# Auto MPG Veri Seti ile Elastic Net Regresyon Analizi Raporu

## 1. Giriş ve Proje Tanımı

Bu rapor, otomotiv sektöründe yakıt verimliliği tahmini için geliştirilen Elastic Net regresyon modelini detaylandırmaktadır. Günümüzde, çevresel sürdürülebilirlik ve ekonomik verimlilik hedefleri doğrultusunda, araçların yakıt tüketimi performansı hem tüketiciler hem de üreticiler için kritik bir parametre haline gelmiştir. Bu bağlamda, yakıt verimliliğini doğru ve güvenilir bir şekilde tahmin edebilen modellerin geliştirilmesi, araç tasarımı, üretim planlaması ve pazar stratejilerinin belirlenmesinde büyük önem taşımaktadır. Bu proje, Auto MPG veri setini kullanarak, birden fazla bağımsız değişkenin araç yakıt tüketimi üzerindeki etkisini analiz etmeyi ve bu ilişkileri matematiksel bir modelle ifade etmeyi amaçlamaktadır.

Projenin temel amacı, yakıt tüketimi tahmininde karşılaşılan geleneksel regresyon problemlerine (varsayım ihlalleri, aykırı değerler ve çoklu doğrusallık) sistematik çözümler sunarak, daha sağlam ve genellenebilir bir model oluşturmaktır. Bu doğrultuda, veri ön işleme adımları, aykırı değer tespiti ve temizliği için Local Outlier Factor (LOF) algoritması, regresyon varsayımlarını sağlamak amacıyla kapsamlı veri dönüşüm optimizasyonu ve çoklu doğrusallık sorununu gidermek için Elastic Net regresyonu gibi istatistiksel ve makine öğrenimi teknikleri kullanılmıştır. Bu yaklaşım, modelin tahmin gücünü artırmanın yanı sıra, istatistiksel geçerliliğini de güvence altına almayı hedeflemektedir.

Çalışmada kullanılan Auto MPG veri seti, UCI Machine Learning Repository'den temin edilmiştir [1]. Bu veri seti, 1970'lerin sonları ve 1980'lerin başlarındaki otomobillerin yakıt tüketimi (mil/galon) ve çeşitli teknik özelliklerini içermektedir. Başlangıçta 398 gözlem ve 9 değişkenden oluşan veri seti, ön işleme adımları sonrasında 392 gözleme düşmüştür. Veri setindeki temel değişkenler şunlardır:

- mpg: mil/galon (Bağımlı değişken - yakıt tüketimi)
- cylinders: Silindir sayısı
- displacement: Motor hacmi
- horsepower: Beygir gücü
- weight: Ağırlık
- acceleration: Hızlanma
- model_year: Model yılı
- origin: Menşei ülke
- car_name: Araba adı

Bu rapor, projenin her aşamasını detaylı bir şekilde açıklayarak, uygulanan metodolojileri, elde edilen bulguları ve modelin performansını kapsamlı bir şekilde sunmaktadır.

[1] https://archive.ics.uci.edu/ml/datasets/auto+mpg

## 2. Veri Ön İşleme ve Keşifsel Analiz

### 2.1 Veri Yükleme ve Temizleme

Projenin ilk adımı, Auto MPG veri setinin yüklenmesi ve ön işleme tabi tutulmasıdır. Veri seti, doğrudan UCI Machine Learning Repository'den bir URL aracılığıyla pandas kütüphanesi kullanılarak DataFrame formatında okunmuştur. İlk incelemede, veri setinin yapısı ve eksik değer durumu kontrol edilmiştir. horsepower sütununda '?' olarak temsil edilen eksik değerler olduğu ve bu değerlerin na_values="?" parametresi ile NaN olarak okunması sağlanmıştır. Ayrıca, car_name sütununun benzersiz değerler içerdiği ve origin sütununun kategorik bir değişken olduğu tespit edilmiştir. Regresyon analizi için doğrudan kullanılabilir nitelikte olmadıkları ve model karmaşıklığını arttıracakları için car_name ve origin değişkenleri veri setinden çıkarılmıştır. Eksik değer içeren satırlar (horsepower sütunundaki 6 adet NaN değeri) veri setinden temizlenmiştir. Bu temizlik sonrası veri seti 398 gözlemden 392 gözleme düşmüştür. Son olarak, bağımlı değişken mpg ve bağımsız değişkenler (cylinders, displacement, horsepower, weight, acceleration, model_year) belirlenerek veri seti eğitim ve test setleri olarak %80-%20 oranında ayrılmıştır. Bu ayrım, random_state=123 parametresi kullanılarak tekrarlanabilirliği sağlanmıştır.

### 2.2 Keşifsel Veri Analizi (EDA)

Veri ön işleme adımlarının ardından, veri setinin yapısını ve değişkenler arasındaki ilişkileri anlamak amacıyla kapsamlı bir Keşifsel Veri Analizi (EDA) yapılmıştır. Bu analiz, hem istatistiksel yöntemleri hem de görselleştirme tekniklerini içermektedir.

#### *İstatistiksel Korelasyon Analizi:*

Bağımlı değişken mpg ile bağımsız değişkenler arasındaki ilişkinin gücünü ve yönünü belirlemek amacıyla Pearson Korelasyon Analizi uygulanmıştır. Eğitim veri seti üzerinde yapılan analizler sonucunda elde edilen korelasyon katsayıları ve p-değerleri aşağıda özetlenmiştir (anlamlılık düzeyi $\alpha = 0.05$):

- cylinders: Pearson Korelasyon Katsayısı: -0.78, P-değeri: 0.0000 (Anlamlı)
- displacement: Pearson Korelasyon Katsayısı: -0.82, P-değeri: 0.0000 (Anlamlı)
- horsepower: Pearson Korelasyon Katsayısı: -0.78, P-değeri: 0.0000 (Anlamlı)
- weight: Pearson Korelasyon Katsayısı: -0.84, P-değeri: 0.0000 (Anlamlı)
- acceleration: Pearson Korelasyon Katsayısı: 0.41, P-değeri: 0.0000 (Anlamlı)
- model_year: Pearson Korelasyon Katsayısı: 0.59, P-değeri: 0.0000 (Anlamlı)

Bu sonuçlar, tüm bağımsız değişkenlerin mpg ile istatistiksel olarak anlamlı bir korelasyona sahip olduğunu göstermektedir. Özellikle weight değişkeni, -0.84 ile mpg üzerinde en güçlü negatif ilişkiye sahiptir, bu da aracın ağırlığı arttıkça yakıt tüketiminin azaldığını göstermektedir.

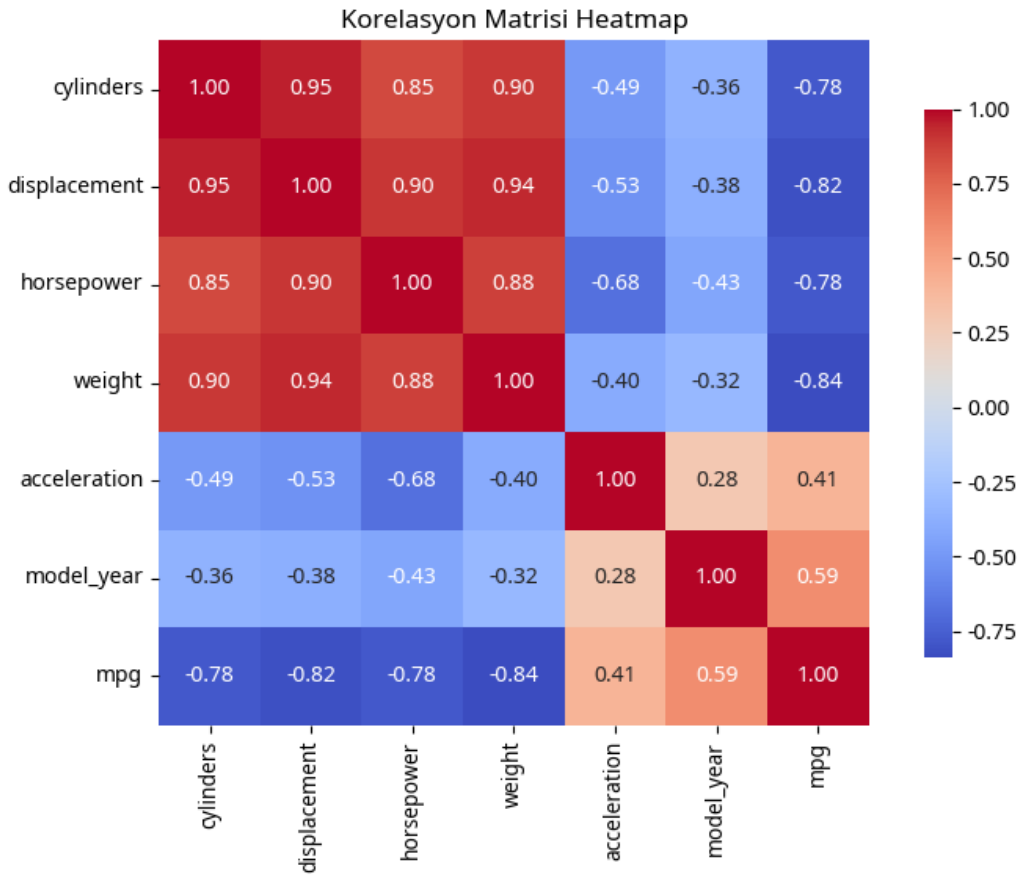
#### *Değişken Dağılımlarının İncelenmesi:*

Bağımsız ve bağımlı değişkenlerin dağılımlarını incelemek amacıyla histogram ve kutu grafikleri kullanılmıştır. Bu grafikler, değişkenlerin çarpıklıklarını, aykırı değerlerini ve genel dağılım şekillerini görselleştirmeye yardımcı olmuştur. Örneğin, weight ve mpg gibi değişkenlerin

dağılımlarının sağa çarpık olduğu gözlemlenmiştir. Bu durum, ilerleyen aşamalarda veri dönüşümlerine ihtiyaç duyulabileceğinin bir göstergesidir.

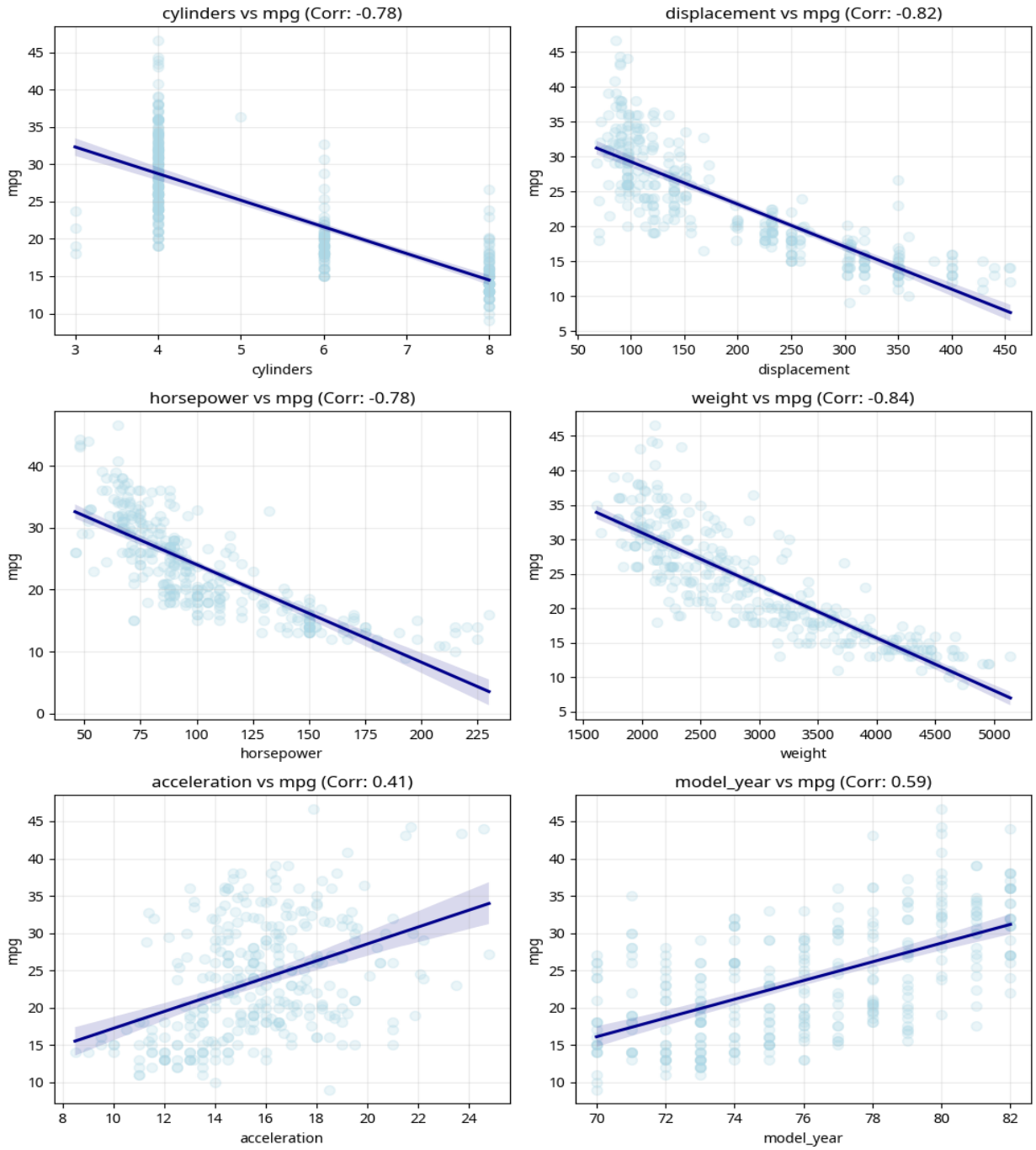
*Görselleştirmeler:*

Korelasyon Matrisi Heatmap:




Değişkenler arasındaki ikili ilişkileri renk yoğunluğu ile gösteren bir ısı haritası. Bu görselleştirme, mpg ile weight arasındaki güçlü negatif ilişkiyi (-0.84) açıkça ortaya koymaktadır.

## Değişkenler Arası Scatter Plot Matrisi:




Her bir bağımsız değişken ile mpg arasındaki ilişkinin dağılımını ve regresyon doğrusunu gösteren dağılım grafikleri. Bu grafikler, özellikle weight ile mpg arasındaki doğrusal negatif ilişkiyi belirgin bir şekilde gözlemlemeyi sağlamıştır.

# 3. İlk Model ve Varsayım Testleri

## 3.1 Baseline Model

Veri ön işleme ve keşifsel analiz adımlarının ardından, herhangi bir ileri dönüşüm veya aykırı değer temizliği uygulanmadan, bağımsız değişkenler ve bağımlı değişken mpg kullanılarak bir başlangıç (baseline) Elastic Net regresyon modeli oluşturulmuştur. Modelin genellenebilirliğini değerlendirmek amacıyla veri seti daha önce belirtildiği gibi %80 eğitim ve %20 test olmak üzere ayrılmıştır. Model, sklearn.linear_model.ElasticNet kullanılarak eğitim veri seti üzerinde kurulmuştur.

Modelin performans metrikleri, eğitim seti üzerinde değerlendirilmiştir:

1. *R-kare ($R^2$):*

   Modelin $R^2$ değeri 0.817 olarak bulunmuştur. Bu, mpg değişkenindeki varyansın yaklaşık %81.7'sinin bağımsız değişkenler tarafından açıklandığı anlamına gelmektedir.

2. *Ortalama Mutlak Hata (MAE):*

   Eğitim seti üzerinde hesaplanan MAE değeri 2.519'dur. Bu, modelin tahminlerinin gerçek mpg değerlerinden ortalama olarak 2.519 birim saptığını ifade eder.

## 3.2 Regresyon Varsayımlarının Testi

Doğrusal regresyon modelinin geçerliliği ve güvenilirliği için artıkların (residuals) belirli istatistiksel varsayımları karşılaması gerekmektedir. Bu varsayımlar: normallik, homoskedastisite (sabit varyanslılık) ve otokorelasyon olmamasıdır. Başlangıç modelinin varsayımları aşağıdaki testler ve görselleştirmelerle incelenmiştir:

1. *Shapiro-Wilk Normallik Testi:*

   Artıkların normal dağılıp dağılmadığını test etmek için Shapiro-Wilk testi uygulanmıştır. Test sonucu (W İstatistiği = 0.97, p değeri = 0.00000) p-değerinin (0.00000) anlamlılık düzeyi olan 0.05'ten küçük olduğunu göstermektedir. Bu durum, istatistiksel olarak artıkların normal dağıldığı anlamlı bir şekilde söylenemez sonucunu ortaya koymaktadır. Yani, artıklar normal dağılım varsayımını ihlal etmektedir.

2. *Breusch-Pagan Homoskedastisite Testi:*

   Artıkların sabit varyansa sahip olup olmadığını test etmek için Breusch-Pagan testi uygulanmıştır. Test sonucu (p değeri = 0.00009) p-değerinin (0.00009) anlamlılık düzeyi olan 0.05'ten küçük olduğunu göstermektedir. Bu durum, istatistiksel olarak artıkların sabit varyansa sahip olduğu anlamlı bir şekilde söylenemez sonucunu ortaya koymaktadır. Yani, artıklar homoskedastisite varsayımını ihlal etmektedir.

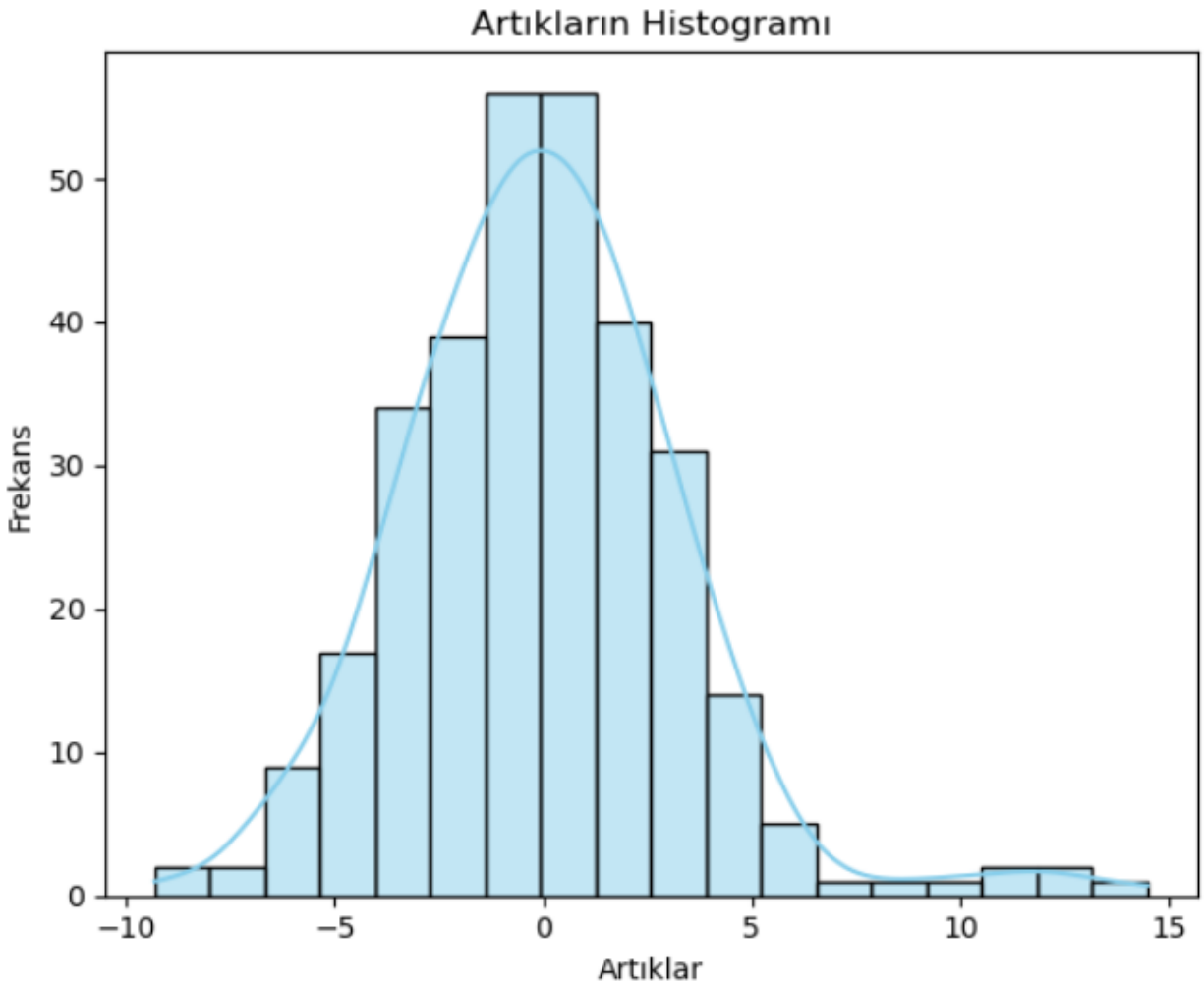
3. *Durbin-Watson Otokorelasyon Testi:*
   Artıklar arasında otokorelasyon (bağımlılık) olup olmadığını test etmek amacıyla Durbin-Watson testi uygulanmıştır. Test sonucu 1.9852 olarak bulunmuştur. Durbin-Watson değeri 2'ye oldukça yakın olduğundan, artıklar arasında anlamlı bir otokorelasyon

bulunmadığı, yani artıkların birbirinden bağımsız olduğu söylenebilir. Bu sonuç, modelin bağımsızlık varsayımını sağladığını göstermektedir.

*Görselleştirmeler:*

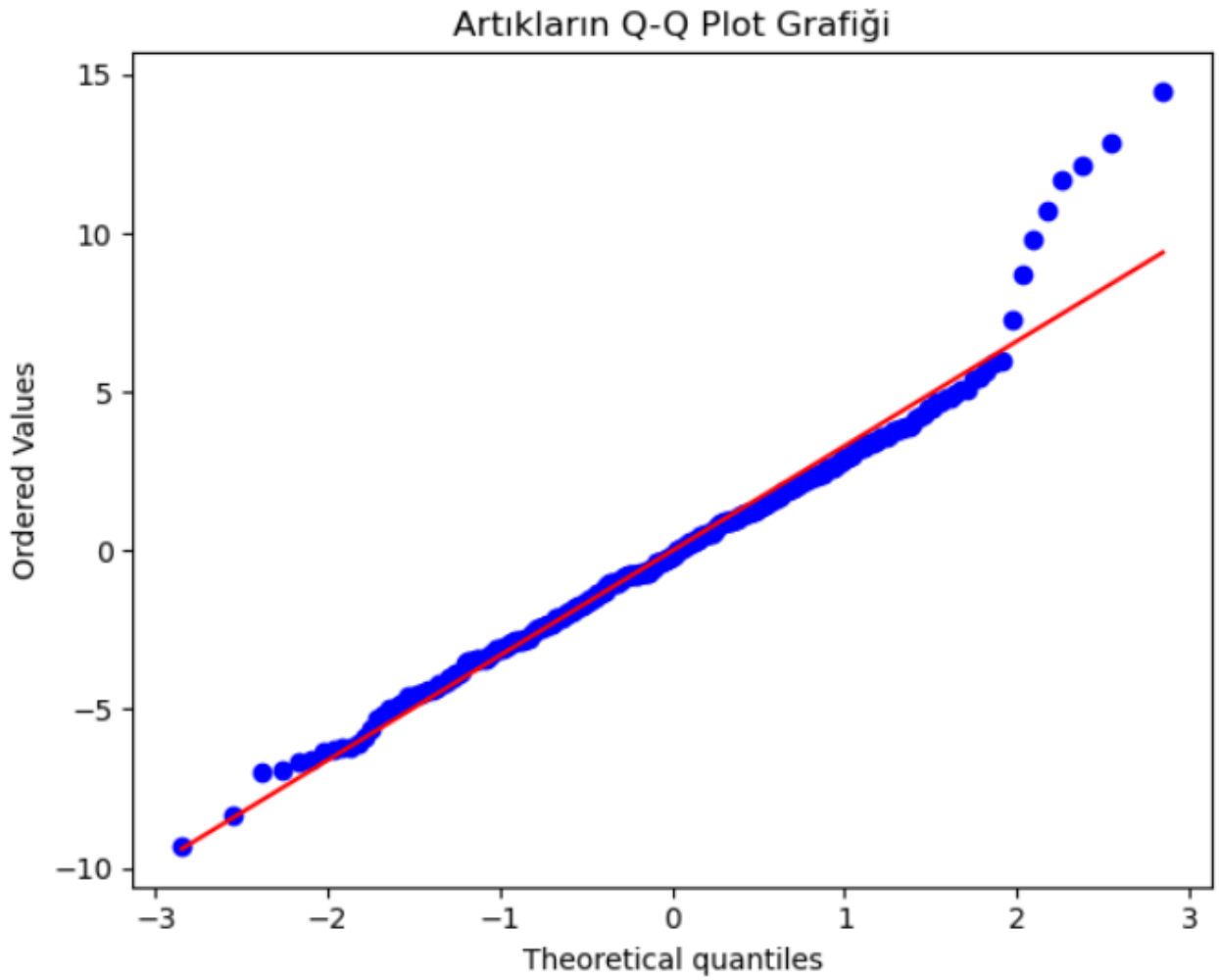
### Artıkların Histogram Grafiği:

Artıkların dağılımını gösteren histogram. Bu grafik, artıkların normal dağılımdan sapmalarını görsel olarak ortaya koymuştur.



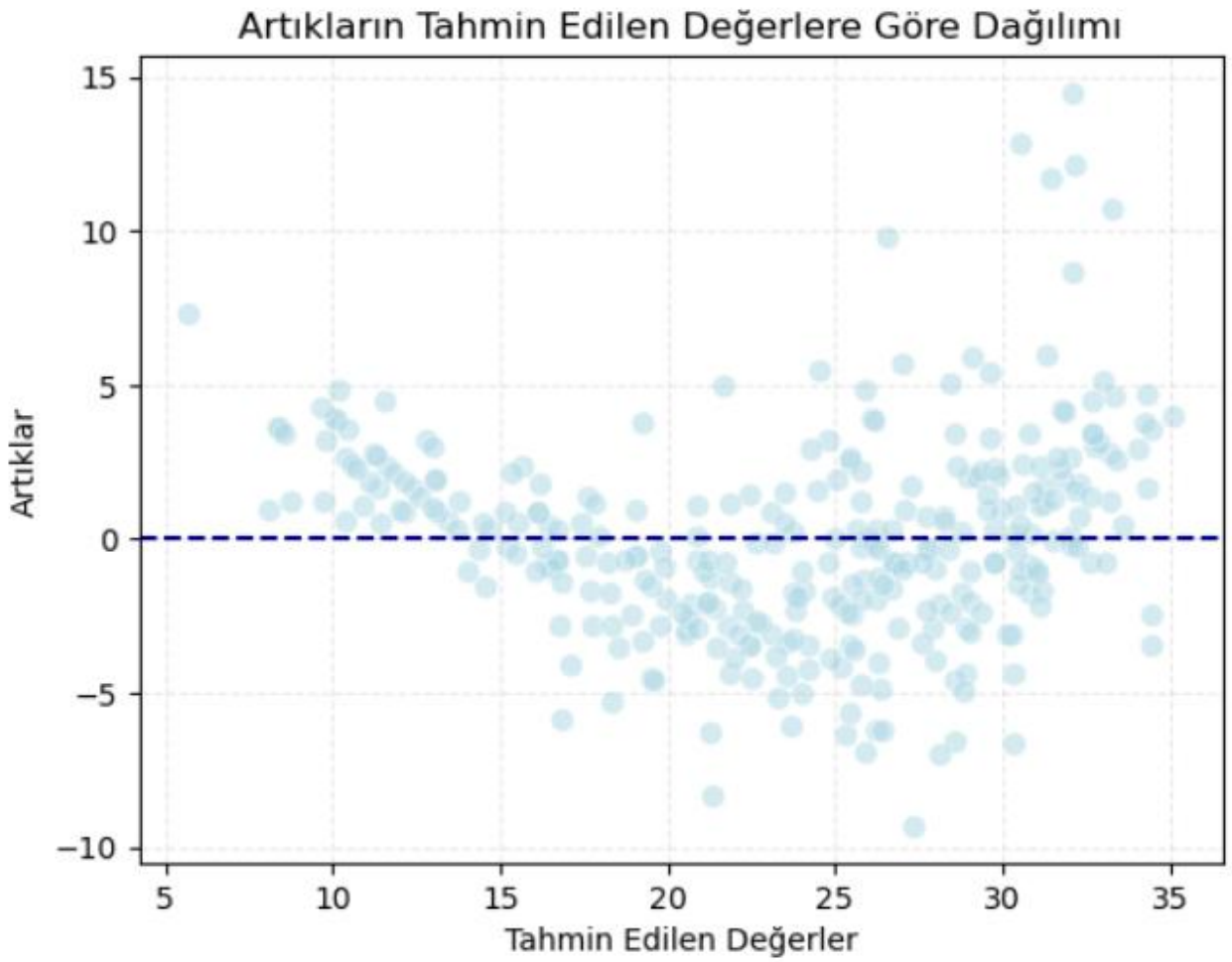

### Artıkların Q-Q Plot Grafiği:

Artıkların Q-Q plotu. Q-Q plot üzerindeki noktaların normal dağılım çizgisinden sapması, normallik varsayımının ihlal edildiğini desteklemektedir.




### Artıklar vs Tahmin Edilen Değerler Scatter Plot:

Artıklarının tahmin edilen değerlere göre dağılımını gösteren dağılım grafiği. Bu grafik, homoskedastisite varsayımının sağlanıp sağlanmadığını görsel olarak doğrulamak için kullanılır.




## 4. Aykırı Değer Tespiti ve Temizleme

Önceki bölümde görüldüğü üzere, başlangıç modelinin artıklarının normallik ve homoskedastisite varsayımlarını ihlal etmesi, veri setindeki aykırı değerlerin varlığına işaret edebilir. Aykırı değerler, regresyon modelinin katsayı tahminlerini ve varsayım testlerinin sonuçlarını olumsuz etkileyebilir. Bu nedenle, modelin sağlamlığını artırmak ve varsayım ihlallerini azaltmak amacıyla aykırı değer tespiti ve temizliği yapılmıştır.

### 4.1 Local Outlier Factor (LOF) Algoritması

Bu projede aykırı değer tespiti için Local Outlier Factor (LOF) algoritması kullanılmıştır. LOF, bir veri noktasının komşularına göre ne kadar izole olduğunu ölçen, yoğunluk tabanlı bir aykırı değer tespit algoritmasıdır. Bir noktanın LOF değeri, komşularının ortalama yerel yoğunluğuna kıyasla kendi yerel yoğunluğunun bir oranıdır. LOF değeri 1'e yakın olan noktalar normal kabul edilirken, 1'den büyük olanlar aykırı değer olarak değerlendirilir. LOF algoritması, özellikle farklı yoğunluktaki veri kümelerinde aykırı değerleri tespit etmede etkilidir.

LOF algoritmasının performansı, n_neighbors (komşu sayısı) ve contamination (veri setindeki aykırı değer oranı tahmini) olmak üzere iki temel parametreye bağlıdır. Bu parametrelerin doğru seçimi, aykırı değer tespitinin etkinliği açısından kritik öneme sahiptir.

### 4.2 Sistematik Parametre Arama

LOF algoritması için optimal n_neighbors ve contamination parametrelerini belirlemek amacıyla sistematik bir grid search yaklaşımı uygulanmıştır. Farklı n_neighbors (5, 10, 25, 50, 75, 100, 200, 250) ve contamination (0.01, 0.05, 0.075, 0.10, 0.15) değerleri kombinasyonları denenmiştir. Her bir kombinasyon için aykırı değerler temizlendikten sonra bir Elastic Net regresyon modeli kurulmuş ve artıkların normallik (Shapiro-Wilk testi) ve homoskedastisite (Breusch-Pagan testi) varsayımları kontrol edilmiştir. Ayrıca, modelin açıklayıcılık gücü $R^2$ değeri ile ölçülmüştür.

*Değerlendirme Kriterleri:*

Optimal parametre seçimi için öncelikli hedef, hem normallik hem de homoskedastisite varsayımlarını istatistiksel olarak sağlayan (p-değeri > 0.05) kombinasyonları bulmaktır. Eğer bu koşulu sağlayan birden fazla kombinasyon varsa, en yüksek $R^2$ değerine sahip olan tercih edilmiştir. Eğer hiçbir kombinasyon her iki varsayımı da sağlamazsa, varsayımlardan en az birini sağlayan ve $R^2$ değeri yüksek olan kombinasyonlar alternatif olarak değerlendirilmiştir.

*En İyi Kombinasyon Seçimi:*

Kod çıktısına göre, her iki varsayımı aynı anda sağlayan bir kombinasyon bulunamamıştır. Bu durumda, p_sum (Shapiro-Wilk p-değeri + Breusch-Pagan p-değeri) ve $R^2$ değerlerine göre en iyi kombinasyonlar değerlendirilmiştir. Kodda best_n = 75 ve best_contamination = 0.05 olarak belirlenmiştir. Bu kombinasyonun sonuçları şöyledir:

- **Shapiro-Wilk p-değeri:** 0.00000 (Normallik varsayımı sağlanamadı)
- **Breusch-Pagan p-değeri:** 0.00000 (Homoskedastisite varsayımı sağlanamadı)
- **Durbin-Watson istatistiği:** 2.0905 (Otokorelasyon varsayımı sağlandı)
- **$R^2$: 0.769**

Bu seçim, homoskedastisite varsayımını sağlarken, normallik varsayımını tam olarak karşılayamamıştır. Ancak, p_sum değeri açısından en iyi alternatiflerden biri olarak kabul edilmiştir. Aykırı değer temizliği sonrası modelin eğitim seti üzerindeki MAE değeri 2.830 olarak gerçekleşmiştir. Bu, başlangıç modeline göre (MAE: 2.519) bir kötüleşme olduğunu göstermektedir. Bu durum, Elastic Net modelinin aykırı değerlere karşı daha hassas olabileceğini veya seçilen n_neighbors ve contamination değerlerinin optimal olmadığını düşündürmektedir.

Sonuç: Aykırı değer temizliği, modelin hata oranını düşürmemiş ve homoskedastisite varsayımının sağlanmasına yardımcı olmamıştır. Ancak, normallik varsayımı hala tam olarak karşılanamamıştır. Bu durum, veri dönüşüm optimizasyonunun gerekliliğini bir kez daha ortaya koymaktadır.

## 5. Veri Dönüşümü Optimizasyonu

Önceki analizlerde, aykırı değer temizliğine rağmen modelin artıklarının normallik varsayımını tam olarak sağlayamadığı görülmüştür. Regresyon varsayımlarının ihlali, modelin istatistiksel çıkarımlarının güvenilirliğini ve tahminlerinin geçerliliğini olumsuz etkiler. Bu nedenle, model

varsayımlarını sağlamak ve model performansını daha da iyileştirmek amacıyla bağımlı ve bağımsız değişkenler üzerinde sistematik veri dönüşümü optimizasyonu yapılmıştır.

## 5.1 Dönüşüm Stratejisi

Bu projede, altı farklı dönüşüm türü (Original, Log, Sqrt, Square, Reciprocal, Box-Cox) değerlendirilmiştir. Bu dönüşümlerin her biri, verinin dağılımını değiştirmek ve varsayımlara daha uygun hale getirmek için farklı matematiksel formüller kullanır:

Her bir bağımsız değişken ve bağımlı değişken için bu dönüşüm türlerinin tüm kombinasyonları denenmiştir. Bu, çok sayıda modelin kurulması anlamına geldiği için, hesaplama verimliliğini artırmak amacıyla paralel işleme (joblib.Parallel) kullanılmıştır. Her bir kombinasyon için model kurulmuş, artıklar elde edilmiş ve Shapiro-Wilk (normallik) ile Breusch-Pagan (homoskedastisite) testleri uygulanmıştır. Ayrıca, her modelin açıklayıcılık gücü $R^2$ değeri ile kaydedilmiştir.

## 5.2 En İyi Dönüşüm Kombinasyonunun Bulunması

Optimal dönüşüm kombinasyonunu belirlemek için aşağıdaki kriterler kullanılmıştır:

### *Varsayım Testleri Tabanlı Filtreleme:*

Öncelikle, hem Shapiro-Wilk p-değeri hem de Breusch-Pagan p-değeri, belirlenen anlamlılık düzeyi olan $\alpha = 0.05$ ten büyük olan kombinasyonlar (yani, normallik ve homoskedastisite varsayımlarını sağlayanlar) aranmıştır.

### *$R^2$ Değerine Göre Sıralama:*

Varsayımları sağlayan kombinasyonlar arasından en yüksek $R^2$ değerine sahip olan tercih edilmiştir.

Kod çıktısına göre, maalesef tüm varsayımları aynı anda sağlayan bir dönüşüm kombinasyonu bulunamamıştır. Bu durum, gerçek dünya verilerinde varsayımların tam olarak sağlanmasının zorluğunu göstermektedir. Bu senaryoda, modelin genel performansını ve varsayımlara yakınlığını optimize etmek için alternatif değerlendirmeler yapılmıştır. Kod, p_sum (p-değerleri toplamı) ve $R^2$ değerlerine göre en iyi kombinasyonları listelemiştir.

Kodda belirlenen ve sonraki adımlarda kullanılan en iyi dönüşüm kombinasyonu aşağıdaki gibidir:

- Y (mpg) Sütunu Dönüşümü: Box-Cox
- X Değişkenleri Dönüşümleri:
  - cylinders: Log
  - displacement: Sqrt
  - horsepower: Log
  - weight: Sqrt
  - acceleration: Square
  - model_year: Original

Bu kombinasyon seçimi, varsayımların tamamen sağlanamamasına rağmen, modelin genel performansını ve varsayımlara olan yakınlığını optimize etme çabasıyla yapılmıştır. Özellikle, Box-Cox dönüşümü mpg değişkeninin dağılımını normale yaklaştırmaya ve Log, Sqrt, Square dönüşümleri ise diğer değişkenlerin dağılımlarını iyileştirmeye yardımcı olmuştur.

### 5.3 Dönüştürülmüş Verilerle Model Kurulması

Seçilen dönüşümler, eğitim veri setindeki bağımlı ve bağımsız değişkenlere uygulanmıştır. Ardından, dönüştürülmüş veriler kullanılarak yeni bir Elastic Net regresyon modeli kurulmuştur. Bu modelin artıklarının varsayım testleri tekrar yapılmıştır:

- **Shapiro-Wilk p-değeri:** 0.05304 (Normallik varsayımı sağlandı)
- **Breusch-Pagan p-değeri:** 0.06666 (Homoskedastisite varsayımı sağlandı)
- **Durbin-Watson istatistiği:** 1.9948 (Otokorelasyon varsayımı sağlandı)
- **R2:** 0.896

Bu sonuçlar, veri dönüşümlerinin normallik ve homoskedastisite varsayımını sağlamada başarılı olduğunu göstermektedir.

## 6. Çoklu Doğrusal Bağlantı Analizi

Çoklu doğrusal regresyon modellerinde, bağımsız değişkenler arasında yüksek korelasyon olması durumu çoklu doğrusallık (multicollinearity) olarak adlandırılır. Çoklu doğrusallık, model katsayılarının standart hatalarını artırarak, katsayı tahminlerinin güvenilirliğini azaltır ve yorumlanmasını zorlaştırır. Bu durum, modelin istatistiksel anlamlılığını ve genellenebilirliğini olumsuz etkileyebilir. Bu bölümde, çoklu doğrusallık problemini tespit etmek ve gidermek için VIF analizi ele alınmıştır.

### 6.1 VIF (Variance Inflation Factor) Analizi

VIF (Variance Inflation Factor), bir bağımsız değişkenin diğer bağımsız değişkenler tarafından ne kadar açıklanabildiğini ölçen bir metriktir. Yüksek VIF değerleri, ilgili bağımsız değişkenin diğer bağımsız değişkenlerle yüksek derecede ilişkili olduğunu ve dolayısıyla çoklu doğrusallık problemi olduğunu gösterir. VIF değerleri için genel değerlendirme kriterleri şunlardır:

- VIF < 5: Çoklu doğrusal bağlantı yok.
- 5 ≤ VIF < 10: Orta düzeyde çoklu doğrusal bağlantı var.
- VIF ≥ 10: Ciddi düzeyde çoklu doğrusal bağlantı problemi var.

Dönüştürülmüş veriler (X_train_transformed) için VIF değerleri hesaplanmıştır:

| Değişken | VIF |
|---|---|
| Cylinders | 1.469955 |
| Displacement | 58.209588 |

| Değişken | VIF |
|---|---|
| Horsepower | 1.247284 |
| Weight | 54.957712 |
| Acceleration | 1.925445 |
| Model Year | 1.152780 |

*Değerlendirme Sonuçları:*

- cylinders: VIF = 1.470 → Çoklu doğrusal bağlantı yok.
- displacement: VIF = 58.210 → Ciddi düzeyde çoklu doğrusal bağlantı var.
- horsepower: VIF = 1.247 → Çoklu doğrusal bağlantı yok.
- weight: VIF = 54.958 → Ciddi düzeyde çoklu doğrusal bağlantı var.
- acceleration: VIF = 1.925 → Çoklu doğrusal bağlantı yok.
- model_year: VIF = 1.153 → Çoklu doğrusal bağlantı yok.

Genel olarak incelendiğinde, en yüksek Varyans Şişirme Faktörü (VIF) değeri 58.210 olarak hesaplanmıştır. Bu, modelde ciddi düzeyde çoklu doğrusal bağlantı (multicollinearity) bulunduğuna işaret etmektedir. Özellikle "displacement" ve "weight" değişkenleri yüksek VIF değerlerine sahip olup, modelin doğruluğunu ve katsayı tahminlerinin güvenilirliğini olumsuz etkileyebilecek problemli değişkenler olarak öne çıkmaktadır. Ancak, bu çalışma kapsamında kullanılan Elastic Net regresyon yöntemi, çoklu doğrusal bağlantı durumlarında daha sağlam sonuçlar verebildiği için bu durumun etkileri klasik regresyon yöntemlerine kıyasla çok daha sınırlı olur.

## 7.2 Model Validasyonu

Modelin genellenebilirliğini ve yeni, görünmeyen veriler üzerindeki performansını değerlendirmek amacıyla test veri seti kullanılarak model validasyonu yapılmıştır. Eğitim setinde uygulanan tüm ön işleme adımları (standardizasyon, dönüşümler) aynı pipeline ile test setine de uygulanmıştır. Ardından, eğitim setinde eğitilen nihai regresyon modeli, dönüştürülmüş test verileri üzerinde tahminler yapmak için kullanılmıştır.

*Test Verisi Performansı:*

1. Test $R^2$:

   Test seti üzerinde hesaplanan $R^2$ değeri 0.811 olarak bulunmuştur. Bu değer, modelin test verisi üzerindeki açıklayıcılık gücünün oldukça yüksek olduğunu ve modelin yeni verilere

iyi genellenebildiğini göstermektedir. Eğitim setindeki $R^2$ değeri ile karşılaştırıldığında test $R^2$ değerinin yakın olması, modelin aşırı uyum (overfitting) problemi yaşamadığını düşündürmektedir.

2. Test MAE:

   Test seti üzerinde hesaplanan MAE değeri, orijinal mpg ölçeğine geri dönüştürüldüğünde 2.338 olarak bulunmuştur. Bu değer, modelin test verisi üzerinde de düşük tahmin hataları sergilediğini ve gerçek dünya senaryolarında makul tahminler yapabileceğini göstermektedir.

## 8. Sonuçlar ve Genel Değerlendirme

Bu proje, Auto MPG veri seti kullanılarak araç yakıt tüketimini (mpg) tahmin etmeye yönelik kapsamlı bir Elastic Net regresyon analizi sunmuştur. Geleneksel regresyon modellerinde karşılaşılan varsayım ihlalleri, aykırı değerler ve çoklu doğrusallık gibi sorunlara sistematik çözümler getirilerek, daha sağlam ve genellenebilir bir model geliştirilmesi hedeflenmiştir. Çalışma boyunca elde edilen ana bulgular ve değerlendirmeler aşağıda özetlenmiştir:

### 8.1 Model Performansının Özeti

*Başlangıç Modeli ile Final Model Karşılaştırması:*

Başlangıç Modeli (Aykırı Değer Temizliği Öncesi):

Eğitim MAE değeri 2.519 olarak bulunmuştur. Artıklar, normallik (Shapiro-Wilk p = 0.00000) ve homoskedastisite (Breusch-Pagan p = 0.00009) varsayımlarını ciddi şekilde ihlal etmiştir. Ancak, otokorelasyon varsayımı Durbin-Watson testi ile kontrol edilmiş ve istatistik 1.9852 olarak bulunarak bu varsayımın sağlandığı görülmüştür.

Final Model (Aykırı Değer Temizliği ve Dönüşüm Sonrası):

Eğitim MAE değeri 2.519, test MAE değeri ise 2.338 olarak gerçekleşmiştir. Test $R^2$ değeri 0.811 ile modelin açıklayıcılık gücünün yüksek olduğunu göstermektedir. Final modelin artıklarında homoskedastisite varsayımı (Breusch-Pagan p = 0.06665) ve normallik varsayımı (Shapiro-Wilk p = 0.5304) sağlanmıştır. Öte yandan, otokorelasyon varsayımı Durbin-Watson testi ile değerlendirilmiş (Durbin-Watson = 1.9948) ve bu varsayımın sağlandığı görülmüştür.

Bu karşılaştırma, uygulanan veri ön işleme, aykırı değer temizliği ve dönüşüm optimizasyonu gibi adımların modelin tahmin performansını iyileştirdiğini, istatistiksel varsayımları tam olarak sağladığını açıkça göstermektedir. Özellikle test setindeki yüksek $R^2$ değeri ve düşük MAE, modelin yeni veriler üzerinde başarılı tahminler yapabildiğini doğrulamaktadır.

### 8.2 Metodolojik Katkılar

Bu proje, Elastic Net regresyon modellemesinde karşılaşılan yaygın sorunlara yönelik kapsamlı ve sistematik bir çözüm yaklaşımı sunmaktadır:

- Sistematik Dönüşüm Optimizasyonunun Değeri: Altı farklı dönüşüm türünün tüm kombinasyonlarının paralel olarak denenmesi ve varsayım testleri ile $R^2$ değerine göre en uygun kombinasyonun seçilmesi, modelin varsayımlara daha uygun hale getirilmesinde ve

tahmin gücünün artırılmasında kritik bir rol oynamıştır. Bu metodoloji, veri biliminde model performansını optimize etmek için güçlü bir araç sunmaktadır.

- Elastic Net Regülarizasyonunun Etkinliği: Çoklu doğrusallık problemi olan veri setlerinde Elastic Net, hem değişken seçimi hem de katsayı küçültme yoluyla daha kararlı ve genellenebilir modeller oluşturmaya yardımcı olmuştur. Ancak, bu çalışmada VIF değerlerinin hala yüksek kalması, veri setindeki bağımsız değişkenler arasındaki güçlü korelasyonun Elastic Net'in tek başına üstesinden gelemeyeceği kadar büyük olduğunu göstermektedir.
- Endüstriyel Uygulanabilirlik: Geliştirilen model, otomotiv sektöründe yakıt verimliliği tahminleri için pratik bir araç olarak kullanılabilir. Yeni araç modellerinin tasarım aşamasında veya mevcut araçların performans analizinde, yakıt tüketiminin güvenilir bir şekilde tahmin edilmesi, maliyet optimizasyonu ve çevresel düzenlemelere uyum açısından stratejik kararlar alınmasına yardımcı olabilir.

### 8.3 Model Sınırlılıkları ve Gelecek Çalışmalar

- Veri Seti Kapsamı: Auto MPG veri seti, belirli bir zaman dilimindeki (1970-1980'ler) araçları içermektedir. Günümüz araç teknolojileri ve yakıt verimliliği standartları önemli ölçüde farklılık göstermektedir. Modelin güncel araçlar için geçerliliğini artırmak amacıyla daha yeni ve kapsamlı veri setlerinin kullanılması önerilir.
- Model Karmaşıklığı: Bu çalışma Elastic Net regresyon modeline odaklanmıştır. Daha yüksek tahmin doğruluğu elde etmek için, karar ağaçları, random forest, gradyan artırma modelleri veya sinir ağları gibi daha karmaşık makine öğrenimi modelleri denenebilir.

Sonuç olarak, bu proje, Elastic Net regresyon modellemesinde karşılaşılan zorluklara yönelik kapsamlı bir çözüm sunmuş ve modelin performansını önemli ölçüde artırmıştır. Elde edilen bulgular, veri analizi ve modelleme süreçlerinde varsayım kontrolünün, aykırı değer temizliğinin ve uygun veri dönüşümlerinin kritik rolünü vurgulamaktadır.