ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZİ

SİGORTA HASAR TUTARININ KESTİRİMİNDE TWEEDIE BİLEŞİK POISSON MODELİ: AŞIRI YAYILIM MODELLEMESİ

Buse CAN

AKTÜERYA BİLİMLERİ ANABİLİM DALI

ANKARA
2026

ÖZET

Yüksek Lisans Tezi

SİGORTA HASAR TUTARININ KESTİRİMİNDE TWEEDIE BİLEŞİK POISSON MODELİ: AŞIRI YAYILIM MODELLEMESİ

Buse CAN

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Aktüerya Bilimleri Anabilim Dalı

Danışman: Prof. Dr. Furkan BAŞER

Sigorta fiyatlandırmasında doğru ve adil primlerin belirlenmesi hem risk yönetimi hem de sigorta şirketlerinin finansal sürdürülebilirliği açısından temel bir gerekliliktir. Bu bağlamda, poliçe sahiplerine ilişkin gelecekte gerçekleşmesi muhtemel hasar tutarlarının güvenilir biçimde tahmin edilmesi, sigortacılık uygulamalarında kritik bir araştırma konusu olarak öne çıkmaktadır.

Bu çalışmada, taşıt sigortası kapsamında gerçekleşen hasar tutarlarının risk faktörlerine bağlı olarak modellenmesi ve kestirimi problemi ele alınmaktadır. Sigorta hasar tutarı verilerinin yarı sürekli yapıda olması; belirgin sıfır yığılma içermesi ve pozitif hasar tutarlarının aşırı sağa çarpık bir dağılım sergilemesi, geleneksel doğrusal modelleme yaklaşımlarının yetersiz kalmasına neden olmaktadır. Bu doğrultuda çalışmada, yarı sürekli veri yapısını tek bir çerçevede modelleyebilen Tweedie bileşik Poisson regresyonu ile modern makine öğrenmesi yöntemleri birlikte kullanılmıştır. Tezde kullanılan veri setleri, araç yaşı, araç değeri, maruziyet süresi, poliçe bilgileri, sürücü özellikleri ve geçmiş hasar kayıtları gibi çok sayıda değişken içermektedir.

Çalışma kapsamında Tweedie modelinin kuramsal yapısı ayrıntılı biçimde incelenmiş, makine öğrenmesi yöntemlerinin ise doğrusal olmayan ilişkileri yakalama ve yüksek boyutlu veri yapılarıyla uyum sağlama avantajları incelenmiştir. Uygulama aşamasında Genelleştirilmiş Doğrusal Regresyon, LASSO Regresyon, Ridge Regresyon, Karar Ağaçları, Rasgele Orman, Destek Vektör Makineleri, Yapay Sinir Ağları, XGBoost, LightGBM ve CatBoost algoritmaları kullanılarak model tahminleri gerçekleştirilmiştir. Elde edilen bulgular, uygun performans ölçütleri yardımıyla karşılaştırmalı olarak değerlendirilmiştir. Yarı sürekli sigorta hasar verilerinin modellenmesinde makine öğrenmesi yaklaşımlarından gradyan artırma algoritmalarının, Tweedie regresyonuna veriye uyum açısından iyi bir alternatif olduğu sonucuna ulaşılmıştır.

Şubat 2026, 113 sayfa

Anahtar Kelimeler: Hasar tutarı, Sıfır yığılmalı veri, Tweedie regresyonu, Aşırı yayılım modellemesi, Makine öğrenmesi

ABSTRACT

Master Thesis

TWEEDIE COMPOUND POISSON MODEL IN PREDICTING INSURANCE CLAIM COST: OVERDISPERSION MODELLING

Buse CAN

Ankara University
Graduate School of Natural and Applied Science
Department of Actuarial Sciences

Supervisor: Prof. Dr. Furkan BAŞER

Determining accurate and fair premiums in insurance pricing is regarded as a fundamental requirement for effective risk management and the financial sustainability of insurance companies. In this context, the reliable estimation of future claim amounts associated with policyholders is considered a critical research topic in insurance practice.

This study addresses the problem of modelling and predicting claim costs arising from motor vehicle insurance based on risk factors. The semi-continuous nature of insurance claim severity data, characterized by pronounced zero inflation and a highly right-skewed distribution of positive claim amounts, renders traditional linear modeling approaches inadequate. Accordingly, Tweedie compound Poisson regression, which allows semi-continuous data structures to be modelled within a single framework, is combined with modern machine learning methods. The datasets used in this thesis include a wide range of variables such as vehicle age, vehicle value, exposure time, policy information, driver characteristics, and historical claim records.

Within the scope of the study, the theoretical structure of the Tweedie model is examined in detail, and the ability of machine learning methods to capture nonlinear relationships and accommodate high-dimensional data structures is investigated. During the empirical analysis, model estimations are performed using Generalised Linear Regression, LASSO Regression, Ridge Regression, Decision Trees, Random Forests, Support Vector Machines, Artificial Neural Networks, XGBoost, LightGBM, and CatBoost. The results are comparatively evaluated using appropriate performance metrics. It is concluded that gradient boosting–based machine learning algorithms provide a strong alternative to Tweedie regression in terms of model fit when modelling semi-continuous insurance claim cost data.

February 2026, 113 pages

Keywords: Claim cost, Zero inflated data, Tweedie regression, Overdispersion modelling, Machine learning

TEŞEKKÜR

Bu tez çalışmasının her aşamasında bilgi birikimi, yönlendirmeleri ve yapıcı eleştirileriyle bana rehberlik eden, sabır ve özveriyle desteğini her zaman hissettiren danışmanım Sayın Prof. Dr. Furkan Başer'e en içten teşekkürlerimi sunarım. Akademik bakış açımın gelişmesinde ve bu çalışmanın bilimsel bir temele oturmasında sağladığı katkılar benim için son derece kıymetlidir. Lisans ve Lisansüstü eğitimim boyunca derslerime girerek bilgi ve deneyimlerini benimle paylaşan, akademik gelişimime katkıda bulunan tüm değerli hocalarıma da içtenlikle teşekkür ederim.

Akademik hayatım boyunca her zaman yanımda olan, desteğini hiçbir zaman esirgemeyen aileme; özellikle zor zamanlarımda bana güç veren kardeşime ve kuzenlerime sonsuz teşekkür ederim. Bu süreçte hem motivasyon kaynağım olan hem de anlayış ve desteğini esirgemeyen arkadaşlarıma da içtenlikle teşekkür ederim. Ayrıca aynı bölümde birlikte tez yazma sürecini paylaştığım, başarılarımı da stresimi de birlikte yaşadığım değerli arkadaşım Beyza Taşbaş'a da içtenlikle teşekkür ederim.

Buse CAN
Ankara, Şubat 2026

İÇİNDEKİLER

KISALTMALAR DİZİNİ

DVM	Destek Vektör Makineleri
EKK	En Küçük Kareler
EM	Expectation Maximization
GAKA	Gradyan Artırımlı Karar Ağaçları
GDM	Genelleştirilmiş Doğrusal Modeller
GTD	Genelleştirilmiş Tahmin Denklemleri
KKT	Karush–Kuhn–Tucker
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
MAE	Mean Absolute Error
MASS	Motorlu Araç Sorumluluk Sigortası
rMAE	Relative Mean Absolute Error
RMSE	Root Mean Squared Error
RSS	Residual Sum of Squares
rRMSE	Relative Root Mean Squared Error
ÜDA	Üstel Dağılım Ailesi
YSA	Yapay Sinir Ağları

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Sigortacılık sektöründe karşılaşılan temel sorunlardan biri, poliçe sahiplerine uygulanacak primlerin doğru ve adil biçimde belirlenmesidir. Rekabetçi piyasa koşullarında sigortacıların her poliçe sahibinin beklenen kaybına uygun bir prim talep etmesi hem fiyatlandırma doğruluğu hem de müşteri portföyünün sürdürülebilirliği açısından kritik önem taşır. Bu nedenle sigortacılar için en hassas aşama, gelecekte gerçekleşme olasılığı bulunan ve belirsizlik içeren hasar tutarlarını güvenilir biçimde tahmin edebilmektir. Ancak sigorta hasar verilerinin çoğunlukla aşırı sağa çarpık olması ve sıfır değerlerinin yüksek bir noktada yoğunlaşması, geleneksel doğrusal modelleme tekniklerini yetersiz kılmakta; sigorta verisinin kendine özgü yapısına uygun özel istatistiksel modellerin kullanılmasını gerektirmektedir (Yang vd. 2018).

Bu gereklilik, günümüzde büyük veri çağının sunduğu imkânlarla daha da belirgin hâle gelmiştir. Veri miktarının artması, gelişmiş hesaplama gücü ve modern modelleme teknikleri sayesinde sigortacılar artık çok daha geniş ve detaylı veri kümelerini analiz edebilmektedir. Bu gelişmeler, gerek risk faktörlerinin daha doğru tanımlanmasına gerekse poliçe sahiplerine ilişkin belirsizliklerin daha isabetli tahmin edilmesine olanak tanımaktadır. Dolayısıyla veri odaklı analiz yöntemlerinin güçlenmesi hem fiyatlandırma süreçlerinin doğruluğunu artırmakta hem de sigorta şirketlerinin risk yönetimi stratejilerini daha sağlam temellere oturtmasına yardımcı olmaktadır (Jain 2018).

Bu bağlamda sigorta verilerinin önemli bir bölümünü oluşturan yarı sürekli veriler, bilimsel, sosyal ve ekonomik araştırmalarda olduğu gibi sigorta uygulamalarında da kritik bir rol üstlenmektedir. Bu tür verilerin karakteristik özelliği, gözlemlerin önemli bir kısmının tam sıfır değerlerinden oluşması ve sıfır dışındaki pozitif değerlerin aşırı sağa çarpık bir dağılım sergilemesidir. Aktüerya biliminde özellikle otomotiv, mülkiyet ve kaza sigortası uygulamalarında yarı sürekli hasar gözlemleri oldukça yaygındır; burada sıfır değerleri herhangi bir hasarın gerçekleşmediği poliçeleri ifade ederken, hasar gerçekleşen poliçelerde birikimli hasar tutarları yüksek düzeyde sağa çarpık bir davranış göstermektedir (Gu 2024). Bu dağılımsal yapı, sigorta şirketlerinin risk

düzeylerini doğru biçimde belirlemesini güçleştirmekte; dolayısıyla primlerin hatalı belirlenmesi hem ters seçim riskini artırmakta hem de adil fiyatlandırma süreçlerini sekteye uğratmaktadır (Dionne vd. 2001).

Bu nedenle yarı sürekli sigorta verilerinin doğru biçimde modellenmesi sigorta uygulamaları açısından büyük önem taşımaktadır. Geleneksel iki aşamalı yaklaşımlar çerçevesinde hasar sıklığı için Poisson, hasar şiddeti için Gamma modellerinin kullanılması yaygın olmakla birlikte, sıfır yoğunluğu ile pozitif değerlerin karmaşık varyans yapısı nedeniyle bu modeller her durumda yeterli performans göstermemektedir. Tweedie'nin (1984) bileşik Poisson–Gamma yapısını Üstel Dağılım Ailesi (ÜDA) çerçevesine oturtması, sıklık ve şiddet bileşenlerinin tek bir Genelleştirilmiş Doğrusal Model (GDM) altında birlikte tahmin edilmesini mümkün kılmıştır. Ayrıca Smyth ve Jørgensen'in (2002) önerdiği çift genelleştirilmiş doğrusal model yaklaşımı hem ortalama hem de varyansın eş zamanlı modellenmesine olanak tanıyarak sigorta hasar verilerinin karmaşık yapısını daha esnek biçimde temsil etmektedir. Bu çerçevede, yarı sürekli hasar verilerinin modellenmesinde Tweedie yaklaşımına dayalı gerçekleştirilen yöntemler yaygın olarak kullanılmaktadır.

Delong vd. (2021), bağımsız ve özdeş dağılımlı Gamma hasar büyüklüklerine dayalı bileşik Poisson modelini yeniden ele alarak modelin iki temel parametrelendirilmesi—Poisson-Gamma ve Tweedie'nin bileşik Poisson parametrelendirilmesi—arasındaki teorik ilişkileri incelemiştir. Çalışma, uygun bağlantı (link) fonksiyonları altında bu iki yaklaşımın GDM çerçevesinde örtüşebileceğini göstermekte ve özellikle Tweedie parametrelendirilmesinde güç-varyans parametresinin verimli biçimde uyarlanmasına yönelik bir teori sunmaktadır. Uygulamalı analizlerinde yazarlar, sigorta sektöründe Poisson-Gamma parametrelendirmesinin neden daha yaygın kullanıldığını tartışmış ve örnek uygulamalarda bu yaklaşımın hem hesaplama açısından daha elverişli hem de sonuçlar bakımından daha istikrarlı olduğunu göstermiştir. Özellikle sinir ağı tabanlı modellerde Tweedie parametrelendirmesinin güç parametresini güvenilir biçimde tahmin edememesi, uygulamada Poisson-Gamma yaklaşımına açık bir üstünlük sağlamaktadır. Delong vd. (2021), sabit şekil parametreli Gamma modelinin bazı portföylerde makul performans gösterdiğini belirtmekle birlikte, birçok uygulamada

şekil parametresinin değişkenlik gösterdiğini vurgulamaktadır. Bu nedenle sektörün, tazminat tutarının hem konum hem de şekil parametresini esnek biçimde modelleyebilen çift GDM yapılarına yöneldiği ifade edilmektedir. Çalışma, ağır kuyruklu dağılımlar için karışım modellerinin de umut verici alternatifler sunduğunu belirtmektedir.

Yang vd. (2018), sigorta talebi modellemesinde doğrusal olmayan risk faktörleri ile etkileşimlerin geleneksel GDM yapılarıyla sınırlı biçimde temsil edilebildiğini vurgulayarak, Tweedie modeli için ağaç tabanlı esnek bir gradyan artırma yöntemi geliştirmiştir. Yang vd. (2018), doğrusal varsayımlara veya önceden belirlenmiş etkileşim yapılarına ihtiyaç duymayan bu yaklaşımı "TDboost" adlı R paketi içinde uygulamış ve yöntemin karmaşık veri kümelerinde yüksek doğrulukta prim tahmini sağladığını göstermiştir. Kişisel otomobil sigortasında poliçe süresinin maruziyet olarak kullanıldığı uygulamalar, TDboost'un saf prim tahmininde başarılı performans sergilediğini ortaya koymaktadır. Ayrıca yöntem, maruziyetin farklı şekilde tanımlandığı ticari sigorta uygulamalarına da kolayca uyarlanabilmektedir. Yang vd. (2018), TDboost'un bağımsız bir tahmin aracı olmasının yanı sıra, doğrusal olmayan etkileri ve etkileşimleri ortaya çıkararak geleneksel GDM modellerini tamamlayıcı bir rol oynadığını da göstermektedir. Elde edilen etkileşim yapılarının GDM'e entegre edilmesi, model doğruluğunu belirgin biçimde artırmaktadır. Çalışma, yöntemin sigorta alanı dışındaki ekoloji, meteoroloji ve siyaset bilimi gibi Tweedie dağılımının yaygın olarak kullanıldığı disiplinlerde de uygulanabilir olduğunu vurgulayarak, esnek ve parametrik olmayan modelleme yaklaşımlarına katkı sağlamaktadır.

Gu (2024), aşırı sıfır oranlarına sahip sigorta talebi verilerinin modellenmesinde sıfır yığılmalı Tweedie yaklaşımının doğal ve esnek bir çerçeve sunduğunu ortaya koymaktadır. Model hem dağılım yapısının hem de sıfır durumuna ilişkin olasılığın birlikte ele alınmasına imkân tanıyarak genel tahmin performansını geliştirmektedir. Çalışmada, Beklenti Maksimizasyonu (EM, Expectation Maximization) algoritmasının LightGBM tabanlı bir yapı ile entegre edilmesiyle yönteminin kolay uygulanabilir hâle geldiği ve pozitif Tweedie modelinden elde edilen başlangıç değerlerinin EM yinelemelerinin sayısını azalttığı belirtilmektedir. Yazar ayrıca, modelin sıfır yığılmalı

Poisson bileşeninin teorik olarak sıfır yığılmalı Negatif Binom dağılımı ile değiştirilebileceğini ifade etmektedir. Bu durumda, aşırı sıfır içeren birleşik hasar büyüklükleri bileşik Negatif Binom–Gamma karışımı ile modellenebilecektir. Bu genişletilmiş yapı, poliçe bazında hasar sayısında belirgin aşırı değişkenliğin gözlendiği portföylerde potansiyel avantajlar sunmaktadır. Bununla birlikte, Gu (2024), bu alternatif modelin hesaplama açısından daha karmaşık ve maliyetli olabileceğini vurgulamaktadır.

Bu çalışmada, sigorta hasar tutarının risk faktörlerine bağlı olarak modellenmesi ve toplam hasar tutarının kestirimi amacıyla Tweedie bileşik Poisson regresyonu ile makine öğrenmesi yöntemleri bir arada kullanılmıştır. Bu çerçevede, farklı modelleme yaklaşımlarından elde edilen bulgular model performans ölçütlerine bağlı olarak karşılaştırılmış ve yorumlanmıştır. Çalışma, yarı sürekli sigorta verilerinin içerdiği dağılımsal ve yapısal güçlüklerin üstesinden gelebilecek yöntemlerin hem istatistiksel hem de makine öğrenmesi perspektifinden sistematik biçimde değerlendirilmesini sağlayarak, sigorta fiyatlamasında daha nesnel, tutarlı ve veri odaklı kararların alınmasına yöntemsel ve uygulamalı düzeyde özgün bir katkı sunmayı amaçlamaktadır.

Bu araştırmada, taşıt sigortası kapsamında gerçekleşen hasar tutarlarının tahminine yönelik istatistiksel ve makine öğrenmesi temelli modelleme yaklaşımları ele alınmaktadır. Bu kapsamda, klasik istatistiksel modelleme yaklaşımının önemli bir bileşeni olan Tweedie dağılımı ile çeşitli makine öğrenmesi algoritmaları incelenmektedir. Tweedie dağılımı hem sıfır hasar içeren gözlemleri hem de pozitif ve sürekli hasar tutarlarını aynı çerçevede modelleyebilme özelliği nedeniyle sigorta verilerinde yaygın olarak kullanılan bir yöntemdir. Tezde kullanılan veri setleri, farklı sigorta şirketlerinden ve farklı kapsamlardan elde edilmiş olup araç yaşı, araç değeri, maruziyet süresi, poliçe bilgileri, sürücü özellikleri, geçmiş hasar kayıtları ve poliçe bazlı risk faktörleri gibi çok sayıda değişken içermektedir. Makine öğrenmesi yöntemleri ise doğrusal olmayan ilişkileri yakalayabilme yeteneği, yüksek boyutlu veri yapılarına uyum sağlayabilmesi ve farklı türdeki değişkenleri birlikte işleyebilmesi nedeniyle sigorta hasar modellemesi bağlamında ele alınmaktadır.

Çalışmanın Birinci Bölümünde araştırmanın konusu, amacı, önemi ve katkısı ortaya konulmakta; ayrıca konuya ilişkin kısa bir literatür değerlendirmesine yer verilmektedir. İkinci Bölümde, önceki çalışmalar incelenmekte ve literatürde Tweedie dağılımına yönelik yapılan çeşitli araştırmalar özetlenmektedir. Üçüncü Bölüm Tweedie modellerine ayrılmış olup, Tweedie yaklaşımı kuramsal temelleriyle birlikte ayrıntılı biçimde ele alınmaktadır. Dördüncü Bölümde ise makine öğrenmesi yöntemleri kapsamında Klasik Doğrusal Regresyon, LASSO Regresyon, Ridge Regresyon, Karar Ağaçları, Rasgele Orman, Destek Vektör Makineleri, Yapay Sinir Ağları, XGBoost, LightGBM ve CatBoost algoritmaları tanıtılmakta ve temel çalışma prensipleri açıklanmaktadır. Beşinci Bölümde uygulamada kullanılan veri setleri ve bu veri setlerinin temel özellikleri sunulmakta; ayrıca kullanılan yöntemlerin yazılım ortamlarındaki parametre ayarları ile tercih edilen yazılım uygulamaları ayrıntılı olarak açıklanmaktadır. Altıncı Bölümde, her bir veri seti üzerinde algoritmalar ayrı ayrı uygulanmakta ve elde edilen sonuçlar performans ölçütleri yardımıyla karşılaştırılmaktadır. Yedinci ve son bölümde ise elde edilen bulgular özetlenmekte ve çalışmanın genel değerlendirmesi yapılarak nihai sonuçlar sunulmaktadır.

2. ÖNCEKİ ÇALIŞMALARIN İNCELENMESİ

Bu bölümde, yarı sürekli verilerin modellenmesine yönelik olarak Tweedie dağılımını temel alan literatürdeki seçilmiş çalışmalar incelenmektedir. Çalışmalarda kullanılan yöntemler, ele alınan veri yapıları ve elde edilen bulgular genel hatlarıyla özetlenmektedir. Böylece Tweedie yaklaşımının yarı sürekli veri modellemesindeki kuramsal ve uygulamalı konumu ortaya konulmaktadır.

Gao (2024), hasar sayısının gözlenmediği durumlarda saf prim tahminine yönelik marjinal Tweedie bileşik Poisson modellerinin performansını iyileştirmeyi amaçlamaktadır. Çalışma, mevcut tek GDM ve çift GDM yaklaşımlarının Tweedie dağılımının üstel dağılım ailesine ait olmasına rağmen teorik ve pratik açıdan çeşitli sınırlılıklar taşıdığını vurgulamaktadır. Bu çerçevede Gao (2024), EM algoritmasına dayanan yeni bir model uyum yöntemi geliştirmekte ve yöntemin genişletilmiş bir veri seti üzerinde tekrarlamalı olarak yeniden ağırlıklandırılmış Poisson–Gamma regresyonlarına denk düştüğünü göstermektedir. Bu yapı hem uygulama kolaylığı hem de yorumlanabilirlik açısından önemli avantajlar sunmaktadır. Önerilen yöntem, tek GDM ve çift GDM'nin temel zayıflıklarını gidermektedir. Gao (2024), yöntemin hem düşük hem de yüksek sıfır hasar oranlarında istikrarlı performans sergilediğini; güç-varyans parametresinin EM yinelemeleri sırasında doğrudan tahmin edilebildiğini; gerçek veri uygulamalarında daha kararlı sonuçlar verdiğini ve hasar oluşumu ile hasar şiddetinin iki ayrı regresyon fonksiyonu aracılığıyla birlikte modellenebilmesine olanak sağladığını ortaya koymaktadır. Sayısal örnekler, dağılım varsayımının ortalama (saf prim) tahmininde sınırlı bir etkiye sahip olduğunu, ancak geleceğe yönelik hasar tutarı dağılımının tahmini söz konusu olduğunda önerilen yöntemin daha üstün bir performans sergilediğini göstermektedir.

Smolárová (2017), temel odak noktası olarak Bileşik Poisson dağılımını ele almakta ve bu dağılımın, Tweedie modellerinde varyans fonksiyonunun gücünü belirleyen parametre olan ξ'nin $(1, 2)$ aralığında yer aldığı durumlarla doğrudan ilişkili olduğunu göstermektedir. Tweedie bileşik Poisson modeli, ÜDA içinde yer aldığı için GDM ve Genelleştirilmiş Tahmin Denklemleri (GTD) çerçevelerine uyum sağlamakta; model

parametreleri standart GDM/GTD yöntemleriyle tahmin edilebilmekte, güç indeksi parametresi ise profil olabilirlik yaklaşımıyla belirlenmektedir. Çalışmanın temel amacı, Tweedie bileşik Poisson modellerinin hayat dışı sigorta fiyatlandırması ve hasar karşılığı bağlamındaki uygulamalarını ortaya koymaktır. Bu doğrultuda model, saf primlerin ve normalleştirilmiş artan tazminatların tahmininde kullanılmak üzere uyarlanmış ve iki farklı gerçek sigorta veri seti üzerinde uygulanmıştır. Elde edilen bulgular karşılaştırmalı olarak değerlendirilmiş ve modelin performansı tartışılmıştır. Smolárová (2017), Tweedie GTD modellerinin uygulanmasında en önemli zorluğun bu modellere yönelik mevcut yazılım eksikliği olduğunu vurgulamaktadır. Bu nedenle çalışma kapsamında, Tweedie GTD modellerinin AR(1) ve değiştirilebilir korelasyon yapıları altında tahmin edilebilmesini sağlayan gerekli R kodu geliştirilmiştir. Geliştirilen kodlar, hayat dışı sigorta fiyatlandırması ve hasar rezervasyonu uygulamalarına yönelik olarak tez eklerinde sunulmuştur.

Jain (2018), sigorta fiyatlandırmasında Poisson–Gamma GDM, Tweedie GDM ve Yapay Sinir Ağlarını (YSA) karşılaştırarak bu modellerin dağılım varsayımlarına yönelik farklı yaklaşımlarını incelemiştir. Poisson–Gamma modeli dağılımı açıkça tanımlarken, Tweedie GDM varyans güç-parametresi aracılığıyla dağılımı çıkarabilmekte; YSA ise herhangi bir dağılım varsayımı olmaksızın tamamen veri odaklı öğrenme gerçekleştirmektedir. Çalışma sonuçları, Poisson–Gamma GDM'nin tahmin doğruluğu açısından güçlü performans gösterdiğini, ancak risk primi oranlarının 1'den büyük olması nedeniyle aktüeryal denge sağlamadığını ortaya koymaktadır. Buna karşılık, Tweedie GDM ve YSA hem benzer öngörü gücü üretmekte hem de tüketicilerden aşırı prim talep etmeyerek daha adil bir yapı sunmaktadır. Ayrıca bu iki model, risk faktörleri arasındaki karmaşık ilişkileri Poisson–Gamma modeline kıyasla daha başarılı biçimde yakalayabilmektedir. Jain (2018), YSA'nın esnek yapısı ve güçlü tahmin performansına karşın hesaplama maliyetinin yüksekliği ve yorumlanabilirlik sınırlılıklarının önemli dezavantajlar olduğunu vurgulamaktadır. Bununla birlikte çalışma, sigorta fiyatlandırmasında makine öğrenmesi yöntemlerinin—özellikle destek vektör makineleri, regresyon ağaçları ve rasgele ormanlar gibi modellerin—gelecekte daha yaygın biçimde araştırılması gerektiğini belirtmektedir.

Smyth ve Jørgensen (2002), sigorta talebi verilerinin analizinde Tweedie'nin bileşik Poisson dağılımına dayanan yaklaşımın yüksek derecede verimli bir yöntem sunduğunu göstermektedir. Smyth ve Jørgensen (2002), dağılım varsayımlarının standart veri analizi teknikleriyle değerlendirilebileceğini ve genişletilmiş yarı-olasılık yapısı sayesinde yöntemin hem hasar sayısı hem de hasar tutarı için varsayılan Poisson ve Gamma dağılımlarından ölçülü sapmalara karşı duyarlı olmadığını vurgulamaktadır. Çalışmada ayrıca, yöntemin yüksek verimliliğinin bir sonucu olarak, yaklaşık yöntemlere veya tek değişkenli modellere kıyasla daha fazla sayıda model teriminin anlamlı bulunabileceği ifade edilmektedir. Bu durum özellikle etkileşim teriminin etkilerinin tespit edilmesinde kendini göstermekte olup, bazı sigorta uygulamalarında modelin pratik kullanımını karmaşıklaştırabilecek düzeyde etkileşimlerin ortaya çıkabileceği belirtilmektedir. Bununla birlikte, Smyth ve Jørgensen (2002), çoğu durumda ana etkilerin baskın olduğunu ve etkileşim terimlerinin—İsveç motorlu taşıt sigortası verilerinde olduğu gibi—sınırlı önem taşıyarak göz ardı edilebileceğini bildirmektedir.

3. TWEEDIE REGRESYONU

Bu bölümde, Tweedie üstel dağılım ailesine (ÜDA) dayalı Genelleştirilmiş Doğrusal Modeller (GDM) ele alınmaktadır. Tweedie dağılımları, Normal, Poisson, Gamma ve Ters Gauss dağılımlarını özel durumlar olarak içeren; dolayısıyla bu dağılımları genelleştiren esnek bir dağılım ailesidir. Ayrıca, bu aile başka birçok olasılık dağılımını da kapsayarak geniş bir modelleme çerçevesi sunar (Dunn ve Smyth 2018).

Bu bölümde öncelikle, Tweedie dağılımının ÜDA içerisindeki matematiksel temeli ortaya konulmakta ve dağılımın yapısını belirleyen indeks parametresi ile güç-varyans fonksiyonu temel alınarak yapılan sınıflandırma sunulacaktır. Devamında, Tweedie dağılımının bileşik Poisson–Gamma yapısı ayrıntılı biçimde ele alınacak ve bu yapıdan varyans fonksiyonunun nasıl türetildiği açıklanacaktır. Sonraki aşamada Tweedie ÜDA'ların farklı veri türlerine göre nasıl sınıflandırıldığı incelenecektir. Bu kapsamda, pozitif sürekli verilerin modellenmesi için kullanılan Tweedie ÜDA'ları (Gamma ve Ters Gauss dağılımlarının özel halleri dâhil) ele alınacak; ardından sıfır değerleri içeren pozitif sürekli verileri modellemeye imkân veren Tweedie dağılımları tartışılacaktır.

Bu teorik yapıların uygulamadaki karşılığı ise Tweedie GDM modelleri üzerinden ortaya konulmaktadır. Bu bölümde Tweedie GDM'lerinin kurulumu, bağlantı fonksiyonları, dağılım parametrelerinin tahmini ve uygulama adımları ayrıntılı biçimde açıklanmaktadır. Ayrıca, Tweedie modelinin uygulanmasında kritik öneme sahip olan indeks parametresi ξ'nin tahmin yöntemleri de ayrıntılı olarak ele alınmaktadır.

3.1 Tweedie Dağılımının ÜDA İçindeki Matematiksel Yapısı

ÜDA, hem sürekli dağılımları (örneğin Normal ve Gamma) hem de kesikli dağılımları (örneğin Binom ve Poisson) kapsayan geniş bir dağılım sınıfıdır. Bununla birlikte, bu aile, kesikli ve sürekli bileşenlerin bir arada bulunduğu karışım (mixture) dağılımları da içermektedir. Sigorta hasar tutarlarının modellenmesinde sıklıkla kullanılan bu tür karışım dağılımlardan biri Tweedie dağılımıdır. Tweedie dağılımının yapısında, sıfır

değeri “hasar yapılmadığı” durumu, sıfırdan büyük sürekli değerler ise “bir veya birden fazla hasar sonucunda oluşan toplam hasar tutarı tutarını” temsil etmektedir (Frees vd. 2014).

Tweedie dağılımı, Gamma dağılımlı rasgele değişkenlerin Poisson toplamı olarak tanımlanır. Başka bir ifadeyle, ortalaması λ olan bir Poisson dağılımına sahip N değişkeninin toplam hasar sayısını temsil ettiği varsayılır. Ayrıca, her biri hasar tutarını gösteren ve α ile γ parametrelerine sahip Gamma dağılımlı, bağımsız ve özdeş dağılmış rasgele değişkenler $Y_j (j = 1,2, \dots, N)$ tanımlanır. Bu durumda toplam hasar tutarı: $S_N = Y_1 + Y_2 + \cdots + Y_N$ şeklinde ifade edilir. Burada S_N, Poisson sayıda Gamma değişkeninin toplamıdır ve Tweedie dağılımının temelini oluşturur. Bu yapı, sigorta verilerinde sıklıkla gözlemlenen “sıfır yığılmalı” ve “sağa çarpık sürekli değer” özelliklerini tek bir modelde temsil edebilmesi nedeniyle, Tweedie dağılımını aktüeryal modelleme açısından oldukça güçlü ve pratik bir araç haline getirmektedir (Frees vd. 2014).

Dunn ve Smyth (2005), literatürdeki uygulamaları inceleyerek Tweedie modellerinin aktüerya çalışmaları, sağkalım analizi, ekoloji ve meteoroloji gibi farklı alanlarda yaygın olarak kullanıldığını göstermiştir. Tweedie modelinin genel gösterimi $y \sim Tw_\xi(\mu, \phi)$ biçimindedir; burada y, ortalaması μ, dağılma parametresi ϕ ve güç indeksi parametresi $\xi \in \mathbb{R}$ olan bir rasgele değişkendir. Jørgensen (1997), Tweedie modellerinin ölçek dönüşümlerine göre kapalı olan, yani ölçekten bağımsız tek üstel dağılım modelleri olduğunu göstermiştir. Dolayısıyla, $y \sim Tw_\xi(\mu, \phi)$ ise, herhangi bir pozitif sabit c için $cy \sim Tw_\xi(c\mu, c^{2-\xi}\phi)$ olur. Bu özellik, ölçü biriminin keyfi olduğu durumlarda Tweedie modellerini veri modellemesi açısından uygun hale getirir.

Tweedie dağılımının karışım yapısı, sıfır hasar olasılığı üzerinden açık biçimde görülebilir. Toplam hasar tutarı S_N olarak tanımlandığında, sıfır hasar olasılığı

$$P(S_N = 0) = P(N = 0) = e^{-\lambda} \tag{3.1}$$

biçiminde hesaplanır. Bu durumda dağılım fonksiyonu koşullu beklenen değer tanımı kullanılarak

$$Pr(S_N \leq y) = e^{-\lambda} + \sum_{n=1}^{\infty} P(N = n) Pr(S_n \leq y),\ y \geq 0 \tag{3.2}$$

biçiminde ifade edilir. Bağımsız ve özdeş dağılıma sahip Gamma rasgele değişkenlerinin toplamı yine bir Gamma dağılımı olduğundan, S_n değişkeni $n\alpha$ ve γ parametrelerine sahip bir Gamma dağılımına uyar. Bu durumda, $y > 0$ için Tweedie dağılımının olasılık yoğunluk fonksiyonu,

$$f_s(y) = \sum_{n=1}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \frac{\gamma^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha-1} e^{-\gamma y} \tag{3.3}$$

biçiminde tanımlanır.

İlk bakışta bu yoğunluk ifadesi ÜDA yapısına uymuyor gibi görünse de momentler incelendiğinde Tweedie dağılımının bu aileye ait olduğu görülmektedir. Beklenen değer ve varyans,

$$E[S_N] = \frac{\lambda\alpha}{\gamma},\ Var[S_N] = \frac{\lambda\alpha}{\gamma^2}(1 + \alpha) \tag{3.4}$$

biçiminde elde edilir. Bu ifadelerden hareketle Tweedie dağılımının parametreleri,

$$\lambda = \frac{\mu^{2-\xi}}{\phi(2-\xi)},\ \alpha = \frac{2-\xi}{\xi-1},\ \frac{1}{\gamma} = \phi(p-1)\mu^{\xi-1} \tag{3.5}$$

biçiminde yeniden tanımlanabilir. Bu parametrelerin yerine konulmasıyla elde edilen yoğunluk fonksiyonu, Tweedie dağılımının ÜDA içerisinde yer aldığını göstermektedir. Dolayısıyla,

$$E[S_N] = \mu,\ Var[S_N] = \phi\mu^{\xi},\ (1 < \xi < 2) \tag{3.6}$$

şeklinde olur. Bu özellikler, Tweedie dağılımını Poisson ve Gamma dağılımları arasında bir geçiş modeli olarak konumlandırır.

Tweedie tabanlı GDM kapsamında, açıklayıcı değişkenler $x_{i,T}$ ve bunlara karşılık gelen regresyon katsayıları β_T ile temsil edilir. Logaritmik bağlantı fonksiyonu kullanıldığında ortalama

$$\mu_i = \exp(x'_{i,T}\beta_T) \tag{3.7}$$

biçiminde ifade edilir. Her ne kadar Tweedie dağılımı için olasılık fonksiyonunun kapalı formda bir ifadesi bulunmasa da uygulamada bu hesaplamalar R yazılımındaki ptweedie fonksiyonu aracılığıyla doğrudan gerçekleştirilebilmektedir. Bu özelliği sayesinde Tweedie dağılımı hem sıfır içeren hem de pozitif sürekli değerler barındıran verilerin modellenmesinde son derece esnek, etkili ve pratik bir yöntem olarak öne çıkmaktadır (Frees vd. 2014).

3.2 Tweedie İndeks Parametresi ve Güç-Varyans Fonksiyonu Temelli Dağılım Sınıflandırması

Binom ve Negatif Binom dağılımları dışında, birçok olasılık dağılımı benzer biçimlerde tanımlanan varyans fonksiyonlarına sahiptir. Örneğin, Normal dağılım için varyans fonksiyonu $V(\mu) = \mu^0 = 1$ biçimindedir; Poisson dağılımı için $V(\mu) = \mu^1$; Gamma dağılımı için $V(\mu) = \mu^2$; ve Ters Gauss dağılımı için $V(\mu) = \mu^3$ olarak ifade edilir (Dunn ve Smyth 2018).

Bu dağılımlar genel olarak, $V(\mu) = \mu^\xi$ biçimindeki güç-varyans fonksiyonu ile temsil edilir. Burada ξ, dağılımın Tweedie indeks parametresi olarak adlandırılır. Bu parametre, $0 < \xi < 1$ aralığı dışındaki herhangi bir reel değeri alabilir (Jørgensen 1997). Varyans fonksiyonu bu biçimi aldığında, söz konusu dağılım Tweedie dağılımı veya Tweedie ÜDA olarak adlandırılır. Güç-varyans ilişkisinin bu türü, doğadaki birçok

popülasyon ve süreçte gözlemlenen istatistiksel bir olgudur (Taylor 1961, Tweedie 1946).

Tweedie modelleri sınıfı, kesikli, sürekli ve karışım dağılımları kapsar. Güç indeksi parametresi ξ'nin değeri, hangi dağılımın elde edileceğini belirler. Normal ($\xi = 0$), Poisson ($\xi = 1$ ve $\phi = 1$), Gamma ($\xi = 2$) ve Ters Gauss ($\xi = 3$) dağılımları bu sınıfın özel durumlarıdır (Smolárová 2017). Diğer dağılımlar daha az bilinmekle birlikte, Jørgensen (1997), Tweedie modellerinin $(0,1)$ aralığı dışındaki tüm ξ değerleri için var olduğunu göstermiştir.

Tweedie dağılımlarında ξ, Tweedie indeks parametresi olarak adlandırılır ve bu parametre, dağılım ailesi içerisindeki belirli bir dağılım türünü tanımlar (Dunn ve Smyth 2018):

$\xi \leq 0$ olduğunda, Tweedie dağılımları $-\infty < y < \infty$ aralığında tanımlı sürekli verilerin modellenmesi için kullanılabilir. Bu durumda Normal dağılım ($\xi = 0$) özel bir örnektir. Ancak $\xi < 0$ için Tweedie dağılımları, değişken (y) tüm reel eksende tanımlanmasına karşın ortalamasının (μ) yalnızca pozitif olabilmesi gibi alışılmadık bir özellik gösterir. Bu tür dağılımların bilinen pratik uygulamaları bulunmadığından, genellikle istatistiksel modellemede dikkate alınmazlar (Dunn ve Smyth 2018).

$\xi = 1$ olduğunda, Tweedie dağılımları kesikli verilerin modellenmesinde kullanılır. Bu durumda y değişkeni $0, \varphi, 2\varphi, 3\varphi, \ldots$ biçiminde kesikli değerler alır. Örneğin, $\varphi = 2$ olduğunda $y = 0,2,4, \ldots$ değerleri için pozitif olasılıklar tanımlıdır. Poisson dağılımı, bu grubun özel bir durumu olup $\varphi = 1$ değerinde elde edilir (Dunn ve Smyth 2018).

$1 < \xi < 2$ aralığında yer alan Tweedie modelleri, Gamma dağılımlarının Poisson karışımı biçimindedir. Bu sözde bileşik Poisson dağılımları, sıfır değerini alma olasılığı pozitif olan, negatif olmayan reel sayılar üzerinde tanımlı karışım dağılımlardır. Sıfır noktasında kesikli bir değerin bulunması, bu dağılımları gözlemlerin sıklıkla sıfır olduğu ancak zaman zaman pozitif değerlerin de gözlendiği uygulamalarda son derece

uygun hale getirir. Bu nedenle Tweedie modelleri, hayat dışı sigortacılıkta fiyatlandırma ve hasar karşılığı ayırma alanlarında yaygın biçimde kullanılmaktadır. Ayrıca bu dağılımlar, literatürde bileşik Gamma veya Poisson-Gamma dağılımları olarak da anılmaktadır (Smolárová 2017).

$\xi \geq 2$ durumunda elde edilen dağılımlar, pozitif reel sayılar üzerinde tanımlı sürekli dağılımlardır. Bu dağılımlar, Gamma dağılımına benzer olmakla birlikte daha belirgin bir sağa çarpıklık gösterirler. Gamma dağılımı ile, genellikle hasar şiddetinin modellenmesinde kullanılmaktadırlar. Buna karşılık, ξ'nin negatif değerleri tüm reel eksende tanımlı sürekli dağılımlar verir; ancak bu tür dağılımların sigortacılık uygulamalarında kullanımı henüz önerilmemiştir (Smolárová 2017). Bu durumda Gamma dağılımı ($\xi = 2$) ve Ters Gauss dağılımı ($\xi = 3$) Tweedie ailesinin özel örnekleridir. Ayrıca ξ arttıkça, dağılımın şekli giderek sağa çarpık hale gelir (Dunn ve Smyth 2018).

3.3 Tweedie Dağılımının Bileşik Poisson–Gamma Yapısı ve Varyans Fonksiyonunun Türetilmesi

Tweedie dağılımı, varyans fonksiyonunun ortalama ile güç biçiminde ilişkilendirildiği bir dağılım ailesine aittir. Bu yapı genel olarak

$$V(\mu) = \mu^{\xi} \tag{3.8}$$

şeklinde ifade edilir. Burada ξ, yanıt değişkeninin dağılım türünü belirleyen bir parametresini; μ ise yanıt değişkeninin ortalamasını ifade etmektedir. Bu yaklaşım, beklenen hasar tutarlarının bileşik Poisson dağılımı izlediği varsayımına dayanmaktadır (Smyth ve Jørgensen 2002).

Bu çerçevede, N_i değişkeninin $i.$ kategoriye ait gözlemlenen hasar sayısını, Z_i değişkeninin ise aynı kategorideki gözlemlenen toplam hasar tutarını temsil ettiği

varsayılır. Her bir kategori için risk altındaki birim sayısı w_i ile gösterilir ve bunun bir poliçe yılı olduğu kabul edilir (Jain 2018). Buna göre gözlemlenen hasar tutarı

$$Y_i = \frac{Z_i}{w_i} \tag{3.9}$$

şeklinde tanımlanır. Eğer $w_i = 1$ olduğu kabul edilirse bu ifade $Y_i = Z_i$ biçimine indirgenir. Böylece sıklık (frekans) ve şiddet bileşenleri ayrı ayrı modellenmek yerine, toplam hasar tutarı tek adımda modellenmiş olur (Jain 2018).

Model varsayımlarına göre, hasar sayısı N_i, ortalaması λ_i olan bir Poisson dağılımına; hasar tutarı ise ortalaması τ_i ve şekil parametresi α olan bir Gamma dağılımına uymaktadır (Smyth ve Jørgensen 2002). Bu varsayımlar altında, N_i verildiğinde ve $N_i > 0$ olduğu durumlarda Y_i'nin koşullu dağılımı, ortalaması $N_i\tau_i$ olan bir Gamma dağılımı olarak ifade edilir (Jain 2018).

Tweedie yaklaşımı, yanıt değişkeninin varyansının ortalama ile üstel bir ilişki gösterdiğini varsayar. Buna göre $\mu_i = E(Y_i) = \lambda_i\tau_i$ şeklinde tanımlanır ve varyans fonksiyonu $V(Y_i) = \phi\mu_i^{\xi}$ biçimini alır. Burada ξ, dağılım parametresi ya da diğer adıyla varyans gücü parametresi olarak adlandırılmaktadır. Bu parametre, Gamma dağılımının şekil parametresi α

$$\xi = \frac{\alpha+2}{\alpha+1} \tag{3.10}$$

bağıntısına sahiptir.

Eş. (3.10), $\alpha > 0$ olduğu durumda dağılımın bileşik Poisson yapıda olması için $1 < \xi < 2$ aralığında bulunması gerektiğini göstermektedir (Jain 2018).

Koşullu varyans tanımı kullanılarak Y_i'nin varyansı,

$$V(Y_i) = E_{N_i}[V(Y_i \mid N_i)] + V_{N_i}[E(Y_i \mid N_i)] = (\frac{1}{\alpha} + 1)\lambda_i \tau_i^2 \quad (3.11)$$

biçiminde elde edilir. Bu sonuç ile varyansın genel formu eşitlenirse,

$$\phi \mu_i^{\xi} = (\frac{1}{\alpha} + 1)\lambda_i \tau_i^2 \quad (3.12)$$

dir. Burada, $\xi = \frac{\alpha+2}{\alpha+1}$ eşitliği kullanıldığında, $\alpha = \frac{2-\xi}{\xi-1}$ bulunur. Bu ifade varyans denklemine yerleştirildiğinde:

$$\phi \mu_i^{\xi} = \frac{1}{2-\xi} \mu_i \tau_i^2 \quad (3.13)$$

sonucuna ulaşılır. Buradan dağılım parametresi ϕ, ortalama parametreleri λ_i ve τ_i ile varyans gücü parametresi ξ cinsinden

$$\phi = (\frac{1}{2-\xi})\, \lambda_i^{1-\xi} \tau_i^{2-\xi} \quad (3.14)$$

biçiminde ifade edilir. Bu sonuç, dağılım parametresi ϕ'nin Poisson ortalaması λ_i, Gamma ortalaması τ_i ve varyans gücü parametresi ξ ile doğrudan ilişkili olduğunu ve dolayısıyla bu parametrelerin kullanılmasıyla tahmin edilebileceğini göstermektedir (Smyth ve Jørgensen 2002).

3.4 Tweedie ÜDA ile Frekans-Şiddet Modeli Arasındaki İlişki

Alternatif bir yaklaşım olarak, sigorta verilerinde toplam hasar tutarının frekans ve şiddet bileşenlerine ayrıldığı iki aşamalı bir modelleme çerçevesi kullanılabilir. Bu yöntem, özellikle hasar sıklığı ile hasar büyüklüğünün birbirinden farklı süreçler tarafından belirlendiği durumlarda oldukça etkili bir modelleme stratejisi sunmaktadır (Frees vd. 2014).

İlk aşamada, frekans bileşeni, birey bazında gerçekleşen hasar sayısını modellemek amacıyla Poisson regresyon modeli kullanılarak ifade edilir. Bu durumda model şu şekilde tanımlanır:

$$N_i \sim Poisson(\lambda_i), \qquad \lambda_i = \exp(x'_{i,F}\beta_F) \tag{3.15}$$

Burada $x_{i,F}$, frekans modellemesinde kullanılan değişkenleri (örneğin yaş, araç tipi, sigorta süresi vb.) temsil ederken; β_F, bu değişkenlere ait regresyon katsayılarını göstermektedir. Modelde logaritmik bağlantı fonksiyonu kullanılması, ortalama hasar sayısının pozitif olmasını garanti etmektedir (Frees vd. 2014).

İkinci aşamada, şiddet bileşeni, bireysel hasar tutarlarını modellemek için Gamma regresyon modeli ile tanımlanır. Her bir hasar için model şu şekilde yazılmaktadır:

$$y_{ij} \sim Gamma(\alpha, \gamma_i) \quad \text{ve} \quad E[y_{ij}] = \frac{\alpha}{\gamma_i} = \exp(x'_{i,S}\beta_S) \tag{3.16}$$

Bu ifadede $x_{i,S}$, şiddet modelinde kullanılan değişkenleri; β_S ise bunlara karşılık gelen regresyon katsayılarını temsil etmektedir. Böylece, frekans ve şiddet modelleri aynı veya farklı değişken setleri üzerinden tanımlanabilir. İki süreç birbirinden bağımsız olmak zorunda olmamakla birlikte, istatistiksel olarak ayrı biçimde modellenmektedir (Frees vd. 2014).

Bu iki bileşenin birleştirilmesiyle toplam hasar tutarı değişkeni elde edilmektedir:

$$S_{N,i} = y_{ij} + \cdots + y_{i,N_i} \tag{3.17}$$

Toplam hasar tutarının beklenen değeri ve varyansı sırasıyla şu şekilde ifade edilmektedir:

$$E[S_{N,i}] = E[N_i] \times E[y_{ij}] = \exp(x'_{i,F}\beta_F + x'_{i,S}\beta_S) \tag{3.18}$$

$$Var[S_{N,i}] = \lambda_i \frac{\alpha}{\gamma_i^2}(1+\alpha) = \exp\left(x'_{i,F}\beta_F + 2x'_{i,S}\beta_S + \ln\left(1+\frac{1}{\alpha}\right)\right) \tag{3.19}$$

Bu sonuçlar, toplam hasar tutarının hem hasar sayısındaki değişkenlikten hem de bireysel hasar tutarındaki değişkenlikten etkilendiğini göstermektedir. Ayrıca, λ_i ve γ_i parametrelerinin her bir gözlem için farklılık gösterebileceği, yani bireye özgü olduğu unutulmamalıdır (Frees vd. 2014).

Tweedie dağılımı, bu iki bileşenli frekans-şiddet modelinin bileşik bir temsilini sunmaktadır. Tweedie parametrizasyonu ile ilişkili dönüşümler

$$\xi = \frac{\alpha+2}{\alpha+1}, \quad \mu_i = \frac{\lambda_i \alpha}{\gamma_i}, \quad \phi_i \mu_i^{\xi} = \lambda_i \frac{\alpha}{\gamma_i^2}(1+\alpha) \tag{3.20}$$

biçiminde verilmektedir. Bu parametrik ilişkiler, Tweedie dağılımının hem Poisson hem de Gamma bileşenlerini içeren bir yapıya sahip olduğunu açıkça göstermektedir. Uygulamada, toplam hasar tutarının dağılım fonksiyonu R yazılımındaki `ptweedie` fonksiyonu kullanılarak doğrudan hesaplanabilmektedir. Sonuç olarak, bu biçimde oluşturulan frekans-şiddet modelleme çerçevesi hem beklenen toplam hasar tutarının tahmininde hem de risk priminin belirlenmesinde sigorta analitiği açısından güçlü ve esnek bir yöntem sunmaktadır (Frees vd. 2014).

3.5 Tweedie ÜDA'ların Yapısı

Tweedie dağılımları, varyans fonksiyonu $V(\mu) = \mu^{\xi}$ olan ÜDA üyeleri olarak tanımlanmaktadır. Bu ilişki kullanılarak, θ ve $\kappa(\theta)$ ifadeleri elde edilmektedir (Dunn ve Smyth 2018). İntegrasyon sabitlerinin sıfır olarak alınmasıyla

$$\theta = \begin{cases} \frac{\mu^{1-\xi}}{1-\xi}, & \xi \neq 1 \\ log\ \mu, & \xi = 1 \end{cases} \quad ve \quad \kappa(\theta) = \begin{cases} \frac{\mu^{2-\xi}}{2-\xi}, & \xi \neq 2 \\ log\ \mu, & \xi = 2 \end{cases} \tag{3.21}$$

biçiminde elde edilir.

Diğer parametre, integrasyon sabitleri farklı değerlere ayarlanarak elde edilebilmektedir. Faydalı bir parametre, θ ve $\kappa(\theta)$'nın ξ'nin sürekli fonksiyonları olmasını sağlamaktadır (Dunn ve Smyth 2008). θ ve $\kappa(\theta)$ ifadeleri ξ'yi içerdiği için, Tweedie dağılımları yalnızca ξ bilindiğinde ÜDA olarak tanımlanabilir. Uygulamalarda ξ değeri genellikle tahmin edilmektedir. Eğer y, ortalaması μ, dağılım parametresi ϕ ve indeks parametresi ξ olan bir Tweedie dağılımına sahipse, $y \sim Tw_{\xi}(\mu, \phi)$ biçiminde gösterilmektedir (Dunn ve Smyth 2018).

Bu ifadeler temel alınarak Tweedie olasılık fonksiyonu kanonik biçimde (Eş. (3.22)) yazılabilmektedir. Ancak, Normal, Poisson, Gamma ve Ters Gauss dağılımları gibi özel durumlar dışında normalleştirme sabiti $a(y, \phi)$ kapalı formda elde edilemez. Bu nedenle genel Tweedie ÜDA'larında olasılık fonksiyonunun hesaplanması sayısal yöntemler gerektirmektedir (Dunn ve Smyth 2005, Dunn ve Smyth 2008).

Birim sapma

$$d(y,\mu) = \begin{cases} 2\left\{\frac{\max(y,0)^{2-\xi}}{(1-\xi)(2-\xi)} - \frac{y\mu^{1-\xi}}{1-\xi} + \frac{\mu^{2-\xi}}{2-\xi}\right\} & \xi \neq 1,2, \\ 2\left\{y\log\frac{y}{\mu} - (y-\mu)\right\} & \xi = 1, \\ 2\left(-log\frac{y}{\mu} + \frac{y-\mu}{\mu}\right) & \xi = 2. \end{cases} \tag{3.22}$$

biçiminde ifade edilir.

$y = 0$ durumunda birim sapma, yalnızca $\xi \leq 0$ ve $1 < \xi < 2$ için sonlu değerler almaktadır. Tweedie olasılık fonksiyonu, bu birim sapma kullanılarak dağılım modeli biçiminde de ifade edilebilmektedir. Bu formda normalleştirme sabiti $b(y, \phi)$, dört özel durum dışında kapalı biçimde yazılamaz. Eyer noktası yaklaşımı kullanıldığında, doğrusal öngörücüde p parametreli bir model için $D(y, \hat{\mu}) \sim \chi^2_{n-p'}$ yaklaşık olarak geçerlidir. Ancak, $1 < \xi < 2$ ve $y = 0$ içeren durumlarda yaklaşımın doğruluğu zayıflayabilir. Ayrıca, $\xi = 3$ değeri, eyer noktası yaklaşımının tam olarak geçerli olduğu Ters Gauss dağılımına karşılık gelmektedir (Dunn ve Smyth 2018).

Tweedie dağılımları için dikkat çekici bir özellik, yeniden ölçeklendirme kimliğidir (Dunn ve Smyth 2008). Eğer $P_\xi(y;\mu,\phi)$, indeks parametresi ξ olan bir Tweedie ÜDA'nın olasılık fonksiyonu ise:

$$P_\xi(y;\mu,\phi) = cP_\xi(cy;c\mu,c^{2-\xi}\phi) \tag{3.23}$$

ifadesi tüm ξ değerleri için geçerlidir; burada $y > 0$ ve $c > 0$'dır (Dunn ve Smyth 2018).

3.6 Pozitif Sürekli Veriler için Tweedie ÜDA'lar

Birçok durumda pozitif sürekli değişkenler, Gamma veya Ters Gauss dağılımları ile yeterli biçimde modellenebilmektedir. Ancak bazı durumlarda, özellikle de verilerin aşırı derecede çarpık olduğu örneklerde bu dağılımlar yetersiz kalmaktadır. Varyans fonksiyonu $V(\mu) = \mu^\xi$ biçiminde tanımlanan ve $\xi \geq 2$ olan tüm Tweedie dağılımları, pozitif sürekli verileri modellemek için uygundur. Gamma ($\xi = 2$) ve Ters Gauss ($\xi = 3$) dağılımları bu grubun yalnızca özel iki örneğidir ve kapalı formda yazılabilen olasılık fonksiyonuna sahip tek Tweedie ÜDA türleridir. Bu bağlamda önemli bir örnek, varyans fonksiyonunun $V(\mu) = \mu^4$ olduğu durumdur. Bu tür bir yapı, doğrusal regresyon modellerinde yanıt değişkenine $1/y$ dönüşümünün uygulanmasıyla yaklaşık olarak aynı etkiye sahiptir (Dunn ve Smyth 2018).

3.7 Sıfır Yığılmalı Pozitif Sürekli Veriler için Tweedie ÜDA'lar

$1 < \xi < 2$ aralığında yer alan Tweedie dağılımları, sıfır değerleri içeren sürekli verilerin modellenmesinde etkili bir şekilde kullanılmaktadır. Bu tür verilere örnek olarak sigorta hasar tutarı verileri gösterilebilmektedir (Jørgensen ve de Souza 1994, Smyth ve Jørgensen 1999). Belirli bir zaman diliminde bir sigorta şirketinde yapılan hasar sayısı N olsun ve bu değişkenin Poisson dağılımına uyduğu varsayılsın: $N \sim Pois(\lambda^*)$. Burada λ^*, söz konusu zaman diliminde beklenen ortalama hasar sayısını ifade etmektedir. Bu durumda, herhangi bir hasar olmadığında $N = 0$ olabilir. Eğer

$N > 0$ ise, her bir hasar tutarı z_i (i $= 1, \ldots,$ N) pozitif bir değer alır ve ortalaması μ^*, dağılım parametresi ϕ^* olan bir Gamma dağılımına uyar: $z_i \sim Gam(\mu^*, \phi^*)$. Toplam hasar tutarı y,

$$y = \sum_{i=1}^{N} z_i \tag{3.24}$$

biçiminde ifade edilir. Burada $N = 0$ olduğunda $y = 0$'dır. Bu durumda toplam hasar tutarı y, $1 < \xi < 2$ olan bir Tweedie dağılımına sahiptir. Başka bir ifadeyle, y, Gamma dağılımlarının bir Poisson toplamı olarak düşünülebilmektedir. Bu nedenle, bu aralıktaki Tweedie dağılımları genellikle Poisson–Gamma dağılımları olarak adlandırılmaktadır (Smyth 1996). Ancak, bu terim bazı kaynaklarda benzer ancak farklı anlamlarda da kullanılabilmektedir (Foster 2013).

Tweedie ÜDA parametreleri olan μ, ϕ ve ξ temel Poisson ve Gamma dağılımlarının parametreleriyle

$$\lambda^* = \frac{\mu^{2-\xi}}{\phi(2-\xi)} \tag{3.25}$$

$$\mu^* = (2-\xi)\phi\mu^{\xi-1} \tag{3.26}$$

$$\mu^* = (2-\xi)(\xi-1)\phi^2\mu^{2(\xi-1)} \tag{3.27}$$

biçiminde ilişkilidir.

$1 < \xi < 2$ aralığındaki Tweedie dağılımları $y > 0$ için sürekli bir yapıya sahiptir; ancak $y = 0$ noktasında pozitif bir olasılık (π_0) bulunur. Bu olasılık

$$\pi_0 = \Pr(y = 0) = \exp(-\lambda^*) = \exp\left\{-\frac{\mu^{2-\xi}}{\phi(2-\xi)}\right\} \tag{3.28}$$

biçiminde ifade edilir.

π_0'ın en çok olabilirlik tahmini elde edilirken, μ, ξ ve ϕ'nin kendi maksimum olabilirlik tahminleri (Eş. (3.28)) kullanılmalıdır (Dunn ve Smyth 2018).

μ, ϕ ve ξ'nin maksimum olabilirlik tahminleri elde edildikten sonra, temel Poisson ve Gamma dağılımlarına karşılık gelen parametreler olan λ^*, μ^* ve ϕ^*'nin tahminleri Eş. (3.25), Eş. (3.26) ve Eş. (3.27) yardımıyla hesaplanabilmektedir. Bu tahminler, Tweedie modelinin yaklaşık bir yorumu olarak değerlendirilebilir ve bazı uygulamalarda modelin anlaşılmasını kolaylaştırıcı bir nitelik taşımaktadır (Brown ve Dunn 2011, Dunn 2004, Dunn ve Smyth 2005).

3.8 Tweedie GDM'leri

Tweedie dağılımlarına dayalı GDM, uygun bir bağlantı fonksiyonu ile ifade edilen Tweedie GDM'leridir. Bu modellerin her iki temel durumu için (yani $\xi > 2$ ve $1 < \xi < 2$) ortalama değerin $\mu > 0$ olması gerekmektedir. Bu nedenle, Tweedie GDM'lerinde genellikle logaritmik bağlantı fonksiyonu kullanılmaktadır. Dağılım parametresi ϕ, genellikle Pearson tahmini yöntemiyle belirlenmektedir (Dunn ve Smyth 2018). Bununla birlikte, $1 < \xi < 2$ durumunda sıfır yığılmalı (zero inflated) gözlemleri için olasılığın en yüksek olabilirlik tahmininin yapılabilmesi adına ϕ'nin de maksimum olabilirlik yöntemiyle tahmin edilmesi gerekmektedir (Dunn ve Smyth 2018).

Tweedie GDM'lerin kurulabilmesi için, modelin ait olduğu Tweedie ailesini tanımlayan indeks parametresi ξ değerinin belirlenmesi gerekir. Ancak bu değer genellikle doğrudan bilinmez ve model kurulmadan önce tahmin edilmesi gerekir. $\hat{\xi}$ ile $\hat{\beta}$ arasındaki korelasyon genellikle zayıf olduğundan, $\hat{\xi}$'nin tahmini kullanılarak yapılan çıkarımlar, ξ'nin gerçek değerinin bilinmesine göre yalnızca küçük farklar yaratmaktadır (Dunn ve Smyth 2018).

Yanıt değişkenine Box–Cox dönüşümü uygulanan doğrusal regresyon modelleri, Tweedie GDM'lerine benzer bir ortalama-varyans yapısı sunmaktadır. Ancak çıkarım açısından değerlendirildiğinde, Box–Cox dönüşümüne dayalı normal yaklaşım özellikle

geniş aralıklı yanıtlar veya sıfır/sıfıra yakın değerler içeren verilerde yetersiz kalabilmektedir. Bu durumlarda, Tweedie GDM yaklaşımı daha sağlam ve güvenilir sonuçlar üretme eğilimindedir (Dunn ve Smyth 2018).

3.9 Tweedie GDM'lerinde Ortalama ve Dağılım Parametrelerinin Eşzamanlı Modellemesi

$\alpha > 0$ olduğundan $\xi \in (1,2)$ aralığında yer aldığını hatırlatmak gerekir. Smyth ve Jørgensen'in (2002) belirttiği üzere, hasar tutarının temel frekansını (λ_i) ve ortalama büyüklüğünü (şiddet; τ_i) etkileyen herhangi bir faktör, hem ortalamayı ($\mu_i = \lambda_i \tau_i$) hem de dağılımı ($\phi_i = \lambda_i^{1-\xi} \tau_i^{2-\xi}/(2-\xi)$) etkileme potansiyeline sahiptir. Ortalamayı etkileyip dağılımı etkilemeyen bir değişken söz konusu olduğunda, λ_i ve τ_i parametrelerinin $\lambda_i^{1-\xi} \tau_i^{2-\xi}$ ifadesinin sabit kalacak şekilde eşzamanlı değişmesi gerekir; bunun pratikte gerçekleşmesi ise pek olası değildir. Bu nedenle, hem μ_i hem de ϕ_i'nin kovaryatların değerlerine bağlı olmasına izin verilmesi makul bir yaklaşımdır (Gu 2024).

Bir regresyon modelindeki ortalama parametrelerin doğru bir biçimde tahmin edilmesinin, dağılımın doğru modellenmesine bağlı olduğu bilinmektedir (Smyth ve Verbyla 1999). Dağılımın değişkenlik gösterdiği durumlarda sabit bir dağılımın kullanılması, önemli ölçüde performans kaybına yol açabilir; bu durum ortalama parametrelerine ilişkin standart hataların ve güven aralıklarının güvenilirliğini zayıflatır. Diğer yandan dağılım, kendi başına da doğrudan ilgi konusu olabilir; özellikle risk yönetimi uygulamalarında veya hasarların hem temel sıklığını hem de şiddetini anlamaya yönelik çalışmalarda olduğu gibi. Bu nedenle, i. poliçe için kovaryat vektörü x_i verildiğinde, $g(\mu_i) = F_\mu(x_i)$ ve $h(\phi_i) = F_\phi(x_i)$ varsayılmaktadır; burada g ve h, bilinen monoton bağlantı fonksiyonları, F_μ ve F_ϕ ise verilerden parametrik olmayan olarak tahmin edilen bilinmeyen fonksiyonlardır (Gu 2024).

Y_i'nin dağılımı, Tweedie $(\mu_i, \phi_i/w_i, \xi)$ ile gösterilir ve yoğunluğu

$$d_{TW}(y; \mu_i, \phi_i/w_i, \xi) = \exp\left(\frac{w_i}{\phi_i}\left[\frac{y\,\mu_i^{1-\xi}}{1-\xi} - \frac{\mu_i^{2-\xi}}{2-\xi}\right] + c\left(y, \frac{\phi_i}{w_i}, \xi\right)\right) \tag{3.29}$$

biçiminde gösterilir.

Eş. (3.29)'da yer alan normalleştirme fonksiyonu $c(y, \phi_i/w_i, \xi)$, ϕ_i'nin karmaşık bir fonksiyonu olabilmekte ve bu durum, F_μ ve F_ϕ'nin eşdeğişkenlerin basit fonksiyonları olarak varsayılması hâlinde dahi bu iki fonksiyonun eşzamanlı tahmin edilmesini zorlaştırmaktadır. Pratikte, ϕ küçük olduğunda başarılı sonuçlar verdiği bilinen eyer noktası yaklaşımı kullanılabilir ve $c(y, \phi, \xi) \approx -\log(2\pi\phi y^{\xi})/2$ olarak alınabilir (Jørgensen 1997). Eyer Noktası yaklaşımının doğruluğu Smyth ve Verbyla (1999) tarafından ayrıntılı biçimde incelenmiştir. Ayrıca bu yaklaşımın, Eş. (3.29)'daki olabilirlik yöntemine kıyasla model yanlış belirtiminden daha az etkilenen genişletilmiş yarı olabilirlik yaklaşımıyla ilişkili olduğu da ifade edilmektedir (Nelder ve Pregibon 1987). Bu prosedür, F_μ ve F_ϕ'nin dönüşümlü olarak güncellendiği ardışık (iteratif) yinelemelere yol açar (Smyth 1996).

Çift GDM çerçevesinde, $F_\mu(x) = x^T\beta$ ve $F_\phi(x) = x^T\lambda$ varsayılmaktadır. Bu durumda, ortalama parametreleri β ve dağılım parametreleri λ, ardışık yinelemeleri aracılığıyla sırayla güncellenmektedir. Özellikle, λ verildiğinde β'nın Tweedie modelinin sapma fonksiyonuna dayalı bir genelleştirilmiş doğrusal model çözülerek; β verildiğinde ise λ'nın bir Gamma GDM çözülerek elde edilebildiği gösterilmiştir (Nelder ve Pregibon 1987, Smyth ve Verbyla 1999, Smyth ve Jørgensen 2002). Doğrusallık varsayımı birçok uygulamada yaygın olsa da gerek doğrusal olmayan gerekse etkileşim içeren karmaşık ilişkileri yakalamakta yetersiz kalabilir (Gu 2024).

3.10 İndeks Parametresi ξ 'nin Tahmini

Bir Tweedie GDM modelinin kurulabilmesi için, kullanılacak özel Tweedie ÜDA tanımlayan indeks parametresi (ξ) değerinin bilinmesi gerekmektedir. Tweedie dağılımları, varyans fonksiyonu $\text{Var}[y] = \phi V(\mu) = \phi\mu^{\xi}$ biçiminde tanımlandığından, bu ifade logaritmik dönüşümle şu hale gelir: $\log(\text{var}[y]) = \log(\phi) + \xi \log \mu$. Bu ilişki, ξ'nin tahmini için basit bir yöntem sunmaktadır. Veriler birkaç gruba ayrılır, ardından grup varyanslarının logaritmaları grup ortalamalarının logaritmalarına karşı çizilerek doğrusal bir ilişki elde edilmektedir. Ancak, elde edilen ξ tahmininin değeri, verinin nasıl gruplandırıldığına bağlı olarak değişiklik gösterebilmektedir (Dunn ve Smyth 2018).

Eğer veri kümesinde sıfır yığılmalı değerler bulunuyorsa, bu durumda $1 < \xi < 2$ olduğu kabul edilmektedir. Veride sıfır gözlemler yer almıyorsa, genellikle $\xi \geq 2$ olur; ancak $1 < \xi < 2$ olasılığı tamamen dışlanmaz. Bu durumda yorum, veride teorik olarak sıfırların mümkün olduğu ancak mevcut örneklemde gözlemlenmediği şeklinde yapılabilmektedir (Dunn ve Smyth 2018).

Her iki yöntem de farklı ξ tahminleri üretse de sonuçlar genellikle $1 \leq \xi \leq 2$ aralığında kalmaktadır. Açıklayıcı değişkenlerdeki bilgiyi doğrudan kullanan ve verilerin keyfi biçimde gruplandırılmasına bağlı olmayan daha sağlam bir yaklaşım, ξ'nin maksimum olabilirlik yöntemiyle tahmin edilmesidir. Bu yöntemin uygulanmasında hesaplamaları düzenlemenin uygun bir yolu, profil olabilirliği kavramını kullanmaktır (Dunn ve Smyth 2018).

Bu yöntemde, ξ için çeşitli aday değerler belirlenir ve her bir ξ değeri sabit kabul edilerek Tweedie GDM modeli uygulanmaktadır. Her uygulama sonrasında log-olabilirlik değeri hesaplanır ve elde edilen değerler ξ'ye karşı çizilerek profil log-olabilirlik eğrisi oluşturulmaktadır. Bu eğrinin en yüksek olduğu noktadaki ξ değeri, profil olabilirlik tahmini olarak kabul edilmektedir. Ayrıca, profil log-olabilirliğin ξ'ye

göre grafiğinin çizilmesi, tahmin sürecinin görsel olarak değerlendirilmesi açısından oldukça yararlıdır (Dunn ve Smyth 2018).

Bu yöntemin temel zorluklarından biri, Tweedie ÜDA için olasılık fonksiyonunun hesaplanmasını gerektirmesidir. Ancak Tweedie dağılımlarında, bilinen özel durumlar dışında (örneğin Normal, Poisson, Gamma ve Ters Gauss) bu fonksiyonun kapalı formda bir ifadesi bulunmamaktadır (Dunn ve Smyth 2018).

Profil olabilirlik grafiği olasılık değerlerinin yalnızca belirli birkaç ξ noktasında dolu dairelerle hesaplandığını ve bu noktalar arasından düzgün bir eğrinin geçtiğini göstermektedir. Grafikte yer alan yatay kesikli çizgi, ξ için yaklaşık %95 güven aralığını temsil eden log-olabilirlik seviyesini belirtmektedir. Burada $\ell\left(\xi; y; \hat{\phi}_{\xi}, \hat{\mu}_{\xi}\right)$ ξ için hesaplanan profil log-olabilirlik değerini, $\ell(\hat{\xi}; y; \hat{\phi}, \hat{\mu})$ ise genel maksimum olabilirlik değerini göstermektedir (Dunn ve Smyth 2018).

Bazı durumlarda, ξ parametresinin tahmin edilmesi sürecinde teknik zorluklarla karşılaşılabilmektedir. Bu durum, literatürde birçok araştırmacı tarafından da gözlemlenmiştir (Gilchrist ve Drinkwater 1999, Jørgensen 1987, Jørgensen ve de Souza 1994). Bölüm 3.2'de belirtildiği gibi, $\xi = 1$ değeri, Tweedie dağılımının kesikli veri (örneğin $y = 0, \phi, 2\phi, 3\phi, ...$) modellemesi için uygun olduğu özel bir durumu temsil etmektedir. Ancak, eğer yanıt değişkeni $y$örneğin bir ondalık basamağa yuvarlanmışsa, log-olabilirlik fonksiyonu genellikle $\varphi = 0.1$ ve $\xi = 1$ değerlerinde maksimuma ulaşabilmektedir. Benzer şekilde, veriler tam sayılara (sıfır ondalık basamak) yuvarlanmışsa, maksimum log-olabilirlik $\varphi = 1$ ve $\xi = 1$ değerlerinde elde edilebilmektedir. Bu durum, verilerin ayrıklaştırılmasının ξ tahminini etkileyebileceğini göstermektedir. Dunn ve Smyth (2005), bu teknik problemi daha ayrıntılı biçimde incelemiş ve ξ tahmininde dikkat edilmesi gereken koşulları tartışmışlardır.

4. MAKİNE ÖĞRENMESİ

Günümüzde "büyük veri çağı" olarak adlandırılan bir dönemde yaşanmaktadır. Geçmişte veri üretimi ve depolanması büyük ölçüde kurumlar tarafından gerçekleştirilmiş ve bu veriler bilgisayar merkezlerinde saklanmıştır. Kişisel bilgisayarların yaygınlaşması ve kablosuz iletişim teknolojilerinin gelişmesiyle birlikte veri üretimi bireyler tarafından da yapılır hâle gelmiştir. Günlük yaşamda bir ürün satın alındığında, bir film kiralandığında, bir web sayfası ziyaret edildiğinde, blog yazıldığında, sosyal medyada paylaşım yapıldığında veya yalnızca hareket edildiğinde dahi veri üretilmektedir. Aynı zamanda bireyler veri tüketicisi konumuna da gelmiştir. Kişiselleştirilmiş ürün ve hizmetlerin sunulması talep edilmekte; bireysel ihtiyaçların anlaşılması ve çıkarların öngörülmesi beklenmektedir (Alpaydın 2014).

Alpaydın (2014)'e göre, makine öğrenmesi yöntemlerinin büyük veri tabanlarına uygulanması "veri madenciliği" olarak adlandırılmıştır. Bu kavram, bir madenden büyük miktarda hammadde çıkarılıp işlendiğinde az miktarda fakat yüksek değere sahip malzeme elde edilmesine benzetilmiştir. Benzer biçimde, veri madenciliğinde de büyük hacimli veriler işlenerek yüksek tahmin doğruluğuna sahip, çeşitli amaçlarla kullanılabilir modeller elde edilmiştir. Veri madenciliğinin uygulama alanları oldukça geniştir. Perakende sektörünün yanı sıra finans sektöründe geçmiş veriler analiz edilerek kredi başvuruları, dolandırıcılık tespiti ve borsa tahmini için modeller oluşturulmuştur. İmalat sektöründe öğrenme modelleri optimizasyon, kontrol ve sorun giderme süreçlerinde; tıp alanında makine öğrenmesi programları tıbbi teşhis amaçlı; telekomünikasyon sektöründe ise arama kalıplarının analizi yoluyla ağ optimizasyonu ve hizmet kalitesinin artırılmasında kullanılmıştır. Bilimsel araştırmalarda ise fizik, astronomi ve biyoloji gibi alanlarda üretilen büyük veriler yalnızca bilgisayarlar tarafından yeterli hızda analiz edilebilmiştir. Ayrıca World Wide Web'in büyüklüğü ve sürekli genişlemesi nedeniyle ilgili bilgilere manuel yöntemlerle ulaşmak mümkün olmamıştır.

Makine öğrenmesi yalnızca bir veri tabanı problemi değil, aynı zamanda yapay zekâ alanının bir parçası olarak değerlendirilmiştir. Değişen ortamlarda zekâdan söz

edilebilmesi için sistemlerin öğrenme yeteneğine sahip olması gerekmektedir. Bu tür değişikliklerin sistem tarafından öğrenilmesi ve bunlara uyum sağlanabilmesi durumunda, tüm olası durumların tasarımcı tarafından önceden tanımlanmasına gerek kalmamaktadır. Ayrıca makine öğrenmesinin görme, konuşma tanıma ve robotik gibi alanlardaki birçok probleme çözüm geliştirilmesine olanak sağladığı belirtilmiştir. Makine öğrenmesi, örnek verilerden veya geçmiş deneyimlerden yararlanarak bir performans ölçütünü optimize edecek şekilde bilgisayarların programlanması olarak tanımlanmıştır. Belirli parametrelere sahip bir model kurulmakta ve öğrenme süreci, eğitim verileri veya geçmiş deneyimler kullanılarak bu parametrelerin optimize edilmesiyle yürütülmektedir. Model, geleceğe yönelik tahminler üretmek amacıyla öngörücü; veriden bilgi çıkarmak amacıyla tanımlayıcı veya her iki işlevi birden üstlenici nitelikte olabilmektedir. Makine öğrenmesi sürecinde istatistik teorisinden yararlanılmasının nedeni, temel amacın örneklerden çıkarım yapılması olmasıdır. Bilgisayar bilimi ise iki temel açıdan rol üstlenmektedir: İlk olarak, eğitim aşamasında optimizasyon probleminin çözümü ve büyük veri hacimlerinin depolanıp işlenmesi için verimli algoritmalara ihtiyaç duyulmaktadır. İkinci olarak, model öğrenildikten sonra çıkarım sürecinin uzay ve zaman bakımından verimli şekilde yürütülmesi gerekmektedir. Bazı uygulamalarda öğrenme veya çıkarım algoritmalarının verimliliği, yani zaman ve bellek karmaşıklığı, tahmin doğruluğu kadar önemli olabilmektedir (Alpaydın 2014).

Bu tez kapsamında, on farklı makine öğrenmesi algoritması incelenmiştir. Bunlar sırasıyla Klasik Doğrusal Regresyon, LASSO Regresyon, Ridge Regresyon, Karar Ağaçları, Rasgele Orman, Destek Vektör Makineleri, Yapay Sinir Ağları, XGBoost, LightGBM ve CatBoost algoritmalarıdır.

4.1 Klasik Doğrusal Regresyon

Makine öğrenmesi yöntemleri, klasik bilgisayar yaklaşımlarıyla çözümlenmesi güç olan problemlerin çözümünde çeşitli alanlarda yaygın olarak kullanılmaktadır. Bu yöntemler arasında doğrusal regresyon, en basit ve en yaygın uygulanan makine öğrenmesi yaklaşımlarından biri olarak kabul edilmektedir. Doğrusal regresyon, tahmine dayalı

analiz için kullanılan bir modelleme yöntemidir ve sürekli değişkenler arasındaki ilişkinin matematiksel olarak ifade edilmesine olanak tanımaktadır. Regresyon kavramı ilk olarak Sir Francis Galton tarafından 1894 yılında ortaya konmuş olup, günümüzde incelenen değişkenler arasındaki ilişkinin ölçülmesi ve nicel olarak değerlendirilmesi amacıyla kullanılmaktadır (Shalev-Shwartz ve Ben-David 2014, Bargarai vd. 2020, Akgün ve Öğüdücü 2015, Abdulqader vd. 2020). Bununla birlikte tek değişkenli yöntemler (ki-kare, Fisher'in kesin testi, t-testi, ANOVA) yalnızca tek bir değişkenin etkisini değerlendirebilmekte ve analizdeki ortak değişkenlerin veya karıştırıcı değişkenlerin etkisini dışarıda bırakmaktadır. Bu nedenle kısmi korelasyon ve regresyon yöntemleri, iki değişken arasındaki ilişkinin karıştırıcı etkilerden arındırılarak analiz edilmesine imkân tanımaktadır (Zeebaree vd. 2019, Zebari vd. 2020, Abdulazeez vd. 2020).

Regresyon analizleri, esas olarak iki amaç için kullanılmaktadır. İlk olarak, regresyon modelleri tahmin ve öngörü amacıyla makine öğrenmesi uygulamalarıyla örtüşen biçimde kullanılmaktadır. İkinci olarak, bazı durumlarda bağımsız değişkenler ile bağımlı değişken arasındaki nedensel ilişkilerin belirlenmesinde başvurulmaktadır. Ancak regresyonun tek başına yalnızca sabit bir veri kümesinde yer alan bağımsız değişkenler ile bağımlı değişken arasındaki ilişkileri ortaya koyduğu göz önünde bulundurulmalıdır (Wu vd. 2019). Regresyon modellerinde bağımsız değişkenler, bağımlı değişkenin tahmin edilmesinde kullanılmakta ve bağımsız değişkenlerin belirli bir x aralığındaki değerlerine göre bağımlı değişken y'nin değeri öngörülmektedir (Roopa ve Asha 2019, Seber ve Lee 2012). Bu yaklaşım, probleme bağlı olarak basit doğrusal regresyon ya da çoklu regresyon biçiminde uygulanabilmekte ve tahmin edilen ilişkilerin birden fazla giriş değişkeni üzerinden modellenmesine olanak tanımaktadır (Kavitha vd. 2016). Basit doğrusal regresyon, tek bir bağımsız değişkenin bulunduğu durumlar için kullanılan bir modeldir (Abdulazeez vd. 2020). Bu modelde bağımlı değişken (y) ile bağımsız değişken (x) arasındaki ilişki

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{4.1}$$

şeklinde ifade edilmektedir. Burada β_0 bağımlı değişkenin açıklayıcı değişken sıfır olduğunda alacağı ortalama değeri, β_1 bağımsız değişkendeki bir birimlik değişimin y üzerindeki doğrusal etkisini, ε ise modelin açıklayamadığı hata bileşenini göstermektedir. Basit regresyon, bağımsız değişkenin etkisini, bağımlı değişkenin diğer olası etkileşimlerinden ayırarak inceleme olanağı sağlamaktadır (Acharya vd. 2019).

Çoklu doğrusal regresyon, birden fazla açıklayıcı değişken kullanılarak bir cevap değişkeninin tahmin edilmesini amaçlayan istatistiksel bir yöntemdir. Bu yöntemin temel amacı, analizde yer alan bağımsız değişkenler x ile bağımlı değişken y arasındaki doğrusal ilişkiyi modellemektir (Zhang vd. 2019). Çoklu doğrusal regresyon için temel model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \tag{4.2}$$

biçiminde ifade edilir. Eş. (4.2)'de yer alan terimler şu şekilde tanımlanmaktadır:

y: Tahmin edilmek istenen bağımlı değişken (cevap değişkeni),

$x_1, x_2, \ldots, x_m$: Modele dahil edilen bağımsız (açıklayıcı) değişkenler,

β_0: Sabit terim (doğrunun y-eksenini kestiği nokta),

$\beta_1, \beta_2, \ldots, \beta_m$: Her bir bağımsız değişkenin bağımlı değişken üzerindeki etki katsayıları,

ε: Modelin açıklayamadığı varyasyonu temsil eden hata terimi.

Regresyon modellemesi, özellikle gözlemsel nitelikteki bilimsel araştırmalarda yaygın biçimde kullanılan bir istatistiksel yaklaşımdır. Analizden elde edilen sonuçların geçerliliği, uygun regresyon modelinin seçilmesine ve modele dâhil edilen değişkenlerin doğruluğuna bağlıdır. Modelin hatalı seçilmesi veya yanlış uygulanması durumunda, analiz bulguları gerçeği yansıtmayarak yanıltıcı sonuçlara neden olabilmektedir (Maulud ve Abdulazeez 2020).

4.2 LASSO Regresyonu

LASSO (Least Absolute Shrinkage and Selection Operator), yani "En Küçük Mutlak Daralma ve Seçim Operatörü", ilk olarak Robert Tibshirani tarafından literatüre kazandırılmıştır (Tibshirani 1996). LASSO, istatistiksel regresyon modellerinde hem değişken seçimi hem de düzenlileştirmeyi aynı anda gerçekleştiren bir regresyon analiz yöntemidir. Bu yönüyle, modelin tahmin doğruluğunu artırmakta ve yorumlanabilirliğini güçlendirmektedir. Günümüzde özellikle genomik, biyoinformatik, ekonomi ve finans gibi büyük veri kümelerinin analiz edildiği yüksek boyutlu veri alanlarında, verimli ve hızlı algoritmalara duyulan ihtiyaç nedeniyle yaygın biçimde kullanılmaktadır (Friedman vd. 2010).

LASSO yöntemi, ℓ_1 cezalı En Küçük Kareler (EKK) kriterine dayanmakta ve çözümünde seyrek yapılar elde etme eğilimi göstermektedir (Emmert-Streib ve Dehmer 2019). LASSO tahmincisi

$$\hat{\beta} = \arg\min \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \right\} \tag{4.3}$$

koşuluyla: $\| \beta \|_1 \leq t$ (4.4)

biçiminde tanımlanmaktadır. Burada t, düzenlileştirme veya ceza parametresi olarak adlandırılan bir ayar parametresidir; $\| \beta \|_1$ ifadesi ise L_1 normudur. LASSO'nun amacı, modelin açıklayıcı gücünü korurken gereksiz değişkenlerin katsayılarını sıfıra indirerek sade ve genellenebilir bir model elde etmektir (Emmert-Streib ve Dehmer 2019).

Eş. (4.3), Lagrange formunda şu şekilde yeniden yazılabilir:

$$\hat{\beta} = \arg min \left\{ \frac{1}{2n} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \right\} \tag{4.5}$$

Burada $\lambda > 0$, modeldeki katsayıların küçülme derecesini belirleyen düzenlileştirme (veya ceza) parametresidir. Eş. (4.4) ile Eş. (4.5) arasındaki ilişki, dualite ve Karush–

Kuhn–Tucker (KKT) koşullarıyla açıklanabilir; her $t > 0$ için aynı çözümü veren bir $\lambda > 0$ bulunmaktadır (Hastie vd. 2015).

LASSO'nun amaç fonksiyonu türevlenebilir olmadığından, analitik (kapalı formda) bir çözüm bulunmamaktadır. Ancak, ortonormal tasarım matrisleri gibi özel durumlarda kapalı form çözümler elde edilebilmektedir. Genel durumda çözüm, döngüsel koordinat iniş algoritması kullanılarak etkin biçimde elde edilir (Friedman vd. 2010). Bu algoritma, farklı λ değerleri için tüm çözüm yollarını hesaplayabilmekte ve LARS (Least Angle Regression) algoritmasından daha hızlı çalışmaktadır (Efron vd. 2004). Bu özellikler, LASSO'yu hem değişken seçimi hem de model basitleştirme açısından güçlü ve yaygın bir yöntem hâline getirmiştir.

Modelin genelleme performansını doğrudan etkileyen λ parametresi, çapraz doğrulama yöntemiyle belirlenmektedir (Emmert-Streib ve Dehmer 2019). Bu süreçte K-katlı çapraz doğrulama yaklaşımı kullanılarak her bir kat F_k için ortalama karesel hata

$$e(\lambda)_k = \frac{1}{\#F_k} \sum_{j \in F_k} (y_j - \hat{y}_j)^2 \tag{4.6}$$

biçiminde hesaplanmaktadır. Burada $\#F_k$, F_k kümesindeki gözlem sayısını göstermektedir. Tüm katların ortalaması alınarak çapraz doğrulama hatası elde edilir:

$$CV(\lambda) = \frac{1}{K} \sum_{k=1}^{K} e(\lambda)_k \tag{4.7}$$

Bu fonksiyondan optimal λ değerinin seçimi için iki yöntem kullanılmaktadır. Birinci yöntemde $CV(\lambda)$ fonksiyonunu en küçük yapan değer seçilir:

$$\hat{\lambda}_{min} = \arg\min CV(\lambda) \tag{4.8}$$

İkinci yöntemde ise, "1-SE kuralı" uygulanarak, $CV(\hat{\lambda}_{min}) + SE(\hat{\lambda}_{min})$ değerinden daha küçük bir hataya sahip maksimum λ seçilir:

$$\hat{\lambda}_{1se} = \underset{CV(\lambda) \leq CV(\hat{\lambda}_{min}) + SE(\hat{\lambda}_{min})}{max} \lambda \tag{4.9}$$

Bu yaklaşım, daha basit ancak genelleme gücü yüksek modellerin seçilmesini sağlamaktadır (Emmert-Streib ve Dehmer 2019).

Bununla birlikte, LASSO yönteminin bazı sınırlılıkları da bulunmaktadır. Özellikle, açıklayıcı değişkenler arasında yüksek korelasyon bulunduğunda model kararsızlık gösterebilir. Bu durumda yöntem, yüksek derecede ilişkili değişkenler arasından keyfi olarak yalnızca birini seçip diğerlerini dışlayabilmekte veya tüm öngörücüler aynı olduğunda çözümsüz hâle gelebilmektedir (Friedman vd. 2010). Ayrıca, LASSO yöntemi "oracle" özelliğinden yoksundur (Zou 2006, Fan ve Li 2001).

"Oracle" özelliğine sahip bir tahmin yöntemi, sıfır katsayıya sahip parametrelerin alt kümesini olasılık 1'e yakın bir doğrulukla tam olarak belirleyebilir; başka bir deyişle, model sanki gerçek değişken alt kümesi önceden biliniyormuş gibi davranır (Fan ve Li 2001). Bu özellik, asimptotik olarak tutarlı, yansız ve etkin parametre tahmini yapılmasını sağlar. Ancak LASSO, sıfır olmayan büyük katsayıları kısmen küçülterek tahmin ettiği için asimptotik sapma üretir ve bu nedenle oracle özelliğini taşımaz (Zou 2006, Fan ve Li 2001). Dolayısıyla yalnızca tasarım matrisinin güçlü koşulları sağladığı durumlarda tutarlı parametre tahmini gerçekleştirebilmektedir (Zou 2006).

Sonuç olarak, LASSO regresyon yöntemi, değişken seçimi, model sadeleştirme ve aşırı uyumun azaltılması açısından oldukça etkili bir araçtır. Ancak, yüksek korelasyonlu verilerde kararsızlık göstermesi, oracle özelliğinden yoksun olması ve değişken sayısının gözlem sayısından fazla olduğu $(p > n)$ durumlarda sınırlı sayıda değişken seçebilmesi nedeniyle dikkatli biçimde uygulanması gerekmektedir (Ogutu vd. 2012).

4.3 Ridge Regresyonu

Ridge regresyon, katsayıları sıfırdan farklı olan ve normal dağılımdan türetilmiş çok sayıda bağımsız değişkenin (öngörücünün) yer aldığı durumlar için uygun bir yöntemdir

(Hoerl ve Kennard 1970, Friedman vd. 2010). Özellikle, her biri küçük etkiye sahip çok sayıda değişkenin bulunduğu modellerde yüksek performans göstermekte ve klasik doğrusal regresyon modellerinde çoklu doğrusal bağlantı nedeniyle katsayı tahminlerinin küçük veri değişikliklerine aşırı duyarlı hâle gelmesi ve yüksek varyans göstermesi sorunlarını azaltmaktadır. Ridge regresyon, yüksek korelasyona sahip öngörücülerin katsayılarını ℓ_2 (L2) normu temelinde orantılı biçimde sıfıra doğru küçültür. Örneğin, k adet tamamen aynı açıklayıcı değişken bulunduğunda, her biri tek başına modele alındığında sahip olacağı katsayının $1/k$ oranında küçültülmüş eşit katsayılara sahip olur (Friedman vd. 2010). Bu nedenle Ridge regresyon, katsayıların tam olarak sıfır olmasını sağlamaz; dolayısıyla yalnızca en önemli değişkenleri seçen bir model oluşturmak için doğrudan değişken seçimi gerçekleştirmez (Ogutu vd. 2012).

EKK yönteminin geliştirilme motivasyonu, bu tür modellerden elde edilen tahminlerin genellikle düşük yanlılık (bias) ancak yüksek varyans göstermesidir. Bu durum, modelin tahmin doğruluğunu olumsuz yönde etkileyebilmektedir. Regresyon katsayılarının değerlerini küçülterek veya bazı katsayıları sıfıra yaklaştırarak tahmin doğruluğunun artırılabileceği bilinmektedir (Hastie vd. 2015). Bunun temel nedeni, modele bir miktar yan (bias) eklenerek varyansın azaltılabilmesidir. Bu sorunu gidermek amacıyla Hoerl ve Kennard (1970) tarafından önerilen Ridge regresyon (Hoerl ve Kennard 1970), klasik EKK formülasyonuna bir düzenlileştirme terimi ekleyerek varyans-bias dengesini optimize etmektedir.

Ridge regresyon tahmincisi, klasik EKK problemine ℓ_2 cezalı bir düzenlileştirme terimi eklenerek

$$\hat{\beta}^{RR} = \arg\min \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \parallel \beta \parallel_2^2 \right\} \tag{4.10}$$

biçiminde tanımlanmaktadır. Bu ifade, Artık Kareler Toplamı (Residual Sum of Squares, RSS) kullanılarak

$$\hat{\beta}^{RR} = \arg\min \left\{ \frac{1}{2n} RSS(\beta) + \lambda \parallel \beta \parallel_2^2 \right\} \tag{4.11}$$

biçiminde tanımlanmaktadır veya matris gösterimiyle:

$$\hat{\beta}^{RR} = \arg\min \left\{ \frac{1}{2n} \parallel y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_2^2 \right\} \tag{4.12}$$

Burada $RSS(\beta)$, modelin kaybı olarak adlandırılan artık kareler toplamını; $\lambda \parallel \beta \parallel_2^2$ ise düzenlileştirme cezasını temsil etmektedir. $\lambda \geq 0$, düzenlileştirme (veya karmaşıklık) parametresi olup katsayıların ne ölçüde küçültüleceğini kontrol eder. Eş. (4.10)'da yer alan L2 cezası, literatürde sıklıkla Tikhonov düzenlemesi olarak da adlandırılmaktadır (Emmert-Streib ve Dehmer 2019).

Ridge regresyonun analitik çözümü

$$\hat{\beta}^{RR}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T y \tag{4.13}$$

biçiminde gösterilir. Burada I_p, $p \times p$ boyutlarında birim matristir. EKK yönteminde, $\text{rank}(X) < p$ olduğunda $X^T X$ matrisinin tersi alınamaz ve çözüm tanımsız hâle gelir. Ridge regresyon ise, sıfırdan farklı bir düzenlileştirme parametresi ($\lambda > 0$) ekleyerek $X^T X + \lambda I_p$ matrisinin tersinin alınabilir olmasını sağlar. Böylece, yüksek korelasyonlu veya çok sayıda öngörücü içeren durumlarda daha kararlı ve genellenebilir katsayı tahminleri elde edilir. Ayrıca, λ parametresi büyüdükçe katsayılar daha fazla küçülmekte, modelin varyansı azalmakta ancak önyargısı artmaktadır. Bu nedenle λ'nın uygun değeri genellikle veri odaklı yöntemler, özellikle çapraz doğrulama kullanılarak belirlenmektedir. Ridge regresyon, bu sayede modelin genelleme yeteneğini artırmakta ve aşırı uyum riskini azaltmaktadır (Ogutu vd. 2012).

4.4 Karar Ağaçları

Karar ağaçları, sinir ağlarıyla birlikte yer bilimlerinde en yaygın kullanılan makine öğrenmesi algoritmalarından biridir (Friedl ve Brodley 1997, Hansen vd. 1996, Lippitt vd. 2008, Pal ve Mather 2003, Rogan vd. 2003, Wessels vd. 2004). Karar ağaçlarının

yaygın kullanımının temel nedenleri; basit yapıları, yüksek yorumlanabilirlikleri, düşük hesaplama maliyetleri ve grafiksel olarak kolay biçimde temsil edilebilmeleridir. Bir karar ağacı, kökten terminal düğümlere (veya yapraklara) kadar hiyerarşik olarak düzenlenmiş bir dizi koşul veya kısıtlamayı ifade eden bir yapıya sahiptir (Breiman vd. 1984, Quinlan 1993). Sınıflandırma kararlarının hiyerarşik bir ağaç yapısıyla alınmasının başlıca avantajı, bu yapının şeffaflığıdır; dolayısıyla, karar ağaçları yapay sinir ağlarına kıyasla çok daha kolay yorumlanabilmektedir (Rodriguez-Galiano vd. 2015).

Bir veri kümesinden karar ağacı oluşturulurken, düğümler arasındaki farklılığı (heterojenliği) en üst düzeye çıkarmak amacıyla her bir açıklayıcı değişken bir değerlendirme ölçütü kullanılarak analiz edilir. Karar ağaçları iki temel yöntem altında incelenmektedir: sınıflandırma ağaçları ve regresyon ağaçları. Bu tezde, regresyon ağaçlarının teorik temelleri ele alınmaktadır. Bir karar ağacının indüksiyon sürecinde, veri seti üzerinde özyinelemeli bölme işlemleri ve çoklu regresyon analizleri gerçekleştirilir. Kök düğümden başlayarak, her bir iç düğümde veri bölme işlemi önceden belirlenmiş bir durma koşuluna ulaşılıncaya kadar tekrarlanır. Her bir terminal düğüme (yaprağa), yalnızca o bölgeye özgü basit bir regresyon modeli atanır. İndüksiyon süreci tamamlandıktan sonra, ağacın yapısal karmaşıklığını azaltmak ve genelleme yeteneğini artırmak amacıyla budama işlemi uygulanabilir. Budama işlemi sırasında düğümlerdeki örnek sayısı, karar kriterlerinden biri olarak kullanılmaktadır (Rodriguez-Galiano vd. 2015).

Breiman vd. (1984) tarafından açıklandığı üzere, karar ağacının oluşturulma süreci (indüksiyon) öncelikle optimal bölme ölçütlerinin belirlenmesini içermektedir. Süreç, bağımlı değişkenin veya ana düğümün (kök düğüm) ikili alt parçalara ayrılmasıyla başlamakta; burada elde edilen alt düğümler, kök düğüme göre daha yüksek "saflık" düzeyine sahip olmaktadır. Bu aşamada karar ağacı algoritması, tüm olası bölmeleri inceleyerek, düğüm saflığında en büyük iyileşmeyi (veya kirlilikteki azalmayı) sağlayan optimal bölmeyi s^* olarak belirlemektedir.

Saflıkta meydana gelen değişim

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{4.14}$$

biçiminde ifade edilir. Eş. (4.14)'te s, t düğümündeki olası bir bölmeyi; p_L ve p_R, s bölmesiyle oluşturulan sol (t_L) ve sağ (t_R) alt düğümlere ait örneklerin oranlarını göstermektedir. $i(t)$, bölme öncesindeki saflık ölçüsünü; $i(t_L)$ ve $i(t_R)$ ise bölme sonrasındaki saflık ölçülerini temsil etmektedir. Buna göre, $\Delta i(s,t)$ değeri, gerçekleştirilen s bölmesinin saflık üzerindeki etkisini (yani kirlilikteki azalmayı) ölçmektedir (Rodriguez-Galiano vd. 2015).

Kirlilik ölçümünde birden fazla yaklaşım kullanılabilmektedir. Literatürde en sık başvurulan ölçütler arasında Bilgi Kazancı Oranı, Gini İndeksi ve Ki-kare Testi yer almaktadır. Bu tezde, yaygın olarak kullanılan Gini indeksi tercih edilmiştir. Gini indeksi, $i(t)$ saflık değerini

$$I_G\left(t_{X(x_i)}\right) = 1 - \sum_{j=1}^{m} f\left(t_{X(x_i)}, j\right)^2 \tag{4.15}$$

biçiminde tanımlar. Burada $f(t_{X(x_i)}, j)$, t düğümünde j sınıfına ait x_i örneklerinin orantısını ifade etmektedir. Karar ağacı algoritmasında bölme kriteri, en düşük Gini saflık indeksine (I_G) sahip özniteliğin seçilmesine dayanmaktadır (Rodriguez-Galiano vd. 2015).

4.5 Rasgele Orman

Rasgele Orman yöntemi, bir değişkenin değerini sınıflandırmak veya tahmin etmek amacıyla birden fazla karar ağacının performansını birleştiren bir regresyon tekniğidir (Breiman 2001, Guo vd. 2011, Rodriguez-Galiano vd. 2012b). Rasgele Orman modeli, belirli bir eğitim alanına ait gözlemlere ilişkin kanıt özelliklerinden oluşan bir x girdi vektörü aldığında, K adet regresyon ağacı üretir ve elde edilen sonuçların ortalamasını

alarak tahmin değerini oluşturur (Rodriguez-Galiano vd. 2015). K adet ağacın $\{\mathrm{T}(x)\}_1^K$ biçiminde tanımlanması durumunda Rasgele Orman regresyon tahmincisi

$$\hat{f}_{rf}{}^{K}(x) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{T}(x) \tag{4.16}$$

biçiminde ifade edilir.

Farklı ağaçlar arasında yüksek korelasyon oluşmasını engellemek amacıyla, Rasgele Orman algoritması torbalama adı verilen bir prosedür kullanmaktadır. Bu yöntem, orijinal veri kümesinden rasgele yeniden örnekleme yaparak her bir ağacın farklı bir eğitim alt kümesi üzerinde büyütülmesini sağlar. Yeniden örnekleme sırasında seçilen veriler kümeden çıkarılmadığından, bazı örnekler eğitim sürecinde birden fazla kez kullanılabilirken bazıları hiç seçilmeyebilir. Bu durum, modelin küçük veri değişikliklerine karşı daha dayanıklı hâle gelmesini ve tahmin doğruluğunun artmasını sağlamaktadır (Breiman 2001).

Her bir ağaç oluşturulurken, model tüm kanıt özelliklerinden rasgele seçilen bir alt küme içerisinden en uygun bölünme noktasını belirler. Bu yaklaşım, bireysel ağaçların tahmin gücünü kısmen azaltabilse de ağaçlar arasındaki korelasyonu düşürerek genel genelleme hatasının azalmasına katkı sağlar (Breiman 2001). Rasgele Orman yönteminin dikkat çekici özelliklerinden biri, oluşturulan ağaçların budama işlemine ihtiyaç duymadan büyümesidir; bu da yöntemi hesaplama açısından daha verimli hâle getirmektedir (Rodriguez-Galiano vd. 2015).

Ayrıca, her bir ağacın eğitimi sırasında seçilmeyen örnekler torba dışı (TD) olarak adlandırılan bir alt kümede tutulur. Bu TD örnekleri, ilgili ağacın doğruluk performansını değerlendirmek için kullanılabilmektedir (Peters vd. 2007). Böylece Rasgele Orman, dışsal bir test veri setine ihtiyaç duymadan genelleme hatasının yansız bir tahminini üretebilmektedir (Breiman 2001). Genelleme hatasının, ağaç sayısı arttıkça belirli bir değere yakınsadığı ve yöntemin aşırı uyum göstermediği bilinmektedir (Rodriguez-Galiano vd. 2015).

Rasgele Orman algoritmasının bir diğer önemli özelliği, farklı kanıt özelliklerinin göreceli önemini değerlendirme yeteneğidir. Bu özellik, özellikle veri boyutunun yüksek olduğu ve çok kaynaklı bilgilerin kullanıldığı çalışmalarda büyük avantaj sağlamaktadır. Her bir değişkenin önemini belirlemek için Rasgele Orman, ilgili özelliğin değerlerini rasgele karıştırarak diğer özellikleri sabit tutmakta ve TD hata tahminine dayalı olarak doğrulukta meydana gelen azalmayı ölçmektedir. Bu doğruluk azalması, değişkenin model tahmini üzerindeki etkisini yansıtmaktadır (Breiman 2001, Gislason vd. 2006, Pal 2005).

4.6 Destek Vektör Makineleri

Destek Vektör Makineleri (DVM), son yıllarda veri odaklı disiplinlerde giderek önem kazanan ve güçlü tahmin yeteneği sayesinde yaygın olarak kullanılan denetimli öğrenme yöntemlerinden biridir. DVM, çok boyutlu özellik vektörlerinin ikili sınıflandırmasını gerçekleştirmek amacıyla geliştirilmiş olup, ilk olarak Vapnik ve Chervonenkis ile Vapnik ve Lerner tarafından önerilmiştir. Başlangıçta doğrusal sınıflandırma problemi için tasarlanan bu yöntem, daha sonra doğrusal olmayan sınıflandırma problemlerine genelleştirilmiş ve sonrasında regresyon analizine uyarlanmıştır (Cortes ve Vapnik 1995).

DVM'nin temel prensibi, girdi özelliklerinin, iki sınıfın hiperdüzlem olarak bilinen bir sınır yüzey ile doğrusal olarak ayrılabileceği daha yüksek boyutlu bir uzaya dönüştürülmesidir (Rodriguez-Galiano vd. 2015). N örnek içeren bir eğitim veri kümesi $\{x_n, y_n\}_{n=1}^{N}$ verildiğinde, burada $x_n \in \mathbb{R}^L$ (L boyutlu girdi vektörü) ve $y_n \in \{-1,1\}$ (çıktı etiketleri) olmak üzere, DVM regresyon modeli

$$f(x) = w^{\top}\theta(x) + b \tag{4.17}$$

biçiminde gösterilir. Burada $\theta: x \to \theta(x) \in \mathbb{R}^H$, girdileri $H \geq L$ olacak şekilde daha yüksek boyutlu bir özellik uzayına eşleyen doğrusal olmayan bir dönüşüm fonksiyonudur. Başlangıçta doğrusal olarak ayrılabilir veriler için bu fonksiyon

$\theta(x) = x$ olarak tanımlanmıştır. Modelin bilinmeyen parametreleri w (hiperdüzleme dik olan ağırlık vektörü) ve b (önyargı terimi) olarak ifade edilir (Rodriguez-Galiano vd. 2015).

DVM regresyonu, verilerin doğrusal olarak ayrılabilir olmadığı durumlarda, sınıflandırma hatalarına izin veren yumuşak marj yaklaşımını kullanır. Bu durumda model

$$y_n - f(x_n) \leq \zeta_n + \varepsilon \tag{4.18}$$

$$f(x_n) - y_n \leq \zeta_n^* + \varepsilon \tag{4.19}$$

$$\varepsilon, \zeta_n, \zeta_n^* \geq 0, \forall n \tag{4.20}$$

biçiminde kısıtlamalara tabidir. Burada ε, modelin izin verdiği maksimum hata (duyarlılık parametresi) olup, ζ_n ve ζ_n^* pozitif ve negatif sapmaları gösteren gevşek değişkenlerdir. Böylece model, eğitim verilerini maksimum marj ile ayıran bir hiperdüzlemi tanımlayacak şekilde optimize edilir (Rodriguez-Galiano vd. 2015).

Optimizasyon problemi Lagrange çarpanları yöntemiyle çözülür (Vapnik 2000). Çözüm sürecinde elde edilen maliyet fonksiyonu şu şekilde yazılabilir:

$$L(\{a_n, a_n^*\}_{n=1}^N) = -\frac{1}{2}\sum_{i,j=1}^N (a_i - a_i^*)(a_j - a_j^*)K(x_i, x_j) - \varepsilon\sum_{i=1}^N (a_i + a_i^*) + \sum_{i=1}^N (a_i - a_i^*)y_i \tag{4.21}$$

Burada $K(x_i, x_j) := \langle \theta(x_i) | \theta(x_j) \rangle$ çekirdek (kernel) fonksiyonudur ve dönüştürülmüş özellik vektörlerinin iç çarpımı olarak tanımlanır. Çekirdek notasyonunun tanıtılması, hesaplamaları büyük ölçüde basitleştirir; çünkü doğrudan yüksek boyutlu uzaya dönüşüm yapılmasına gerek kalmadan, iç çarpımlar çekirdek fonksiyonu aracılığıyla hesaplanabilir (Rodriguez-Galiano vd. 2015).

DVM uygulamalarında en yaygın kullanılan çekirdek fonksiyonları

$$K_{\text{linear}}(x, x') = x, x' \tag{4.22}$$

$$K_{\text{polynomial}} = (\gamma x x' + r)^{\rho} \tag{4.23}$$

$$K_{\text{RBF}}(x, x') = \exp(-\gamma \|x - x'\|^2) \tag{4.24}$$

$$K_{\text{sigmoid}}(x, x') = \tanh(\gamma x x' + r) \tag{4.25}$$

biçiminde gösterilir.

Lagrange çarpanlarının $\{\hat{a}_n, \hat{a}_n^*\}_{n=1}^N$ tahmini sonrasında, ağırlık vektörü $\hat{w}$ ve regresyon fonksiyonu $f(x)$

$$\hat{w} = \sum_{n=1}^{N} (\hat{a}_n - \hat{a}_n^*)\theta(x_n) \tag{4.26}$$

$$\hat{f}(x) = \sum_{n=1}^{N} (\hat{a}_n - \hat{a}_n^*)K(x_i, x) + \hat{b} \tag{4.27}$$

biçiminde hesaplanır. Bu formülasyon, modelin doğrusal olmayan ilişkileri de yakalayabilmesini sağlamaktadır. DVM yöntemi, yüksek boyutlu verilerde dahi genel performansını koruyan sağlam bir yaklaşımdır ve aşırı öğrenmeye karşı güçlü bir denge sunmaktadır (Rodriguez-Galiano vd. 2015).

4.7 Yapay Sinir Ağları

Parametrik olmayan ve doğrusal olmayan sınıflandırma ya da regresyon modellerinin oluşturulmasında en yaygın kullanılan yöntemlerden biri yapay sinir ağlarıdır. Literatürde, farklı mimarilere ve öğrenme stratejilerine sahip çok sayıda yapay sinir ağı modeli bulunmaktadır. Yapay sinir ağlarında, tıpkı biyolojik sinir sistemlerinde olduğu gibi temel işlem birimleri nöronlardır (birim veya düğüm). Bu nöronlar katmanlar hâlinde düzenlenmekte ve bilgi akışı giriş katmanından başlayarak gizli katmanlara, oradan da çıkış katmanına doğru tek yönlü olarak ilerlemektedir. Giriş katmanındaki birimler girdileri alarak bunları gizli katmandaki birimlere iletmektedir. Bir nöronun

gerçekleştirdiği işlem temelde doğrusal bir regresyonun ardından doğrusal olmayan bir dönüştürücü fonksiyon $f(\cdot)$ uygulanması şeklinde özetlenebilir. Farklı katmanlardaki nöronlar, aralarındaki bağlantı ağırlıkları aracılığıyla birbirine bağlanmaktadır (Rodriguez-Galiano vd. 2015).

Bu tezde kullanılan standart çok katmanlı yapay sinir ağı (YSA) modelinde, $(l+1)$. katmandaki j nöronunun çıktısı

$$x_j^{l+1} = f\left(\sum_i w_{ij}^l x_i^l + w_{bj}^l\right) \tag{4.28}$$

şeklinde ifade edilmektedir. Burada w_{ij}^l, l. katmandaki i nöronunu $(l+1)$. katmandaki j nöronuna bağlayan ağırlığı; w_{bj}^l, j nöronuna ait önyargı terimini; f ise lojistik aktivasyon fonksiyonunu göstermektedir. Modele verilen x_i girdisi için üretilen tahmin $f(x_i)$ ile gösterilmektedir. Algoritmanın amacı, her bir giriş vektörü için ağ tarafından üretilen çıktının istenen çıkışa eşit veya ona yeterince yakın olmasını sağlayan ağırlıkların bulunmasıdır. Giriş-çıkış çiftlerinin (desenlerin) sayısı sonlu olduğunda, belirli bir ağırlık kümesi için ağın toplam hatası, örneğin EKK yöntemi ile her desenin gerçek ve hedef çıktıları karşılaştırılarak hesaplanabilmektedir (Rodriguez-Galiano vd. 2015).

Bir yapay sinir ağının eğitilmesi; ağ yapısının (gizli katman sayısı ve katman başına nöron sayısı), ağırlıkların başlangıç değerlerinin, öğrenme oranının ve aşırı öğrenmeyi önlemeye yönelik düzenleme parametrelerinin uygun şekilde belirlenmesini gerektirmektedir (Rodriguez-Galiano vd. 2015).

4.8 XGBoost

Karar ağaçlarından türetilen XGBoost (Extreme Gradient Boosting) algoritması, yapay zekâ ve makine öğrenmesi alanlarında önemli bir gelişme olarak değerlendirilmektedir. XGBoost'un işlevselliğini tam olarak anlayabilmek için, bu yöntemin ortaya çıkışına zemin hazırlayan önceki modellerin incelenmesi gerekmektedir. XGBoost'un

geliştirilmesinin ardındaki temel motivasyon, ölçeklenebilir, yüksek performanslı ve optimize edilmiş bir ağaç tabanlı güçlendirme modeline duyulan gereksinimdir. Mevcut ağaç tabanlı yöntemlerin özellikle hesaplama hızı ve ölçeklenebilirlik açısından belirli sınırlılıklar taşıdığı fark edilmiştir (Niazkar vd. 2024). Bu doğrultuda Chen ve Guestrin (2016), yalnızca bu sınırlamaları aşmayı değil, aynı zamanda makine öğrenmesi modellerinin genel performansını artırmayı da amaçlayan bir yaklaşım geliştirmiştir. XGBoost'un, Kaggle ve KDD Cup gibi uluslararası makine öğrenmesi yarışmalarında elde ettiği yüksek başarılar, yöntemin etkinliğini ve alandaki önemini açık biçimde ortaya koymuştur (Niazkar vd. 2024).

XGBoost algoritmasının çekirdeğini, Quinlan (1986) tarafından sınıflandırma ve regresyon görevleri için geliştirilen ve denetimli öğrenmede yaygın olarak kullanılan karar ağacı yöntemi oluşturmaktadır. Bu yapı, veriyi girdi özelliklerinin belirli eşik değerlerine göre tekrarlı biçimde bölerek daha küçük alt gruplara ayırır. Böylece veri, hiyerarşik bir biçimde kök, iç düğümler ve yapraklar olmak üzere üç temel bileşenden oluşan ağaç benzeri bir yapıya dönüştürülür (Bisong 2019). Regresyon görevlerinde, bir yaprağa atanan örneklerin ortalama değeri genellikle o yaprağın tahmin edilen çıktısını temsil ederken, sınıflandırma problemlerinde en yaygın yaklaşım, yapraktaki örnekler arasında en sık görülen sınıf etiketinin ilgili yaprağa atanmasıdır. Bu yapı, XGBoost'un karar ağaçlarının güçlü yanlarını optimize edilmiş bir öğrenme mekanizmasıyla birleştirerek yüksek doğruluk, genelleme gücü ve hesaplama verimliliği elde etmesini sağlamaktadır (Niazkar vd. 2024).

Karar Ağaçları, aşırı öğrenmeye yatkın modellerdir; yani eğitim verilerine yüksek düzeyde uyum sağlarken, görülmemiş veriler üzerinde düşük performans gösterebilirler. Topluluk yöntemleri ise, karar ağaçlarının basitliğini esneklik ve uyarlanabilirlikle birleştirerek modelin doğruluğunu önemli ölçüde artırır ve bu sayede aşırı öğrenme sorununu belirli ölçüde azaltır (Niazkar vd. 2024). Rasgele Orman, Breiman tarafından geliştirilen topluluk yöntemlerinden biridir. Rasgele Orman, başlangıçtaki veri setinden aynı popülasyon boyutunu koruyarak rasgele alt kümeler oluşturmak için önyükleme tekniğini kullanır. Bu alt kümelerin her biri için bağımsız bir karar ağacı inşa edilir. Genellikle, oluşturulan yaklaşık yüz ağacın kombinasyonu Rasgele Orman modelinin

nihai yapısını oluşturur. Regresyon görevlerinde, her bir ağacın yaptığı tahminlerin ortalaması alınarak modelin çıktısı elde edilirken; sınıflandırma görevlerinde, en sık gözlenen sınıf etiketi nihai sonuç olarak seçilir. Sonuç olarak Rasgele Orman, tek başına karar ağaçlarına kıyasla aşırı öğrenmeyi daha etkili biçimde yönetebilir ve yüksek boyutlu veri setleri üzerinde verimli bir şekilde uygulanabilir (Lu ve Ma 2020).

Freund ve Schapire (1997) tarafından geliştirilen bir diğer topluluk öğrenme yöntemi ise Uyarlanabilir Artırma (Adaptive Boosting – AdaBoost) algoritmasıdır. Rasgele Orman ile AdaBoost karşılaştırıldığında, her iki yöntem de zayıf öğrenicileri bir araya getirerek güçlü bir tahmin modeli oluşturmayı amaçlamaktadır. Ancak Rasgele Orman'da, her ağacın düğüm sayısı için varsayılan bir sınırlama bulunmazken; AdaBoost, tek bir düğüm ve iki yapraktan oluşan, sınırlı öngörü gücüne sahip kütük yapılar üretir (Niazkar vd. 2024).

AdaBoost algoritmasında zayıf öğrenicilerin oluşturulma sırası büyük önem taşır; çünkü her bir öğrenicinin yaptığı hatalar, bir sonrakinin oluşturulma sürecini doğrudan etkiler. Algoritma, verilerin farklı alt kümeleri üzerinde zayıf öğrenicileri yinelemeli biçimde eğitir. Başlangıçta, her veri noktası yeni veri setine eşit olasılıkla dâhil edilmekte ve alt kümeler Rasgele Orman'a benzer biçimde rasgele oluşturulmaktadır. Ancak sonraki adımlarda, önceki öğrenicilerde yüksek hata oranına sahip veri noktaları, yeni alt kümelere dâhil edilme ve tekrarlanma açısından daha yüksek olasılığa sahip olur (Niazkar vd. 2024). Zayıf öğrenicilerin oluşturulma sürecini etkilemenin yanı sıra, AdaBoost algoritması her bir öğreniciye yaptığı hatalara göre ağırlık atar. Modelin nihai çıktısı, Rasgele Orman'daki gibi elde edilmekle birlikte, AdaBoost'ta tahminler ağırlıklı ortalama biçiminde birleştirilir (Schapire, 2013).

Friedman (2001), kayıp fonksiyonunu optimize etmek amacıyla gradyan inişi kullanan bir algoritma geliştirmiştir. Bu çalışma sonucunda, Gradyan Artırma olarak adlandırılan, daha esnek ve daha doğru modeller ortaya çıkmıştır. Önceki topluluk yöntemlerinden farklı olarak, Gradyan Artırma algoritması yalnızca başlangıçtaki bir tahmini temsil eden tek bir yaprakla sürece başlar. Regresyon görevlerinde bu başlangıç tahmini genellikle verilerin ortalamasına, sınıflandırma görevlerinde ise hedef sınıf

olasılıklarının logaritmik oranlarına karşılık gelmektedir. AdaBoost algoritmasına benzer biçimde, ilk yapraktan sonra oluşturulan sonraki ağaçlar, önceki ağaçların hatalarını düzeltmek üzere eğitilir. Ancak Gradyan Artırma'daki ağaçlar genellikle AdaBoost'taki kütüklerden daha fazla sayıda düğüm ve yaprağa sahip olacak şekilde, kullanıcı tarafından belirlenmiş sınırlı bir derinlikle oluşturulur. Ayrıca Gradyan Artırma'da ağaçlara ağırlık atanır; ancak AdaBoost'tan farklı olarak tüm ağaçlar eşit ağırlığa sahiptir. Ağaç oluşturma süreci, kullanıcı tarafından tanımlanan bir eşik değerine ulaşılana veya ek ağaç üretmenin model doğruluğunu artırmadığı gözlenene kadar devam eder (Bisong 2019).

Rasgele Orman modelinden farklı olarak, Gradyan Artırma'nın nihai çıktısı ağaçların basit ortalamasıyla değil, her ağacın ağırlıklı tahminlerinin birleştirilmesiyle elde edilir. Başka bir ifadeyle, başlangıç yaprağı ilk tahmini temsil ederken, sonraki ağaçlar bu tahminin kalıntılarını ekleyip çıkararak modelin doğruluğunu artırır. Gradyan Artırma modelinin hiperparametreleri doğru biçimde ayarlandığında, Rasgele Orman ve AdaBoost algoritmalarına kıyasla daha yüksek doğruluk elde edilir. Bununla birlikte, Gradyan Artırma modellerinde de aşırı öğrenme sorunu görülebilmektedir (Niazkar vd. 2024). Bu sorunu azaltmak amacıyla Chen ve Guestrin (2016), düzenlileştirme ve ağaç budama gibi çeşitli ek özellikleri dâhil ederek Gradyan Artırma algoritmasını geliştirmiş ve XGBoost modelini önermiştir. Bu gelişmeler, XGBoost'un aşırı öğrenmeyi azaltan, sağlam ve uyarlanabilir bir model yapısı sunmasını sağlamış ve yöntemi literatürdeki diğer makine öğrenmesi modelleri arasında öne çıkarmıştır (Niazkar vd. 2024).

4.9 LightGBM

LightGBM (Light Gradient Boosting Machine), 2016 yılında Microsoft Research Asia (MSRA) tarafından geliştirilen açık kaynaklı, hızlı ve verimli bir Gradyan Artırımlı Karar Ağaçları (GAKA) algoritmasıdır. LightGBM, sıralama, sınıflandırma, regresyon gibi çok sayıda makine öğrenmesi görevinde kullanılmakta olup, verimli paralel eğitim desteği sunmaktadır. GAKA'nın bir türevi olan LightGBM, özellikle büyük veri kümeleriyle çalışırken GAKA'nın karşılaştığı ölçeklenebilirlik ve hız sorunlarını

çözmek amacıyla tasarlanmış, böylece GAKA modellerinin daha hızlı ve pratik bir biçimde uygulanmasına olanak sağlamıştır (Ma vd. 2018).

LightGBM'de karar ağacı modeli, düğümleri yaprak bölme yöntemiyle böler. Bu yaklaşım, XGBoost ile karşılaştırıldığında hesaplama maliyetini azaltır. Bununla birlikte, modelin aşırı öğrenme eğilimini önlemek için ağacın maksimum derinliği ve her yaprak düğümündeki minimum örnek sayısı kullanıcı tarafından kontrol edilmektedir. LightGBM, karar ağacı algoritmasını histogram tabanlı bir yapıya dönüştürür; bu yöntemde özellik değerleri çok sayıda küçük "kova (bin)" içine bölünür. Böylece, bölünme noktaları bu kovalar üzerinden seçilir ve hem hesaplama hem de bellek kullanımı açısından önemli ölçüde verimlilik sağlanır. LightGBM ayrıca kategorik değişkenlerin işlenmesi konusunda da güçlü bir yapıya sahiptir. Bu özellik, özellikle kategorik veri içeren büyük ölçekli veri kümelerinde LightGBM'yi diğer GAKA tabanlı algoritmalara kıyasla daha avantajlı hâle getirir (Ma vd. 2018).

LightGBM'in paralel öğrenme mimarisi üç ana bileşenden oluşmaktadır: özellik paralelliği**,** veri paralelliği ve oylama paralelliği**.** Özellik paralelliği, çok sayıda özelliğe sahip veri kümelerinde eşzamanlı işlem yapmayı sağlar. Veri paralelliği, büyük hacimli veri setleriyle çalışırken verimliliği artırmak amacıyla kullanılır. Oylama paralelliği ise hem çok sayıda özelliğin hem de model oylamasının gerektiği durumlarda tercih edilir. Bu yapı sayesinde LightGBM, farklı veri özelliklerine sahip senaryolarda ölçeklenebilir, esnek ve yüksek performanslı bir biçimde çalışabilmektedir. 2016 yılında piyasaya sürülmesinden bu yana LightGBM, büyük veri makine öğrenmesi alanında yaygın biçimde kullanılmaktadır. XGBoost ile makine öğrenmesi uygulamalarında güçlü bir araç olarak kabul edilmektedir. Deneysel çalışmalar, LightGBM'in mevcut diğer güçlendirme algoritmalarına göre daha verimli, doğru ve hızlı olduğunu göstermektedir. Ayrıca LightGBM, daha az bellek kullanımı gerektirmekte ve birden fazla makine kullanıldığında doğrusal hızlanma sağlamaktadır (Ma vd. 2018).

4.10 CatBoost

CatBoost, üstün performansıyla öne çıkan ve diğer kamuya açık güçlendirme yöntemlerinden daha başarılı sonuçlar elde eden bir topluluk öğrenme algoritmasıdır (Prokhorenkova vd. 2018). Bu başarısı, modelin performansını önemli ölçüde artıran yenilikçi algoritmalar ve yaklaşımlar sayesinde sağlanmaktadır. CatBoost'un öne çıkan en önemli özelliği, kategorik değişkenleri etkin bir şekilde işleyebilmesidir (Zhang ve Jánošík 2024).

Geleneksel GAKA, kategorik özelliklerin modellenmesinde zorluklarla karşılaşır ve genellikle one-hot encoding adımı gerektirir. Bu yöntem, her kategori için ayrı bir ikili değişken oluşturduğundan, çok sayıda kategoriye sahip özelliklerde değişken sayısının önemli ölçüde artmasına neden olur. Buna karşılık CatBoost, kategorik değişkenleri one-hot kodlamaya ihtiyaç duymadan doğrudan işleyebilme yeteneğine sahiptir (Zhang ve Jánošík 2024).

Bu yöntem, kategorik özelliklerin hedef istatistikleri adı verilen sayısal temsillere dönüştürülmesine dayanır. Hedef istatistikleri, her kategorinin beklenen hedef değerini tahmin etmeyi amaçlar. CatBoost, bu dönüşümü gerçekleştirmek için "sıralı hedef istatistiği" olarak adlandırılan özel bir strateji uygular. Bu strateji, kategorik değişkenleri sayısal değerlere dönüştürürken örneklerin sıralı biçimde değerlendirilmesini sağlar (Zhang ve Jánošík 2024). Bu dönüşüm süreci

$$X_{\sigma_{p,k}} = \frac{\sum_{j=1}^{p-1}\left[X_{\sigma_{j,k}} = X_{\sigma_{p,k}}\right] Y_{\sigma_s} + \varrho P}{\sum_{j=1}^{p-1}\left[X_{\sigma_{j,k}} = X_{\sigma_{p,k}}\right] + \varrho} \tag{4.29}$$

biçiminde ifade edilir. Burada, P önsel değeri ve ϱ yumuşatma parametresi olmak üzere, ϱP terimi düşük frekanslı kategoriler için hedef istatistiğini dengelemekte önemli bir rol oynar. Bu formülün uygulanması ve eğitim örneklerinin rasgele permütasyonlarının dikkate alınması sayesinde, CatBoost kategorik değişkenleri etkili biçimde sayısal

değerlere dönüştürür. Böylece model, bu bilgiyi güçlendirme sürecine entegre ederek yüksek doğruluk ve kararlılık elde eder (Zhang ve Jánošík 2024).

CatBoost algoritması, hedef sızıntısından kaynaklanan tahmin kayması sorununu ortadan kaldırmak için özel bir çözüm yaklaşımı kullanmaktadır. Bu yöntem, birkaç temel adımdan oluşmaktadır. İlk olarak, eğitim örnekleri üzerinde σ ile gösterilen rasgele bir permütasyon uygulanır. Bu işlem, gradyan artırma sürecinin farklı aşamalarında birbirinden farklı örnek sıralamalarının kullanılmasına olanak tanır (Zhang ve Jánošík 2024).

Daha sonra CatBoost, $M_1, M_2, \dots, M_n$ olarak etiketlenen ve her biri permütasyondaki ilk i örnek üzerinde eğitilen n adet yardımcı modelden oluşan bir model koleksiyonu oluşturur. Bu aşamalı öğrenme stratejisi, modellerin giderek daha geniş alt kümelerden bilgi edinmesini sağlar. Gradyan artırma sürecinin her adımında, $j.$ örnek için artık (residual) değeri, bir önceki M_{j-1} modeli tarafından hesaplanır. Artık, belirli bir örneğe ait tahmin edilen değer ile gerçek değer arasındaki farkı ifade eder. Bu yapı sayesinde CatBoost, hedef sızıntısının neden olduğu sistematik kaymaları azaltır ve tahmin doğruluğunu korur (Dorogush vd. 2018).

Buna ek olarak CatBoost, kategorik özellikler arasındaki etkileşimleri yakalayabilmek için bu özelliklerin kombinasyonlarını da modelleme sürecine dâhil eder. Yani algoritma, kategorik özellikleri yalnızca tek başına değil, diğer kategorik özelliklerle birleştirerek de değerlendirir. Bu yaklaşım, modelin kategorik değişkenler arasındaki karmaşık ilişkileri daha etkin biçimde öğrenmesine olanak tanır (Zhang ve Jánošík 2024).

Birleştirme işlemi, mevcut karar ağacında yeni bir bölünme yapılacağı zaman açgözlü bir şekilde gerçekleştirilir. İlk bölünmede herhangi bir kombinasyon yer almazken, sonraki bölünmelerde mevcut ağaçtan elde edilen tüm kombinasyonlar ve kategorik özellikler birlikte değerlendirilir. Ayrıca CatBoost, ağaçta seçilen bölümleri iki değerli (binary) kategoriler olarak ele alır; böylece hem sayısal hem de kategorik özelliklerin

kombinasyonlarını oluşturur ve bu değerleri yeni bölünme adayları arasında değerlendirir. Bu yenilikçi teknikler sayesinde CatBoost, geleneksel GAKA yöntemlerine kıyasla daha yüksek tahmin doğruluğu elde etmekte ve aşırı öğrenme riskini önemli ölçüde azaltmaktadır (Zhang vd. 2020).

4.11 Algoritmaların Karşılaştırılması

Bu alt bölümde, anlatılan algoritmaların güçlü ve zayıf yönleri tablo şeklinde karşılaştırmalı olarak verilmiştir.

Çizelge 4.1 Algoritmaların güçlü ve zayıf yönlerinin karşılaştırılması

Algoritma	Güçlü Yönleri	Zayıf Yönleri
Klasik Doğrusal Regresyon	Kolayca yorumlanabilir.	Bağımsız değişkenler yüksek derecede doğrusal ise aşırı uyum eğilimi gösterir. Optimal fonksiyonel form ve etkileşimler önceden belirlenmelidir.
LASSO Regresyon	Cezalandırma, değişkenler yüksek oranda ilişkili olduğunda aşırı uyum riskini azaltır. Değişken seçimi için kullanılabilir.	LASSO hangi katsayıların küçültüleceğini keyfi olarak seçebilir (örneğin, kavramsal olarak önemli öngörücüler). Küçültmeden sonra regresyon katsayıları yorumlanamayabilir. Katsayılar için standart hatalar hesaplanamaz.
Ridge Regresyon	Cezalandırma, değişkenler yüksek oranda ilişkili olduğunda aşırı uyum riskini azaltır. Değişken seçimi için kullanılabilir.	Küçültmeden sonra regresyon katsayıları yorumlanamayabilir. Katsayılar için standart hatalar hesaplanamaz.
Karar Ağaçları	Karar ağaçları görsel olarak yorumlanabilir. Değişken önem ölçütleri sağlanır.	Bireysel sınıflandırma ağaçları verilere aşırı uyum sağlayabilir.
Rasgele Orman	Rasgele ormanlar, tek bir ağaçtan daha düşük aşırı uyum riski taşır. Değişken önem ölçütleri sağlanır.	Bireysel sınıflandırma ağaçları verilere aşırı uyum sağlayabilir. Ormanlar görsel olarak yorumlanamaz (örneğin, değişken önem ölçütleri etkilerin yönünü göstermez).
Destek Vektör Makineleri	Yüksek derecede karmaşık verilerde (örneğin metin, görüntüler) iyi performans gösterir; bu, tahmin edici sayısının gözlem sayısından fazla olduğu durumlarda da geçerlidir "yüksek boyutlu" veriler).	Diğer algoritmalara göre aşırı uyuma daha yatkındır (örneğin, düzenleme). "Kara kutu" algoritması; hiperdüzlemi optimize etmek için tahmincilerin nasıl birleştirildiğine dair ölçütler sağlanmamıştır. (tahminlerin neden doğru olduğu belirsiz olabilir).

Çizelge 4.1 Algoritmaların güçlü ve zayıf yönlerinin karşılaştırılması (devam)

Algoritma	Güçlü Yönleri	Zayıf Yönleri
Yapay Sinir Ağları	Karmaşık doğrusal olmayan ilişkileri yüksek doğrulukla modelleyebilir. Büyük veri setlerinde güçlü genelleme kapasitesine sahiptir. Özellik mühendisliğine duyulan gereksinimi azaltır.	Eğitim süreci yüksek hesaplama gücü ve zaman gerektirir. Modelin iç yapısı "kara kutu" niteliğinde olduğundan yorumlanabilirliği düşüktür. Aşırı öğrenme riski yüksektir.
XGBoost	XGBoost, düzenlileştirme ile aşırı uyumu azaltır, paralel işlemle hızlı çalışır ve kayıp verileri otomatik işler. Ayrıca budama, çapraz doğrulama ve artımlı öğrenme özellikleriyle esnek ve güçlü bir model sunar.	Küçük veri setlerinde aşırı uyum riski görülebilir. Hiperparametre ayarları karmaşık ve performansa duyarlıdır. Modelin iç yapısı karmaşık olduğu için yorumlanabilirliği düşüktür. Büyük veriyle çalışırken bellek kullanımı artabilir.
LightGBM	LightGBM, hızlı eğitim süreci ve düşük bellek kullanımıyla öne çıkan bir algoritmadır. Yüksek model doğruluğu sağlar, paralel öğrenmeyi destekler ve büyük veri setleriyle verimli şekilde çalışabilir.	Küçük veri setlerinde aşırı uyum riski artabilir. Modelin yorumlanabilirliği düşüktür. Hiperparametre seçimi model performansını önemli ölçüde etkiler.
CatBoost	Kategorik değişkenleri doğrudan işleyebilme yeteneğine sahiptir. Aşırı öğrenmeyi azaltan düzenlileştirme ve sıralı hedef istatistiği kullanır. Diğer GAKA yöntemlerine kıyasla daha yüksek doğruluk ve kararlılık sağlar.	Büyük veri setlerinde eğitim süresi uzun olabilir. Modelin iç yapısı karmaşık olup yorumlanabilirliği sınırlıdır. Hiperparametre ayarları modele duyarlıdır ve dikkatle yapılmalıdır.

5. YÖNTEM

Bu bölümde, sigorta hasar tutarlarının tahminine yönelik olarak kullanılan veri setleri, modelleme yaklaşımları ve performans değerlendirme yöntemleri ele alınmaktadır. Tezde, hasar tutarlarının öngörülmesi amacıyla Tweedie Regresyonu ile çeşitli Makine Öğrenmesi algoritmaları (Klasik Doğrusal Regresyon, LASSO, Ridge, Karar Ağaçları, Rasgele Orman, Destek Vektör Makineleri, Yapay Sinir Ağları, XGBoost, LightGBM ve CatBoost) uygulanmıştır. Tüm modeller, aynı veri yapısı üzerinde %80 eğitim ve %20 test ayrımı korunarak karşılaştırılmış; her bir algoritmanın hiperparametre optimizasyonu, kendi parametre uzayı içerisinde gerçekleştirilmiştir. Böylece yöntemlerin kendi potansiyelleri doğrultusunda nesnel bir performans değerlendirmesi yapılmıştır. Analizler Google Colab ortamında, Python ve R programlama dilleri kullanılarak yürütülmüştür.

5.1 Veri Setleri ve Özellikler

Bu tezde dört farklı sigorta hasar veri seti kullanılmıştır: Otomobil Sigortası, Araç Sigorta Hasarı, MASS ve Ohlsson. Veri setlerinin seçilme amacı, farklı kaynaklardan elde edilmiş olmalarına rağmen, ortak yapısal özellikleri bakımından sigorta hasar modellemesinde kritik önem taşıyan sıfır yığılmalı ve ağır kuyruklu dağılımları incelemektir. Bu kapsamda, her bir veri seti hem hasar frekansı (sıfır veya pozitif hasar oluşumu) hem de hasar şiddeti (pozitif hasar tutarlarının büyüklüğü) açısından farklı veri özellikleri sunmakta ve böylece kullanılan algoritmaların bu koşullar altındaki performanslarının karşılaştırılmasına olanak tanımaktadır. Bu çeşitlilik, modellerin hem düşük hasar sıklığı hem de yüksek varyanslı hasar tutarları gibi sigorta verilerinin karakteristik zorluklarıyla başa çıkma yeteneklerinin değerlendirilmesi açısından önemli bir avantaj sağlamaktadır. Bu tez kapsamında kullanılan dört veri setine ilişkin temel erişim bilgileri Çizelge 5.1'de sunulmuştur. Çizelge 5.2'de ise tezde kullanılan dört veri setine ait gözlem sayıları ile sürekli ve kategorik değişkenlerin dağılımı özetlenmiştir. Veri setleri, R ortamında açık erişimli istatistiksel paketler aracılığıyla elde edilmiştir. Her bir veri seti, farklı sigorta türleri ve modelleme yaklaşımları için literatürde yaygın olarak kullanılan referans veri kaynaklarını temsil etmektedir.

Çizelge 5.1 Kullanılan veri setlerinin kaynağı ve erişim bilgileri

Veri Seti Adı	**Gerçek Adı**	**R Paketinde Bulunduğu Yer**	**Kaynak**
Otomobil Sigortası	dataCar	insuranceData	R ortamında insuranceData paketi aracılığıyla erişilebilir.
Araç Sigorta Hasarı	AutoClaim	cplm	R ortamında cplm paketi aracılığıyla erişilebilir.
MASS	MTPL	insurancerating	R ortamında insurancerating paketi aracılığıyla erişilebilir.
Ohlsson	dataOhlsson	insuranceData	R ortamında insuranceData paketi aracılığıyla erişilebilir.

Çizelge 5.2 Veri setlerinin kısaca açıklaması

Veri Seti	**Gözlem Sayısı**	**Değişken Sayısı**	**Sürekli Değişkenlerin Sayısı**	**Kategorik Değişkenlerin Sayısı**
Otomobil Sigortası	67,856	11	6	5
Araç Sigorta Hasarı	10,296	29	15	14
MASS	30,000	7	6	1
Ohlsson	64,548	9	8	1

Otomobil Sigortası veri seti, aktüerya literatüründe en sık kullanılan örnek veri setlerinden biridir. Veri seti yaklaşık 67,856 gözlemden ve 11 değişkenden oluşmaktadır. Bu tez kapsamında seçilen hedef değişken hasar_tutarı (hasar_tutari) olup, poliçeler için gerçekleşen toplam hasar maliyetini temsil etmektedir. Veri setinin en dikkat çekici özelliği, aşırı derecede sıfır yığılmalı bir yapıya sahip olmasıdır; gözlemlerin yaklaşık %93'ünde herhangi bir hasar gerçekleşmemiştir. Pozitif hasar gözlemleri ise yüksek varyansa ve ağır kuyruklu bir dağılıma sahiptir. Bu durum, modelleme sürecinde önemli zorluklar yaratmaktadır. Söz konusu özellikler, hem sıfır/pozitif hasar frekansını hem de pozitif hasar şiddetini açıklamada belirleyici bir rol oynamaktadır. Özellikle bölge ve gövde tipi değişkenleri hasar olasılığı üzerinde, araç değeri değişkeni ise hasar tutarı üzerinde ayırt edici bir etkiye sahiptir.

Çizelge 5.3 Otomobil veri setindeki tüm değişkenler

Değişken	Açıklama
bölge	Sigortalının bulunduğu bölgeyi gösteren kategorik değişken; 6 seviyelidir. A, B, C, D, E ve F
hasar_tutarı	Hasar tutarlarını temsil eden sürekli değişken.
yaş	Araç sahibinin yaşı; 6 seviyelidir. 1, 2, 3, 4, 5 ve 6 (1 en genç sahipleri ve 6 en yaşlı sahipleri temsil etmektedir.)
arac_değeri	Aracın göreli değerini temsil eden sürekli değişken.
araç_yaşı	Aracın yaşını temsil eden sürekli değişken.
araç_tipi	Araç gövde tipi; aşağıdaki seviyeleri içeren kategorik bir değişkendir: Otobüs, Cabrio, Spor Otomobil, Hatchback, Hardtop, Mini Araç, Minibüs, Panelvan, Roadster, Sedan, Bagajlı Araç, Kamyonet ve Arazi tipi kamyonet
cinsiyet	2 gruba sahip bir kategorik değişken. K (Kadın) ve E (Erkek); sürücünün/araç sahibinin cinsiyetini temsil eder.
maruziyet	Poliçenin yıl cinsinden geçerlilik süresi; 0 ile 1 arasında değer alan sürekli bir değişken.
hasar	2 gruba sahip bir kategorik değişken. 0 (Yok) ve 1 (Var); O yıl içerisinde hasar olup olmadığını temsil eder.
hasar_sayısı	Bir yıl içinde açılan toplam hasar sayısı; 0 ile 2 arasında değer alan sürekli bir değişken.
x_sütunu	İşlem sütunu; sabit değer içerir ve tek seviyelidir.

Araç Sigorta Hasarı veri seti, araç sigortalarında hasar maliyetlerinin modellenmesi amacıyla kullanılan önemli bir örnek veri setidir. Veri seti yaklaşık 10,296 gözlem ve 29 değişkenden oluşmaktadır. Bu tez kapsamında hasar_tutarı değişkeni bağımlı değişken olarak seçilmiştir. Veri seti, sıfır yığılmalı bir yapıya sahiptir; poliçelerin büyük çoğunluğunda herhangi bir hasar gerçekleşmezken, az sayıdaki poliçede yüksek maliyetli hasarlar gözlemlenmektedir. Veri setinde yer alan değişkenler hem sürücüye hem de araca ilişkin risk faktörlerinin analiz edilmesine olanak tanımaktadır. Bu özellikleriyle Araç Sigorta Hasarı veri seti, özellikle hasar tahmini ve riske göre fiyatlama çalışmalarında sıklıkla başvurulan referans veri setlerinden biridir.

Çizelge 5.4 Araç Sigorta Hasarı veri setindeki tüm değişkenler

Değişken	Açıklama
poliçe_numarası	Poliçe numarası (ID).
poliçe_başlangıcı	Poliçe başlangıç tarihi.
hasar_sayısı_5	Son 5 yıldaki toplam hasar sayısı; sürekli değişken.
hasar_tutarı_5	Son 5 yıldaki toplam hasar tutarı; sürekli değişken.
hasar_tutarı	Hasar tutarlarını temsil eden sürekli değişken.
araç_kullanan_çocuk_sayısı	Araç kullanan çocuk sayısı, 2 seviyelidir; 0 ve 1.
seyahat_süresi	Sigortalının işe/okula olan uzaklığının dakika cinsinden süresi; sürekli değişken.
araç_kullanım_amacı	Sigortalının araç kullanım amacını gösteren kategorik değişken; Özel ve Ticari.
araç_değeri	Aracın göreli değerini temsil eden sürekli değişken.
süreklilik	Sigortalının şirkette kaldığı yıl sayısı; sürekli değişken.
poliçe_sayısı	Sigortalının sahip olduğu poliçe sayısı; sürekli değişken.
araç_tipi	Araç gövde tipi; aşağıdaki seviyeleri içeren kategorik bir değişkendir: Kamyonet, Pikap, Sedan, Spor Otomobil ve Minivan
kırmızı_araba	2 gruba sahip bir kategorik değişken. Evet ve Hayır; sigortalının aracının kırmızı olup olmadığını temsil eder.
ehliyet_iptali	2 gruba sahip bir kategorik değişken. Evet ve Hayır; son 7 yıl içinde ehliyetin iptal edilip edilmediğini temsil eder.
ceza_puanı	Sürücünün trafik ceza puanı; sürekli değişken.
hasar_durumu	2 gruba sahip bir kategorik değişken. Evet ve Hayır; sigortalının aracında hasar olup olmadığını temsil eder.
yaş	Araç sahibinin yaşı; sürekli değişken.
çocuk_sayısı	Çocuk sayısı; sürekli değişken.
iş_yeri_yıl	Sigortalının işyerindeki yıl sayısı; sürekli değişken.
gelir	Sigortalının yıllık geliri; sürekli değişken.
cinsiyet	2 gruba sahip bir kategorik değişken. K (Kadın) ve E (Erkek); sürücünün/araç sahibinin cinsiyetini temsil eder.
medeni_durum	2 gruba sahip bir kategorik değişken. Evli ve Bekar; sigortalının medeni durumu temsil eder.
ebeveyn	2 gruba sahip bir kategorik değişken. Evet ve Hayır; sigortalının çocuğunun olup olmadığını temsil eder.
meslek	Sigortalının mesleğinin sınıfı; aşağıdaki seviyeleri içeren kategorik bir değişkendir: Bilinmiyor, Mavi Yaka, Ofis Çalışanı, Doktor, Ev Hanımı, Avukat, Yönetici, Profesör ve Öğrenci
eğitim	Sigortalının eğitim seviyesi; aşağıdaki seviyeleri içeren kategorik bir değişkendir: Lise veya daha aşağı, Lise, Lisans, Yüksek Lisans ve Doktora
evin_değeri	Sigortalının evinin değeri; sürekli değişken.
aynı_evde_yıl	Sigortalının aynı evde geçirdiği yıl sayısı; sürekli değişken.
bölge	Sigortalının bulunduğu bölgeyi gösteren kategorik değişken; Kırsal ve Kentsel.
x_sütunu	2 gruba sahip bir kategorik değişken. Doğru ve Yanlış; veri setinde işaret.

MASS (Motorlu Araç Sorumluluk Sigortası) veri seti, sigorta ve aktüerya literatüründe klasik bir karşılaştırma veri seti olarak kullanılmaktadır. Veri seti yaklaşık 30,000 poliçe bilgisinden oluşmakta ve 7 değişken içermektedir. Bu tez kapsamında hedef değişken hasar_tutarı olarak tanımlanmıştır. MASS veri seti, sıfır yığılmalı ve ağır kuyruklu bir dağılım yapısına sahiptir. Özellikle risk_sınıfı (risk_sinifi) değişkeni, sürücülerin risk seviyelerini ayırt etmede belirleyici bir rol oynamaktadır. Bu özellikleriyle MASS veri seti, aktüerya literatüründe hem hasar frekansı hem de hasar şiddeti modellemeleri için yaygın biçimde kullanılan standart bir test veri seti konumundadır.

Çizelge 5.5 MASS veri setindeki tüm değişkenler

Değişken	Açıklama
motor_gücü	Araç motor gücü / beygir gücü; sürekli değişken.
hasar_tutarı	Hasar tutarlarını temsil eden sürekli değişken.
bölge	Sigortalının bulunduğu bölgeyi gösteren kategorik değişken; 4 seviyelidir. 0, 1, 2 ve 3
yaş	Sigortalının yaşı; sürekli değişken.
risk_sınıfı	Sigortalının risk sınıfı; hasar geçmişine dayalı olarak belirlenir. Daha yüksek bm sınıfları genellikle düşük risk anlamına gelir.
hasar_sayısı_toplam	Bir yıl içinde açılan toplam hasar sayısı; 0 ile 2 arasında değer alan sürekli bir değişken.
maruziyet	Poliçenin yıl cinsinden geçerlilik süresi; sürekli değişken.

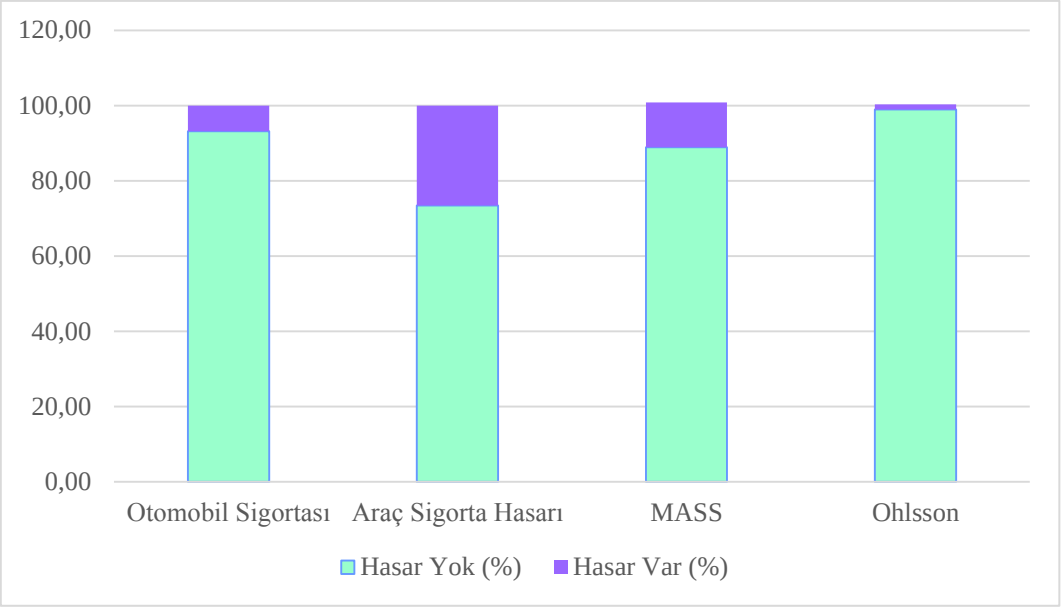
Ohlsson veri seti, sigorta aktüeryası literatüründe Ohlsson ve Johansson tarafından sunulmuştur. Veri seti yaklaşık 64,548 gözlem ve 9 değişkenden oluşmaktadır. Bu tez kapsamında hedef değişken hasar_tutarı olarak tanımlanmıştır. Ohlsson veri seti, özellikle İsveç motor sigortası pazarı için hasar modellemesi amacıyla oluşturulmuş olup, aktüeryal uygulamalarda hem hasar frekansı hem de hasar şiddeti modellemelerinde referans niteliği taşımaktadır.

Çizelge 5.6 Ohlsson veri setindeki tüm değişkenler

Değişken	Açıklama
maruziyet	Poliçenin yıl cinsinden geçerlilik süresi; sürekli değişken.
hasar_tutarı	Hasar tutarlarını temsil eden sürekli değişken.
araç_sınıfı	Aracın sınıfı; $(Motor\ gücü * 100) / (Araç\ ağırlığı + 75)$ formülünden türetilen oran ile sınıflandırılır. 7 seviyelidir. 1, 2, 3, 4, 5, 6 ve 7
araç_yaşı	Aracın yaşı; sürekli değişken.
cinsiyet	2 gruba sahip bir kategorik değişken. K (Kadın) ve M (Erkek); sürücünün/araç sahibinin cinsiyetini temsil eder.
bölge	Sigortalının yaşadığı bölge; 7 seviyelidir. 1, 2, 3, 4, 5, 6 ve 7
yaş	Sigortalının yaşı; sürekli değişken.
risk_sınıfı	Sigortalının risk sınıfı; hasar geçmişine dayalı olarak belirlenir. Daha yüksek risk sınıfları genellikle düşük risk anlamına gelir.
hasar_sayısı	Poliçe döneminde gerçekleşen hasarların sayısı; sürekli değişken.

Çizelge 5.7 Dört veri setinde hasar_tutarı değişkeninde sıfır değerlerin analizi

Veri Seti	Değişken	Sıfırların Sayısı	Gözlem Sayısı	Sıfırların Yüzdeliği
Otomobil Sigortası	hasar_tutarı	63,232	67,856	%93,20
Araç Sigorta Hasarı	hasar_tutarı	7,556	10,296	%73,43
MASS	hasar_tutarı	26,674	30,000	%88,91
Ohlsson	hasar_tutarı	63,878	64,548	%98,96



Şekil 5.1 Dört veri setinde hasar_tutarı değişkeninin sıfır değerlerinin dağılımı

Çizelge 5.7 ve Şekil 5.1, dört veri setinde hasar_tutarı değişkeninde sıfır değerlerinin dağılım özelliklerini karşılaştırmalı olarak göstermektedir. Çizelge 5.7'de, veri setlerinde yer alan hasar tutarı değişkeninde sıfır değerinin sıklığı ve yüzdelik değeri verilmiştir. Buna göre, sıfır oranı Ohlsson veri setinde %98,96 ile en yüksek düzeydedir; bu durum veri setinin neredeyse tamamının sıfır hasarlardan oluştuğunu göstermektedir. Otomobil Sigortası veri seti %93,20 ile benzer biçimde oldukça yüksek bir sıfır oranına sahiptir. MASS veri seti %88,91 ile orta düzeyde, Araç Sigorta Hasarı veri seti ise %73,43 ile diğerlerine kıyasla daha düşük düzeyde sıfır oranı göstermektedir.

Şekil 5.1, bu oranların görsel bir karşılaştırmasını sunarak veri setlerinin sıfır yığılmalı yapısını açık biçimde ortaya koymaktadır. Grafikten görüldüğü üzere, dört veri setinde de sıfır (hasar yok) gözlemleri baskın durumdadır. Bu yapı, sigorta verilerinde sıkça karşılaşılan seyrek ve dengesiz dağılım problemini yansıtmaktadır.

Çizelge 5.8'de Otomobil Sigortası veri setinde yer alan hasar_tutarı değişkenine ilişkin frekans dağılımı görülmektedir. Tabloya göre, gözlemlerin çok büyük bir kısmı 0–10.000 aralığında yoğunlaşmıştır. Bu aralıkta yer alan %96,76'lık oran, veri setindeki hasarların çoğunlukla düşük tutarlarda gerçekleştiğini göstermektedir. Daha yüksek tutarlardaki hasarlar oldukça nadirdir; örneğin 10.000–20.000 aralığında yer alan

hasarların oranı yalnızca %2,55, 20.000–30.000 aralığındaki hasarların oranı ise %0,52 düzeyindedir. 40.000 üzerindeki hasar tutarlarının ise toplam içindeki payı ihmal edilebilir düzeydedir.

Çizelge 5.8 Otomobil Sigortası veri setinde hasar_tutarı değişkeni için frekans tablosu

Hasar Tutarı Aralığı	**Sıklık**	**Yüzde**
$0 - 10.000$	4473	96,76
$10.000 - 20.000$	118	2,55
$20.000 - 30.000$	24	0,52
$30.000 - 40.000$	6	0,13
$40.000 - 50.000$	2	0,04

Bu durum, veri setinde küçük ölçekli hasarların baskın, yüksek tutarlı hasarların ise istisnai nitelikte olduğunu ortaya koymaktadır. Dolayısıyla, hasar_tutarı değişkeni sağa çarpık bir dağılım sergilemektedir.

Çizelge 5.9'da Araç Sigorta Hasarı veri setinde hasar_tutarı değişkenine ilişkin frekans dağılımı yer almaktadır. Tabloya göre, hasarların büyük çoğunluğu 0–10.000 aralığında toplanmıştır. Bu aralıktaki %92,86'lık oran, veri setindeki hasarların önemli bir kısmının düşük tutarlı olduğunu göstermektedir. 10.000–20.000 aralığında gerçekleşen hasarlar %3,43, 20.000–30.000 aralığındakiler ise %1,64 oranındadır. 30.000 üzerindeki hasarların frekansı ve oranı kademeli olarak azalarak toplamda oldukça düşük bir seviyede kalmaktadır.

Çizelge 5.9 Araç Sigorta Hasarı veri setinde hasar_tutarı değişkeni için frekans tablosu

Hasar Tutarı Aralığı	Sıklık	Yüzde
$0 - 10.000$	2381	92,86
$10.000 - 20.000$	88	3,43
$20.000 - 30.000$	42	1,64
$30.000 - 40.000$	21	0,82
$40.000 - 50.000$	14	0,55
$50.000 - 60.000$	9	0,35
$60.000 - 70.000$	4	0,16
$70.000 - 80.000$	3	0,12
$80.000 - 90.000$	1	0,04
$100.000 - 110.000$	1	0,04

Bu dağılım, veri setinde küçük ve orta ölçekli hasarların baskın, yüksek tutarlı hasarların ise oldukça nadir olduğunu göstermektedir. Dolayısıyla, hasar tutarı değişkeni sağa çarpık bir yapıya sahiptir.

Çizelge 5.10 MASS veri setinde hasar_tutarı değişkeni için frekans tablosu

Hasar Tutarı Aralığı	Sıklık	Yüzde
$0 - 1.000.000$	3298	99,19
$1.000.000 - 2.000.000$	22	0,66
$2.000.000 - 3.000.000$	2	0,06
$3.000.000 - 4.000.000$	1	0,03
$5.000.000 - 6.000.000$	1	0,03
$9.000.000 - 10.000.000$	1	0,03

Çizelge 5.10'da MASS veri setinde yer alan hasar_tutarı değişkenine ilişkin frekans dağılımı verilmiştir. Tabloya göre, hasarların çok büyük bir kısmı 0–1.000.000 aralığında toplanmış olup, bu aralık toplam gözlemlerin %99,19'unu oluşturmaktadır. Bu durum, veri setindeki hasarların neredeyse tamamının düşük tutarlı olduğunu göstermektedir. 1.000.000–2.000.000 aralığındaki hasarların oranı %0,66, 2.000.000

üzerindeki hasarların oranı ise oldukça düşüktür (her biri %0,03–0,06 düzeyindedir). Bu değerler, yüksek tutarlı hasarların oldukça nadir gerçekleştiğini ve veri setinde uç değerlerin sınırlı sayıda bulunduğunu ortaya koymaktadır. Genel olarak dağılım, sağa çarpık bir yapı sergilemektedir; yani düşük tutarlı hasarlar yoğunlaşırken, yüksek tutarlı hasarlar giderek azalmaktadır.

Çizelge 5.11 Ohlsson veri setinde hasar_tutarı değişkeni için frekans tablosu

Hasar Tutarı Aralığı	Sıklık	Yüzde
$0 - 100.000$	638	95,37
$100.000 - 200.000$	29	4,33
$200.000 - 300.000$	2	0,30

Çizelge 5.11'de Ohlsson veri setinde yer alan hasar_tutarı değişkenine ilişkin frekans dağılımı sunulmuştur. Çizelge incelendiğinde, hasarların büyük çoğunluğunun 0–100.000 aralığında toplandığı görülmektedir. Bu aralık, toplam gözlemlerin %95,37'sini oluşturmakta olup, veri setinde düşük tutarlı hasarların baskın olduğunu göstermektedir. 100.000–200.000 aralığındaki hasarların oranı %4,33, 200.000–300.000 aralığındaki hasarların oranı ise yalnızca %0,30 düzeyindedir. Bu bulgular, yüksek tutarlı hasarların oldukça nadir gerçekleştiğini, dolayısıyla veri setinde küçük ölçekli hasarların ağırlıkta, büyük hasarların ise istisnai nitelikte olduğunu ortaya koymaktadır. Genel olarak dağılım sağa çarpık bir yapı sergilemekte, bu da pozitif ve asimetrik dağılımların hâkim olduğunu göstermektedir.

5.2 Veri Temizleme ve Özellik Seçimi

Analiz öncesinde veri setleri üzerinde kapsamlı bir veri temizleme süreci uygulanmıştır. İlk olarak, veri setinde kayıp gözlemler bulunup bulunmadığı kontrol edilmiş, kayıp değerlere sahip kayıtlar tespit edildiğinde bu gözlemler veri setinden çıkarılmıştır. Bu işlem, analizlerin güvenilirliğini artırmak ve modelleme aşamasında oluşabilecek tutarsızlıkları önlemek amacıyla gerçekleştirilmiştir. Ardından, veri setinde aykırı değerler incelenmiş; bu tür gözlemlerin model performansını veya eğitim-test dengesini

olumsuz etkilediği durumlarda, ilgili gözlemler veri setinden çıkarılmıştır. Böylece modellerin uç değerlerden kaynaklanabilecek sapmalardan arındırılması sağlanmıştır.

Veri setinin genellenebilirliğini artırmak amacıyla, veri beş farklı eğitim-test parçalanması ile rasgele biçimde ayrılmıştır. Bu yaklaşım, modellerin tek bir veri bölünmesine bağımlı kalmaksızın farklı örneklemler üzerinde benzer performans göstermesini sağlamayı hedeflemektedir. Her bir bölünme için performans ölçütleri ($RMSE$, MAE, $rRMSE$ ve $rMAE$) ayrı ayrı hesaplanmış; ardından beş denemenin sonuçları ortalama alınarak nihai performans değerleri elde edilmiştir.

Değişken seçimi aşamasında, ilk olarak tüm değişkenleri içeren klasik bir doğrusal regresyon modeli kurulmuştur. Daha sonra, modeldeki değişkenlerin istatistiksel anlamlılık düzeyleri değerlendirilmiş ve yalnızca anlamlı bulunan değişkenler nihai modele dahil edilmiştir. Modelleme sürecinde yer alan kategorik değişkenler, uygun biçimde faktör tipine dönüştürülerek analize dahil edilmiştir. Bu dönüşüm, değişkenlerin kategorik yapısının doğru biçimde temsil edilmesini ve algoritmaların bu değişkenleri uygun şekilde işlemesini sağlamıştır. Son aşamada, modellerin performansını artırmak amacıyla algoritmalara ait hiperparametreler belirli aralıklar içinde sistematik olarak taranmış ve her model için en iyi sonuç veren hiperparametre kombinasyonu seçilmiştir.

Tüm bu adımlar sonucunda, analizlerde kullanılmak üzere temizlenmiş, dengelenmiş ve optimize edilmiş veri setleri elde edilmiştir. Bu süreç, modelleme aşamasında hem doğruluk hem de genellenebilirlik açısından sağlam bir temel oluşturmuştur. Değişken seçimi sürecinin tamamlanmasının ardından, modelleme aşamasında kullanılacak değişkenler nihai hâlini almıştır. Anlamlılık düzeyi düşük veya modele katkısı sınırlı bulunan değişkenler analiz dışında bırakılmış; böylece modellerin açıklayıcı gücü, istatistiksel tutarlılığı ve genellenebilirliği artırılmıştır. Bu süreç sonunda, her bir veri setinde analize dâhil edilen değişkenler belirlenmiş ve bu değişkenlere ilişkin açıklamalar sistematik biçimde düzenlenmiştir. Aşağıda, dört farklı veri setinde (Otomobil Sigortası, Araç Sigorta Hasarı, MASS ve Ohlsson) kullanılan değişkenlerin nihai hâlleri ve açıklamaları tablo hâlinde sunulmaktadır.

Çizelge 5.12 Otomobil Sigortası veri setinin bağımsız değişkenleri

Değişken	Açıklama
bölge	Sigortalının bulunduğu bölgeyi gösteren kategorik değişken; 6 seviyelidir. A, B, C, D, E ve F
yaş	Araç sahibinin yaşı; 6 seviyelidir. 1, 2, 3, 4, 5 ve 6 (1 en genç sahipleri ve 6 en yaşlı sahipleri temsil etmektedir.)
arac_değeri	Aracın göreli değerini temsil eden sürekli değişken.
araç_yaşı	Aracın yaşını temsil eden sürekli değişken.
araç_tipi	Araç gövde tipi; aşağıdaki seviyeleri içeren kategorik bir değişkendir: Otobüs, Cabrio, Spor Otomobil, Hatchback, Hardtop, Mini Araç, Minibüs, Panelvan, Roadster, Sedan, Bagajlı Araç, Kamyonet ve Arazi tipi kamyonet
cinsiyet	2 gruba sahip bir kategorik değişken. K (Kadın) ve E (Erkek); sürücünün/araç sahibinin cinsiyetini temsil eder.
maruziyet	Poliçenin yıl cinsinden geçerlilik süresi; 0 ile 1 arasında değer alan sürekli bir değişken.

Çizelge 5.13 Araç Sigorta Hasarı veri setinin bağımsız değişkenleri

Değişken	Açıklama
araç_değeri	Aracın göreli değerini temsil eden sürekli değişken.
poliçe_sayısı	Sigortalının sahip olduğu poliçe sayısı; sürekli değişken.
ehliyet_iptali	2 gruba sahip bir kategorik değişken. Evet ve Hayır; son 7 yıl içinde ehliyetin iptal edilip edilmediğini temsil eder.
hasar_durumu	2 gruba sahip bir kategorik değişken. Evet ve Hayır; sigortalının aracında hasar olup olmadığını temsil eder.
cinsiyet	2 gruba sahip bir kategorik değişken. K (Kadın) ve E (Erkek); sürücünün/araç sahibinin cinsiyetini temsil eder.
medeni_durum	2 gruba sahip bir kategorik değişken. Evli ve Bekar; sigortalının medeni durumu temsil eder.
aynı_evde_yıl	Sigortalının aynı evde geçirdiği yıl sayısı; sürekli değişken.

Çizelge 5.14 MASS veri setinin bağımsız değişkenleri

Değişken	Açıklama
motor_gücü	Araç motor gücü / beygir gücü; sürekli değişken.
bölge	Sigortalının bulunduğu bölgeyi gösteren kategorik değişken; 4 seviyelidir. 0, 1, 2 ve 3
yaş	Sigortalının yaşı; sürekli değişken.

Çizelge 5.15 Ohlsson veri setinin bağımsız değişkenleri

Değişken	Açıklama
maruziyet	Poliçenin yıl cinsinden geçerlilik süresi; sürekli değişken.
araç_sınıfı	Aracın sınıfı; $(Motor\ gücü * 100) / (Araç\ ağırlığı + 75)$ formülünden türetilen oran ile sınıflandırılır. 7 seviyelidir. 1, 2, 3, 4, 5, 6 ve 7
araç_yaşı	Aracın yaşı; sürekli değişken.
bölge	Sigortalının yaşadığı bölge; 7 seviyelidir. 1, 2, 3, 4, 5, 6 ve 7
yaş	Sigortalının yaşı; sürekli değişken.
risk_sınıfı	Sigortalının risk sınıfı; hasar geçmişine dayalı olarak belirlenir. Daha yüksek risk sınıfları genellikle düşük risk anlamına gelir.

5.3 Hata Ölçütleri

Bu tezde kullanılan modellerin performanslarının değerlendirilmesinde birden fazla hata ölçütünden yararlanılmıştır. Amaç, farklı yönleriyle model doğruluğunu ve genelleme yeteneğini ortaya koymaktır. Öncelikle, Kök Ortalama Kare Hata (Root Mean Squared Error, $RMSE$) kullanılmıştır. $RMSE$'nin yanında kullanılan bir diğer temel ölçüt ise Ortalama Mutlak Hata (Mean Absolute Error, MAE)'dır. Bu iki temel ölçüt, farklı ölçeklerdeki veri setlerinde karşılaştırma yapmayı zorlaştırabileceğinden, tez kapsamında ayrıca göreli hata ölçütleri de kullanılmıştır.

5.3.1 Kök ortalama kare hata ($RMSE$)

Kök Ortalama Kare Hata ($RMSE$), regresyon ve tahmin modellerinin başarımını değerlendirmede yaygın olarak kullanılan bir hata ölçütüdür. Gerçek ve tahmin değerleri arasındaki farkların karelerinin ortalamasının karekökü alınarak

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.2}$$

biçiminde ifade edilir. Burada:

n: gözlem sayısını,

y_i: i. gözlemin gerçek değerini,

$\hat{y}_i$: modelin i. gözlem için tahminini,

$y_i - \hat{y}_i$: tahmin hatasını ifade eder.

$RMSE$, büyük hatalara karşı daha duyarlıdır; yani birkaç büyük hata $RMSE$ değerini ciddi şekilde yükseltebilir. $RMSE$, modelin tahmin performansını değerlendirirken sık kullanılan bir ölçüttür ve modelin genel uyumunu özetleyen güçlü bir göstergedir.

5.3.2 Ortalama mutlak hata (MAE)

Ortalama Mutlak Hata (MAE), gerçek değerler ile model tahminleri arasındaki farkların mutlak değerlerinin ortalamasıdır. Gerçek değer ile tahmin arasındaki farkın mutlak değeri alınarak, tüm gözlemler üzerinden ortalaması

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5.3}$$

biçiminde gösterilir. Burada:

n: gözlem sayısını,

y_i: i. gözlemin gerçek değerini,

$\hat{y}_i$: modelin i. gözlem için tahminini,

$y_i - \hat{y}_i$: tahmin hatasını ifade eder.

MAE, tipik hatanın büyüklüğünü doğrudan gösterir ve özellikle aykırı değerlere karşı görece dayanıklıdır. Küçük MAE değerleri, modelin genel olarak daha doğru tahminler yaptığını ifade eder.

5.3.3 Göreli kök ortalama kare hata (relative $RMSE$, $rRMSE$)

Göreli Kök Ortalama Kare Hata (relative $RMSE$, $rRMSE$), $RMSE$ değerinin bir referans büyüklüğe (çoğunlukla hedef değişkenin ortalaması) bölünmesiyle

$$rRMSE = \frac{RMSE}{\bar{y}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} y_i} \tag{5.4}$$

biçiminde elde edilir. Burada:

n: gözlem sayısını,

$\bar{y}$: hedef değişkenin ortalamasını,

y_i: i. gözlemin gerçek değerini,

$\hat{y}_i$: modelin i. gözlem için tahminini,

$y_i - \hat{y}_i$: tahmin hatasını ifade eder.

$RMSE$ tabanlı olduğu için büyük hatalara duyarlıdır. Farklı veri setleri veya modeller arası kıyaslarda kullanışlıdır. Birimi yoktur ve boyutsuzdur. $rRMSE$'nin küçük olması, modelin veriye göre daha başarılı bir uyum gösterdiğini belirtir.

5.3.4 Göreli ortalama mutlak hata (relative MAE, $rMAE$)

Göreli Ortalama Mutlak Hata (relative MAE, $rMAE$), MAE değerinin bir referans büyüklüğe (genellikle hedef değişkenin ortalaması) bölünmesiyle

$$rMAE = \frac{MAE}{\bar{y}} = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\frac{1}{n}\sum_{i=1}^{n} y_i} \tag{5.5}$$

biçiminde ifade edilir. Burada:

n: gözlem sayısını,

$\bar{y}$: hedef değişkenin ortalamasını,

y_i: i. gözlemin gerçek değerini,

$\hat{y}_i$: modelin i. gözlem için tahminini,

$y_i - \hat{y}_i$: tahmin hatasını ifade eder.

Bu oranlama, hata değerlerini ölçekten bağımsız hale getirir ve farklı veri setleri veya farklı modeller arasında karşılaştırma yapmayı kolaylaştırır. $rMAE$'nin küçük olması, modelin veriye göre daha başarılı bir uyum gösterdiğini belirtir.

5.4 Hiperparametre Seçimi

Bu tezde kullanılan Tweedie Regresyonu ve Makine Öğrenmesi algoritmalarının performansları, yalnızca uygun model seçimine değil, aynı zamanda hiperparametrelerin doğru belirlenmesine de bağlıdır. Hiperparametreler, modelin öğrenme kapasitesini, genelleme yeteneğini ve aşırı uyum riskini doğrudan etkileyen; kullanıcı tarafından önceden tanımlanması gereken değerlerdir. Bu nedenle hiperparametre uzayı, literatürde önerilen genel aralıklar dikkate alınarak tanımlanmış; arama stratejisi olarak çapraz doğrulama ile ızgara araması yöntemleri tercih edilmiştir. Gerektiğinde rasgele arama ve erken durdurma stratejileri de uygulanmıştır. Seçim ölçütleri, tezin amaçları

doğrultusunda belirlenmiş; model doğruluğu, genellenebilirlik ve hata ölçütleri dengeli biçimde değerlendirilmiştir. Ayrıca, veri sızıntısı riskini önlemek amacıyla tüm veri dönüşümleri çapraz doğrulama katları içinde gerçekleştirilmiş; standartlaştırma ve kategorik değişkenlerin faktör dönüşümü eğitim-test ayrımı korunarak yapılmıştır.

Tweedie Regresyonunda en önemli hiperparametre güç parametresi (ξ)'dir. Bu parametre dağılımın doğasını belirler: $\xi = 1$ Poisson, $\xi = 2$ Gamma dağılımına karşılık gelirken, $1 < \xi < 2$ aralığındaki değerler sıfır değerleriyle sürekli pozitif değerleri birlikte modellemeye olanak tanır. Sigorta uygulamalarında genellikle bu aralıkta değerler tercih edilmektedir. Klasik Doğrusal Regresyon herhangi bir hiperparametreye sahip değildir; parametreler doğrudan en küçük kareler yöntemiyle tahmin edilir. LASSO Regresyonunda ise $L1$ düzenleme katsayısı (α) değişken seçimini etkiler; büyük değerler modeli sadeleştirirken, küçük değerler daha karmaşık bir yapı oluşturur. Her iki modelde de uygun katsayı, geniş aralıklar taranarak ve doğrulama sonuçları değerlendirilerek belirlenmiştir. Ridge Regresyonunda temel hiperparametre $L2$ düzenleme katsayısı (α)'dır. Bu katsayının büyümesi katsayıları cezalandırarak modelin basitleşmesini sağlarken, çok küçük değerler modelin karmaşıklaşmasına, çok büyük değerler ise yetersiz uyuma neden olabilir.

Karar Ağaçlarında modelin karmaşıklığını belirleyen başlıca hiperparametreler ağaç derinliği ve yaprak başına minimum gözlem sayısıdır. Derinlik arttıkça model ayrıntı kazanır ancak aşırı uyum riski yükselir; yaprak başına düşen gözlem sayısı arttıkça model basitleşir ve genelleme yeteneği artar. Ayrıca budama parametreleri de ağaçların denge ve kararlılığını sağlar. Rasgele Orman yönteminde önemli hiperparametrelerden biri ağaç sayısıdır; bu sayı arttıkça model kararlılığı artar fakat hesaplama maliyeti yükselir. Her düğümde kullanılacak değişken sayısını belirleyen parametre modelin çeşitliliğini doğrudan etkilerken, derinlik ve yaprak büyüklüğü parametreleri tekil ağaçların karmaşıklığını sınırlandırır.

Destek Vektör Makinelerinde üç temel hiperparametre bulunmaktadır: düzenleme katsayısı (C), hata tolerans bandı (ε) ve çekirdek parametresi (γ). C değeri büyüdükçe model esnek, küçük oldukça daha kısıtlayıcı hale gelir; ε değeri hataların tolere

edilebileceği sınırı belirler; gamma ise özellikle RBF çekirdeğinde karar yüzeyinin esnekliğini kontrol eder. Yapay Sinir Ağlarında kapasiteyi belirleyen en önemli hiperparametreler katman yapısı, düzenleme katsayısı ve öğrenme oranıdır. Katman ve nöron sayısı arttıkça modelin öğrenme kapasitesi artar ancak aşırı uyum riski de yükselir. Düzenleme katsayısı büyüdükçe ağırlıklar daha fazla sınırlandırılır; öğrenme oranı küçük olduğunda model yavaş fakat kararlı, büyük olduğunda ise hızlı ancak dengesiz öğrenir. Ayrıca erken durdurma stratejisi, doğrulama setinde iyileşme durduğunda eğitimi sonlandırarak aşırı uyumu önler.

XGBoost algoritmasında öne çıkan hiperparametrelerden biri öğrenme hızı (η)'dır. Küçük değerler yavaş fakat güvenilir öğrenme sağlarken, büyük değerler hızlı ancak aşırı uyuma eğilimli modeller oluşturabilir. Ağaç derinliği, minimum gözlem sayısı ve düzenleme katsayıları modelin doğruluk–genelleme dengesini belirlerken, örnekleme oranları modele çeşitlilik kazandırır. LightGBM algoritmasında kapasiteyi belirleyen temel parametre yaprak sayısıdır; yaprak başına minimum gözlem sayısının artırılması aşırı uyumu önler. Özellik ve gözlem örnekleme oranları modelin çeşitliliğini artırır; öğrenme hızı ve iterasyon sayısı ise erken durdurma stratejisiyle dengelenerek genelleme yeteneği korunur. CatBoost algoritmasında ise öne çıkan hiperparametreler derinlik, öğrenme oranı ve düzenleme katsayısıdır. Derinlik arttıkça model esnekliği yükselir, düşük öğrenme oranı daha kararlı bir öğrenme süreci sağlar, düzenleme katsayısı aşırı uyumu sınırlandırır. Ayrıca rasgelelik parametreleri modelin çeşitliliğini artırarak genelleme performansını güçlendirir. Çizelge 5.16'da algoritmalarda kullanılan hiperparametrelerin ve arama aralığının tablosu verilmiştir.

Çizelge 5.16 Algoritmalarda kullanılan hiperparametrelerin ve arama aralığının tablosu

Algoritma	Hiperparametre	Açıklama	Arama Aralığı
Tweedie Regresyon	Güç parametresi	Dağılımın yapısını belirler	$1.1 - 1.9$
LASSO Regresyon	$L1$ düzenleme katsayısı	Büyüdükçe katsayıları sıfıra iterek değişken seçimi yapar	$0.01 - 10.00$
Ridge Regresyon	$L2$ düzenleme katsayısı	Büyüdükçe katsayılar daha çok cezalanır, model sadeleşir	$0.01 - 10.00$
Karar Ağaçları	Karmaşıklık parametresi	Modelin genel uyumsuzluğunu cp katsayısı oranında azaltmayan herhangi bir düğüm bölünmesi gerçekleştirilmez	$[0.001, 0.01, 0.1]$
	Maksimum derinlik	Tek bir ağacın maksimum derinliği	$[3, 5, 7]$
Rasgele Orman	Tahmin edicinin sayısı	Ormandaki ağaç sayısı	$[300, 800, 1200]$
	Maksimum derinlik	Tek bir ağacın maksimum derinliği	$[10, 15]$
Destek Vektör Makineleri	Çekirdek fonksiyonu	Verileri daha yüksek boyutlu uzaylara dönüştürme işlevi	$\{Radyal, Doğrusal\}$
	C	Tek bir eğitim örneğinin hiperdüzlem üzerindeki etkisini kontrol eder	$[1, 10, 100]$
Yapay Sinir Ağları	Gizli birim sayısı	Gizli katmandaki nöron sayısı	20
	$L2$ düzenleme katsayısı (ağırlık çürümesi)	Düzenleme parametresi	$0.0001 - 0.05$
XGBoost	Tahmin edicinin sayısı	XGBoost'taki ağaç sayısı	$[1, 5]$
	Maksimum derinlik	Tek bir ağacın maksimum derinliği	$[4, 6]$
	Alt örnekleme oranı	Tek bir ağacı eğitmek için kullanılan örneklerin oranı	0.8
	Öğrenme oranı	Her ağacın katkısını küçültür	$[0.01, 0.05, 0.10]$
	Sütun örnekleme oranı	Tek bir ağacı eğitmek için kullanılan özelliklerin oranı	0.8
	Gamma	Daha fazla bölünme için minimum kayıp azaltma	$[0, 1]$
LightGBM	Tahmin edicinin sayısı	LightGBM'deki ağaç sayısı	$[31, 63]$
	Maksimum derinlik	Tek bir ağacın maksimum derinliği	$[3, 5]$
	Öğrenme oranı	Her ağacın katkısını küçültür	$[0.01, 0.05]$
CatBoost	Döngü sayısı	CatBoost'taki ağaç sayısı	200
	Maksimum derinlik	Tek bir ağacın maksimum derinliği	5

6. UYGULAMA

Bu bölümde, tez kapsamında kullanılan dört farklı veri seti üzerinde gerçekleştirilen modelleme uygulamaları sunulmaktadır. Veri setleri, sigorta hasar modellemesinde yaygın olarak karşılaşılan sıfır yığılmalı ve ağır kuyruklu yapılara sahiptir. Bu doğrultuda, farklı veri özellikleri altında algoritmaların performanslarını karşılaştırmak ve modelleme yaklaşımlarının genellenebilirlik düzeyini değerlendirmek amacıyla analizler gerçekleştirilmiştir. Ayrıca, veri setlerinin temel istatistiksel özellikleri bu bölümde ayrıntılı biçimde incelenmiştir. Her bir veri setine ait hasar tutarlarının dağılımları, sürekli ve kategorik değişkenlerin tanımlayıcı istatistikleri tablolar ve grafikler yardımıyla sunulmuştur. Bu kapsamda, veri setlerinin dağılım yapıları sütun ve histogram grafikleriyle desteklenmiş; böylece veri setlerinin genel yapısal özellikleri ile sıfır yığılmalarının düzeyi görsel olarak ortaya konulmuştur.

Tüm veri setleri üzerinde Tweedie Regresyonu ile Klasik Doğrusal Regresyon, LASSO, Ridge, Karar Ağaçları, Rasgele Ormanlar, Destek Vektör Makineleri, Yapay Sinir Ağları, XGBoost, LightGBM ve CatBoost algoritmaları uygulanmıştır. Her bir algoritma için hiperparametreler, literatürde önerilen aralıklar dikkate alınarak optimize edilmiş; modellerin performansları eğitim ve test verileri üzerinde $RMSE$, MAE, $rRMSE$ ve $rMAE$ ölçütleri kullanılarak değerlendirilmiştir. Ayrıca, her algoritmanın tüm veri setlerindeki performansları karşılaştırmalı olarak incelenmiş ve elde edilen sonuçlardan hareketle genel ortalama sıralamaları oluşturulmuştur. Böylece, algoritmaların hem bireysel veri setleri üzerindeki başarımı hem de genel modelleme performansı ortaya konulmuştur. Kodlamalar ağırlıklı olarak Google Colab (Python) ortamında yapılmış, yalnızca Araç Sigorta Hasarı veri seti R programlama dili kullanılarak analiz edilmiştir.

6.1 Sürekli Değişkenlere Ait Özet İstatistikler

Bu alt bölümde, dört veri setinde yer alan sürekli değişkenlere ilişkin temel tanımlayıcı istatistikler sunulmaktadır. Her veri seti için ortalama, medyan, minimum, maksimum ve standart sapma değerleri ayrı ayrı hesaplanmış ve tablolar hâlinde özetlenmiştir.

Çizelge 6.1 Otomobil Sigortası veri setindeki sürekli değişkenlere dair özet istatistik

Değişken	Ortalama	Medyan	Minimum	Maksimum	Standart Sapma
araç_değeri	1,78	1,50	0	34,56	1,21
araç_yaşı	2,67	3	1	4	1,07
maruziyet	0,47	0,45	0	1	0,29

Çizelge 6.2 Araç Sigorta Hasarı veri setindeki sürekli değişkenlere dair özet istatistik

Değişken	Ortalama	Medyan	Minimum	Maksimum	Standart Sapma
araç_değeri	15676,76	14400	1500	69740	8416,92
poliçe_sayısı	1,69	1	1	9	0,93
aynı_evde_yıl	8,30	8	-3	28	5,71

Çizelge 6.3 MASS veri setindeki sürekli değişkenlere dair özet istatistik

Değişken	Ortalama	Medyan	Minimum	Maksimum	Standart Sapma
motor_gücü	56,09	54	13	235	19,06
yaş	47,04	46	18	95	14,88

Çizelge 6.4 Ohlsson veri setindeki sürekli değişkenlere dair özet istatistik

Değişken	Ortalama	Medyan	Minimum	Maksimum	Standart Sapma
maruziyet	1,01	0,83	0	31,34	1,31
araç_yaşı	12,54	12	0	99	9,73
yaş	42,42	44	0	92	12,98

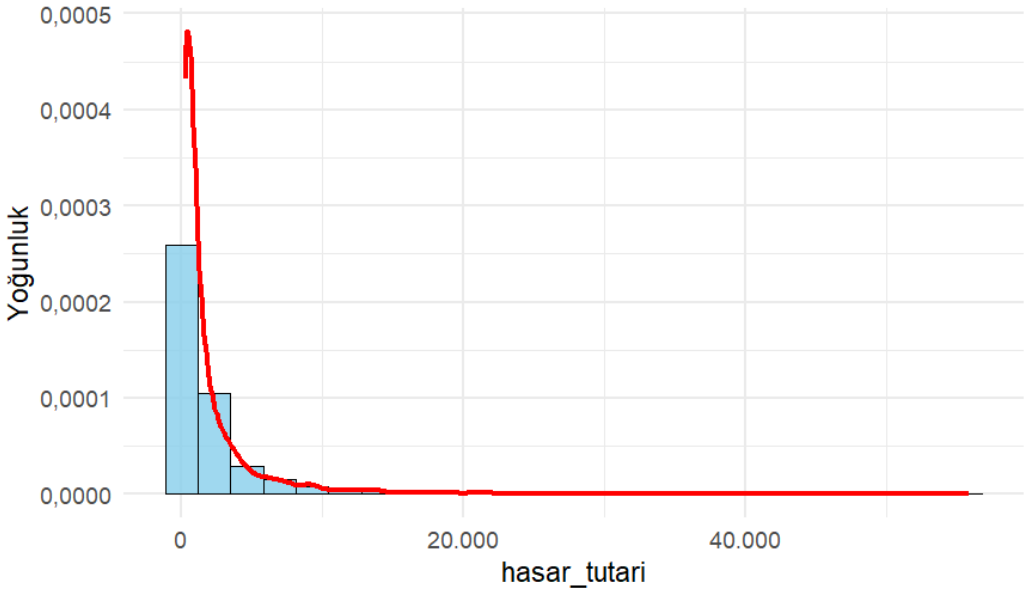
Elde edilen özet istatistikler, veri setlerinin dağılım özelliklerini genel hatlarıyla ortaya koymakta ve modelleme aşamasında kullanılacak değişkenlerin yapısal farklılıklarını değerlendirmek açısından temel bir referans oluşturmaktadır.

6.2 Bağımsız Değişkenlerin İncelenmesi

Bu bölümde, modelleme sürecinde kullanılan bağımsız değişkenlerin genel dağılım özellikleri ve veri setleri içindeki yapısal davranışları grafiksel olarak incelenmiştir. Değişkenlerin hem sürekli hem de kategorik türleri dikkate alınarak farklı grafik türlerinden yararlanılmıştır. Bu görseller, veri setlerinde yer alan değişkenlerin dağılım biçimlerini, sınıf dengesini ve değişkenlerin genel karakteristiklerini değerlendirmeye olanak sağlamaktadır. Grafikler, veri temizleme ve değişken seçimi aşamalarında belirlenen nihai değişken yapısı temel alınarak hazırlanmış; her bir veri setine ait görseller ilgili alt bölümlerde sunulmuştur.

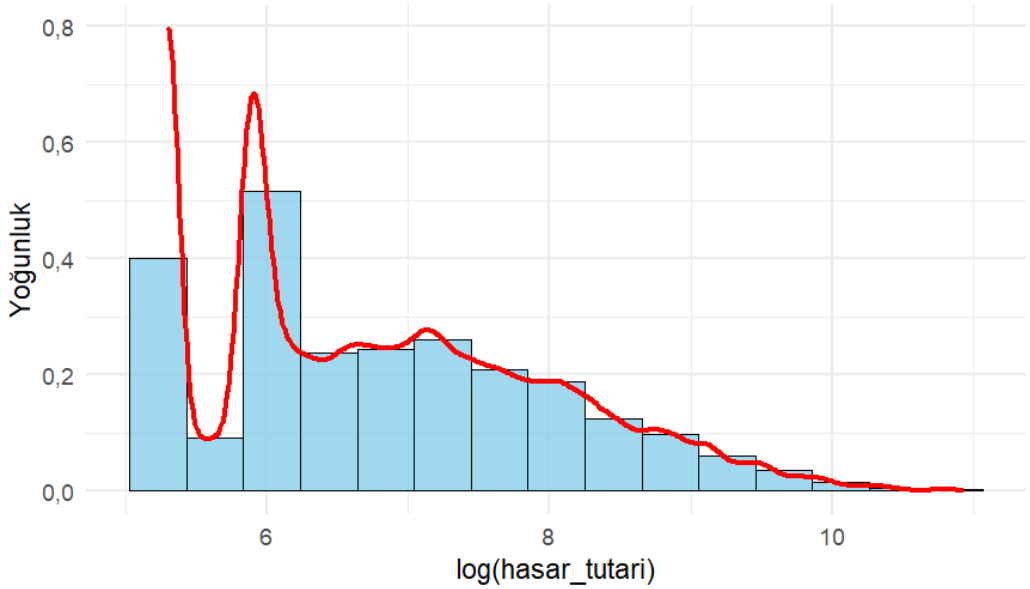
6.2.1 Otomobil veri seti

Bu alt bölümde, Otomobil Sigortası veri setinde yer alan bağımsız değişkenlerin genel dağılım özellikleri ve veri yapısı grafikler aracılığıyla incelenmiştir. Sürekli değişkenlerin yoğunluk ve histogram grafikleri, kategorik değişkenlerin ise frekans ve oran grafikleri sunularak veri setinin temel yapısı görsel olarak ortaya konulmuştur.



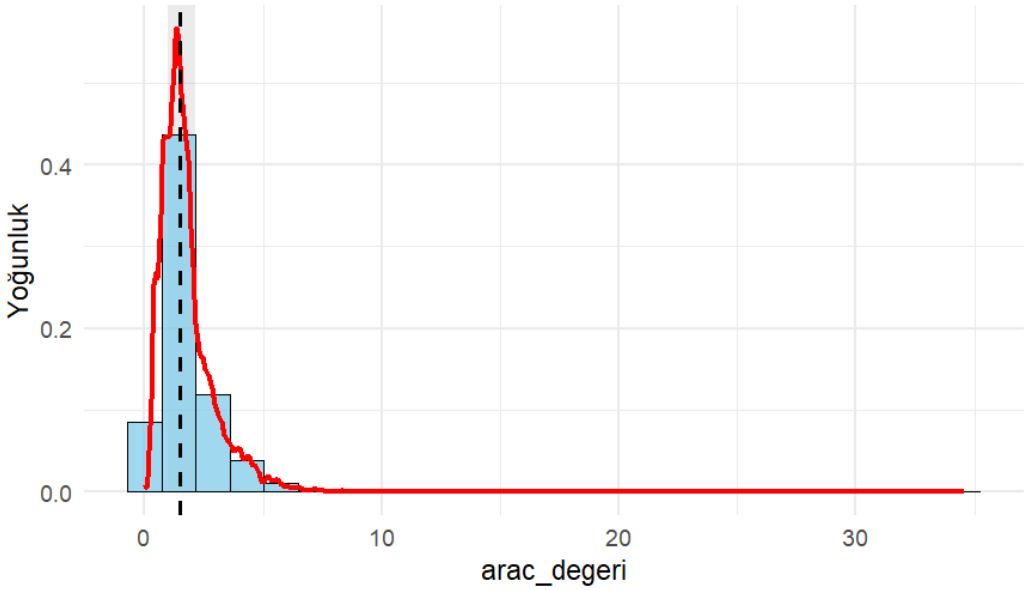
Şekil 6.1 Otomobil Sigortası veri setinde pozitif hasar tutarlarının dağılımı

Şekil 6.1'de Otomobil Sigortası veri setindeki pozitif hasar tutarlarının dağılımı gösterilmektedir. Grafik incelendiğinde, hasar tutarlarının büyük çoğunluğunun düşük değerlerde yoğunlaştığı, yüksek hasar tutarlarının ise oldukça seyrek gerçekleştiği görülmektedir. Dağılım belirgin biçimde sağa çarpık bir yapıdadır. Bu durum, sigorta hasar verilerinin tipik özelliği olan yüksek varyans ve ağır kuyruk davranışını yansıtmaktadır. Başka bir ifadeyle, poliçelerin büyük bir kısmında düşük tutarlı hasarlar meydana gelirken, az sayıda poliçede yüksek maliyetli hasarlar gözlenmektedir.



Şekil 6.2 Otomobil Sigortası veri setinde pozitif hasar tutarlarının logaritmik dönüşümlü dağılımı

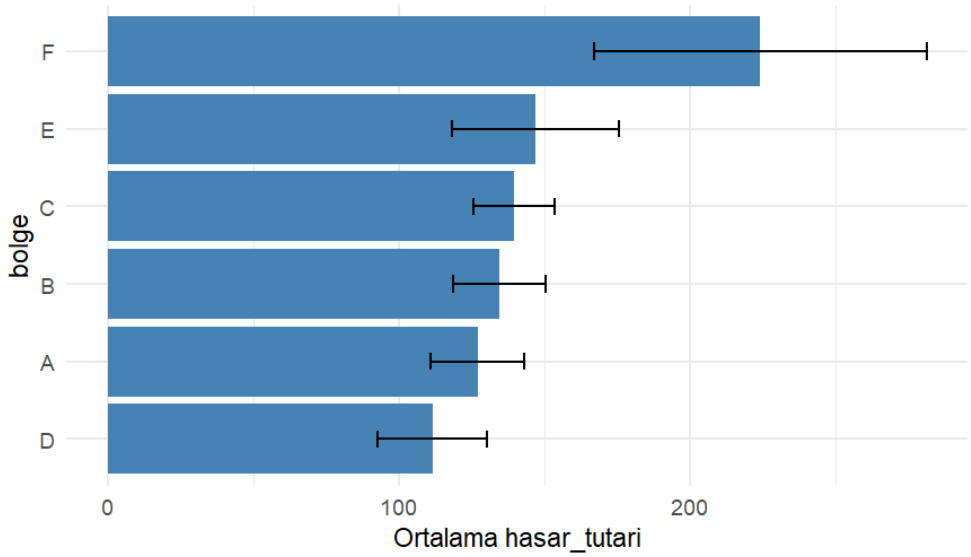
Şekil 6.2'de ise aynı değişkenin logaritmik dönüşüm uygulanmış biçimi yer almaktadır. Log dönüşümüyle veri ölçeği sıkıştırılmış ve aşırı sağ çarpıklık önemli ölçüde azalmıştır. Bu dönüşüm sonucunda yoğunluk eğrisi üzerinde birden fazla küçük tepe noktası ortaya çıkmıştır. Bu durum, farklı türde hasar gruplarının (örneğin küçük hasarlar, orta büyüklükte hasarlar ve nadir büyük hasarlar) varlığına işaret edebilir. Sonuç olarak, pozitif hasar tutarlarının ham dağılımı oldukça çarpık ve ağır kuyruklu bir yapı sergilerken, logaritmik dönüşüm bu çarpıklığı azaltmıştır.



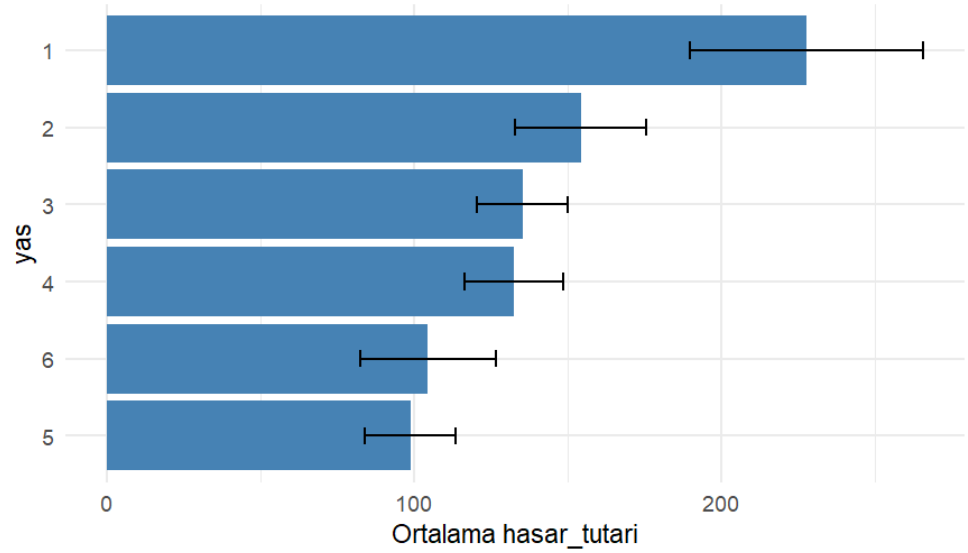
Şekil 6.3 Otomobil Sigortası veri setinde araç değeri değişkeninin dağılımı

Şekil 6.3'te Otomobil Sigortası veri setinde araç değerlerinin histogramı ve buna karşılık gelen yoğunluk eğrisi görülmektedir. Dağılımın belirgin biçimde sağa çarpık olduğu, düşük ve orta değerli araçların yoğunlaştığı, yüksek değerli araçların ise seyrek biçimde gözlendiği görülmektedir. Siyah kesikli çizgi medyan değeri göstermekte olup, medyanın dağılımın sol tarafına yakın konumlanması çarpıklığı desteklemektedir. Gri gölgeli alan birinci ve üçüncü çeyrekler (IQR) aralığını temsil ederek gözlemlerin büyük bölümünün bu bantta toplandığını göstermektedir.

Şekil 6.4'te Otomobil Sigortası veri setinde bölge değişkenine göre ortalama hasar tutarları ile bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafik incelendiğinde, F bölgesi diğer bölgelere kıyasla en yüksek ortalama hasar tutarına sahiptir. Ayrıca, bu bölgedeki güven aralığının genişliği, hasar tutarlarında daha yüksek bir değişkenliğe işaret etmektedir. D bölgesi ise hem ortalama değeri hem de güven aralığı bakımından en düşük seviyededir. Genel olarak bölgeler arasındaki farkların belirgin olması, coğrafi faktörlerin hasar maliyetleri üzerinde etkili olabileceğini düşündürmektedir.



Şekil 6.4 Otomobil Sigortası veri setinde bölgelere göre ortalama hasar tutarları ve %95 güven aralıkları

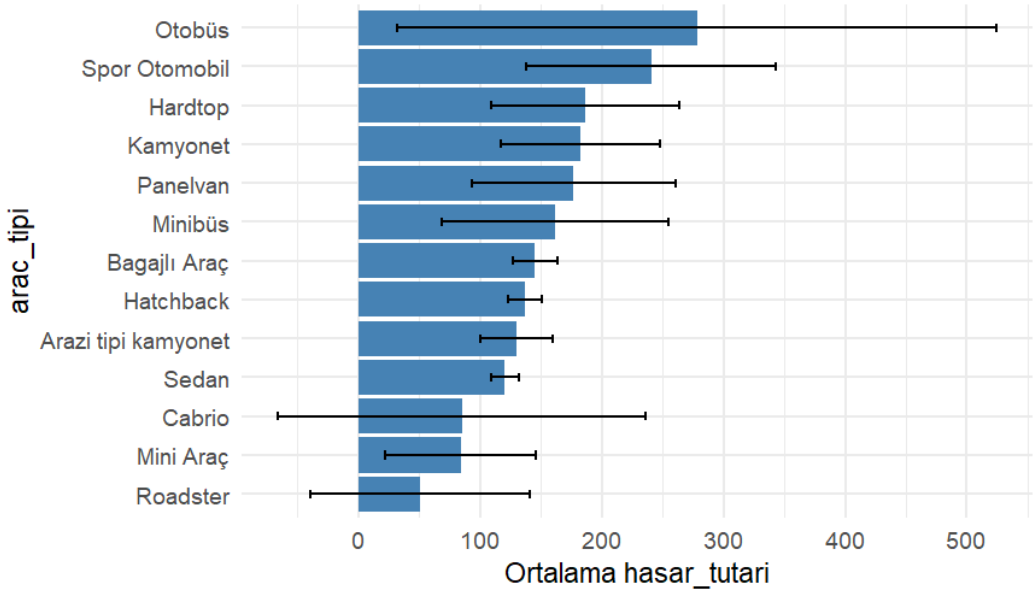


Şekil 6.5 Otomobil Sigortası veri setinde araç sahibinin yaşına göre ortalama hasar tutarları ve %95 güven aralıkları

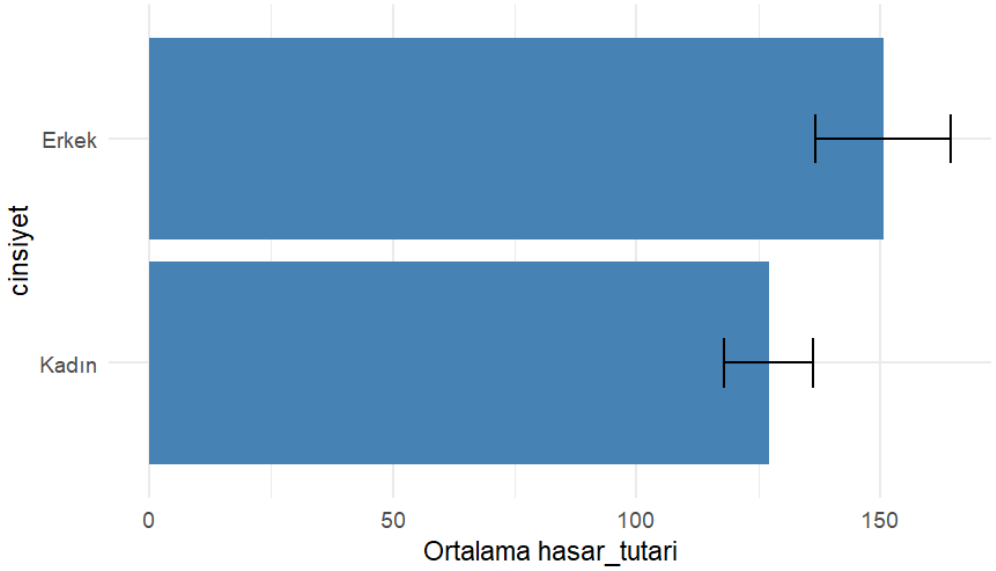
Şekil 6.5'te Otomobil Sigortası veri setinde araç sahibinin yaşı değişkenine göre ortalama hasar tutarları ve bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafik incelendiğinde, genç sürücü gruplarının (özellikle 1 ve 2. yaş kategorileri) daha yüksek ortalama hasar tutarlarına sahip olduğu, yaş ilerledikçe hasar ortalamalarının

belirgin biçimde azaldığı görülmektedir. Bu durum, genç sürücülerde risk düzeyinin ve dolayısıyla hasar maliyetlerinin daha yüksek olabileceğini göstermektedir.

Şekil 6.6'da Otomobil Sigortası veri setinde araç gövde tiplerine göre ortalama hasar tutarları ve bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafik incelendiğinde, otobüs ve spor otomobil gruplarının diğer araç tiplerine göre belirgin biçimde daha yüksek ortalama hasar tutarlarına sahip olduğu görülmektedir. Bu iki araç türünde güven aralıklarının da geniş olması, hasar tutarlarının daha değişken olduğunu göstermektedir. Buna karşın, mini araç, roadster ve sedan gruplarında ortalama hasar tutarlarının düşük olduğu ve güven aralıklarının dar seyrettiği gözlenmiştir. Bu durum, bu araç türlerinde hasar maliyetlerinin daha sınırlı bir aralıkta gerçekleştiğini ve daha öngörülebilir bir dağılım sergilediğini göstermektedir. Genel olarak araç tipi değişkeni, hasar tutarları üzerinde belirgin bir ayrıştırıcı etkiye sahiptir.



Şekil 6.6 Otomobil Sigortası veri setinde araç gövde tipine göre ortalama hasar tutarları ve %95 güven aralıkları

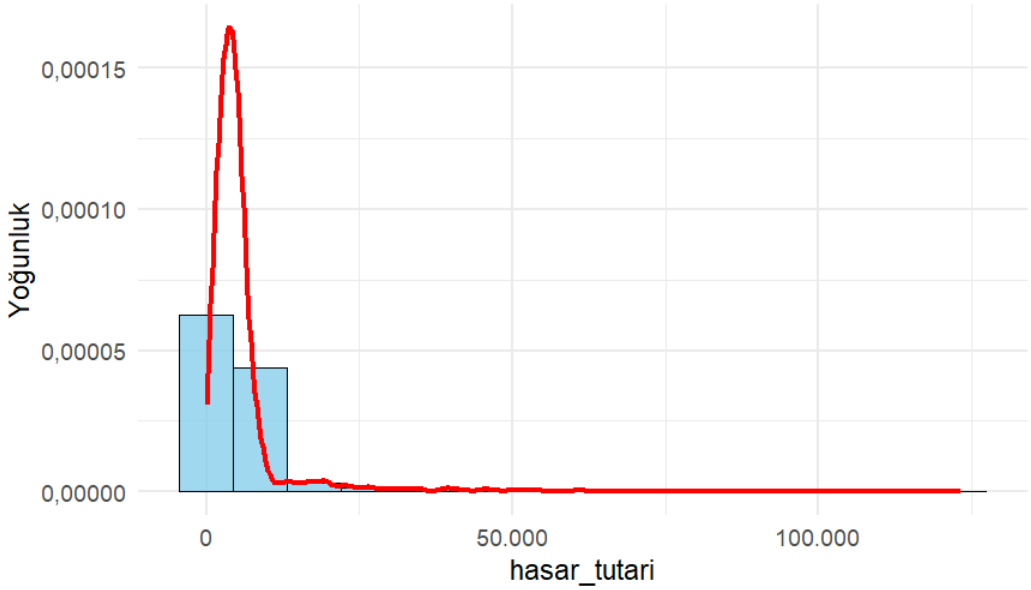


Şekil 6.7 Otomobil Sigortası veri setinde cinsiyet değişkenine göre ortalama hasar tutarları ve %95 güven aralıkları

Şekil 6.7'de Otomobil Sigortası veri setinde cinsiyet değişkenine göre ortalama hasar tutarları ve bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafikten görüldüğü üzere, erkek sürücülerin ortalama hasar tutarları kadın sürücülere göre bir miktar daha yüksektir. Bununla birlikte, iki grup arasındaki farkın mutlak değeri sınırlı olup, güven aralıklarının kısmen örtüştüğü gözlenmektedir. Bu durum, cinsiyet değişkeninin hasar tutarları üzerinde belirgin ancak istatistiksel olarak sınırlı bir etkiye sahip olabileceğini göstermektedir.

6.2.2 Araç sigortası hasarı veri seti

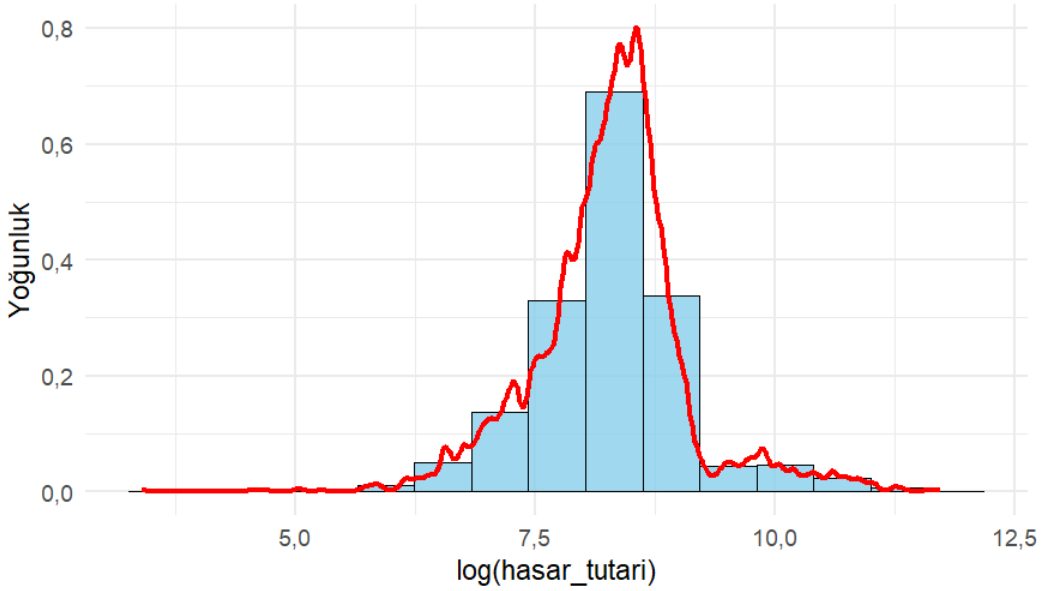
Bu kısımda, Araç Sigorta Hasarı veri setindeki değişkenlerin dağılım biçimleri ve veri yapısı görsel olarak değerlendirilmiştir. Grafikler, veri setinde yer alan sürekli ve kategorik değişkenlerin genel özelliklerini yansıtmakta ve değişkenlerin sınıf dengeleri ile olası uç değerleri göstermektedir.



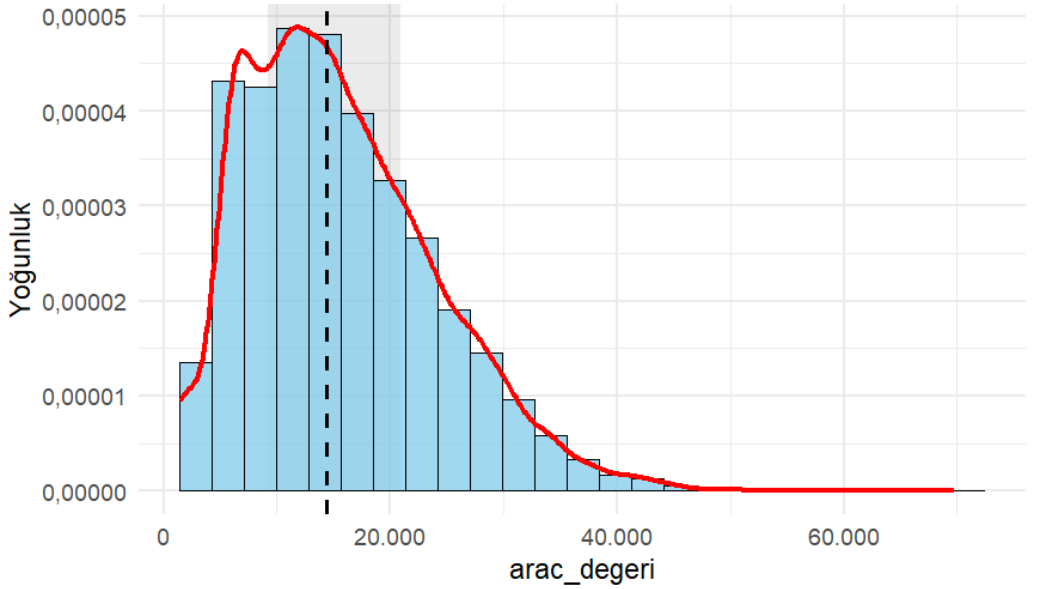
Şekil 6.8 Araç Sigorta Hasarı veri setinde pozitif hasar tutarlarının dağılımı

Şekil 6.8'de Araç Sigorta Hasarı veri setinde pozitif hasar tutarlarının dağılımı görülmektedir. Grafik incelendiğinde, hasar tutarlarının büyük bir kısmının düşük değerlerde yoğunlaştığı, yüksek tutarlı hasarların ise oldukça az sayıda gerçekleştiği dikkat çekmektedir. Dağılımın belirgin biçimde sağa çarpık olması, sigorta hasar verilerinin kendine özgü ağır kuyruklu yapısını ortaya koymaktadır. Diğer bir ifadeyle, çoğu poliçede küçük tutarlı hasarlar gözlemlenirken, yalnızca sınırlı sayıda poliçede yüksek maliyetli hasarlar meydana gelmektedir.

Şekil 6.9'da ise aynı değişkenin logaritmik dönüşüm uygulanmış hali yer almaktadır. Log dönüşümü, veri setindeki aşırı sağ çarpıklığı azaltmıştır. Bu dönüşüm sonucunda yoğunluk eğrisi, ortalama etrafında daha dengeli bir görünüm kazanmış ve uç değerlerin etkisi azalmıştır. Sonuç olarak, ham hasar tutarları değişkeni sağa çarpık ve ağır kuyruklu bir dağılım sergilerken, logaritmik dönüşüm uygulanmasıyla veri dağılımı daha normal bir forma yaklaşmıştır.



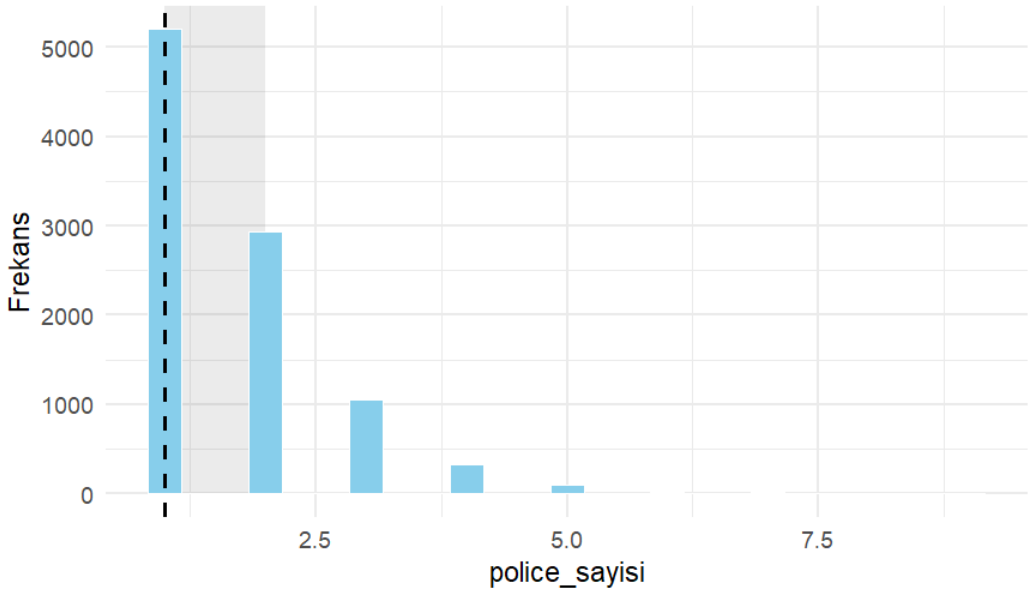
Şekil 6.9 Araç Sigorta Hasarı veri setinde pozitif hasar tutarlarının logaritmik dönüşümlü dağılımı



Şekil 6.10 Araç Sigorta Hasarı veri setinde araç değerinin histogram grafiği

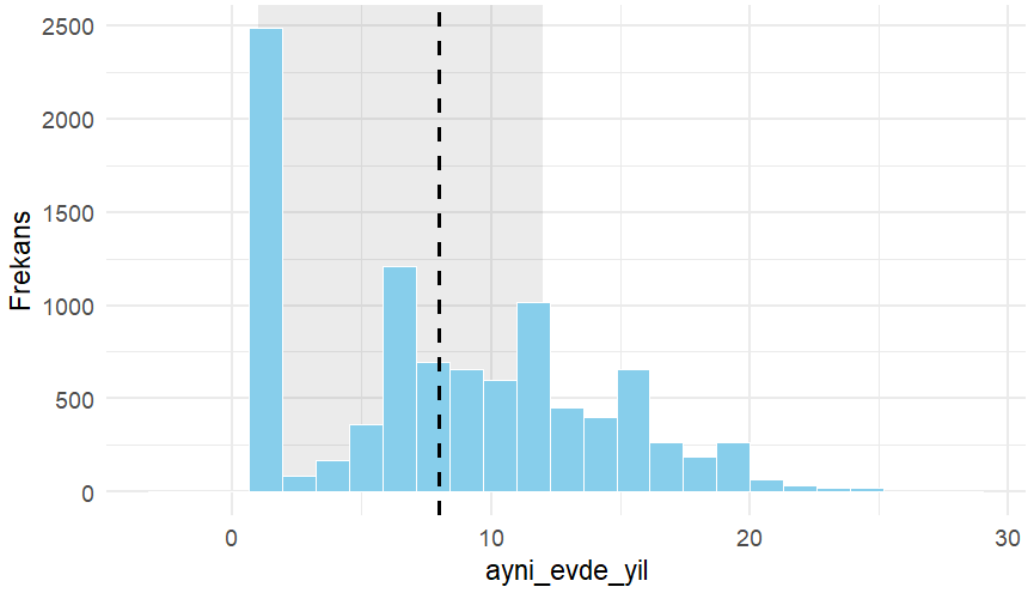
Şekil 6.10'da Araç Sigorta Hasarı veri setinde araç değerinin histogramı ve buna karşılık gelen yoğunluk eğrisi gösterilmektedir. Grafik incelendiğinde, dağılımın sağa çarpık olduğu, düşük ve orta değerli araçların yoğun biçimde gözlendiği, yüksek değerli araçların ise nadir olarak dağılımın kuyruğunda yer aldığı görülmektedir. Siyah kesikli çizgi medyan değeri göstermekte olup, bu konum çarpıklık yapısı ile uyumludur. Gri

gölgelendirilmiş bölge birinci ve üçüncü çeyrekler aralığını (IQR) ifade ederek gözlemlerin büyük bölümünün bu bantta toplandığını göstermektedir. Yoğunluk eğrisinin uzun bir sağ kuyruk göstermesi, veri setinde uç (yüksek) araç değerlerinin az sayıda fakat etkili biçimde bulunduğuna işaret etmektedir.



Şekil 6.11 Araç Sigorta Hasarı veri setinde poliçe sayısının histogram grafiği

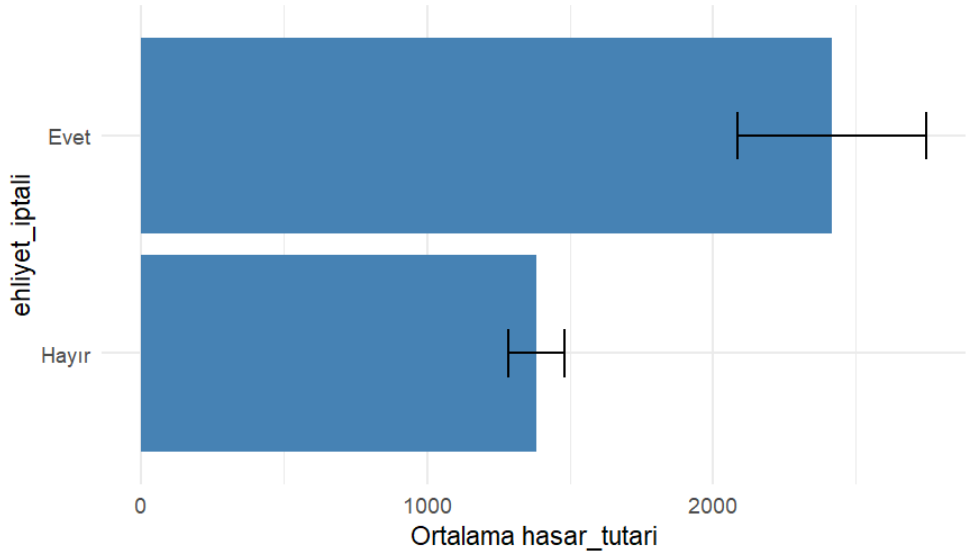
Şekil 6.11'de Araç Sigorta Hasarı veri setinde poliçe sayısının histogram grafiği gösterilmektedir. Dağılımın büyük ölçüde 1 poliçe seviyesinde yoğunlaştığı, daha yüksek poliçe sayılarına doğru gidildikçe gözlem frekansının hızla azaldığı görülmektedir. Siyah kesikli dikey çizgi medyan değeri belirtmekte olup, medyanın 1 değeri üzerine denk gelmesi dağılımın ağırlık merkezinin düşük değerlerde toplandığını doğrulamaktadır. Gri gölgelendirilmiş alan birinci ve üçüncü çeyrekler (IQR) aralığını temsil ederek, gözlemlerin çok büyük kısmının düşük poliçe sayılarında gerçekleştiğini göstermektedir.



Şekil 6.12 Araç Sigorta Hasarı veri setinde aynı evde geçirilen yıl sayısının histogram grafiği

Şekil 6.12'de Araç Sigorta Hasarı veri setinde aynı evde geçirilen yıl sayısının histogram grafiği gösterilmektedir. Gözlemlerin önemli bir bölümünün kısa süreli ikametlerde toplandığı, ikamet süresi arttıkça frekansın kademeli olarak azaldığı izlenmektedir. Siyah kesikli çizgi medyan değeri belirtmekte olup medyanın düşük bir değer civarında yer alması dağılımın ağırlığının kısa ikamet sürelerinde olduğunu doğrulamaktadır. Gri gölgelendirilmiş bölge birinci ve üçüncü çeyrekler (IQR) aralığını ifade ederek gözlemlerin büyük kısmının bu aralıkta kümelendiğini göstermektedir.

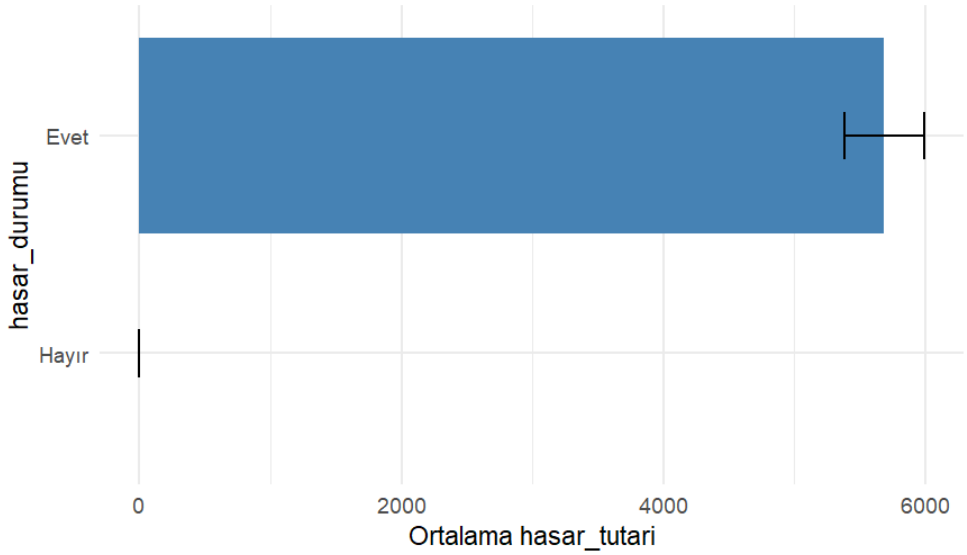
Şekil 6.13'te, sigortalıların ehliyet iptali durumuna göre ortalama hasar tutarları ve bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafik incelendiğinde, ehliyeti iptal edilmiş sürücülerin ortalama hasar tutarlarının ehliyeti iptal edilmemiş sürücülere kıyasla oldukça yüksek olduğu görülmektedir. Bu durum, geçmişte ehliyet iptali yaşamış sürücülerin daha yüksek risk profiline sahip olabileceğini ve dolayısıyla daha yüksek hasar maliyetleriyle ilişkilendirilebileceğini göstermektedir.



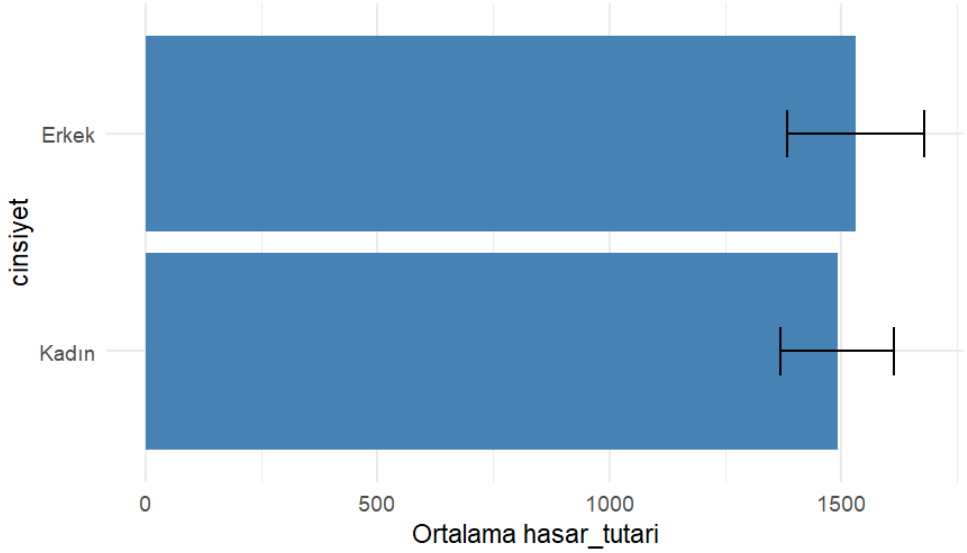
Şekil 6.13 Araç Sigorta Hasarı veri setinde ehliyet iptaline göre ortalama hasar tutarları ve %95 güven aralıkları

Ayrıca ehliyeti iptal edilen grup için güven aralığının geniş olması, bu gruptaki hasar tutarlarının değişkenliğinin daha fazla olduğunu göstermektedir. Buna karşın, ehliyeti iptal edilmemiş sürücülerde ortalama hasar tutarları düşük olup, güven aralığının dar olması bu grubun daha homojen bir risk yapısına sahip olduğunu düşündürmektedir.

Şekil 6.14'te Araç Sigorta Hasarı veri setinde hasar durumuna göre ortalama hasar tutarları ve bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafik incelendiğinde, hasar gerçekleşen poliçelerde ("Evet") ortalama hasar tutarının oldukça yüksek olduğu, buna karşın hasar gerçekleşmeyen poliçelerde ("Hayır") ortalama değerin sıfıra yakın olduğu görülmektedir. Bu durum, veri setinin sıfır yığılmalı yapısını açık biçimde yansıtmaktadır. Gözlemlerin büyük bir kısmında hasar oluşmamakta, ancak hasar gerçekleştiğinde tutarlar yüksek seviyelere ulaşmaktadır. Ayrıca, hasar oluşan grup için güven aralığının geniş olması, bu gruptaki hasar tutarlarının yüksek varyansa ve ağır kuyruklu bir dağılıma sahip olduğunu göstermektedir.



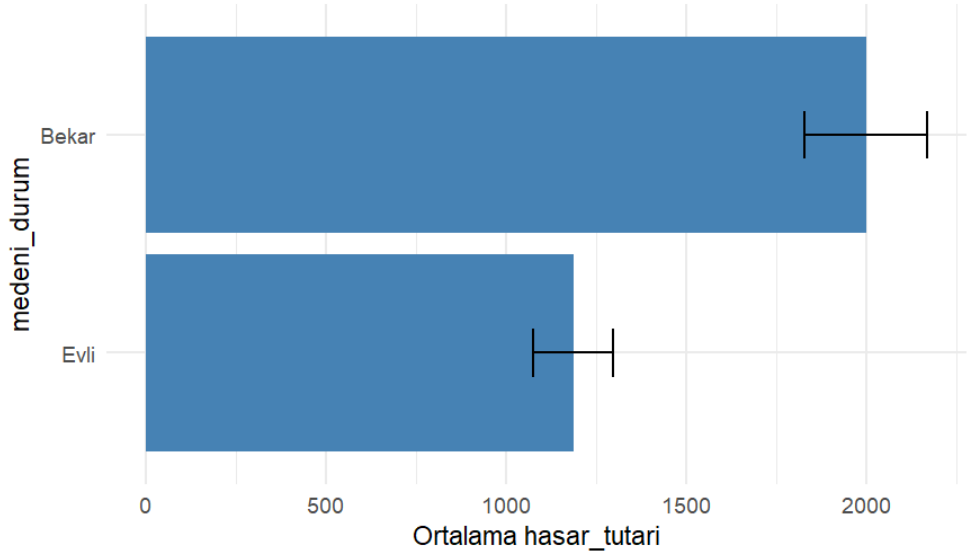
Şekil 6.14 Araç Sigorta Hasarı veri setinde hasar durumuna göre ortalama hasar tutarları ve %95 güven aralıkları



Şekil 6.15 Araç Sigorta Hasarı veri setinde cinsiyete göre ortalama hasar tutarları ve %95 güven aralıkları

Şekil 6.15'te Araç Sigorta Hasarı veri setinde cinsiyete göre ortalama hasar tutarları ve bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafiğe göre, erkek sürücülerin ortalama hasar tutarları kadın sürücülerden biraz daha yüksektir. Bununla birlikte, iki grup arasındaki fark görece sınırlıdır ve güven aralıkları büyük ölçüde

çakışmaktadır. Bu durum, cinsiyet değişkeninin hasar tutarı üzerindeki etkisinin zayıf veya istatistiksel olarak anlamlı olmayabileceğini düşündürmektedir.

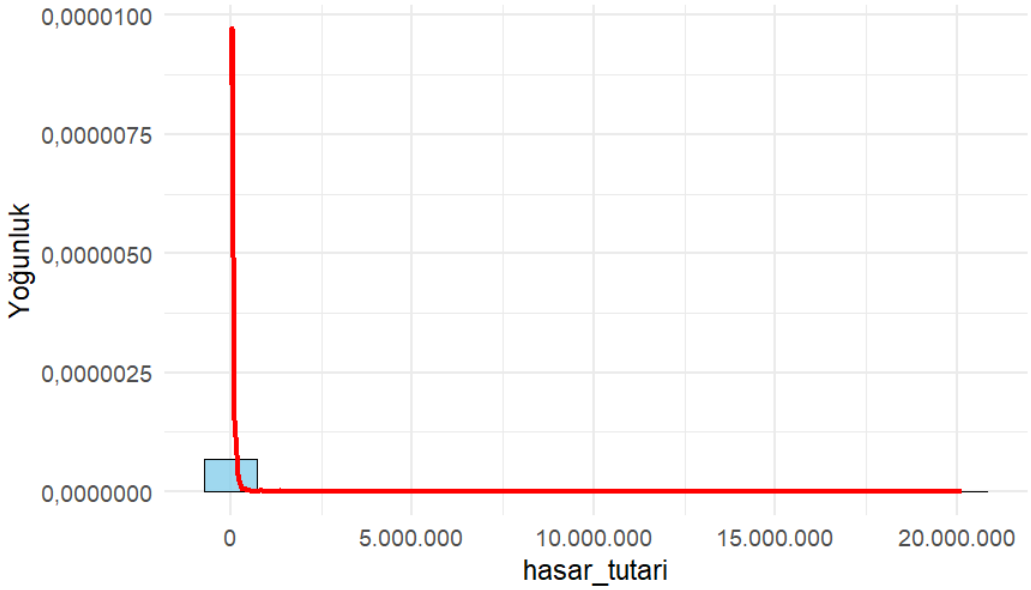


Şekil 6.16 Araç Sigorta Hasarı veri setinde medeni duruma göre ortalama hasar tutarları ve %95 güven aralıkları

Şekil 6.16'da Araç Sigorta Hasarı veri setinde medeni duruma göre ortalama hasar tutarları ve bu ortalamalara ilişkin %95 güven aralıkları sunulmaktadır. Grafik incelendiğinde, bekar sigortalıların ortalama hasar tutarlarının evli sigortalılara kıyasla belirgin şekilde daha yüksek olduğu görülmektedir. Bu sonuç, bekar sürücülerin risk alma eğilimlerinin veya sürüş davranışlarının daha yüksek hasar maliyetleriyle ilişkili olabileceğini düşündürmektedir. Evli sigortalılarda ortalama hasar tutarının daha düşük olması, bu grubun daha temkinli sürüş alışkanlıklarına sahip olabileceğini ya da sigorta kapsamı açısından daha istikrarlı bir profil sergilediğini göstermektedir. Ayrıca, bekar gruba ait güven aralıklarının görece geniş olması, bu grubun hasar tutarlarında daha yüksek bir varyans bulunduğunu, dolayısıyla risk dağılımının daha heterojen olduğunu ortaya koymaktadır. Bu bulgular, medeni durum değişkeninin sigorta risk profili üzerinde açıklayıcı bir değişken olabileceğini desteklemektedir.

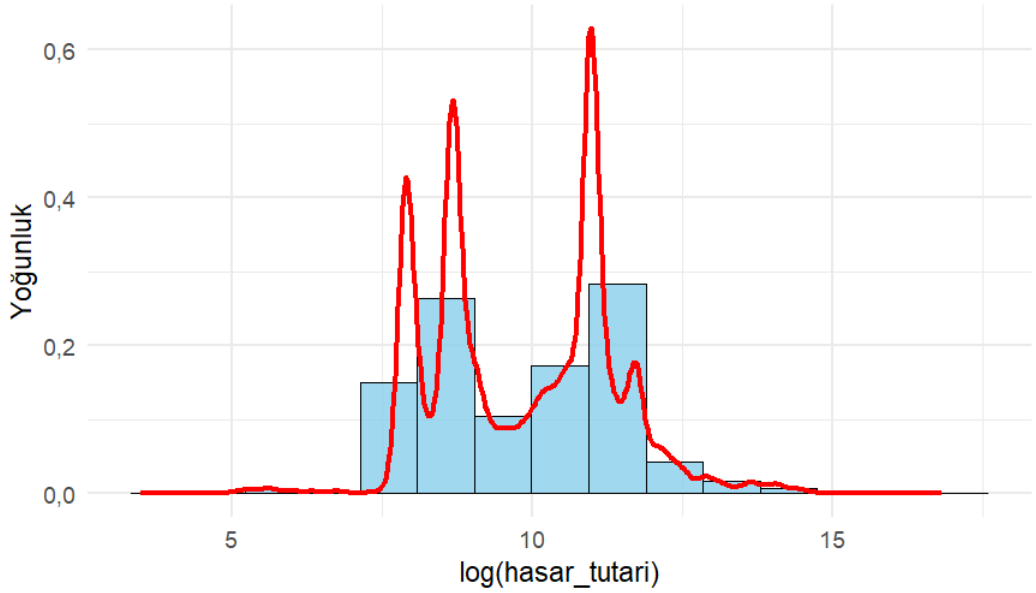
6.2.3 MASS veri seti

Bu bölümde, MASS veri setine ait bağımsız değişkenlerin dağılım yapıları incelenmiştir. Sürekli değişkenler için histogram ve yoğunluk grafikleri, kategorik değişkenler içinse sütun grafikleri kullanılmıştır. Bu görseller, veri setinin genel yapısını ve değişkenlerin istatistiksel özelliklerini özetlemektedir.



Şekil 6.17 MASS veri setinde pozitif hasar tutarlarının dağılımı

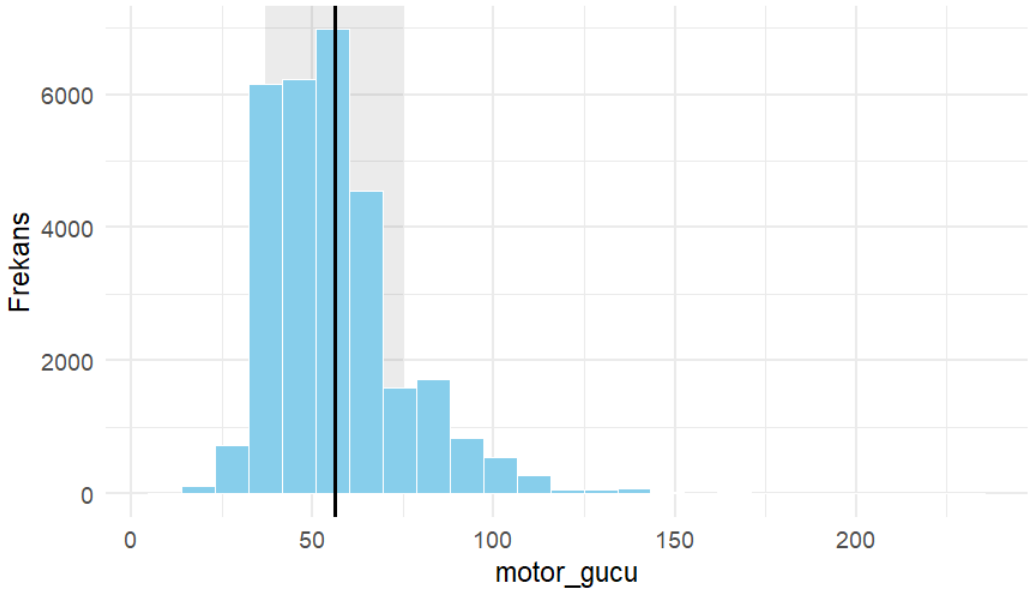
Şekil 6.17'de MASS veri setindeki pozitif hasar tutarlarının dağılımı gösterilmektedir. Grafik incelendiğinde, hasar tutarlarının büyük çoğunluğunun düşük değerlerde yoğunlaştığı, yüksek tutarlı hasarların ise oldukça seyrek gerçekleştiği görülmektedir. Dağılım belirgin biçimde sağa çarpık bir yapı sergilemektedir. Bu durum, sigorta hasar verilerinin tipik özelliği olan ağır kuyruklu dağılım biçimini yansıtmaktadır.



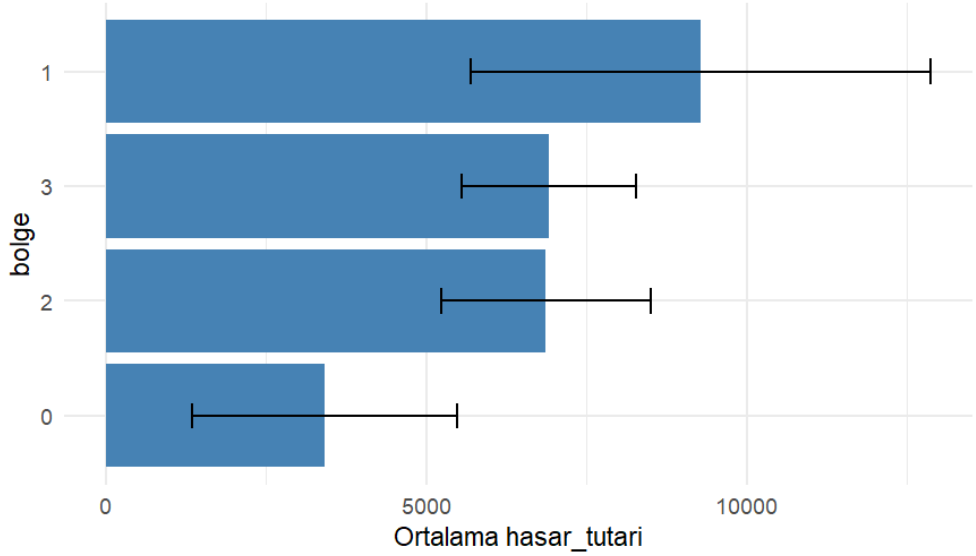
Şekil 6.18 MASS veri setinde pozitif hasar tutarlarının logaritmik dönüşümlü dağılımı

Şekil 6.18'de ise aynı değişkenin logaritmik dönüşüm uygulanmış hali sunulmuştur. Log dönüşümü sonucunda dağılımın çarpıklığı belirgin biçimde azalmış, veri daha dengeli bir görünüme kavuşmuştur. Ayrıca, grafikte birden fazla tepe noktası gözlemlenmektedir. Bu durum, veri setinde birden fazla hasar tipi veya farklı büyüklükteki olay gruplarının varlığına işaret etmektedir. MASS veri setinde yer alan pozitif hasar tutarları, ham biçimleriyle belirgin bir sağa çarpıklık ve ağır kuyruk özelliği gösterirken; logaritmik dönüşüm uygulandığında, dağılımın istatistiksel açıdan daha uygun ve dengeli bir forma yaklaştığı gözlemlenmektedir.

Şekil 6.19'da MASS veri setinde araç motor gücünün histogram grafiği gösterilmektedir. Grafik incelendiğinde, dağılımın orta değerlerin çevresinde yoğunlaştığı ve sağ tarafa doğru giderek azalan bir seyir izlediği görülmektedir. Dikey çizgi ortalama motor gücünü temsil etmekte olup gözlemlerin önemli bir kısmının bu değerin yakınlarında toplandığı izlenmektedir. Dağılımın ortalamadan belirli bir sapma içinde yoğunlaşması, gözlemlerin büyük bölümünün benzer güç aralığında kümelendiğini, ancak daha yüksek motor gücüne sahip araçların nispeten seyrek olduğunu göstermektedir. Sağ kuyruk yapısı, yüksek motor gücüne sahip araçların istisnai nitelikte olduğunu işaret etmektedir.



Şekil 6.19 MASS veri setinde araç motor gücünün histogram grafiği



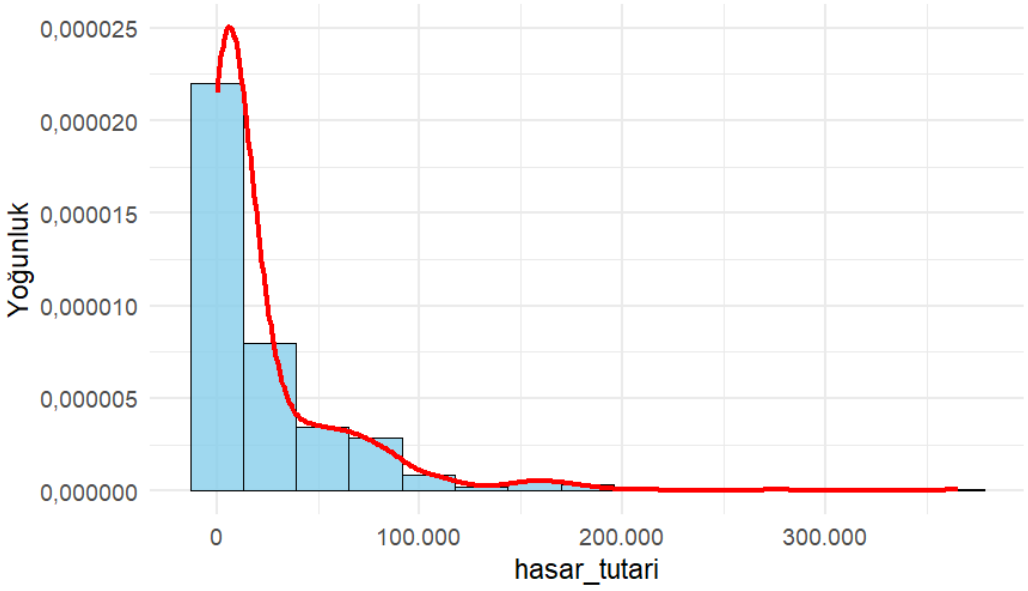
Şekil 6.20 MASS veri setinde bölgeye göre ortalama hasar tutarları ve %95 güven aralıkları

Şekil 6.20'de MASS veri setinde bölge değişkenine göre ortalama hasar tutarları ve bu ortalamalara ilişkin %95 güven aralıkları sunulmaktadır. Grafik incelendiğinde, 1. bölgedeki sigortalıların ortalama hasar tutarlarının diğer bölgelere göre belirgin biçimde daha yüksek olduğu görülmektedir. Bu durum, 1. bölgedeki sürüş koşulları, araç yoğunluğu veya sosyoekonomik faktörlerin hasar maliyetlerini artırabileceğini

düşündürmektedir. Diğer bölgelerde ise ortalama hasar tutarlarının birbirine daha yakın olduğu ve 0. bölgenin en düşük ortalama hasara sahip olduğu dikkat çekmektedir. Ayrıca, 1. bölgeye ait güven aralıklarının geniş olması, bu bölgede hasar tutarlarının yüksek değişkenlik gösterdiğini, dolayısıyla risk seviyesinin daha heterojen olduğunu göstermektedir.

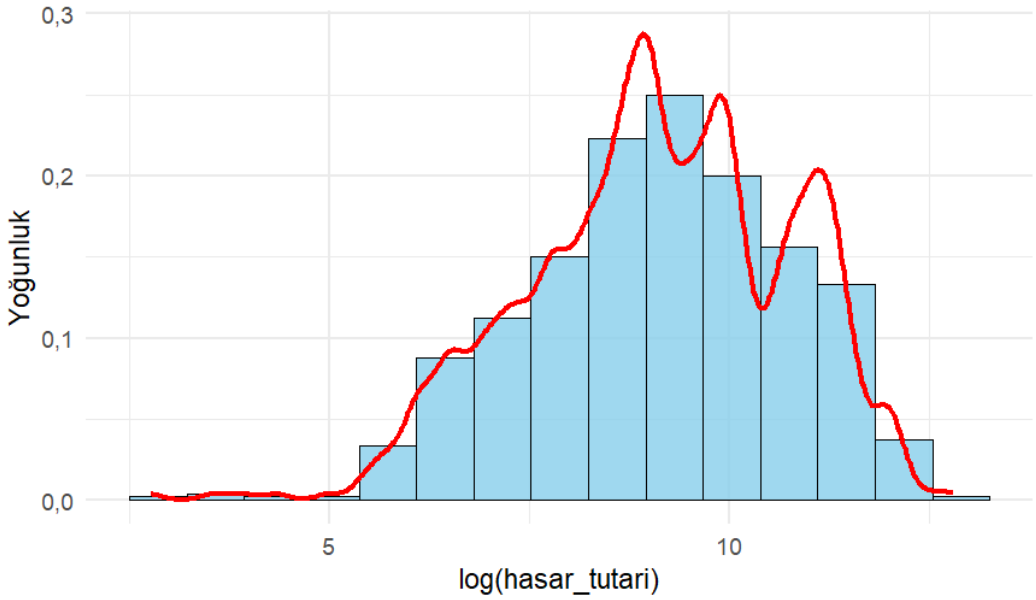
6.2.4 Ohlsson veri seti

Bu alt bölümde, Ohlsson veri setinde yer alan değişkenlerin genel dağılım özellikleri görsel olarak sunulmuştur. Grafikler, veri setinin yapısal karakteristiklerini ve değişkenlerin genel davranışlarını ortaya koymakta; veri setinin modelleme sürecine temel oluşturan yapısını desteklemektedir.



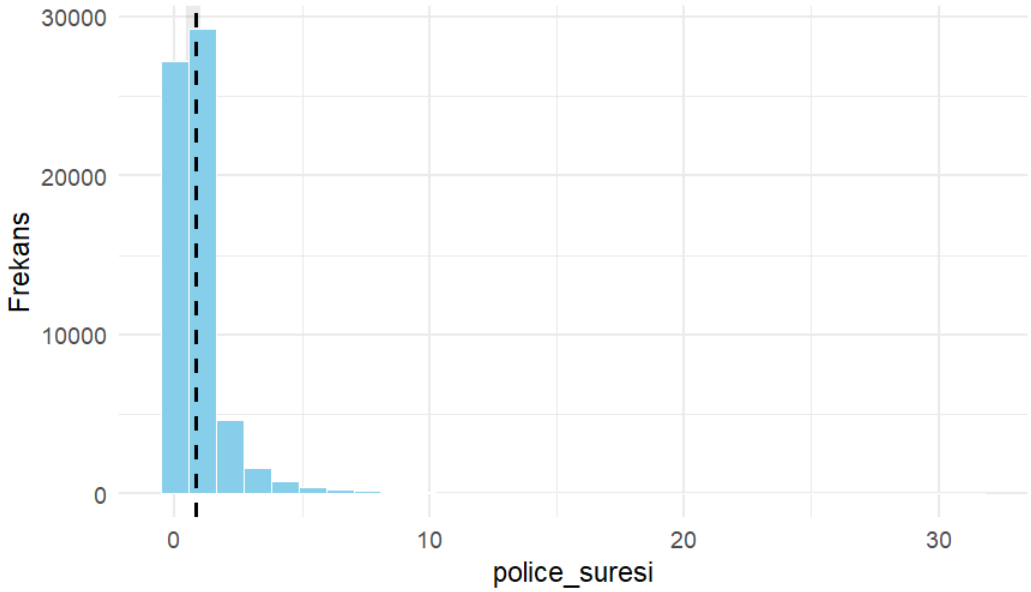
Şekil 6.21 Ohlsson veri setinde pozitif hasar tutarlarının dağılımı

Şekil 6.21'de Ohlsson veri setindeki pozitif hasar tutarlarının dağılımı gösterilmektedir. Grafik incelendiğinde, hasar tutarlarının büyük çoğunluğunun düşük değerlerde yoğunlaştığı ve yüksek tutarlı hasarların oldukça seyrek olduğu görülmektedir. Dağılım açık biçimde sağa çarpık bir özellik göstermektedir. Bu durum, sigorta hasar verilerinde karakteristik olarak görülen ağır kuyruklu dağılım yapısının bir göstergesidir.



Şekil 6.22 Ohlsson veri setinde pozitif hasar tutarlarının logaritmik dönüşümlü dağılımı

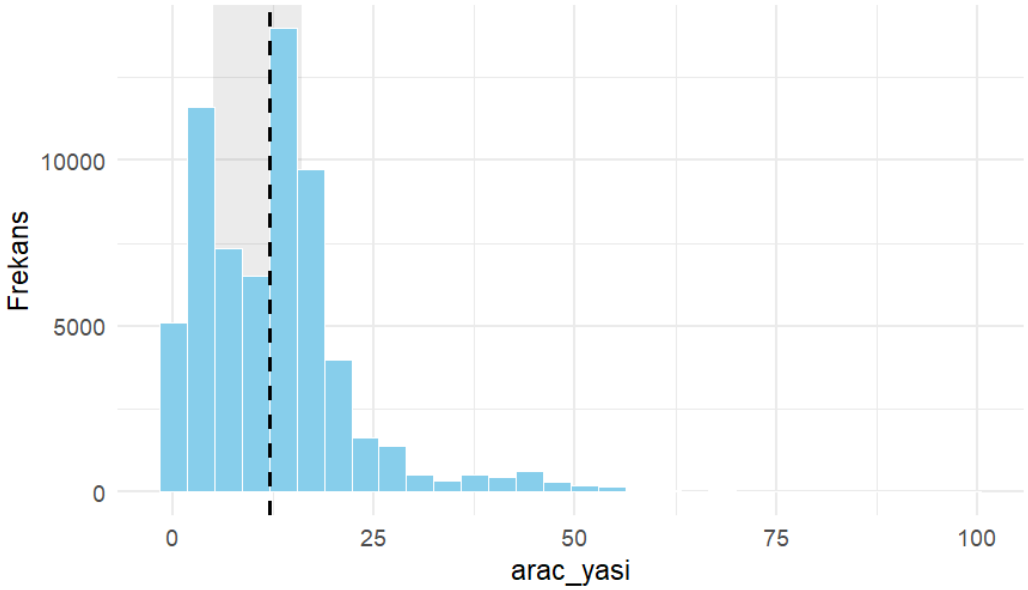
Şekil 6.22'de ise aynı değişkenin logaritmik dönüşüm uygulanmış hali gösterilmektedir. Log dönüşümü sonucunda dağılımın çarpıklığı azalmıştır. Ayrıca yoğunluk eğrisi üzerinde birden fazla tepe noktası dikkat çekmektedir; bu durum farklı büyüklükteki hasar gruplarının veya çeşitli hasar türlerinin veri içinde yer aldığını göstermektedir.Pozitif hasar tutarları, dönüştürülmemiş halleriyle belirgin biçimde sağa çarpık ve ağır kuyruklu bir dağılım gösterirken; logaritmik dönüşüm sonrasında dağılımın daha dengeli ve istatistiksel olarak uygun bir şekle yaklaştığı görülmektedir.



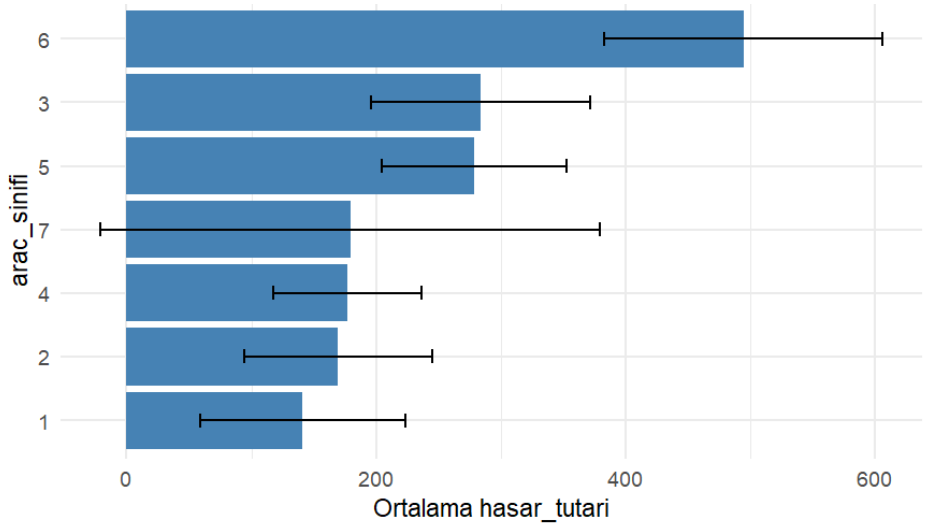
Şekil 6.23 Ohlsson veri setinde poliçe süresinin histogram grafiği

Şekil 6.23'te Ohlsson veri setinde poliçe süresinin histogram grafiği gösterilmektedir. Gözlemlerin çok büyük kısmının kısa süreli poliçelerde yoğunlaştığı, süre arttıkça frekansın hızla azaldığı izlenmektedir. Siyah kesikli çizgi medyan değeri temsil etmekte olup medyanın düşük bir değer civarında yer alması dağılımın ağırlıklı olarak kısa poliçe sürelerinde toplandığını doğrulamaktadır. Gri gölgelendirilmiş bölge birinci ve üçüncü çeyrekler (IQR) aralığını ifade ederek gözlemlerin çoğunluğunun bu aralıkta kümelendiğini göstermektedir. Genel olarak dağılım sağa çarpık olup, uzun süreli poliçeler istisnai niteliktedir.

Şekil 6.24'te Ohlsson veri setinde araç yaşının histogram grafiği sunulmaktadır. Gözlemlerin büyük kısmının düşük ve orta yaşlı araçlarda yoğunlaştığı, araç yaşı arttıkça frekansın hızla azaldığı izlenmektedir. Siyah kesikli dikey çizgi medyan değeri temsil etmekte olup medyanın dağılımın sol tarafında konumlanması, veri setinde görece genç araçların ağırlıkta olduğunu göstermektedir. Gri gölgeli alan birinci ve üçüncü çeyrekler (IQR) aralığını ifade ederek gözlemlerin çoğunluğunun bu yaş aralığında toplandığını ortaya koymaktadır. Genel olarak dağılım sağa çarpık yapı sergilemekte olup ileri yaşlı araçlar az sayıda ve dağılımın kuyruğunda yer almaktadır.

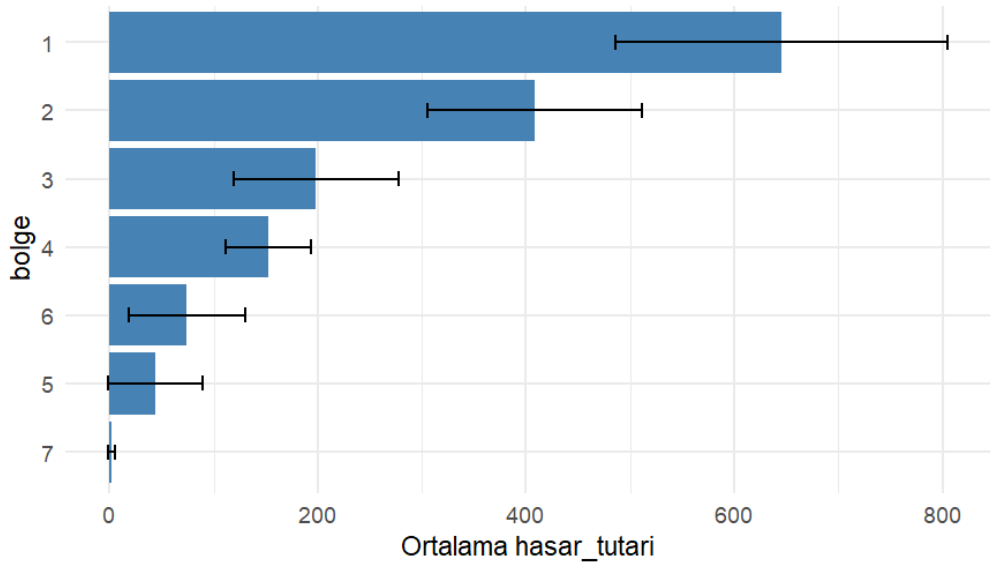


Şekil 6.24 Ohlsson veri setinde araç yaşının histogram grafiği



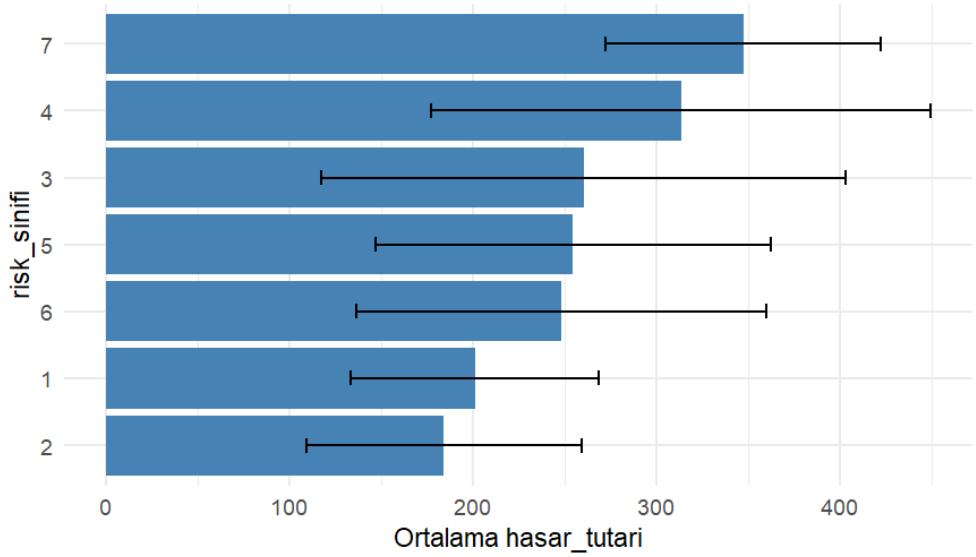
Şekil 6.25 Ohlsson veri setinde araç sınıfına göre ortalama hasar tutarları ve %95 güven aralıkları

Şekil 6.25'de Ohlsson veri setinde araç sınıfına göre ortalama hasar tutarları ve bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafik incelendiğinde, 6. sınıf araçların ortalama hasar tutarının diğer sınıflara göre belirgin biçimde daha yüksek olduğu dikkat çekmektedir. 1. ve 2. sınıf araçlar ise görece düşük ortalama hasar düzeyine sahiptir. Ayrıca, bazı sınıflarda (örneğin 7. sınıf) güven aralıklarının geniş olması, bu gruplarda hasar tutarlarının yüksek değişkenlik gösterdiğini ve risk seviyesinin heterojen bir yapıya sahip olduğunu göstermektedir.



Şekil 6.26 Ohlsson veri setinde bölgeye göre ortalama hasar tutarları ve %95 güven aralıkları

Şekil 6.26'da Ohlsson veri setinde bölge değişkenine göre ortalama hasar tutarları ve bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafik incelendiğinde, 1. bölgedeki sigortalıların ortalama hasar tutarlarının diğer bölgelere göre en yüksek düzeyde olduğu görülmektedir. Bu durum, 1. bölgede yer alan sigortalıların daha yoğun trafik koşullarına, yüksek araç değerlerine veya farklı sosyoekonomik özelliklere sahip olabileceğini düşündürmektedir. Buna karşılık, 5., 6. ve 7. bölgelerde ortalama hasar tutarları oldukça düşük olup, bu bölgelerin daha düşük risk profiline sahip olabileceği anlaşılmaktadır. Ayrıca, 1. ve 2. bölgelere ait güven aralıklarının geniş olması, bu bölgelerdeki hasar tutarlarının değişkenliğinin yüksek olduğunu, dolayısıyla risk dağılımının heterojen bir yapıda olduğunu göstermektedir.



Şekil 6.27 Ohlsson veri setinde risk sınıfına göre ortalama hasar tutarları ve %95 güven aralıkları

Şekil 6.27'de Ohlsson veri setinde risk sınıfına göre ortalama hasar tutarları ve bu ortalamalara ait %95 güven aralıkları gösterilmektedir. Grafik incelendiğinde, yüksek risk sınıflarına (özellikle 7. sınıf) ait ortalama hasar tutarlarının daha düşük risk sınıflarına göre belirgin biçimde yüksek olduğu görülmektedir. Bu durum, risk sınıflandırmasının sigortalıların hasar olasılığı ve maliyetleriyle tutarlı bir şekilde ilerlediğini göstermektedir. Düşük risk sınıflarındaki sigortalıların (örneğin 1. ve 2. sınıf) ortalama hasar tutarlarının daha düşük olması, bu grupların daha güvenli sürüş davranışlarına sahip olduğunu veya geçmiş hasar geçmişlerinin daha temiz olduğunu düşündürmektedir. Ayrıca, bazı yüksek risk sınıflarında (örneğin 4. ve 7. sınıflar) güven aralıklarının geniş olması, bu gruplarda hasar tutarlarının yüksek varyansa sahip olduğunu, dolayısıyla riskin daha değişken bir yapıda bulunduğunu ortaya koymaktadır.

6.3 Modellerin Performans Değerlendirmesi ve Karşılaştırmalı Analizi

Bu bölümde, dört farklı sigorta veri seti üzerinde kurulan modellerin performans sonuçları sunulmakta ve karşılaştırmalı olarak değerlendirilmektedir. Her bir veri seti için, eğitim ve test verilerinde elde edilen $RMSE$, MAE, $rRMSE$ ve $rMAE$ performans ölçütleri tablolar hâlinde verilmiş ve her modelin doğruluk düzeyi yorumlanmıştır.

Ardından, tüm veri setlerinden elde edilen performans ölçütlerinin ortalamaları alınarak algoritmaların genel başarı sıralamaları oluşturulmuştur. Bu analiz, modellerin farklı veri özellikleri altında gösterdikleri tutarlılığı ve genellenebilirlik düzeylerini değerlendirmek açısından önemli bir karşılaştırma zemini sunmaktadır.

6.3.1 Otomobil sigortası veri seti sonuçları

Bu alt bölümde, Otomobil Sigortası veri seti üzerinde uygulanan modellerin eğitim ve test performansları sunulmakta ve karşılaştırmalı olarak değerlendirilmektedir.

Çizelge 6.5 Otomobil Sigortası veri setinde eğitim veri seti için model performansı sonuçları

	$RMSE$	MAE	$rRMSE$	$rMAE$
Tweedie Regresyon	**1042,04804**	**250,72838**	**7,66268**	1,84366
Klasik Doğrusal Regresyon	1059,54244	252,31657	7,72409	1,83959
LASSO Regresyon	1059,61812	252,52124	7,72464	1,84108
Ridge Regresyon	1059,54262	252,32075	7,72409	1,83962
Karar Ağaçları	1058,65265	252,84403	7,71763	1,84343
Rasgele Orman	**1044,01242**	**249,42537**	**7,61087**	**1,81849**
Destek Vektör Makineleri	1059,54251	252,34751	7,72409	1,83981
Yapay Sinir Ağları	1059,37806	251,30302	7,72284	**1,83260**
XGBoost	**1043,56331**	**251,00155**	**7,64117**	**1,83801**
LightGBM	1056,10701	253,49576	7,69150	1,84637
CatBoost	1053,68779	253,15278	7,68147	1,84572

Çizelgede yer alan performans ölçütleri dikkate alındığında, Rasgele Orman, XGBoost ve Tweedie Regresyon modelleri eğitim veri setinde en iyi sonuçları veren algoritmalar olarak öne çıkmaktadır; bu modeller özellikle $RMSE$, MAE ve $rRMSE$ değerleri bakımından daha düşük hata oranları sergileyerek diğer yöntemlere göre daha başarılı bir tahmin performansı göstermiştir. Buna karşılık, genel hata düzeyleri açısından değerlendirildiğinde LASSO Regresyon modeli en zayıf performansı ortaya koymuş,

hem $RMSE$ hem de $rRMSE$ değerlerinin yüksekliği nedeniyle diğer algoritmalara kıyasla daha düşük bir tahmin gücüne sahip olduğu gözlenmiştir.

Çizelge 6.6 Otomobil Sigortası veri setinde test veri seti için model performansı sonuçları

	$RMSE$	MAE	$rRMSE$	$rMAE$
Tweedie Regresyon	1103,40404	256,24052	7,75682	**1,80248**
Klasik Doğrusal Regresyon	**1029,77853**	**252,71928**	**7,46399**	1,83700
LASSO Regresyon	1029,90890	252,94999	7,46490	1,83866
Ridge Regresyon	**1029,77587**	**252,72078**	**7,46397**	1,83701
Karar Ağaçları	1032,34387	253,65373	7,48291	1,84404
Rasgele Orman	1031,15230	253,09745	7,47397	1,83991
Destek Vektör Makineleri	**1029,77906**	252,75096	**7,46399**	1,83722
Yapay Sinir Ağları	1029,83240	**251,98856**	7,46442	**1,83016**
XGBoost	1066,49776	255,26045	7,60287	**1,82364**
LightGBM	1036,96545	253,63046	7,54703	1,85114
CatBoost	1030,34203	254,11062	7,46813	1,84704

Test veri seti sonuçları incelendiğinde, Klasik Doğrusal Regresyon, Ridge Regresyon ve Destek Vektör Makineleri modelleri $RMSE$ ve $rRMSE$ değerleri açısından en düşük hata oranlarını göstererek öne çıkmış, ayrıca MAE sonuçlarında da benzer biçimde güçlü performans sergilemişlerdir. Bu durum, söz konusu modellerin test verisi üzerinde oldukça istikrarlı ve güvenilir tahminler üretebildiğini göstermektedir. Buna karşılık, Tweedie Regresyon modeli yüksek $RMSE$ ve $rRMSE$ değerleri nedeniyle test setinde en zayıf performansı ortaya koymuş, bu da genel tahmin başarısının diğer algoritmalara kıyasla daha düşük seviyede kaldığını göstermektedir.

Otomobil Sigortası veri setinin hem eğitim hem de test sonuçları birlikte değerlendirildiğinde, XGBoost modeli hem eğitim hem test aşamasında istikrarlı biçimde düşük hata değerleri elde ederek genel anlamda en başarılı algoritma olmuştur. XGBoost'u Rasgele Orman ve Ridge Regresyon izlemiştir; bu iki model de farklı veri bölünmelerinde dengeli performans sergileyerek genelleme kabiliyetini korumuştur.

Buna karşılık, LASSO Regresyon her iki aşamada da yüksek hata değerleriyle en zayıf modeli oluşturmuştur. Ayrıca Tweedie Regresyon, eğitimde iyi performans göstermesine rağmen test aşamasında belirgin bir performans düşüşü sergileyerek aşırı uyum eğilimi göstermiştir.

6.3.2 Araç sigorta hasarı veri seti sonuçları

Bu kısımda, Araç Sigorta Hasarı veri seti üzerinde elde edilen modelleme sonuçları yer almakta olup, her algoritmanın eğitim ve test verilerindeki performans ölçütleri karşılaştırılmıştır.

Çizelge 6.7 Araç Sigorta Hasarı veri setinde eğitim veri seti için model performansı sonuçları

	$RMSE$	MAE	$rRMSE$	$rMAE$
Tweedie Regresyon	4045,79280	969,19504	2,66650	0,63876
Klasik Doğrusal Regresyon	4077,44820	1128,38120	2,68733	0,74355
LASSO Regresyon	4077,80420	1109,84360	2,68756	0,73133
Ridge Regresyon	4078,02640	1104,28298	2,68771	0,72768
Karar Ağaçları	4043,63540	967,70498	2,66506	0,63775
Rasgele Orman	**3441,79160**	**884,23814**	**2,26865**	**0,58281**
Destek Vektör Makineleri	4163,93500	958,14854	2,74429	0,63132
Yapay Sinir Ağları	4033,54680	1046,45426	2,65818	0,68962
XGBoost	**3798,10500**	**886,93658**	**2,50209**	**0,58497**
LightGBM	3962,80000	994,98642	2,61184	0,65569
CatBoost	**3884,03340**	**942,15284**	**2,55999**	**0,62098**

Araç Sigorta Hasarı veri setinin eğitim sonuçları incelendiğinde, Rasgele Orman, XGBoost ve CatBoost modelleri dört performans ölçütü açısından da öne çıkarak en iyi sonuçları veren algoritmalar olmuştur; özellikle Rasgele Orman en düşük $RMSE$, MAE, $rRMSE$ ve $rMAE$ değerleriyle en güçlü performansı sergilemiştir. Buna karşılık, Destek Vektör Makineleri, Ridge Regresyon ve LASSO Regresyon modelleri yüksek hata

değerleriyle zayıf sonuçlar üretmiş ve diğer yöntemlere kıyasla daha düşük bir tahmin gücü ortaya koymuştur.

Çizelge 6.8 Araç Sigorta Hasarı veri setinde test veri seti için model performansı sonuçları

	$RMSE$	MAE	$rRMSE$	$rMAE$
Tweedie Regresyon	**3803,10260**	**950,02012**	**2,54583**	**0,63961**
Klasik Doğrusal Regresyon	3812,41900	1102,22040	2,55103	0,74308
LASSO Regresyon	3811,79720	1082,61700	2,55058	0,72985
Ridge Regresyon	3811,30260	1078,02390	**2,55007**	0,72659
Karar Ağaçları	3823,22720	951,95368	2,55929	0,64119
Rasgele Orman	3834,74700	1004,36936	2,56683	0,67620
Destek Vektör Makineleri	3882,75260	**930,20502**	2,59683	**0,62607**
Yapay Sinir Ağları	3857,67060	1057,48980	2,58312	0,71282
XGBoost	**3770,40600**	**910,92052**	**2,52308**	**0,61043**
LightGBM	3823,73360	999,32280	2,55868	0,67319
CatBoost	**3810,66600**	955,67340	2,55048	0,64340

Araç Sigorta Hasarı veri setinin test sonuçları incelendiğinde, XGBoost ve Tweedie Regresyon modelleri en düşük hata değerleriyle öne çıkarak en başarılı algoritmalar olarak belirlenmiştir; özellikle XGBoost tüm performans ölçütlerinde ($RMSE$, MAE, $rRMSE$ ve $rMAE$) en düşük hata değerlerini elde ederek test verisinde en yüksek genelleme başarısını göstermiştir. Buna karşılık, Klasik Doğrusal Regresyon ve Destek Vektör Makineleri modelleri yüksek hata değerleriyle en zayıf performansı sergilemiştir. Özellikle Destek Vektör Makineleri'nin $rRMSE$ ve $RMSE$ değerlerinin yüksekliği, modelin karmaşık veri yapısını yeterince temsil edemediğini göstermektedir.

Araç Sigorta Hasarı veri setinin hem eğitim hem de test sonuçları birlikte değerlendirildiğinde, Rasgele Orman, XGBoost ve CatBoost modelleri genel olarak en iyi performansı sergileyen algoritmalar olmuştur. Rasgele Orman eğitim setinde tüm ölçütlerde açık ara üstünlük sağlamış, XGBoost ise test setinde en düşük hata değerleriyle öne çıkarak güçlü bir genelleme yeteneği ortaya koymuştur. CatBoost her

iki veri setinde de istikrarlı sonuçlar üreterek en iyi üç algoritma arasına girmiştir. Buna karşılık, Klasik Doğrusal Regresyon, LASSO Regresyon ve Destek Vektör Makineleri her iki aşamada da yüksek hata değerleri üretmiş ve en zayıf performansı göstermiştir

6.3.3 MASS veri seti sonuçları

Bu bölümde, MASS veri seti için elde edilen model performansları incelenmekte ve algoritmaların tahmin doğrulukları değerlendirilmiştir.

Çizelge 6.9 MASS veri setinde eğitim veri seti için model performansı sonuçları

	$RMSE$	MAE	$rRMSE$	$rMAE$
Tweedie Regresyon	16891,81924	6804,34909	4,45144	1,79313
Klasik Doğrusal Regresyon	16893,18322	6805,63906	4,45180	1,79347
LASSO Regresyon	16893,21352	6805,88799	4,45180	1,79354
Ridge Regresyon	16893,18343	6805,64688	4,45180	1,79347
Karar Ağaçları	16871,82339	6792,20697	4,44617	1,78993
Rasgele Orman	**16819,27394**	**6777,82117**	**4,43232**	**1,78614**
Destek Vektör Makineleri	16893,20700	6804,03489	4,45180	1,79305
Yapay Sinir Ağları	16892,80920	6805,76793	4,45170	1,79350
XGBoost	**16810,11495**	**6767,25454**	**4,42993**	**1,78337**
LightGBM	16862,93347	6794,18409	4,44383	1,79046
CatBoost	**16859,16619**	**6791,40215**	**4,44283**	**1,78972**

MASS eğitim veri seti sonuçlarına göre en iyi performansı XGBoost, Rasgele Orman ve CatBoost algoritmaları göstermiştir. Bu üç yöntem, tüm performans ölçütlerinde en düşük hata değerlerini üreterek veri setindeki karmaşık ilişkileri başarılı biçimde modellemiş ve güçlü tahmin kapasitesi sergilemiştir. Buna karşılık, Klasik, LASSO ve Ridge Regresyon yöntemleri daha yüksek $RMSE$ ve MAE değerleriyle görece zayıf sonuçlar elde etmiştir.

Çizelge 6.10 MASS veri setinde test veri seti için model performansı sonuçları

	$RMSE$	MAE	$rRMSE$	$rMAE$
Tweedie Regresyon	17176,23958	6839,20178	4,48045	1,78447
Klasik Doğrusal Regresyon	17176,94316	6840,63882	4,48063	1,78484
LASSO Regresyon	17176,89587	6840,79359	4,48062	1,78489
Ridge Regresyon	17176,89773	6840,61932	4,48062	1,78484
Karar Ağaçları	17182,26705	**6834,51762**	4,48207	**1,78332**
Rasgele Orman	**17173,48832**	6840,54465	**4,47974**	1,78485
Destek Vektör Makineleri	17176,69883	6838,89153	4,48056	1,78439
Yapay Sinir Ağları	17176,91106	6841,24755	4,48062	1,78500
XGBoost	**17169,83576**	**6830,24190**	**4,47879**	**1,78213**
LightGBM	17175,70332	6840,00680	4,48033	1,78471
CatBoost	**17172,22126**	**6837,42489**	**4,47942**	**1,78405**

MASS veri setinin test sonuçları değerlendirildiğinde, XGBoost, CatBoost ve Rasgele Orman modelleri dört performans ölçütünde de en düşük hata değerlerini göstererek en başarılı algoritmalar olarak öne çıkmıştır; özellikle XGBoost tüm performans ölçütlerinde ($RMSE$, MAE, $rRMSE$ ve $rMAE$) en düşük hata değerlerini elde ederek test verisinde en yüksek genelleme başarısını göstermiştir. Buna karşılık, Yapay Sinir Ağları modeli daha yüksek hata değerleriyle diğer yöntemlerin gerisinde kalmış ve en zayıf performansı sergilemiştir.

MASS veri setinin hem eğitim hem de test sonuçları birlikte ele alındığında, XGBoost, CatBoost ve Rasgele Orman modelleri en başarılı algoritmalar olarak öne çıkmaktadır. XGBoost her iki veri setinde de en düşük $RMSE$ ve $rMAE$ değerleriyle dikkat çekerken, CatBoost istikrarlı biçimde güçlü performans sergilemiş ve Rasgele Orman özellikle eğitim setinde öne çıkmıştır. Bu üç modelin dengeli ve düşük hata oranları, yüksek genelleme kabiliyetine işaret etmektedir. Buna karşılık, LASSO Regresyon modeli eğitim setinde en yüksek hata değerlerini alarak zayıf bir uyum göstermiştir. Karar Ağaçları ve Yapay Sinir Ağları ise test aşamasında hata değerlerinin yükselmesiyle model başarısında düşüş yaşamıştır. Bu durum, bu iki algoritmanın test verisinde genelleme gücünü koruyamadığını göstermektedir.

6.3.4 Ohlsson veri seti sonuçları

Ohlsson veri seti üzerinde yürütülen modelleme uygulamalarına ait eğitim ve test performans ölçütleri bu alt bölümde sunulmuş ve sonuçlar karşılaştırmalı olarak yorumlanmıştır.

Çizelge 6.11 Ohlsson veri setinde eğitim veri seti için model performansı sonuçları

	$RMSE$	MAE	$rRMSE$	$rMAE$
Tweedie Regresyon	4579,25047	506,60110	17,86218	1,97665
Klasik Doğrusal Regresyon	4584,96802	580,66153	17,88466	2,26351
LASSO Regresyon	4585,21456	571,73217	17,88562	2,22875
Ridge Regresyon	4584,96806	580,55723	17,88466	2,26310
Karar Ağaçları	4568,14446	501,65088	17,81866	1,95572
Rasgele Orman	**4451,65519**	**482,54129**	**17,39103**	**1,88414**
Destek Vektör Makineleri	4585,44756	571,94115	17,88657	2,22906
Yapay Sinir Ağları	4556,30921	505,42513	17,77294	1,97113
XGBoost	**4354,01195**	**480,76807**	**16,97943**	**1,87369**
LightGBM	**4503,90692**	**494,52552**	**17,56859**	**1,92804**
CatBoost	4504,51060	495,03077	17,56988	1,92993

Ohlsson veri setinin eğitim sonuçları incelendiğinde, XGBoost, Rasgele Orman ve CatBoost modelleri en düşük $RMSE$, MAE, $rRMSE$ ve $rMAE$ değerleriyle öne çıkarak en iyi performansı sergileyen algoritmalar olmuştur; özellikle XGBoost tüm ölçütlerde birinci sırada yer alarak belirgin şekilde üstün sonuçlar elde etmiştir. Buna karşılık, Klasik Doğrusal Regresyon yüksek hata değerleriyle diğer yöntemlere kıyasla daha zayıf bir performans göstermiştir.

Çizelge 6.12 Ohlsson veri setinde test veri seti için model performansı sonuçları

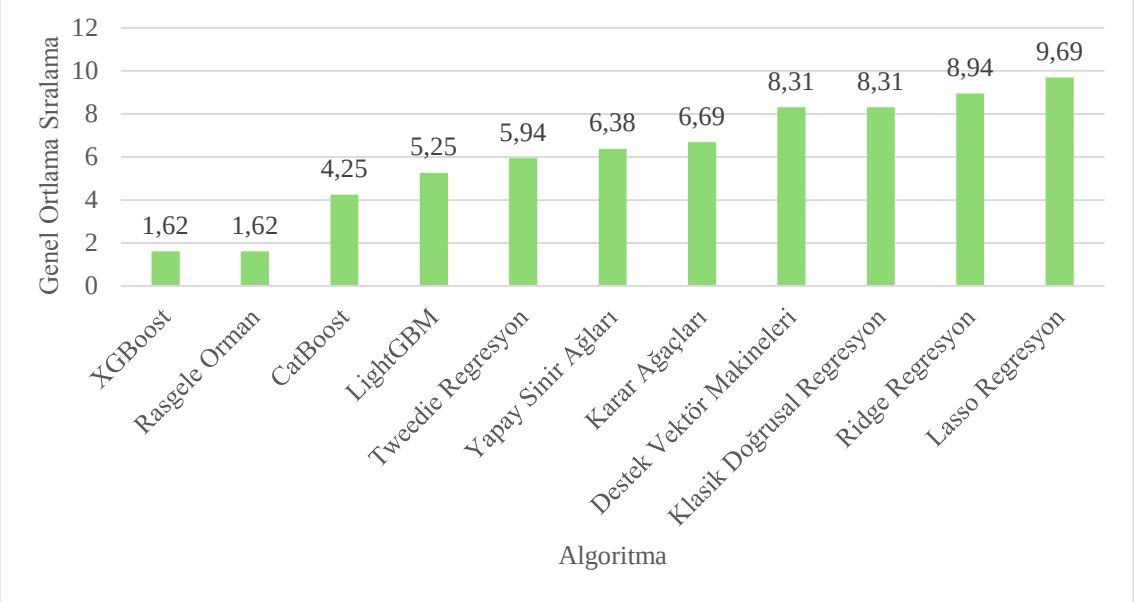
	$RMSE$	MAE	$rRMSE$	$rMAE$
Tweedie Regresyon	5036,96166	544,84477	17,23531	1,86207
Klasik Doğrusal Regresyon	5040,24018	619,14678	17,24659	2,12443
LASSO Regresyon	5040,17652	610,26191	17,24622	2,09356
Ridge Regresyon	5040,23877	619,04141	17,24658	2,12406
Karar Ağaçları	5052,23127	540,55172	17,28453	1,85088
Rasgele Orman	5036,18809	542,64358	**17,13490**	**1,84860**
Destek Vektör Makineleri	5040,10818	610,11046	17,24615	2,09464
Yapay Sinir Ağları	**5021,49341**	544,91301	**17,18288**	1,86453
XGBoost	**5025,09731**	**531,77244**	**17,19326**	**1,82210**
LightGBM	**5030,59424**	**538,06226**	17,21290	**1,84327**
CatBoost	5033,66572	**540,16595**	17,22151	1,84998

Ohlsson veri setinin test sonuçları incelendiğinde, tüm performans ölçütleri dikkate alındığında XGBoost modeli en düşük hata değerlerini elde ederek test aşamasında en başarılı algoritma olmuştur. LightGBM ve CatBoost modelleri ise benzer biçimde düşük hata düzeyleriyle XGBoost'un hemen ardından gelmiş ve güçlü bir genelleme performansı sergilemiştir. Buna karşılık, Klasik Doğrusal Regresyon ve Ridge Regresyon modelleri özellikle MAE ve $rMAE$ değerlerinde en yüksek hataları üreterek test verisinde en zayıf performansı göstermiştir.

Eğitim ve test sonuçları birlikte değerlendirildiğinde, Ohlsson veri setinde her iki aşamada da tutarlı biçimde yüksek performans sergileyen algoritmaların XGBoost, LightGBM ve CatBoost olduğu görülmektedir. Bu üç ağaç tabanlı topluluk yöntemi hem eğitim hem de test verilerinde en düşük $RMSE$, MAE, $rRMSE$ ve $rMAE$ değerlerini elde etmiş; dolayısıyla hem öğrenme başarısı hem de genelleme yeteneği açısından güçlü bir denge kurmuştur. Özellikle XGBoost, her iki aşamada da hata değerleri bakımından birinci sırada yer alarak genel olarak en başarılı model olmuştur. Buna karşılık, Klasik Doğrusal Regresyon, Ridge Regresyon ve LASSO Regresyon modelleri hem eğitim hem test setlerinde yüksek MAE ve $rMAE$ değerleriyle en zayıf sonuçları vermiştir.

6.3.5 Genel ortalama performans karşılaştırması

Dört veri setinden elde edilen performans ölçütlerinin ortalamaları alınarak, tüm modellerin genel başarı sıralamaları oluşturulmuştur. Bu kısımda, algoritmaların genel performans düzeyleri tablo ve grafikler aracılığıyla karşılaştırmalı olarak değerlendirilmektedir.

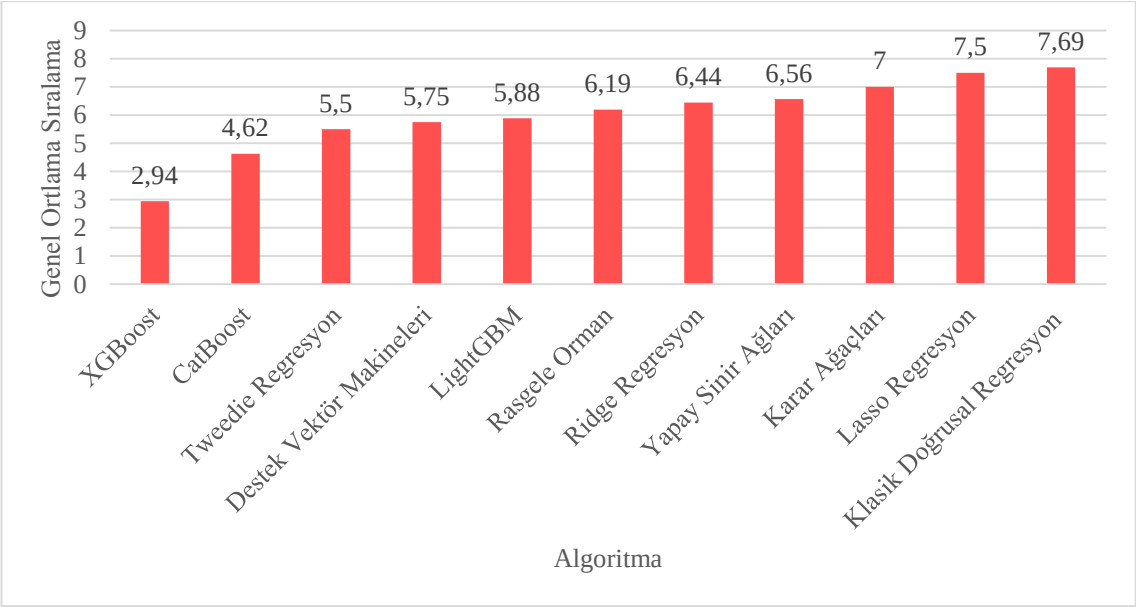


Şekil 6.28 Veri setlerinin tamamı için eğitim aşamasındaki model performans sıralamaları

Çizelge 6.13 Tüm eğitim veri setleri ve performans ölçütleri genelinde XGBoost, Rasgele Orman ve CatBoost algoritmalarının ortalama sıralama sonuçları

	RMSE	***MAE***	***rRMSE***	***rMAE***
XGBoost	1.5	1.75	1.5	1.75
Rasgele Orman	2.0	1.5	1.5	1.5
CatBoost	3.5	5.0	3.5	5.0

Şekil 6.28'e göre, dört taşıt sigortası veri setinin eğitim aşamasındaki ortalama performans sıralamaları incelendiğinde, XGBoost ve Rasgele Orman modelleri en düşük ortalama sıralama değerlerini elde ederek en başarılı algoritmalar olmuştur. Bu iki modelin ardından CatBoost da güçlü bir performans sergileyerek üçüncü sırada yer almıştır. Buna karşılık, LASSO Regresyon ve Ridge Regresyon yüksek ortalama değerleriyle en düşük başarıyı göstermiştir. Bu sonuçlar, ağaç tabanlı topluluk yöntemlerinin eğitim verisi üzerinde doğrusal modellere göre çok daha yüksek tahmin performansı sergilediğini göstermektedir.



Şekil 6.29 Veri setlerinin tamamı için test aşamasındaki model performans sıralamaları

Çizelge 6.14 Tüm test veri setleri ve performans ölçütleri genelinde XGBoost, CatBoost ve Tweedie Regresyon algoritmalarının ortalama sıralama sonuçları

	$\boldsymbol{RMSE}$	$\boldsymbol{MAE}$	$\boldsymbol{rRMSE}$	$\boldsymbol{rMAE}$
XGBoost	3.5	3.25	3.75	1.25
CatBoost	3.75	5.0	4.25	5.5
Tweedie Regresyon	6.0	6.25	6.0	3.75

Şekil 6.29'da dört taşıt sigortası veri seti üzerindeki test performansları karşılaştırıldığında, XGBoost en düşük ortalama sıralama değeriyle en başarılı algoritma olmuştur. XGBoost'u sırasıyla CatBoost ve Tweedie Regresyon takip etmiş, bu üç model test verisi üzerinde en yüksek genelleme başarısını göstermiştir. Buna karşılık, Karar Ağaçları ve LASSO Regresyon modelleri yüksek ortalama sıralama değerleriyle en zayıf performansı sergilemiştir. Bu sonuçlar, karmaşık ve farklı dağılımlara sahip verilerde topluluk yöntemlerinin ve gelişmiş regresyon yaklaşımlarının daha tutarlı ve başarılı sonuçlar ürettiğini göstermektedir.

7. SONUÇ

Bu çalışma, motorlu kara taşıtları sigortasında hasar tutarının tahmin edilmesine yönelik farklı istatistiksel ve makine öğrenmesi yöntemlerini karşılaştırmalı olarak incelemiş ve çeşitli veri setleri üzerinden kapsamlı bir değerlendirme sunmuştur. Sigorta hasar verilerinin yapısal özellikleri doğrusal modellerin varsayımlarını büyük ölçüde sınırlamakta, bu nedenle daha esnek dağılımsal modellerin ve güçlü makine öğrenmesi algoritmalarının kullanımını gerekli kılmaktadır. Bu kapsamda tezde Tweedie Modeli ile Klasik Doğrusal Regresyon, LASSO Regresyon, Ridge Regresyon, Karar Ağaçları, Rasgele Orman, Destek Vektör Makineleri, Yapay Sinir Ağları, XGBoost, LightGBM ve CatBoost algoritmaları değerlendirilmiştir. Tezin temel amacı, farklı modelleme yaklaşımlarının sigorta hasar tutarının tahmindeki performanslarını karşılaştırmaktır.

Tezde kullanılan Otomobil Sigortası, Araç Sigorta Hasarı, MASS ve Ohlsson veri setleri, farklı örneklem yapıları ve değişken türleri sunarak modellerin genellenebilirliğini değerlendirmek açısından uygun bir zemin oluşturmuştur. Veri setlerinde sıfır hasar oranının yüksek olması ve pozitif hasar tutarlarının sağa çarpık bir dağılım göstermesi, model seçiminin performans açısından kritik olduğunu ortaya koymuştur. Elde edilen sonuçlar hem eğitim hem de test aşamalarında gradyan artırma tabanlı algoritmaların sigorta hasar tutarının tahmininde diğer yöntemlere belirgin şekilde üstünlük sağladığını göstermektedir. Özellikle XGBoost, eğitimde ve testte açık ara en yüksek performansı elde ederek tüm modeller arasında istikrarlı biçimde ilk sırada yer almıştır. Eğitim aşamasında Rasgele Orman da XGBoost ile aynı ortalama sıralama değerine (1.62) sahip olmuş, ancak test aşamasında performansı bir miktar düşerek orta-üst seviyede konumlanmıştır. Test sonuçlarında ise XGBoost'un ardından CatBoost ikinci sırayı almış ve gradyan artırma yöntemlerinin test performansında da avantajını koruduğu görülmüştür. LightGBM her iki aşamada da güçlü bir performans sergilemekle birlikte, XGBoost ve CatBoost'un gerisinde kalmıştır.

Tweedie regresyonu, eğitim ve test aşamalarında sırasıyla 5.25 ve 5.50 ortalama sıralama değerleriyle genel sıralamada orta düzey bir performans sergilemiş; bununla birlikte bazı veri setlerinde ilk üçte yer alarak rekabetçi sonuçlar üretebilmiştir. Bu

bulgu, çalışmada ele alınan hasar tutarı verilerinin yapısıyla tutarlı biçimde, Tweedie'nin yarı sürekli (yüksek oranda sıfır değer içeren ve sıfır dışındaki pozitif tutarları aşırı sağa çarpık dağılan) veri yapısını tek bir model çatısı altında ele alabilme avantajını desteklemektedir. Tweedie regresyonu, sıfır gözlemlerini ve pozitif hasar tutarlarını aynı anda modelleyerek sigorta verisindeki temel mekanizmaya uygun, tutarlı ve teorik olarak sağlam bir çerçeve sunmaktadır. Buna ek olarak, Tweedie modeli makine öğrenmesi yaklaşımlarının aksine kapalı kutu (black-box) bir yapı sunmamakta; model katsayıları aracılığıyla değişkenlerin hasar tutarı üzerindeki yönü ve göreli etkisi doğrudan yorumlanabilmektedir. Bu özellik, hangi risk faktörlerinin hasar tutarı üzerinde daha belirleyici olduğunu ortaya koymaya, değişken öneminin istatistiksel olarak değerlendirilmesine ve modelin istatistiksel açıdan geliştirilmesine olanak tanımaktadır. Bununla birlikte Tweedie modelinin yorumlanabilirliği, mevzuat uyumu ve uygulamada tek modelle uygulanabilmesi gibi güçlü yönleri, onu çok sıfırlı hasar tutarı verilerinde hâlâ önemli ve tercih edilebilir bir seçenek hâline getirmektedir.

Model sıralamalarının genel görünümü, LASSO, Klasik Doğrusal Regresyon ve bazı durumlarda Ridge Regresyonu gibi lineer yöntemlerin hem eğitim hem de test aşamasında en düşük performansa sahip olduğunu göstermektedir. Bu bulgu, sigorta hasar verilerindeki doğrusal olmayan ilişki ve karmaşık etkileşim yapılarını yakalayamayan klasik modellerin sınırlılıklarını desteklemektedir.

Bu tez kapsamında geliştirilen modeller ve elde edilen performans sonuçları, sigorta şirketleri açısından doğrudan karar destek aracı olarak kullanılabilir niteliktedir. Özellikle en yüksek tahmin başarısını gösteren gradyan artırma tabanlı algoritmalar, poliçe bazında beklenen hasar tutarının daha hassas tahmin edilmesine imkân sağlayarak teknik prim hesaplamalarında daha doğru risk fiyatlaması yapılmasına katkı sunabilir. Bu durum, risk segmentasyonunun iyileştirilmesi, portföy dengesinin sağlanması ve sermaye planlamasının daha etkin yürütülmesi açısından operasyonel değer üretmektedir. Ayrıca model çıktıları, yüksek tutarlı hasar potansiyeline sahip poliçelerin erken tespiti ve reasürans stratejilerinin gözden geçirilmesi gibi alanlarda da kullanılabilir.

Buna karşılık, Tweedie regresyonu gibi yorumlanabilir yapıya sahip modeller, düzenleyici uyum, iç denetim süreçleri ve risk faktörlerinin etkisinin analiz edilmesi bakımından önemli avantajlar sağlamaktadır. Dolayısıyla bu çalışma, sigorta şirketlerinin yüksek tahmin performansı ile model şeffaflığı arasında denge kurarak bağlama özgü bir model stratejisi geliştirmesine katkı sunmaktadır.

KAYNAKLAR

Abdulazeez, A., Sulaiman, M.A. and Zeebaree, D.Q. 2020. Evaluating Data Mining Classification Methods Performance in Internet of Things Applications. Journal of Soft Computing and Data Mining, 1; 11–25.

Abdulazeez, A.M., Salim, B.W., Zeebaree, D.Q. and Doghramachi, D. 2020. Comparison of VPN Protocols at Network Layer Focusing on Wire Guard Protocol. International Journal of Interactive Mobile Technologies (iJIM), 14(18); 157–177.

Abdulqader, D.M., Abdulazeez, A.M. and Zeebaree, D.Q. 2020. Machine Learning Supervised Algorithms of Gene Selection: A Review. Machine Learning, 62, 233–244.

Acharya, M.S., Armaan, A. and Antony, A.S. 2019. A comparison of regression models for prediction of graduate admissions. 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 1–5.

Akgün, B. and Öğüdücü, Ş.G. 2015. Streaming linear regression on Spark MLlib and MOA. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 1244–1247.

Alpaydin, E. 2014. Introduction to machine learning, 3rd edn. The MIT Press, 616 p., Cambridge.

Bargarai, F., Abdulazeez, A., Tiryaki, V. and Zeebaree, D. 2020. Management of Wireless Communication Systems Using Artificial Intelligence-Based Software Defined Radio. International Journal of Interactive Mobile Technologies (iJIM), 14(13); 107–133.

Bisong, E. 2019. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, 1st edn. Apress. 709 p., New York.

Breiman, L. 2001. Random forests. Machine Learning, 45, 5–32.

Breiman, L., Friedman, J., Olshen, R.A. and Stone, C.J. 1984. Classification and Regression Trees. Chapman & Hall/CRC, 368 p., New York.

Brown, J.E. and Dunn, P.K. 2011. Comparisons of Tobit, linear, and PoissonGamma regression models: an application of time use data. Sociological Methods & Research, 40(3); 511–535.

Cortes, C. and Vapnik, V. 1995. Support-vector networks. Machine Learning, 20, 273–297.

Delong, Ł., Lindholm, M. and Wüthrich, M.V. 2021. Making Tweedie's compound Poisson model more accessible. European Actuarial Journal, 11, 185–226.

Dionne, G., Gouri´ eroux, C. and Vanasse, C. 2001. Testing for evidence of adverse selection in the automobile insurance market: A comment. Journal of Political Economy, 109; 444–453.

Dorogush, A.V., Ershov, V. and Gulin, A. 2018. CatBoost: Gradient boosting with categorical features support. ArXiv Preprint, ArXiv:1810.11363.

Dunn, P.K. 2004. Precipitation occurrence and amount can be modelled simultaneously. International Journal of Climatology, 24, 1231–1239.

Dunn, P.K. and Smyth, G.K. 2005. Series evaluation of Tweedie exponential dispersion models. Statistics and Computing, 15(4); 267–280.

Dunn, P.K. and Smyth, G.K. 2008. Evaluation of Tweedie exponential dispersion models using Fourier inversion. Statistics and Computing, 18(1); 73–86.

Dunn, P.K. and Smyth G.K. 2018. Generalized Linear Models with Examples in R. Springer Texts in Statistics. Springer Nature, 573 p., New York.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. 2004. Least angle regression. Annals of Statistics, 32, 407–451.

Emmert-Streib, F. and Dehmer, M. 2019. High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. Machine Learning and Knowledge Extraction, 1(1); 359–383.

Fan, J. and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96, 1348–1360.

Foster, S.D. and Bravington, M.V. 2013. A Poisson–Gamma model for analysis of ecological data. Environmental and Ecological Statistics, 20(4); 533–552.

Frees, E.W., Derrig, R.A. and Meyers, G. 2014. Predictive Modeling Applications in Actuarial Science: Volume 1, Predictive Modeling Techniques. Cambridge University Press, 563 p., Cambridge.

Friedl, M.A. and Brodley, C.E. 1997. Decision tree classification of land cover from remotely sensed data. Remote Sensing of Environment, 61, 399–409.

Friedman, J., Hastie, T. and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33, 1–22.

Gao, G. 2024. Fitting Tweedie's compound Poisson model to pure premium with the EM algorithm. Insurance Mathematics and Economics, 114, 29–42.

Gilchrist, R. and Drinkwater, D. 1999. Fitting Tweedie models to data with probability of zero responses. In: Friedl, H., Berghold, A. and Kauermann G. (eds.), Statistical Modelling: Proceedings of the 14th International Workshop on Statistical Modelling, 207–214. International Workshop on Statistical Modelling, Gräz.

Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R. 2006. Random forests for land cover classification. Pattern Recognition Letters, 27, 294–300.

Gu, Y. 2024. Dispersion Modeling in Zero-inflated Tweedie Models with Applications to Insurance Claim Data Analysis. *arXiv preprint,* arXiv:2405.14990.

Guo, L., Chehata, N., Mallet, C. and Boukir, S. 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. ISPRS Journal of Photogrammetry and Remote Sensing, 66, 56–66.

Hansen, M., Dubayah, R. and Defries, R. 1996. Classification trees: an alternative to traditional land cover classifiers. International Journal of Remote Sensing, 17, 1075–1081.

Hastie, T., Tibshirani, R. and Wainwright, M. 2015. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, 367 p., Boca Raton.

Hoerl, A.E. and Kennard, R.W. 1970. Ridge regression: biased estimation for nonorthogonal problems. Technomterics, 12, 54–67.

Jain, N. 2018. Towards Machine Learning: Alternative Methods for Insurance Pricing – Poisson-Gamma GLM's, Tweedie GLM's and Artificial Neural Networks. Honor Thesis. UC Berkeley, Statistics, Berkeley.

Jørgensen, B. 1987. Exponential dispersion models (with discussion). Journal of the Royal Statistical Society, Series B, 49, 127–162.

Jørgensen, B. 1997. The Theory of Dispersion Models. Monographs on Statistics and Applied Probability. Chapman and Hall, 256 p., London.

Jørgensen, B. and de Souza, M.C.P. 1994. Fitting Tweedie's compound Poisson model to insurance claims data. Scandinavian Actuarial Journal, 1, 69–93.

Kavitha, S., Varuna, S. and Ramya, R. 2016. A comparative analysis on linear regression and support vector regression. 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 1–5.

Lippitt, C.D., Rogan, J., Li, Z., Eastman, J.R. and Jones, T.G. 2008. Mapping selective logging in mixed deciduous forest: a comparison of machine learning algorithms. Photogrammetric Engineering & Remote Sensing, 74, 1201–1211.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q. and Niu, X. 2018. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electronic Commerce Research and Applications, 31, 24–39.

Maulud, H. and Abdulazeez, A.M. 2020. A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends, 1(2); 140–147.

Nelder, J.A. and Pregibon, D. 1987. An extended quasi-likelihood function. Biometrika, 74, 221–232.

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P. and Righetti, M. 2024. Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). Environmental Modelling and Software, 174, 105971.

Ogutu, J.O., Schulz-Streeck, T. and Piepho, H.-P. 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. BMC Proceedings, 6(Suppl 2); 10.

Pal, M. 2005. Random forest classifier for remote sensing classification. International Journal Remote Sensing, 26, 217–222.

Pal, M. and Mather, P.M. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment, 86, 554–565.

Peters, J., De Baets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., De Becker, P. and Huybrechts, W. 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling, 207, 304–318.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. 2018. CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6639–6649.

Quinlan, J.R. 1993. C4.5 Programs for Machine Learning. 1st edn, Morgan Kaufmann Publishers Inc., 312 p., San Francisco.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. and Rigol-Sánchez, J.P. 2012b. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93–104.

Rodriguez–Galiano V.F., Sanchez–Castillo M., Chica–Olmo M. and Chica–Rivas M. 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, 804–818.

Rogan, J., Miller, J., Stow, D., Franklin, J., Levien, L. and Fischer, C. 2003. Land-cover change monitoring with classification trees using Landsat TM and ancillary data. Photogrammetric Engineering & Remote Sensing, 69, 793–804.

Roopa, H. and Asha, T. 2019. A linear model based on principal component analysis for disease prediction. IEEE Access, 7, 105314–105318.

Seber, G.A. and Lee, A.J. 2003. Linear regression analysis, 2nd edn. Wiley, 592 p., Hoboken.

Shalev-Shwartz, S. and Ben-David, S. 2014. Understanding machine learning: From theory to algorithms. Cambridge University Press, 414 p., Cambridge.

Smolárová, T. 2017. Tweedie models for pricing and reserving. Master Thesis. Charles University, Department of Probability and Mathematical Statistics, Prague.

Smyth, G.K. 1996. Regression analysis of quantity data with exact zeros. Technology Management Centre, 572–580, Brisbane.

Smyth, G.K. 1996. Partitioned algorithms for maximum likelihood and other non-linear estimation. Statistics and Computing, 6, 201–216.

Smyth, G.K. and Jørgensen, B. 1999. Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling. In: Proceedings of the 52nd Session of the International Statistical Institute, Paper Meeting 68: Statistics and Insurance, Helsinki.

Smyth, G.K. and Jørgensen, B. 2002. Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modeling. ASTIN Bulletin: The Journal of the IAA, 32(1); 143–157.

Smyth, G.K. and Verbyla, A.P. 1999. Adjusted likelihood methods for modelling dispersion in generalized linear models. Environmetrics, 10, 695–709.

Taylor, L.R. 1961. Aggregation, variance and the mean. Nature, 189, 732–735.

Tibshirani, R. 1996. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, Series B (Statistical Methodology), 58, 267–288.

Tweedie, M.C.K. 1946. The regression of the sample variance on the sample mean. Journal of the London Mathematical Society, 21, 22–28.

Tweedie, M.C.K. 1984. An index which distinguishes between some important exponential families. In: Ghosh, J.K. and Roy, J. (eds.), Statistics: applications and new directions. Proceeding of the Indian statistical golden jubilee international conference, Indian Statistical Institute, Calcutta, 579–604.

Vapnik, V.N. 2000. The Nature of Statistical Learning Theory. 2nd edn, Springer, 334p., New York.

Wessels, K.J., De Fries, R.S., Dempewolf, J., Anderson, L.O., Hansen, A.J., Powell, S.L. and Moran, E.F. 2004. Mapping regional land cover with MODIS data for biological conservation: examples from the Greater Yellowstone Ecosystem, USA and Pará State, Brazil. Remote Sensing of Environment, 92, 67–83.

Wu, J., Liu, C., Cui, W. and Zhang, Y. 2019. Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression, 2019 IEEE International Conference on Power Data Science (ICPDS), 139–142.

Yang, Y., Qian, W. and Zou, H. 2018. Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models. Journal of Business & Economic Statistics, 36(3); 456–470.

Zebari, D.A., Zeebaree, D.Q., Abdulazeez, A.M., Haron, H., and Hamed, H.N.A. 2020. Improved Threshold Based and Trainable Fully Automated Segmentation for Breast Cancer Boundary and Pectoral Muscle in Mammogram Images. IEEE Access, 8, 203097–203116.

Zeebaree, D.Q., Haron, H., Abdulazeez, A.M. and Zebari, D.A. 2019. Machine learning and Region Growing for Breast Cancer Segmentation. 2019 International Conference on Advanced Science and Engineering (ICOASE), 88–93.

Zhang, C., Chen, Z. and Zhou, J. 2020. Research on Short-Term Load Forecasting Using K-means Clustering and CatBoost Integrating Time Series Features. 2020 39th Chinese Control Conference (CCC), 6099–6104.

Zhang, L. and Jánošík, D. 2024. Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches. Expert Systems With Applications, 241, 122686.

Zhang, Z., Li, Y., Li, L., Li, Z. and Liu, S. 2019. Multiple linear regression for high efficiency video intra coding. ICASSP 2019–IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1832–1836.

Zou, H. 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101, 1418–1429.