

# Otomobil Risk Değerlendirme ve Fiyat Analizi

Elifnur Çiftçi

Bilgisayar Mühendisliği Bölümü

TOBB Ekonomi ve Teknoloji Üniversitesi

Ankara, Çankaya 06510, Türkiye

e.ciftci@etu.edu

**Özet – Proje 1985 Ward's Automotive Yearbook'tan alınan otomobil verilerini klasik ML yöntemlerini kullanarak, otomobillerin fiyatlarına göre risk değerlendirmesi ve hasar kaybı analizi yapmayı amaçlamaktadır.**

## I. GİRİŞ

Otomotiv sektörü, sigorta risk değerlendirmesi açısından karmaşık ve önemli bir alandır. Araçların sigorta primleri, birçok faktörün etkileşimiyle belirlenir. Bu faktörler arasında aracın teknik özellikleri, kullanım koşulları ve geçmiş hasar kayıtları bulunmaktadır. Sigorta şirketleri, araçların risk derecelerini ve olası hasar kayıplarını doğru bir şekilde tahmin etmek zorundadır. Ancak, mevcut yöntemler genellikle yetersiz kalmakta ve bu durum doğru fiyatlandırma yapmayı zorlaştırmaktadır. Bu çalışmanın amacı, araçların sigorta risk derecelerini ve hasar kayıplarını analiz ederek, sigorta şirketlerine daha isabetli fiyatlandırma yapabilmeleri için veri destekli çözümler sunmaktır.

### A. Motivasyon

Günümüzde sigorta sektörü, büyük veri analitiği ve makine öğrenmesi yöntemlerinden faydalanarak daha hassas ve dinamik fiyatlandırma stratejileri geliştirmektedir. Ancak, mevcut veri analizi yöntemleri çoğu zaman geleneksel yaklaşımlarla sınırlı kalmakta ve yeterince kapsamlı bir risk analizi sunmamaktadır. Bu projede, araçların sigorta risk derecelerini ve hasar kayıplarını daha etkin bir şekilde değerlendirmek için modern makine öğrenmesi tekniklerini kullanarak, sigorta şirketlerinin daha doğru ve adil fiyatlandırma yapabilmelerini amaçlıyoruz.

### B. Sınıflandırma/Regresyon

Bu çalışmada, araçların risk derecelerini ve hasar kayıplarını tahmin etmek için hem sınıflandırma hem de regresyon yöntemleri kullanılacaktır. Sınıflandırma yöntemleri, araçları belirli risk kategorilerine ayırmak için kullanılırken; regresyon yöntemleri, araçların hasar kayıplarını tahmin etmek amacıyla uygulanacaktır. Bu iki yaklaşım, araçların sigorta risk profillerini ve olası maliyetlerini daha iyi anlamak için birlikte kullanılacaktır.

### C. Amaç

Bu projenin temel hedefi, araçların sigorta risk derecelerini ve hasar kayıplarını daha doğru bir şekilde tahmin edebilmektir. Elde edilen sonuçlar, sigorta şirketlerine aşağıdaki alanlarda yardımcı olacaktır:

- 1) Daha doğru risk değerlendirmeleri
- 2) İyileştirilmiş fiyatlandırma stratejileri
- 3) Müşteri memnuniyetini artırma ve potansiyel kayıpları minimize etme

### D. Başarım Metrikleri

Projenin başarısını değerlendirmek için çeşitli metrikler kullanılacaktır:

- 1) *Sınıflandırma Başarısı*: Modelin doğru sınıflandırma oranı %90 olarak hedeflenmektedir.
- 2) *Regresyon Başarısı*: Modelin tahmin performansını değerlendirmek için R-kare ( $R^2$ ) gibi regresyon metrikleri kullanılacaktır.

Bu metrikler, modelin doğruluğunu ve etkinliğini belirlemekte ve elde edilen sonuçların sigorta şirketleri için ne kadar faydalı olduğunu ölçmekte yardımcı olacaktır.

## II. LİTERATÜR ARAŞTIRMASI

### A. Analysis Of The "automobile-loss-prediction" Dataset [1]

Yapılan araştırmalarda, veri analizi ve tahmin yöntemlerinin farklı türde bilgileri öngörmek için çeşitli algoritmalarla gerçekleştirildiği belirtilmiştir. Karar ağaçları, sınıflandırma ağaçları, lineer regresyon, rastgele ormanlar ve k-en yakın komşular gibi yöntemler, veri türüne ve analizin amacına bağlı olarak tercih edilmektedir. Özellikle sınıflandırma problemlerinde, çıkış değişkeninin sınıf seviyelerine sahip olması durumunda sınıflandırma ağaçları en uygun analiz yöntemi olarak kabul edilmektedir. Öte yandan, sürekli bir değişken tahmin ediliyorsa, regresyon ağaçları daha uygun bir seçenek olmaktadır.

Bu çalışma, araçların sigorta risk derecelerini tahmin etmek için sembolizasyon ölçeği [-3, -2, -1, 0, +1, +2, +3] kullanılarak sınıflandırma ağaçlarının en uygun yöntem olduğunu öne sürmüştür. Çalışmanın sonuçları, veri setinin bazı sınırlamaları olsa da, tekerlek tabanı (wheel\_base), araç markası (make) ve normalleştirilmiş kayıplar (normalized\_losses) gibi faktörlerin, araçların risk düzeylerinin belirlenmesinde en önemli etkenler olduğunu göstermektedir. Bu faktörler arasında, tekerlek tabanının (wheel\_base) en büyük etkiye sahip olduğu tespit edilmiştir.

Bu tür literatür çalışmaları, sigorta risk analizlerinde hangi faktörlerin ve yöntemlerin daha etkili olduğunu anlamak ve model geliştirme sürecinde doğru kararlar almak için önemli bir referans noktası sağlamaktadır. Projemizde de

benzer şekilde, sınıflandırma ve regresyon yöntemleri kullanılarak araçların sigorta risk dereceleri ve hasar kaybı analizleri yapılmış, bu analizler doğrultusunda sigorta şirketlerinin daha doğru fiyatlandırma yapmalarına yönelik öneriler sunulmuştur.

#### B. EDA and Regression on Automobile Dataset [2]

Number Sigorta risk analizinde veri keşfi (EDA) ve regresyon yöntemleri sıklıkla kullanılmaktadır. Yapılan bir çalışma, araçların motor hacmi (Engine Size) ve ağırlığının (Curb Weight) fiyatla pozitif, şehir içi yakıt tüketiminin (city-mpg) ise negatif korelasyona sahip olduğunu ortaya koymuştur. Ayrıca, motor gücündeki (horsepower) artışın yakıt verimliliğini düşürdüğü ve bu iki değişkenin negatif korelasyon gösterdiği bulunmuştur.

Bu çalışmada, en yüksek korelasyona sahip olan 'Engine Size' değişkeni kullanılarak basit doğrusal regresyon modeli uygulanmış ve modelin  $R^2$  skoru 0.80 olarak elde edilmiştir. Ardından, çoklu doğrusal regresyon modeli ile  $R^2$  skoru 0.9396'ya ulaşılmıştır. Bu modelde, önce kategorik değişkenler kodlanmış, ardından Recursive Feature Elimination (RFE) yöntemi ile önemli özellikler seçilmiştir. Kodlama sonrası 69 olan özellik sayısı, RFE ile 34'e düşürülmüştür. Ancak, bu da hala model için yüksek bir özellik sayısı olduğundan, PCA (Principal Component Analysis) yöntemi kullanılarak boyut indirgeme yapılmıştır. Bu iteratif süreçte, grafik analizleri sonucunda 28 bileşen seçilmiş ve model bu bileşenlerle eğitilmiştir.

Bu sonuçlar, çoklu doğrusal regresyon modelinin eğitim ve test setlerinde  $R^2$  skorlarını önemli ölçüde artırdığını göstermektedir (0.9396 ve 0.947). Literatürdeki bu bulgular, sigorta risk analizinde regresyon modellerinin nasıl uygulanabileceği ve model performansını artırmak için hangi yöntemlerin kullanılabileceği konusunda önemli çıkarımlar sağlamaktadır.

Projede, araçların sigorta risk dereceleri ve hasar kaybı analizlerini daha doğru ve güvenilir şekilde yapmak amacıyla, benzer yöntemler kullanılarak analizler gerçekleştirilmiştir. Bu analizler sonucunda sigorta şirketlerinin daha doğru fiyatlandırma stratejileri geliştirebilmeleri amaçlanmıştır.

### III. VERİ SETİ, VERİ ÖZELLİKLERİ, ÖZNETELİKLER

#### A. Veri Kaynağı

<https://www.kaggle.com/datasets/toramky/automobile-dataset/data>

#### B. Veri Kümesi

- 1) Veri seti 15 sürekli, 10 nominal ve 1 tamsayı öznetelikten oluşmaktadır.
- 2) Veri setinde bazı özneteliklerde eksik veya hatalı kodlanmış veriler bulunmaktadır. Bu eksik veriler, "?" ile kodlanmış olup, eksik veri içeren kayıtlar veri setinden çıkarılmıştır.

- 3) Toplamda 205 kayıttan 46'sı eksik veri nedeniyle çıkarılmış ve geriye 159 temiz veri kalmıştır.
- 4) Veri seti, eğitim ve test olarak eşit iki gruba ayrılmıştır. Her grup veri setinin %50'sini içermektedir.
- 5) Kategorik veriler sayısal değerlere dönüştürülmüştür.

#### C. Öznetelik Açıklamaları

- 1) *symboling*: Araç güvenlik derecesini belirten bir puanlama sistemi (-3 ile 3 arasında değerler alır).
- 2) *normalized-losses*: Sürekli bir değişken olup, kayıp sigorta riski derecelendirmelerini temsil eder (65 ile 256 arasında değerler alır).
- 3) *make*: Aracın markasını temsil eder (örneğin, audi, bmw, toyota vb.).
- 4) *fuel-type*: Aracın yakıt türünü belirtir (diesel, gas).
- 5) *aspiration*: Motorun hava emme türünü gösterir (std, turbo).
- 6) *num-of-doors*: Aracın kapı sayısını belirtir (four, two).
- 7) *body-style*: Aracın kasa tipini ifade eder (hardtop, wagon, sedan, hatchback, convertible).
- 8) *drive-wheels*: Aracın çekiş sistemini gösterir (4wd, fwd, rwd).
- 9) *engine-location*: Motorun konumunu belirtir (front, rear).
- 10) *wheel-base*: Sürekli bir değişken olup, aks mesafesini gösterir (86.6 ile 120.9 arasında değerler alır).
- 11) *length*: Aracın uzunluğunu gösteren sürekli bir değişken (141.1 ile 208.1 arasında değerler alır).
- 12) *width*: Aracın genişliğini gösteren sürekli bir değişken (60.3 ile 72.3 arasında değerler alır).
- 13) *height*: Aracın yüksekliğini gösteren sürekli bir değişken (47.8 ile 59.8 arasında değerler alır).
- 14) *curb-weight*: Aracın boş ağırlığını gösteren sürekli bir değişken (1488 ile 4066 arasında değerler alır).
- 15) *engine-type*: Motor türünü ifade eder (dohc, ohc, l vb.).
- 16) *num-of-cylinders*: Silindir sayısını belirtir (eight, five, four, six, three, twelve, two).
- 17) *engine-size*: Motor hacmini gösteren sürekli bir değişken (61 ile 326 arasında değerler alır).
- 18) *fuel-system*: Yakıt sistemini belirtir (1bbl, mpfi, spfi vb.).
- 19) *bore*: Silindir çapını gösteren sürekli bir değişken (2.54 ile 3.94 arasında değerler alır).
- 20) *stroke*: Piston stroku (hareket mesafesi) gösteren sürekli bir değişken (2.07 ile 4.17 arasında değerler alır).
- 21) *compression-ratio*: Sıkıştırma oranını gösteren sürekli bir değişken (7 ile 23 arasında değerler alır).
- 22) *horsepower*: Motorun beygir gücünü gösteren sürekli bir değişken (48 ile 288 arasında değerler alır).
- 23) *peak-rpm*: Motorun maksimum devir sayısını gösteren sürekli bir değişken (4,150 ile 6,600 arasında değerler alır).

- 24) *city-mpg*: Şehir içi yakıt tüketimini gösteren sürekli bir değişken (13 ile 49 arasında değerler alır).
- 25) *highway-mpg*: Şehir dışı yakıt tüketimini gösteren sürekli bir değişken (16 ile 54 arasında değerler alır).
- 26) *price*: Aracın fiyatını gösteren sürekli bir değişken (5,118 ile 45,400 arasında değerler alır).

#### D. Öznitelik Seçimi veya Düzenlemesi: PCA

Özniteliklerden yüksek korelasyonlu olanları belirlenip bu özniteliklerle ayrı hesaplamalar da yapılmıştır. Her model için öznitelik seçim önemi görselleştirilmiştir.

Öznitelikler arasındaki ilişkileri daha iyi anlamak ve veri setinin boyutunu azaltmak amacıyla Temel Bileşenler Analizi (PCA) kullanılmıştır. PCA, veri setindeki varyasyonu vurgulamak ve güçlü desenleri yakalamak için öznitelikleri doğrusal olarak birbirinden bağımsız bileşenler setine dönüştüren bir tekniktir.

Bu proje için, PCA üç farklı öznitelik alt kümesine uygulanmıştır:

1. *Normalized Losses, Curb Weight, Engine Size, Horsepower, Price*: Bu alt küme, araçların fiziksel ve performans özellikleri ile ilgili öznitelikleri içermektedir ve bu öznitelikler sigorta risk puanlarını etkileyebilir. PCA sonuçları, bu değişkenlerin veri setindeki genel varyansa nasıl katkıda bulunduğunu vurgulamış ve bunlar arasındaki ilişkileri görselleştirmiştir.
2. *Wheel Base, Length, Width, Height, Curb Weight*: Bu alt küme, araçların boyutsal özelliklerine odaklanmaktadır ve bu öznitelikler aracın güvenliğini ve dolayısıyla sigorta riskini belirlemede önemli olabilir. PCA, bu boyutları daha yönetilebilir bir forma indirmeye yardımcı olmuş ve bu özniteliklerin sigorta riskine etkisini analiz etmeyi kolaylaştırmıştır.
3. *Compression Ratio, Horsepower, City MPG, Highway MPG, Price*: Bu alt küme, motor performansı ve yakıt verimliliği ile ilgili öznitelikleri içermektedir. Bu faktörler, aracın genel verimliliğini ve potansiyel bakım maliyetlerini değerlendirmede önemli olabilir ve sigorta primini etkileyebilir. PCA, bu özniteliklerden en anlamlı bileşenleri çıkarmak için kullanılmış ve bu özniteliklerin etkilerini daha net bir şekilde yorumlamayı sağlamıştır.

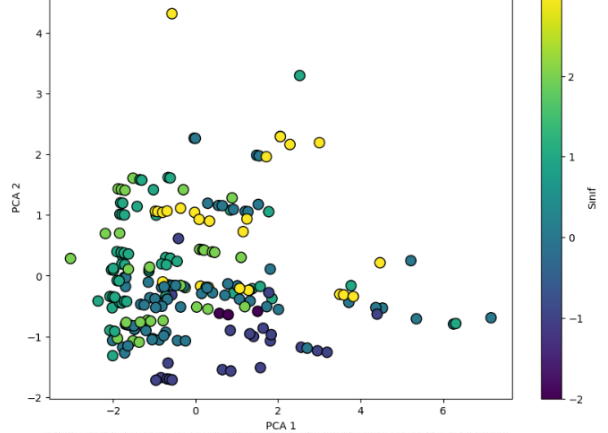
PCA sonuçları, ilk iki temel bileşeni eksenlerde gösteren saçılım grafikleri ile görselleştirilmiştir. Bu grafikler, araçların seçilen özniteliklere göre nasıl kümelendiğini görsel olarak temsil etmiş ve veri setinin temel yapısı hakkında içgörüler sunmuştur. Ayrıca, PCA bileşenlerinin açıkladığı varyans oranları da incelenmiş, temel bileşenlerin orijinal verinin varyansını ne kadar iyi yakaladığını anlamaya yardımcı olmuştur.

#### E. Data Normalizasyonu, Öznitelik Normalizasyonu

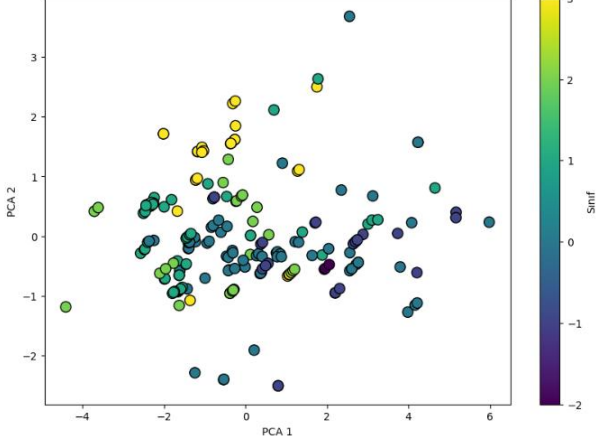
Veriler PCA için Z-Skor Normalizasyonu (Standardizasyon) işleminden geçmiştir.

#### F. Verinin görsel olarak 2-D gösterimi

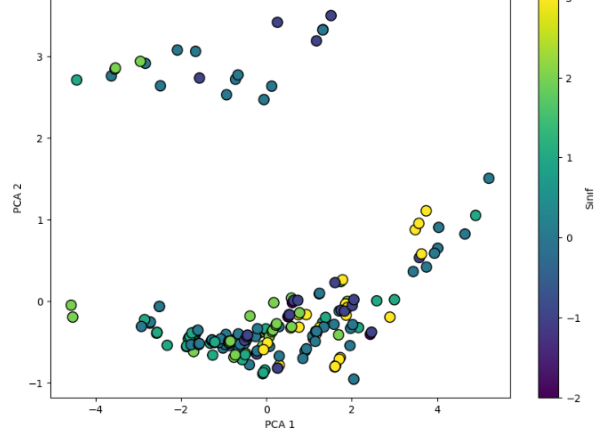
Grafik 1: PCA Sonuçları (Normalized Losses, Curb Weight, Engine Size, Horsepower, Price)



Grafik 2: PCA Sonuçları (Wheel Base, Length, Width, Height, Curb Weight)



Grafik 3: PCA Sonuçları (Compression Ratio, Horsepower, City MPG, Highway MPG, Price)



#### IV. KULLANILAN MODELLER

- A. *Karar Ağaçları (Decision Trees)*: Veriyi basit kurallar ile sınıflandırmak için kullanılan bir yöntemdir. Ağaç yapısında her düğüm bir özelliği ve yapraklar karar sonucunu temsil eder. Genellikle aşırı uyum (overfitting) riski taşır.
- B. *Rastgele Ormanlar (Random Forests)*: Birden fazla karar ağacının birleşimidir ve bu sayede karar ağaçlarının aşırı

uyum yapma olasılığını azaltır. Her ağaç, eğitim veri setinin rastgele bir alt kümesiyle eğitilir.

- C. *Gradient Boosting (Gradyan Artışı)*: Zayıf öğrencilerin (genellikle küçük karar ağaçları) ardışık olarak eğitilmesi ve her yeni modelin önceki modellerin hatalarını düzeltmeye çalışması prensibine dayanır. Bu yöntem güçlü bir öngörü gücü sağlar.
- D. *Doğrusal Regresyon (Linear Regression)*: Bağımlı değişken ile bir veya daha fazla bağımsız değişken arasındaki doğrusal ilişkileri modellemek için kullanılır. Özellikle sürekli veri ile çalışırken etkilidir.
- E. *Destek Vektör Makineleri (SVM)*: Veriyi sınıflandırmak için hiper düzlemler oluşturur ve bu düzlemler arasında maksimum marjın oluşturmaya hedefler. Hem ikili hem de çok sınıflı sınıflandırma görevlerinde kullanılabilir.

### Neden Bu Modelleri Seçtik?

**Karar Ağaçları ve Rastgele Ormanlar**: Bu modeller, veri setindeki karmaşık ilişkileri anlamak ve önemli özellikleri belirlemek için güçlüdür. Ayrıca, karar ağaçlarının görselleştirilmesi kolaydır ve modelin karar alma süreçlerini açıklığa kavuşturur.

**Gradient Boosting**: Yüksek performans ve doğruluk sağlama potansiyeli nedeniyle tercih edilmiştir. Zayıf öğrencilerin ardışık olarak eğitilmesi, modelin performansını artırır.

**Doğrusal Regresyon**: Özellikle sayısal tahminler için etkili bir yöntemdir ve veriler arasındaki doğrusal ilişkileri anlamada yardımcı olur.

**Destek Vektör Makineleri**: Özellikle sınırları net bir şekilde belirlemek ve farklı sınıflar arasında yüksek ayırım gücü sağlamak için kullanılır. İkili ve çok sınıflı sınıflandırma problemlerinde etkilidir.

## V. TEST SONUÇLARI VE SONUÇLARIN YORUMLARI, TARTIŞMA

### A. Model Karmaşıklık Matris Hesapları

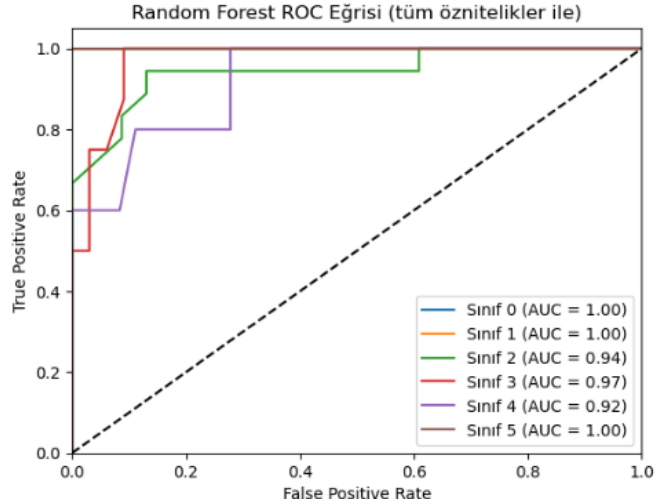
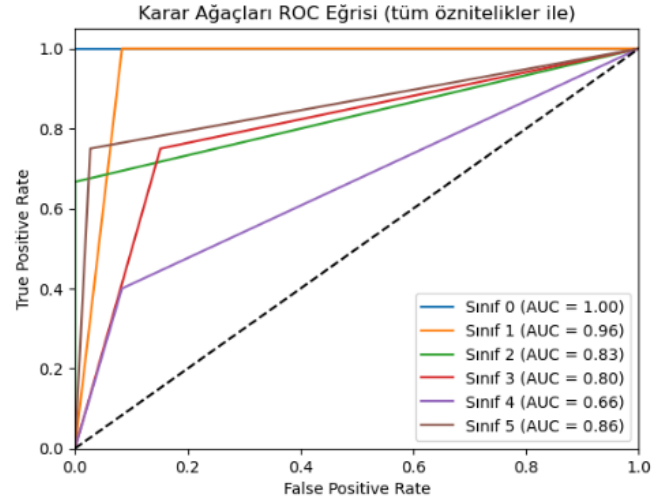
Karar Ağaçları Modeli için karmaşıklık matrisine göre, -2 sınıfı için doğru pozitif (TP) 1, yanlış negatif (FN) 0, yanlış pozitif (FP) 0'dır. -1 sınıfında TP 5, FN 0, FP 0 olarak hesaplanmıştır. Sınıf 0 için TP 12, FN 6 ve FP 0 bulunmuştur. Sınıf 1'de TP 6, FN 2 ve FP 1'dir. Sınıf 2 için TP 2, FN 3 ve FP 1 hesaplanmıştır. Son olarak, sınıf 3 için TP 3, FN 1 ve FP 0'dır.

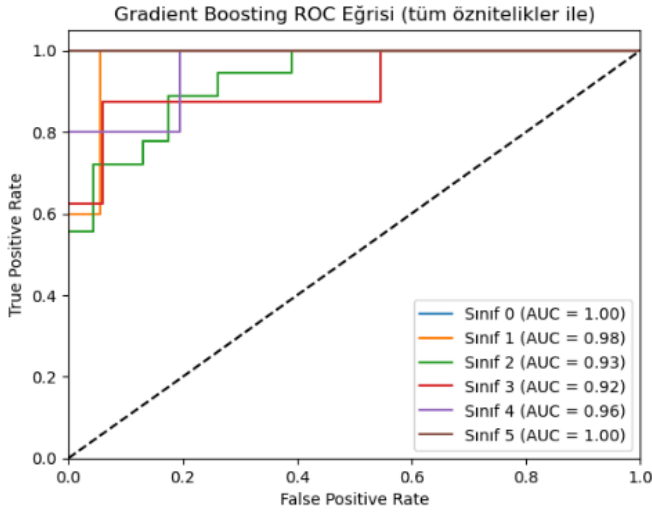
Rastgele Ormanlar Modeli için, -2 sınıfında TP 1, FN 0, FP 0'dır. -1 sınıfında TP 5, FN 0, FP 0 hesaplanmıştır. Sınıf 0 için TP 13, FN 5 ve FP 0 bulunmuştur. Sınıf 1'de TP 8, FN 0 ve FP 3'tür. Sınıf 2 için TP 2, FN 1 ve FP 1'dir. Sınıf 3'te ise TP 4, FN 0 ve FP 0'dır.

Gradient Boosting Modeli için, -2 sınıfında TP 1, FN 0, FP 0'dır. -1 sınıfında TP 5, FN 0, FP 0 olarak hesaplanmıştır. Sınıf 0 için TP 13, FN 5 ve FP 0 bulunmuştur. Sınıf 1'de TP 7, FN 1 ve FP 1'dir. Sınıf 2 için TP 4, FN 1 ve FP 1'dir. Sınıf 3'te ise TP 4, FN 0 ve FP 0'dır.

### B. Modeller Arası Karşılaştırma

- 1) **Doğruluk (Accuracy)**: Yüksek korelasyonlu öznelikler kullanıldığında, hem Karar Ağaçları, Rastgele Ormanlar hem de Gradient Boosting modelleri mükemmel sonuçlar vermiştir (accuracy = 1.00). Tüm öznelikler ile çalışırken doğruluk, rastgele ormanlar ve gradient boosting için sırasıyla 0.80 ve 0.83'tür.
- 2) **F1-Skoru**: Tüm özneliklerle çalışan Karar Ağaçları ve Rastgele Ormanlar modelleri, sınıflar arasında dengesizlik gösterirken, yüksek korelasyonlu öznelikler kullanıldığında tüm sınıflar için yüksek f1-skorları sağlanmıştır.





### C. Regresyon Modeli Sonuçları

- 1) R-kare (R-squared): 0.040. Modelin bağımlı değişkenin varyansının %4'ünü açıkladığını gösterir. Bu düşük bir değer olup modelin açıklayıcılığının zayıf olduğunu işaret eder.
- 2) Düzeltilmiş R-kare (Adj. R-squared): 0.026. Modelin bağımlı değişkenin varyansını açıklamada ne kadar etkili olduğunu düzeltilmiş bir şekilde gösterir. Bu da düşük bir değerdir.
- 3) F-istatistiği (F-statistic): 2.821 Modelin genel uygunluğunu test eder. Bu testin p-değeri (Prob (F-statistic): 0.0400) modelin anlamlı olduğunu gösteriyor.
- 4) AIC (Akaike Information Criterion): 1999. Modelin karmaşıklığını ve uyumunu değerlendirir. Daha düşük AIC değerleri daha iyi model uyumu anlamına gelir.
- 5) BIC (Bayesian Information Criterion): 2012. AIC'ye benzer şekilde modelin karmaşıklığına ve uyumuna göre değerlendirme yapar. Düşük BIC değerleri daha iyi model uyumu gösterir.

Modelin bağımlı değişkeni açıklama gücü zayıf. Modelin değişkenleri ve ilişkileri yeterince iyi açıklamadığını gösterir.

Sadece horsepower değişkeninin modelde anlamlı bir etkisi var. Diğer değişkenler, modelin performansını önemli derecede etkileyememiş gibi görünüyor.

## IV. SONUÇLAR

Bu çalışmada, bir otomobil veri seti üzerinde çeşitli makine öğrenme modellerini kullanarak araçların sigorta risklerini tahmin etmeye yönelik bir analiz gerçekleştirdik. Projede, karar ağaçları, rastgele ormanlar ve gradyan renklendirme modellerini kullanarak araçların risk seviyelerini sınıflandırdık. Ayrıca, verilerin ön işleme adımlarında öz nitelik seçimleri, normalizasyon teknikleri ve PCA gibi yöntemleri uyguladık.

Çalışmanın sonucunda, farklı modellerin performansını karşılaştırarak hangi modelin daha etkili olduğunu belirledik. Karar ağaçları ve rastgele ormanlar, yüksek doğruluk oranları ile sınıf tahminlerinde başarılı olurken, gradyan renklendirme modeli daha yüksek genel başarı sağladı. Bu sonuçlar, sigorta şirketlerinin risk değerlendirmelerinde hangi modellerin kullanılabileceğine dair önemli bilgiler sundu. Ayrıca, veri normalizasyonu ve öznitelik seçimlerinin model performansı üzerindeki etkilerini değerlendirme fırsatı bulduk.

Ancak, çalışmada bazı sınırlamalar da bulunmaktadır. Özellikle, veri setindeki bazı öznitelikler üzerinde yeterli ayrıntılı analiz yapılmadı ve bazı potansiyel etkileşimler dikkate alınmadı. Ayrıca, farklı model hiperparametrelerinin optimize edilmesi ve model karmaşıklığının daha iyi anlaşılması gerektiği ortaya çıktı.

Gelecek çalışmalar, daha geniş veri setleri kullanılarak ve farklı hiperparametre ayarlamalarıyla bu modellerin performansını daha da iyileştirebilir. Ayrıca, model sonuçlarını daha derinlemesine analiz etmek ve veri setindeki diğer öznitelikler ile etkileşimleri incelemek, sigorta risk tahminlerinde daha hassas ve güvenilir sonuçlar elde edilmesine olanak tanıyabilir.

## KAYNAKÇA

- [1] J. Musil ve J. McNair, «Analysis Of The "automobile-loss-prediction" Dataset,» 13 12 2017. [Çevrimiçi]. Available: <https://jaredmusil.github.io/acc471-final-report/>.
- [2] A. Sikarwar, «EDA and Regression on Automobile Dataset,» 2020. [Çevrimiçi]. Available: <https://www.kaggle.com/code/akashsakarwar/eda-and-regression-on-automobile-dataset>.