azaltarak, modelin optimize edilmiş haliyle elde edilmesini sağlar.

- Ridge Regresyon → ***alpha değeri:*** Alpha, düzenlemeyi kontrol eder ve 0'a yaklaştıkça düzenleme etkisi azalır. Bu, her özellik için bir miktar düzenleme uygulayarak aşırı uyum riskini azaltır. ***lambda değeri:*** Lambda, düzenlemenin miktarını kontrol eder. Lambda değeri ne kadar büyükse, regresyon katsayıları o kadar kısıtlanır. Grid search ile birden fazla lambda değeri denemek, modelin aşırı uyum (overfitting) ve düşük uyum (underfitting) arasında bir denge bulmasına yardımcı olur.
- Robust Regresyon → ***method değeri:*** (Minimum Volume Elipsoid) Veri noktalarını içine alan ve hacmi mümkün olduğunca küçük olan bir elipsoid çizerek aykırı değerlere dayanıklı model oluşturur.
- Gradient Boosting Regresyon → ***n.trees değeri:*** Gradient boosting ağaçlarının sayısını belirler. Daha fazla ağaç, modelin karmaşıklığını artırabilir, ancak aynı zamanda aşırı uyuma yol açabilir. Grid araması ile farklı ağaç sayıları denemek, modelin en iyi performansı elde etmesine yardımcı olur. ***interaction.depth değeri:*** Her ağaç için en fazla kaç düğüm (node) katmanına sahip olunacağını belirler. Daha derin ağaçlar, veriyi daha ayrıntılı bir şekilde öğrenebilir, ancak aynı zamanda aşırı uyuma yol açabilir. Yine, grid araması ile farklı derinlik değerleri denemek önemlidir. ***shrinkage değeri:*** Her ağacın katkısını düzenlemeye yardımcı olacak bir faktörü kontrol eder. Daha düşük bir shrinkage değeri, her ağacın daha az ağırlığına sahip olacağı anlamına gelir. Düşük shrinkage, daha fazla ağaç kullanmanın etkisini dengelemeye yardımcı olabilir. ***n.minobsinnode değeri:*** Bir düğümde minimum kaç gözlem olması gerektiğini belirtir. Düşük değerler modelin ayrıntıları öğrenmesine yardımcı olabilir, ancak aynı zamanda aşırı uyuma yol açabilir. Bu değerin optimal bir değeri bulunmalıdır.

Modelleri Eğitme ve Tahmin: Verilerin eğitim aşamasında ve her bir modelin tahmininde Çapraz Doğrulama kullanıldı.

Çapraz Doğrulama, makine öğrenimi modelinin performansını değerlendirmek ve genelleme yeteneğini tahmin etmek için kullanılan bir yöntemdir. Bu yöntem, veri setini k parçaya böler ve her bir parçayı sırayla test verisi olarak kullanırken diğerlerini eğitim verisi olarak kullanmaya yardımcı olur. Bu işlem, k defa tekrarlanır.

Modellerin Ağırlıklarını Elde Etme:

- Öncelikle, her bir katlamada (fold) Ridge Regresyon, Robust Regresyon ve Gradient Boosting Regresyon modellerinin ayrı ayrı tahminlerinin ortalama karesel hatası (MSE) hesaplanır.
- Tüm modellerin tahminlerinin toplam karesel hata (total_mse) hesaplanır.
- Her bir modelin ağırlığını hesaplamak için, o modelin tahmininin ortalama karesel hatasının toplam karesel hataya oranı alınır ve bu değer 1'den çıkarılır. Bu, her model için bir ağırlık faktörü oluşturur.
- Elde edilen ağırlık faktörleri her bir katlama için ayrı ayrı bir diziye (weights) kaydedilir.
- Tüm katlamaların ağırlık faktörlerinin ortalamasını alarak, final_weights adlı bir vektörde birleştirilir. Bu, her modelin genel ağırlığını temsil eder.

Bu adımlar, modellerin K-Fold Çapraz Doğrulama ile elde edilen tahminlerinin performansını dikkate alarak en iyi ağırlık kombinasyonunu bulmaya çalışır. Böylece, her modelin katkısını dengeler ve daha iyi bir tahmin elde edilmesine yardımcı olur.

Ridge-Robust-Boosting Topluluk Regresyon Modeli Oluşturma: Oluşturulan ağırlık faktörleri kullanılarak Ridge Regresyon, Robust Regresyon ve Gradient Boosting Regresyon modellerine ait tahminler ağırlıklı birleştirilir. Yani her bir modelin tahminlerini birleştirmek için ağırlıklı ortalama yöntemi kullanılır.

Ağırlıklı Ortalama Yöntemi, farklı modellerin veya tahminlerin performansını dengelemek ve en iyi sonucu elde etmek için kullanılır. Her bir modelin veya tahminin sonucu, belirlenen ağırlıkla çarpılır ve bu ağırlıklı sonuçlar toplanarak birleştirilir. Ağırlıklar, modellerin veya tahminlerin performansına dayalı olarak atanır. Genellikle, daha iyi performans gösteren modeller veya tahminler daha yüksek ağırlıklarla çarpılırken, daha zayıf performans gösterenler daha düşük ağırlıklarla çarpılır. Bu yöntem sayesinde, farklı modellerin veya tahminlerin farklı güçlü yönleri birleştirilerek daha güvenilir ve genelde daha iyi sonuçlar elde edilebilir.

Performans Değerlendirmesi:

- **MSE (Ortalama Kare Hatası):** Bu ölçüt, gerçek değerler ile tahmin edilen değerler arasındaki farkların karelerinin ortalamasını temsil eder. Düşük MSE değeri, daha iyi bir tahmin performansını ifade eder.

- **RMSE (Ortalama Kare Hatası Karekökü):** MSE'nin karekökü dür. Gerçek ve tahmin edilen değerler arasındaki hata karelerinin ortalama değerinin karekökü alınarak hesaplanır.
- **MAE (Ortalama Mutlak Hata):** Gerçek değerler ile tahmin edilen değerler arasındaki mutlak farkların ortalamasını ifade eder. Kare yerine mutlak değer kullanarak hesaplandığı için büyük hataların etkisi daha dengeli olur. Düşük MAE değeri, daha iyi bir tahmin performansını gösterir.
- **R^2 (Belirlilik Katsayısı):** Gerçek verilerin ne kadar iyi tahmin edildiğini gösterir. 0 ile 1 arasında değer alır. 1'e yakın bir R^2 değeri, tahmin modelinin veriyi iyi açıkladığını ifade eder.

3.3. Simülasyon Sonuçları

Doğrusal Regresyon, Ridge Regresyon, Robust Regresyon, Gradient Boosting Regresyon ve Ridge-Robust-Boosting Topluluk Regresyon modellerinin performanslarının karşılaştırılması amacıyla yapılan analiz sonuçları aşağıdaki gibidir:

Tablo 3.1. Model Performanslarının Karşılaştırılması

Aykırı değer oranı	ÇDB oranı	Regresyon modeli	MSE	RMSE	MAE	R^2
0.2	0.3	Doğrusal Regresyon	26.8399	5.1807	4.3204	0.5121
		Ridge Regresyon	7.760037	2.785684	1.679655	0.6535
		Robust Regresyon	24.42249	4.9419	3.681452	0.5369
		Gradient Boosting Regresyon	7.048867	2.65497	1.585067	0.6852
		Ridge-Robust-Boosting Topluluk Regresyon	4.84638	2.20144	0.569556	0.8943
0.3	0.3	Doğrusal Regresyon	26.31422	5.1297	3.7376	0.4721
		Ridge Regresyon	8.201194	2.863773	1.780463	0.6966
		Robust Regresyon	6.968197	2.639734	1.634661	0.7422
		Gradient Boosting Regresyon	7.260733	2.694575	1.651395	0.7314
		Ridge-Robust-Boosting Topluluk Regresyon	1.120874	1.058713	0.814284	0.9164
0.4	0.3	Doğrusal Regresyon	31.23186	5.58675	3.7376	0.4121
		Ridge Regresyon	8.886158	2.980966	1.801169	0.6029
		Robust Regresyon	6.77806	2.603471	1.546245	0.6971
		Gradient Boosting Regresyon	6.810515	2.609696	1.547999	0.6956
		Ridge-Robust-Boosting Topluluk Regresyon	1.371867	1.171267	0.943827	0.89779
0.2	0.6	Doğrusal Regresyon	29.52289	5.43349	5.01396	0.4944
		Ridge Regresyon	13.25323	3.640499	2.368202	0.60847
		Robust Regresyon	10.65491	3.264186	2.085876	0.68523
		Gradient Boosting Regresyon	10.67055	3.266581	2.088342	0.68477
		Ridge-Robust-Boosting Topluluk Regresyon	1.862164	1.364611	0.898927	0.86393
0.3	0.6	Doğrusal Regresyon	30.6651	5.5376	4.9997	0.4415
		Ridge Regresyon	3.502763	1.871567	0.841495	0.73662

		Robust Regresyon	3.222367	1.795095	0.87334	0.75771
		Gradient Boosting Regresyon	2.74115	1.655642	1.315582	0.80985
		Ridge-Robust-Boosting Topluluk Regresyon	1.433018	1.197087	0.949876	0.89864
		Doğrusal Regresyon	32.54929	5.70519	4.11107	0.3957
		Ridge Regresyon	15.22005	3.901288	2.429865	0.57379
0.4	0.6	Robust Regresyon	14.81448	3.848958	2.340057	0.58515
		Gradient Boosting Regresyon	16.58986	4.073066	2.491982	0.53543
		Ridge-Robust-Boosting Topluluk Regresyon	1.808604	1.344844	0.906015	0.86785
		Doğrusal Regresyon	37.5492	6.12774	4.2685	0.3593
		Ridge Regresyon	8.511809	2.917501	1.71485	0.58349
0.2	0.9	Robust Regresyon	6.66697	2.582048	1.49851	0.67377
		Gradient Boosting Regresyon	6.791796	2.606108	1.50909	0.66766
		Ridge-Robust-Boosting Topluluk Regresyon	1.862164	1.364611	0.89892	0.86393
		Doğrusal Regresyon	40.66516	6.377	4.9979	0.3415
		Ridge Regresyon	12.67465	3.560147	2.32718	0.62962
0.3	0.9	Robust Regresyon	10.50828	3.241648	2.04285	0.69292
		Gradient Boosting Regresyon	10.67724	3.267604	2.05726	0.68799
		Ridge-Robust-Boosting Topluluk Regresyon	1.515991	1.231256	0.99233	0.88705
		Doğrusal Regresyon	43.1352	6.5677	5.0139	0.3216
		Ridge Regresyon	14.43827	3.799773	2.4319	0.60664
0.4	0.9	Robust Regresyon	13.30511	3.647617	2.28724	0.63751
		Gradient Boosting Regresyon	14.54715	3.814072	2.3555	0.60367
		Ridge-Robust-Boosting Topluluk Regresyon	1.399427	1.182974	0.95443	0.89573

Simülasyon sonuçlarına göre, çoklu doğrusallığın düşük (0.3), orta (0.6) ve yüksek (0.9) olduğu senaryolarda, farklı aykırı değer oranları (0.2, 0.3 ve 0.4) dikkate alarak yapılan analizlerde, Ridge-Robust-Boosting Topluluk Regresyon modelinin en etkili performansı gösterdiği görülmüştür. Bu model, diğer regresyon yöntemlerine göre daha düşük MSE, RMSE ve MAE değerleri elde ederek, tahminlerde minimum hata ile daha kesin sonuçlar sunmuştur. Ayrıca, yüksek R^2 değerleri ile verilere güçlü bir şekilde uyum sağlamıştır.

Bununla birlikte, analiz sonuçları doğrusal regresyon modelinin aykırı değer ve çoklu doğrusal bağlantı sorununa karşı hassas olduğunu göstermektedir. Özellikle yüksek aykırı değer oranları ve çoklu doğrusal bağlantı durumlarında, doğrusal regresyon modelinin hata değerleri yüksek ve R^2 değerleri düşüktür. Bu sonuçlar, aykırı değer ve çoklu doğrusal bağlantı sorunu olan veri setlerinde doğrusal regresyonun güvenilir tahminler sunma kapasitesinin sınırlı olduğunu göstermektedir.

4. UYGULAMA

Bu bölümde, aykırı değer ve çoklu doğrusal bağlantı sorununa karşı dirençli sonuçlar elde etmek amacıyla, Ridge regresyon, Robust regresyon ve Gradient Boosting regresyon modellerinin birleşimi ile oluşturulan Ridge-Robust-Boosting Topluluk Regresyon modelinin performansı gerçek veriler ile incelenmiştir.

4.1. Veri Seti ve Yöntem

Bu çalışmada, 24.11.2013 ile 16.07.2023 arasındaki haftalık veriler kullanılarak dolar, euro ve BİST100 endeksinin Bitcoin fiyatları üzerindeki etkisi incelenmiştir. Veri setinde aşağıdaki değişkenler yer almaktadır;

Tablo 4.1. Çalışmada Kullanılan Değişkenler

Değişkenler	Kaynak
Bitcoin Fiyatları	https://tr.investing.com/
Euro Dolar	https://evds2.tcmb.gov.tr/
BİST100	https://tr.investing.com/

Dolar, euro ve BİST100 endeksinin Bitcoin fiyatları üzerindeki etkisini analiz etmek için beş farklı regresyon modeli (Doğrusal Regresyon, Ridge Regresyon, Robust Regresyon, Gradient Boosting Regresyon ve Ridge-Robust-Boosting Topluluk Regresyon) kullanılmıştır. Her modelin performansı, MSE, RMSE, MAE ve R^2 performans kriterleri kullanılarak değerlendirilmiştir. Bu ölçütler, her bir modelin doğruluğunun ve karmaşıklığının kapsamlı bir şekilde değerlendirilmesine olanak tanımaktadır.

4.2. Bulgular

Euro, dolar ve BİST100 endeksinin Bitcoin fiyatları üzerindeki etkisini incelemek amacıyla yapılan Ridge-Robust-Boosting Topluluk Regresyon yöntemi ile Doğrusal Regresyon, Ridge Regresyon, Robust Regresyon ve Gradient Boosting Regresyon modellerine ait performans sonuçları aşağıdaki gibidir;

Tablo 4.2. Bitcoin Değişkenine Etki Eden Faktörlerin İncelenmesi

	MSE	RMSE	MAE	R^2
Doğrusal Regresyon	11.2318	3.3513	2.7376	0.4064
Ridge Regresyon	5.75522	2.399	1.48524	0.69273
Robust Regresyon	6.1109	2.47185	1.81021	0.62593
Gradient Boosting Regresyon	2.51516	1.58592	1.53989	0.74681
Ridge-Robust-Boosting Topluluk Regresyon	1.76308	1.32781	0.22895	0.92743

Analiz sonuçlarına göre, en iyi performansı gösteren modelin Ridge-Robust-Boosting Topluluk Regresyon modeli olduğu görülmektedir. Bu model, diğer modellere göre en düşük MSE, RMSE ve MAE değerlerine sahiptir. Yani, tahminlerinde en düşük hata payına sahip ve daha kesin sonuçlar elde etmektedir. Ayrıca, yüksek bir R^2 değeri (0.92) ile verilere yüksek derecede uyum sağlamaktadır.

Doğrusal regresyon modeli ise bu analizde en düşük performansı göstermektedir. R^2 değeri 0.4064 olduğundan, modelin bağımsız değişkenlerin bağımlı değişkeni açıklamada kısıtlı olduğunu görülmektedir.

Diğer modeller olan Ridge Regresyon, Robust Regresyon ve Gradient Boosting Regresyon modelleri ise ortalama bir performans sergilemektedir. Ancak, Ridge-Robust-Boosting Topluluk Regresyon modeline kıyasla hata ölçütlerinde (MSE, RMSE, MAE) daha yüksek sonuçlar elde etmektedirler ve R^2 değerleri de daha düşüktür. Bu da bu modellerin Ridge-Robust-Boosting Topluluk Regresyon modeli kadar iyi tahminler yapamadığını göstermektedir.

Sonuç olarak, Ridge-Robust-Boosting Topluluk Regresyon modelinin en iyi performansı sergileyen model olduğu görülmektedir. Düşük hata değerleri ve yüksek R^2 değeri, bu modelin diğer regresyon modellerine kıyasla daha iyi tahminler sunduğunu göstermektedir.

5. SONUÇ

Bu çalışmada, çoklu doğrusal bağlantı ve aykırı değer gibi iki önemli istatistiksel sorun ele alınarak, bu sorunlara karşı çözüm önerisi sunmak amaçlanmıştır. Çoklu doğrusal bağlantı, bağımsız değişkenler arasındaki güçlü ilişkileri ifade ederken, aykırı değerler ise diğer verilere göre belirgin şekilde farklı olan veri noktalarını temsil eder. Bu iki sorun, regresyon analizini yanıltıcı hale getirebilir.

Bu bağlamda, Ridge Regresyon, Robust Regresyon ve Gradient Boosting Regresyon modellerini birleşimi ile oluşturulan Ridge-Robust-Boosting Topluluk Regresyon modeli önerilmiştir. Bu model, veri setinde hem çoklu doğrusal bağlantı hem de aykırı değer sorunu olduğu durumlarda daha iyi sonuçlar elde etmeyi amaçlamaktadır.

Çalışma kapsamında, modellerin performanslarını karşılaştırmak amacıyla Monte Carlo simülasyonu kullanılmıştır. Bu simülasyon, Gibbs örnekleme algoritması aracılığıyla rassal bir veri setinin üretilmesini içermektedir. Veri setine farklı çoklu doğrusallık düzeyleri ve aykırı değer oranları eklenerek, modellerin performansları ölçmek üzere MSE, RMSE, MAE ve R^2 ölçütleri kullanılmıştır. Simülasyon sonuçları, Ridge-Robust-Boosting Topluluk Regresyon modelinin diğer regresyon modellerine kıyasla farklı çoklu doğrusallık düzeyleri ve aykırı değer oranlarında daha üstün performans sergilediğini göstermiştir. Bu bulgular, Ridge-Robust-Boosting Topluluk Regresyon modelinin hem çoklu doğrusallık hem de aykırı değerlerin bulunduğu durumlarda daha iyi tahminler yapma yeteneğine sahip olduğunu ortaya koymaktadır.

Gerçek verilerle yapılan ikinci uygulama sonucunda da Ridge-Robust-Boosting Topluluk Regresyon modelinin diğer modellere göre en iyi performansı gösterdiği görülmüştür. Bu sonuçlar, önerilen Ridge-Robust-Boosting Topluluk Regresyon modelinin tahmin gücünü ve etkinliğini doğrulamaktadır.

Gelecekteki çalışmalarda, bu çalışmanın metodolojik yaklaşımını daha da zenginleştirmek için doğrusal olmayan regresyon modellerinin kullanımı incelenebilir. Doğrusal olmayan ilişkilerin daha iyi anlaşılması ve modellemesi, belirli veri setlerinin özelliklerine daha iyi uyan tahminlerin elde edilmesine yardımcı olabilir. Bu tür bir genişleme, regresyon analizi alanında daha kapsamlı bir bakış sunarak gelecekteki araştırmalara rehberlik edebilir. Ayrıca, farklı veri setleri üzerinde yapılan bu tür çalışmalar, regresyon model seçiminde ve analizlerde karar verirken daha fazla esneklik ve bilgi sağlayabilir.

KAYNAKÇA

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Aktaş, C. ve Yılmaz, V. (2003). Çoklu Bağıntılı Modellerde Liu ve Ridge Regresyon Kestiricilerinin Karşılaştırılması. *Anadolu Üniversitesi Bilim ve Teknoloji Dergisi*, 4(2), 189-194.

Albayrak, A. S. (2005). Çoklu Doğrusal Bağlantı Halinde En Küçük Kareler Tekniğinin Alternatifi Yanlı Tahmin Teknikleri ve Bir Uygulama. *ZKÜ Sosyal Bilimler Dergisi*, 1(1), 105-126, 2005.

Albayrak, A. S. (2008). Değişen Varyans Durumunda En Küçük Kareler Tekniğinin Alternatifi Ağırlıklı Regresyon Analizi ve Bir Uygulama. *Afyon Kocatepe Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 10(2), 111-134. https://dergipark.org.tr/tr/pub/akuiibfd/issue/1628/20402.

Arkes, J. (2019). *Regression Analysis A Practical Introduction*. (1. ed.). Routledge, New York.

Basu, A. (2005). Extended Generalized Linear Models: Simultaneous Estimation of Flexible Link and Variance Functions. *The Stata Journal*, 5(4), 501-516. doi:10.1177/1536867X0500500402.

Bates, D. M. ve Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons.

Breiman, L. (1996). Bagging predictors, *Machine Learning*, 24(2), 123-140, doi: 10.1007/BF00058655.

Caruana, R., Niculescu-Mizil, A., Crew, G. ve Ksikes, A. (2004). Ensemble selection from libraries of models, in *International Conference on Machine Learning* (ICML), doi: 10.1145/1015330.1015432.

Chatfield, C. (2019). *The Analysis of Time Series: An Introduction* (7th ed.). CRC Press.

Chen, T. ve Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

Çokluk, Ö. (2010). Lojistik Regresyon Analizi: Kavram ve Uygulama. *Kuram ve Uygulamada Eğitim Bilimleri*, 10(3), 1357-1407.

Dasarathy, B. V. ve Sheela, B. V. (1979). A Composite Classifier System Design: Concepts and Methodology, *IEEE*, 67(5), 708-713, doi: 10.1109/PROC.1979.11321.

Deniz, Ö. (2005). Poisson Regresyon Analizi. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, *4*(7), 59-72.

Donoho, D. L. ve Johnstone, I. M. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association*, 90(432), 1200-1224. https://doi.org/10.1080/01621459.1995.10476626.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine, *Annals Statistics*, 29(5), 1189-1232.

Friedman, J. H. (2002). Stochastic Gradient Boosting, *Computational Statistics & Data Analysis*, 38(4), 367–378, doi: 10.1016/S0167-9473(01)00065-2.

Frisch, R. (1934). *Statistical Confluence Analysis by Means of Complete Regression Systems*, Norway, Institute of Economics Oslo.

Frost, J. (2019). *Regression Analysis An Intuitive Guide for Using and Interpreting Linear Models*. (1. ed.). Statistics By Jim Publishing.

Greene, W. H. (2003). *Econometric Analysis.* New Jersey: Pearson Education.

Gujarati, D. N. (2004). *Basic Econometrics* (4th ed.). McGraw-Hill Companies.

Gujarati, D. N. (2011). *Econometrics by Example*. New York: Palgrave Macmillan.

Hansen, L. K. ve Salamon, P. (1990). Neural Network Ensembles, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10), 993–1001, doi: 1494 10.1109/34.58871.

Hastie, T., Tibshirani, R. ve Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hoerl, A. (1962). Application of Ridge Analysis to Regression Problems. *Chemical Engineering Progress*, 58, 54-59.

Huang, Q., Zhang, H., Chen, J. ve He, M. (2017). Quantile Regression Models and Their Applications: A Review. *Journal of Biomimetics, Biomaterials and Biomedical Engineering*, 8, 354. https://doi.org/10.4172/2155-6180.1000354.

Huber, P.J. (1973) Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Annals of Statistics*, 1, 799-821. http://dx.doi.org/10.1214/aos/1176342503.

James, G., Witten, D., Hastie, T. ve Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag, NewYork.

Kariya, T. ve Kurata, H. (2004). *Generalized Least Squares*. John Wiley & Sons, Ltd.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

LeSage, J. P. ve Pace, R. K. (2009). *Introduction to Spatial Econometrics*. CRC Press.

Li, Y. ve Chen, W. (2020). A Comparative Performance Assessment of Ensemble Learning for Credit Scoring, *Mathematics*, 8(10), 1756, doi: 10.3390/math8101756.

Liang, W., Luo, S., Zhao, G. ve Wu, H. (2020). Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms, *Mathematics*, 8(5), 765, doi: 10.3390/math8050765.

Liu, H., Gegov, A. ve Cocea, M. (2016). Ensemble Learning Approaches, in *Rule Based Systems for Big Data: A Machine Learning Approach*. Switzerland: Springer, 63-73, doi: 10.1007/978-3-319-23696-4_6.

Mienye, I. D. ve Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects, in *IEEE Access*, 10, 99129-99149, doi: 10.1109/ACCESS.2022.3207287.

Montgomery, D. C., Peck, E. A. ve Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). John Wiley & Sons.

Nagalla, R., Pothuganti, P. ve Pawar, D. S. (2017). Analyzing Gap Acceptance Behavior at Unsignalized Intersections Using Support Vector Machines, Decision Tree and Random Forests. *Procedia Computer Science*, 109, 474-481. https://doi.org/10.1016/j.procs.2017.05.312.

Natekin, A. ve Knoll, A. (2013). Gradient Boosting Machines, a Tutorial, *Frontiers Neurorobot*, 7, doi: 10.3389/fnbot.2013.00021.

Nobre, J. ve Neves, R. F. (2019). Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to Trade in the Financial Markets, *Expert Systems with Applications*, 125, 181-194, doi: 10.1016/j.eswa.2019.01.083.

Oğuzlar, A. (2005). Lojistik Regresyon Analizi Yardımıyla Suçlu Profilinin Belirlenmesi. *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi,* 19(1), 21-35.

Özen, N. S., Saraç, S. ve Koyuncu, M. (2021). Covid-19 Vakalarının Makine Öğrenmesi Algoritmaları ile Tahmini: Amerika Birleşik Devletleri Örneği. *Avrupa Bilim ve Teknoloji Dergisi*, 22, 134-139.

Öztürk, L. (2003). Doğrusal Regresyonda Sağlam Kestirim Yöntemleri ve Karşılaştırılmaları. *Yayınlanmamış Doktora Tezi*. Mimar Sinan Üniversitesi, İstanbul.

Panik, M. J. (2005). *Advanced Statistics from an Elementary Point of View*. Elsevier Academic Press.

Polat, B. ve Kızılkan, Ö. (2022). Yenilenebilir Enerji Tüketiminin İşsizlik Üzerindeki Etkisi: OECD Ülkeleri için Örnek Bir Çalışma. *İşletme Araştırmaları Dergisi*, 14(3), 1983-1992.

Polikar, R. (2012). *Ensemble Learning, in Ensemble Machine Learning: Methods and Applications*, C. Zhang Y. Ma, Eds. Boston, MA, USA: Springer, pp. 1–34, doi: 10.1007/978-1-4419-9326-7_1.

Powell, J. L. (1984). Least Absolute Deviations Estimation for the Censored Regression Model. *J Econometrics*, 25, 303-325. http://dx.doi.org/10.1016/0304-4076(84)90004-6.

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, 79, 871–880.

Schapire, R. E. (1990). The Strength of Weak Learnability, *Machine Learning*, 5(2), 197–227, doi: 10.1007/BF00116037.

Sen, P. K. (1968). Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, 63(324), 1379-1389.

Smola, A. J. ve Schölkopf, B. (2004). A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88.

Tanaka, H., Uejima, S. ve Asai, K. (1982). Linear Regression Analysis with Fuzzy Model. *IEEE Transactions on Systems, Man and Cybernetics*, 12(6), 903-907. https://doi.org/10.1109/tsmc.1982.4308925.

Theil, H. (1950). A Rank-Invariant Method of Linear and Polynomial Regression Analysis. *Proceedings of the Koninklijke Nederlandse Akademie Wetenschappen*, 53, 386-392, 521-525, 1397-1412.

Türkay, H. (2004). Doğrusal Regresyon Modellerinin Robust (Dayanıklı) Yöntemlerle Tahmini ve Karşılaştırmalı Uygulamaları. *Yayınlanmamış Doktora Tezi*, İstanbul Üniversitesi, İstanbul.

Uyanık, T., Karatuğ, Ç. ve Arslanoğlu, Y. (2020). Machine Learning Approach to Ship Fuel Consumption: A Case of Container Vessel. *Transportation Research Part D: Transport and Environment*, 84, 102389. https://doi.org/10.1016/j.trd.2020.102389.

Ünver, Ö. ve Gamgam, H. (1996), *Uygulamalı İstatistik Yöntemler*, İkinci Baskı, Siyasal Kitabevi, Ankara.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York.

Wang, H. F. ve Tsaur, R. C. (2000). Resolution of Fuzzy Regression Model. *European Journal of Operational Research*, 124(3), 516-533. https://doi.org/10.1016/S0377-2217(99)00317-3.

Wilcox, R. R. (2017). *Introduction to Robust Estimation and Hypothesis Testing* (4th ed.). Academic Press.

Zadeh, L. A. (1996). Fuzzy Logic-Computing With Words. *IEEE Transactions on Fuzzy Systems*. https://doi.org/10.1109/91.493904.

Zhang, B., Ren, J., Cheng, Y., Wang, B. ve Wei, Z. (2019). Health Data Driven on Continuous Blood Pressure Prediction Based on Gradient Boosting Decision Tree Algorithm, *IEEE Access*, 7, doi: 10.1109/ACCESS.2019.2902217.

Zou, H. (2020). Comment: Ridge Regression-Still Inspiring After 50 Years. *Technometrics*, 62(4), 456-458. https://doi.org/10.1080/00401706.2020.180125.

Zou, H. ve Hastie, T. (2005). Regularization and Variable Selection via the ElasticNet. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.