
GAZİ ÜNİVERSİTESİ
Gazi
1926

SAYMA VERİLERİNDE REGRESYON MODELLERİ VE REGRESYON AĞAÇLARI UYGULAMASI

Mine Fulya GÜRSEL

YÜKSEK LİSANS TEZİ
İSTATİSTİK ANA BİLİM DALI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

HAZİRAN 2024

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Mine Fulya GÜRSEL

11/06/2024

SAYMA VERİLERİNDE REGRESYON MODELLERİ VE REGRESYON AĞAÇLARI UYGULAMASI

(Yüksek Lisans Tezi)

Mine Fulya GÜRSEL

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Haziran 2024

ÖZET

Belirli bir zaman, alan veya hacim başına meydana gelen olayların sayımına bağlı olarak elde edilen veriler sayma verisi olarak tanımlanmaktadır. Regresyon modelleri de bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiyi açıklamaktadır. Bu çalışmada sayıma dayalı verilerin bağımlı değişken olduğu regresyon modellerinden bazıları incelenmiştir. Buna ek olarak heterojen bir popülasyonu daha homojen alt popülasyonlara ayırabilen ağaç tabanlı yöntemlerden, düğüm modelinde Poisson regresyon yöntemini kullanan, CART, GUIDE ve MOB regresyon ağacı algoritmaları incelenmiştir. Algoritmaların değişken seçim yanlılığı, tip 1 hataları ve güçleri simülasyon çalışmaları ile karşılaştırılmıştır. Ayrıca 2017-2022 yılları Ankara ilinde gerçekleşen trafik kaza sayıları ve doktora öğrencileri makale sayıları veri setleri ile bu ağaç modellerinin uygulamaları yapılmıştır. Bu veri setlerinden yola çıkarak bir simülasyon çalışması daha gerçekleştirilmiş ve algoritmaların tahmin hataları karşılaştırılmıştır. Simülasyon çalışmaları sonucunda değişkenlerin yansız olarak seçme, tip 1 hata ve güç bakımından karşılaştırılan üç algoritma arasında en iyi sonuç GUIDE algoritması ile elde edilmiştir. CART algoritmasının tip 1 hatası, tahmin performansı ve gücü yüksek elde edilmiştir. MOB algoritmasının tip 1 hatası ve tahmin performansı iyi olmasına rağmen gücü düşük elde edilmiştir.

Bilim Kodu	:	20514
Anahtar Kelimeler	:	Sayma verisi, sayma regresyon modelleri, regresyon ağaçları, GUIDE, MOB, CART
Sayfa Adedi	:	71
Danışman	:	Doç. Dr. Hatice Tül Kübra AKDUR

APPLICATION OF REGRESSION MODELS AND REGRESSION TREES ON COUNT DATASETS

(M. Sc. Thesis)

Mine Fulya GÜRSEL

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

June 2024

ABSTRACT

Count data is defined as data obtained based on a count of events occurring per specific time, area, or volume. Regression models explain the relationship between dependent and independent variables. In this study, some of the regression models where count data was the dependent variable were examined. In addition, CART, GUIDE, and MOB regression tree algorithms, which utilize the Poisson regression method in the node model, among tree-based methods that can segment a heterogeneous population into more homogeneous subpopulations, were examined. Variable selection bias, type 1 error, and power of the algorithms were compared using simulation studies. These tree models were also applied to the data sets of traffic accident numbers in Ankara, which occurred between 2017 and 2022, and the number of articles written by PhD students. Based on these data sets, another simulation study was conducted, and the prediction errors of the algorithms were compared. As a result of the simulation studies, the best result among the three compared algorithms in terms of unbiased selection of variables, type 1 error, and power was obtained with the GUIDE algorithm. The CART algorithm's type 1 error prediction performance and power were high. Although the MOB algorithm’s type 1 error and prediction performance were good, its power was low.

Science Code	:	20514
Key Words	:	Count data, count regression models, regression trees, GUIDE, MOB, CART
Page Number	:	71
Supervisor	:	Assoc. Prof. Dr. Hatice Tül Kübra AKDUR

TEŞEKKÜR

Yüksek lisans tez çalışmam boyunca öneri, yardım ve desteğini esirgemeyen sayın danışman hocam Doç. Dr. Hatice Tül Kübra AKDUR'a teşekkür ederim. Her türlü manevi ve maddi destekleri için minnettar olduğum aileme çok teşekkür ederim. Yüksek lisans eğitimim için 2211- Yurt İçi Lisansüstü Burs Programı kapsamında beni destekleyen TÜBİTAK Bilim İnsanı Destek Programları Başkanlığı'na (BİDEB) teşekkür ederim.

İÇİNDEKİLER

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	Açıklamalar
μ	Ortalama
β	Regresyon katsayıları

Kısaltmalar	Açıklamalar
AIC	Akaike bilgi kriteri
AID	Otomatik etkileşim tespiti
CART	Sınıflandırma ve regresyon ağaçları
CORE	Sayma regresyon ağaçları
GUIDE	Genelleştirilmiş yansız etkileşim tespiti ve tahmini
LASSO	En az mutlak küçültme ve seçim operatörü
MOB	Model tabanlı özyinelemeli bölümleme
SUPPORT	Düzleştirilmiş ve düzleştirilmemiş parçalı polinom regresyon ağaçları
YAEKK	Yeniden ağırlıklandırılmış en küçük kareler
ZIP	Sıfır yığılmalı Poisson
ZNB	Sıfır yığılmalı negatif binom
ZTP	Sıfır kesikli Poisson

1. GİRİŞ

Sayma verileri, bir olayın meydana gelme sayısını ve negatif olmayan tamsayı değerlerine sahip gözlemleri ifade eder. Bir olasılık dağılımı varsayımı altında, regresyon analizi, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi belirlemek ve bu ilişkiyi özetleyen bir model oluşturmak için kullanılmaktadır. Bağımlı değişken sayma verisi olduğunda genellikle Poisson veya negatif binom regresyon modelleri kullanılmaktadır. Ancak, veride çok fazla bağımsız değişkenin bulunması veya geniş bir gözlem birimi yelpazesinden alınan verilerde, tüm gözlemlere uygulanan tek bir tahmin modeli bağımlı ve bağımsız değişkenler arasındaki ilişkiyi yeterli derecede açıklayamayabilir. Bu durumlarda, ağaç yapısına sahip modellerin kullanımı önerilmektedir (Yang, 1993).

Karar ağaçları, değişkenler arasındaki ilişkiyi tanımlamak için tek bir model yeterli oluncaya kadar veriyi yinelemeli olarak bölümleyen yöntemlerdir. Karar ağaçları akış şemasına benzer bir yapıdadır ve sıklıkla kullanılan öğrenme modellerinden biridir. Bu modellerde her gruba düğüm adı verilmektedir ve her düğüm evet veya hayır cevabı içeren bir mantık sorusu ile alt düğümlere ayrılmaktadır. Her gözlem kök düğümde başlayan yaprak düğümde sonlanan bir yol izler. Ağaç tabanlı modellerin karışık veri tipleri ve eksik veriler ile çalışabilmesi, ilgisiz değişkenlerle baş edebilme özelliği, büyük veri setlerine uygulanabilmesi, kolay anlaşılır ve kolay yorumlanabilir olması gibi birçok avantajı bulunmaktadır (Greenwell, 2022: 10-12). Karar ağaçları oluşturulurken temelde özyinelemeli bölümleme yolunu izlemektedir ancak ayırma değişkenini seçim yaklaşımı, ağacın boyutunu belirleme yöntemleri ve düğümlere uydurulan modeller açısından farklılık göstermektedir (Murthy, 1998; Loh, 2014; Costa ve Pedreira, 2023). Bağımlı değişkenin sayısal olması durumunda ağaçlar regresyon ağacı olarak adlandırılmaktadır.

Regresyon ağacı algoritmalarının, anket verilerini analiz ederken değişkenler arasındaki etkileşimi ortaya çıkarmayı amaçlayan ve bağımlı değişkeni tahmin etme yeteneğini en üst düzeye çıkarmak için veriyi alt gruplara ayıran Otomatik Etkileşim Tespiti (Automatic Interaction Detection- AID) algoritması ile başladığı bilinmektedir (Morgan ve Sonquist, 1963).

Breiman, Friedman, Olshen ve Stone (1984) Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees- CART) algoritmasını geliştirmiştir. AID

algoritmasında olduğu gibi düğüm safsızlık kriteri olarak düğüme ait tahmin edilen bağımlı değişkenin değerleri ile gerçek bağımlı değişkenin değerleri arasındaki farkların kareleri toplamını almaktadır. Safsızlık, ağacın bir düğümündeki bağımlı değişkenin homojenliğini ölçmek için kullanılmaktadır. Algoritmaların amacı safsızlığı en aza indirmektir. CART algoritması iki alt düğümün safsızlık toplamını en küçük yapan ayırmayı seçmektedir. Ciampi (1991) düğüm modellerini genelleştirilmiş lineer modellere uyarlayarak CART algoritmasını genişletmiş, Poisson regresyon ağaçları için safsızlık kriteri olarak sapmayı (deviance) kullanmıştır.

Chaudhuri, Huang, Loh ve Yao (1994) Düzleştirilmiş ve Düzleştirilmemiş Parçalı Polinom Regresyon Ağaçları (Smoothed and Unsmoothed Piecewise-Polynomial Regression Trees-SUPPORT) algoritmasını geliştirmiştir. Bu algoritma öncekilerden farklı olarak ayırma değişkenini artıkların dağılımını analiz ederek seçmektedir. Ayırma noktası olarak ayırma değişkeninin ortalamasını alan algoritma, düğümlere sabit bir değer atamak yerine bir lineer model uydurmakta ve budama yaparken çapraz doğrulamalı çok adımlı ileriye dönük durdurma kuralı uygulamaktadır.

Chaudhuri, Lo, Loh ve Yang (1995) SUPPORT algoritmasını Poisson regresyon ağaçlarına uyacak şekilde genişletmiştir. Düğümlere uydurulan Poisson modeli ile düzeltilmiş Anscombe artıklarının işaretleri kullanılarak gözlemler iki gruba ayrılmakta ortalama ve varyans farklılıklarına göre ayırma değişkeni seçilmektedir.

Loh (2002) Genelleştirilmiş, Yansız Etkileşim Tespiti ve Tahmini (Generalized, Unbiased Interaction Detection and Estimation- GUIDE) algoritmasını geliştirmiştir. Her bir bağımsız değişken ve değişkenlerin ikili kombinasyonu için regresyon modelinde belirlenen artıkların işaretlerine göre oluşturulan çapraz tablolar ile χ^2 dağılımına göre hesaplanan en küçük p değerine sahip değişken ayırma değişkeni olarak seçilmektedir. Alt düğümlere uydurulan regresyon modellerinde toplam hata kareler toplamını veya artık işaretlerine göre gruplandırıldığında iki alt düğümdeki varyansların toplamını en aza indiren ayırma değişkeninin değeri de ayırma noktası olarak seçilmektedir. Değişken seçim yanlılığını kontrol etmek için bootstrap kalibrasyonu yapmaktadır. Loh (2006) GUIDE algoritmasını Poisson regresyon ağacı için genişletmiştir. SUPPORT algoritmasında olduğu gibi düğümlere uydurulan Poisson modelinin Anscombe artıklarını ayırma değişkeni seçiminde kullanmaktadır.

Sayma verilerinde aşırı yayılım durumları ile baş edebilmek için Choi, Ahn ve Chen (2005) GUIDE ve SUPPORT metodolojisini kullanarak yarı en çok olabilirlik yaklaşımı ile düğümlere ekstra Poisson modelinin uydurulduğu, ayırma kriteri olarak sapma fonksiyonunun kullanıldığı ve çok adımlı ileriye dönük durdurma kuralının uygulandığı bir prosedür önermiştir.

Lee ve Jin (2006) sıfır yığılmalı Poisson (Zero-Inflated Poisson- ZIP) için CART algoritması ile düğüm safsızlık kriteri olarak ZIP regresyon olabilirliğini kullanan Sıfır yığılmalı Poisson ağaçları (ZIP tree) metodunu önermiştir. Model karmaşıklığını belirlemek için ön budama yapmaktadır.

Zeileis, Horthorn ve Hornik (2008a) düğümlere parametrik bir model uydurarak oluşturulan Model Tabanlı Özyinelemeli Bölümleme (Model-based Recursive Partition-MOB) algoritmasını geliştirmiştir. Düğümlere Poisson, yarı-Poisson ve negatif binom da dahil olmak üzere birçok genelleştirilmiş lineer model uydurulabilmektedir (Rusch ve Zeileis, 2013). Ayırma değişkenini parametre kararsızlığı testi ile belirlemektedir. Hata kareler ortalamasını veya negatif log olabilirliği yerel olarak optimize eden ayırma noktası seçilmektedir. Anlamlı bir kararsızlık tespit edilmeyinceye kadar prosedür tekrarlanmakta ve son budama işlemi yapılmamaktadır.

Cao (2019), GUIDE algoritmasının ayırma kuralını kullanarak sıfır yığılmalı ve aşırı yayılmış Poisson modelleri için CounTree algoritmasını önermiştir. Düğüm modelinde LASSO düzenlemesini kullanmaktadır. Ayırma değişkeni seçiminde, düğüm modeli Poisson veya negatif binom ise düzeltilmiş Anscombe artıklarının işaretlerini; ZIP ise, Pearson artıklarının işaretlerini kullanmaktadır. Ayrıca GUIDE algoritmasından farklı olarak belirli bir eşiğin üzerindeki test istatistiğine sahip değişkenleri önem sırasına dizerek en büyük istatistiğe sahip değişkeni seçmektedir.

Liu, Lin ve Shih (2020) düğümlere Poisson, negatif binom, engel (hurdle) veya ZIP regresyon modelinin uydurulduğu Sayma Regresyon Ağaçları (Count Regression Tree-CORE) metodunu önermiştir. CORE metodu ayırma değişken seçiminde GUIDE algoritmasını temel almaktadır.

Ağaç tabanlı yöntemler sıklıkla kullanılan makine öğrenmesi yöntemlerinden biridir. Regresyon ağaçları da düğümlere istatistiksel modellerin uydurulmasının mümkün olduğu ağaç tabanlı yöntemlerdir. Sayma verileri için regresyon ağaçlarında düğümlere Poisson modelinin uydurulması uygun olmaktadır. Değişken seçim yanlılığı bir ağaç algoritmasının modelde bulunan herhangi bir değişkeni yansız olarak seçmesi ile ifade edilmektedir. Seçim yanlılığı hatalı çıkarımlara yol açabilmektedir. Bir ağacın bölünmemesi gerekirken bölünmesi tip 1 hata, bölünmesi gerekirken bölünmemesi ise güç olarak tanımlanabilir. Bu çalışmada, Poisson regresyon modeli kullanan ağaç tabanlı yöntemlerden üç algoritmanın (CART, GUIDE ve MOB) performanslarını değerlendirmek ve karşılaştırmak amacıyla simülasyon verileri üzerinden değişken seçim yanlılığı, tip 1 hata ve güçleri incelenmiştir. Gerçek veriler üzerinden sayma verilerine uygun olan regresyon modellerinin ve regresyon ağaçlarının uygulanması yapılmıştır. Ayrıca gerçek veriler temel alınarak bir simülasyon çalışması daha gerçekleştirilmiş, benzer bir gruptan alınan yeni verilere ne kadar uyum sağladığını görmek adına üç algoritmanın tahmin performansları ve güçleri karşılaştırılmıştır.

2. SAYMA VERİSİ

Sayma verisi, bir olayın belirli bir süre içerisinde kaç kez gerçekleştiğini ifade eder. Sayma verilerine; uçak kazalarının sayısı, bir müşterinin poliçe talep sayısı, ödenmemiş kredi taksitlerinin sayısı, epilepsi nöbeti sayısı, iki dakikalık konuşma sırasındaki duraklamaların sayısı gibi örnekler verilebilir. Olay sayısı, negatif olmayan tam sayı değerli bir rastgele değişkendir (Cameron ve Trivedi, 1998: 1).

Sayma verileri için temel model Poisson dağılımıdır. Temel Poisson modelinin varsayımları aşağıdaki gibi sıralanmaktadır.

1. Dağılım, ortalaması μ ile gösterilen tek bir parametreye sahip kesikli bir olasılık dağılımıdır. Bu parametre her bir zaman, alan veya hacim birimi başına meydana gelmesi beklenen olay sayısıdır.
2. Bağımlı değişken *Y* negatif olmayan tam sayı değerlerini alır.
3. Gözlemler birbirinden bağımsızdır.
4. Modelin ortalaması ve varyansı tam olarak veya hemen hemen aynıdır. $\mathrm{E}(Y) = V(Y) = \mu t$

   Burada t, olayların meydana geldiği süre, alan veya hacim birimini belirtmektedir.
5. Gözlenen değerlerin hiç birisi dağılımın parametresine bağlı olarak beklenenden önemli ölçüde daha fazla ya da daha az değildir.
6. Gözlenen varyans ve tahmin edilen varyans ile Pearson χ^2 yayılım istatistiği 1'e yakın bir değere sahiptir (Hilbe, 2014: 37).

Kesikli rastgele *Y* değişkeni μ parametresi ile Poisson dağıldığında *Y* değişkeninin olasılık fonksiyonu

$$\mathrm{P}(Y = y) = \frac{e^{-\mu t}(\mu t)^{y}}{y!}, \qquad y = 0,1,2,... \tag{2.1}$$

ile gösterilmektedir.

Periyot uzunluğu *t*, 1'e eşit olduğunda *Y* değişkeninin olasılık fonksiyonu

$$\mathrm{P}(Y=y)=\frac{e^{-\mu}\mu^{y}}{y!}, \qquad y=0,1,2,... \tag{2.2}$$

şeklinde elde edilmektedir.

2.1. Aşırı Yayılım

Poisson dağılımının ortalama ve varyansının eşit olma özelliği eşit yayılım olarak adlandırılmaktadır. Ancak gerçek hayat verilerinde sıklıkla bu özellik gözlenmemektedir. Varyans ortalamayı aştığında aşırı yayılım, varyansın ortalamadan daha az olduğu durumda ise az yayılım görülmektedir.

Pearson artıklarının karesinin toplamının serbestlik derecesine bölünmesiyle elde edilen değer Pearson yayılım istatistiği olarak adlandırılır. Gözlenen ve tahmin edilen varyanslar eşit olduğunda yayılım istatistiği 1'e yakın bir değer alır. Pearson yayılım istatistiği 1'den küçükse az yayılım, 1'den büyükse aşırı yayılım olduğu ifade edilebilir. Ancak aşırı yayılım olduğunu belirtmek için 1'in ne kadar üzerinde olması gerektiğine dair bir kriter bulunmamaktadır. Gözlem sayısının büyüklüğüne göre yayılım istatistiğinin aldığı değerler farklı yorumlanabilir. Aşırı yayılım; bağımlı değişkenler arasındaki korelasyondan, bağımlı değişkenlerin olasılıkları veya sayıları arasındaki aşırı varyasyondan veya gözlemlerin bağımsızlık varsayımının ihlalinden kaynaklanabilir. Aşırı yayılım, standart hataların beklenenden daha düşük tahmin edilmesine sebep olabilir.

Temel Poisson dağılım varsayımları ihlali açıkça gözükmemesine rağmen yayılım istatistiği 1'den önemli ölçüde farklı ise model görünürde aşırı yayılım gösteriyor olabilir. Regresyon modeli önemli bağımsız değişkenleri veya yeterli etkileşim terimini içermediği ya da aykırı değerleri içerdiği zaman görünürde aşırı yayılım meydana gelebilir. Görünürde aşırı yayılım bir bağımsız değişkenin farklı bir ölçekle dönüştürülmesi gerektiği, toplanan verinin yetersiz olduğu veya eksik değerlerin rastgele dağılmadığı durumlarda da ortaya çıkmaktadır. Görünürde aşırı yayılım çeşitli yollarla telafi edilebilir. Ancak gerçek aşırı yayılım hem model parametre tahminlerinin hem de genel uyumun güvenilirliğini etkileyen bir sorundur (Hilbe, 2011: 141-158).

Poisson modelinde bağımlı değişkenin ortalama ve varyansın yaklaşık aynı değerde olduğu varsayılmaktadır. Bu durum eş yayılım olarak adlandırılmaktadır. Bunun tersi ise aşırı yayılım olarak adlandırılmaktadır. Aşırı yayılımın tespiti için Pearson yayılım istatistiğinin yanı sıra skor ve Lagrange çarpanı testi de kullanılabilir. Lagrange çarpanı testi

$$LM = \frac{\left(\sum_{i=1}^{n} \mu_i^2 - n\hat{y}_i \right)^2}{2\sum_{i=1}^{n} \mu_i^2} \tag{2.3}$$

1 serbestlik dereceli Ki-kare testidir (Hilbe, 2014: 84-87).

$\mathrm{g}(.)$ bir fonksiyon ve $\mathrm{V}(Y) = \mu + \alpha\,\mathrm{g}(\mu)$ olmak üzere; ortalama ve varyans eşitliği,

$$\begin{aligned} H_0 &: \alpha = 0 \\ H_1 &: \alpha \neq 0 \end{aligned} \tag{2.4}$$

hipotezleri altında aşırı yayılım için $\alpha = 0$ olup olmadığı test edilir. Student *t* dağılımına uygun olduğu varsayılan skor testi

$$z = \frac{(y-\mu)^2 - y}{\mu} \tag{2.5}$$

ile tanımlanmaktadır (Cameron ve Trivedi, 1990). Uygulama bölümünde uygulanan test, R programı (R Development Core Team, 2023) "AER" paketinde (Kleiber ve Zeileis, 2022; Kleiber ve Zeileis, 2008) bulunan aşırı yayılım testi, Eş. 2.5' teki formülü kullanmaktadır.

2.2. Sıfır Yığılma

Verilerde gözlenen sıfır değerleri beklenen sıfır değerlerinden fazla olduğunda sıfır yığılma durumu ortaya çıkmaktadır. Poisson dağılımının özelliklerinden biri olan, bir olayın çok sayıda denemenin herhangi birinde gerçekleşme olasılığının küçük olduğu nadir olaylar yasası nedeniyle sıfır olay sayısının baskın olduğu durumlar sıklıkla gözlenmektedir (Cameron ve Trivedi, 1998: 97).

Veride, elde edilmesi mümkün olmayan durumlarda da sıfır değerleri gözlemlenmektedir. Bu durumda gözlem değerinin aldığı sıfır yapısal sıfır olarak adlandırılır. Örneğin bir alkol araştırmasında bir hafta boyunca alkol kullanılan günler ele alındığında içki içmekten kaçınan alt gruptan gelen sıfır yapısal sıfırı; diğer gözlem birimlerinden gelen sıfır örnekleme varyasyonu sonucu elde edilen sıfırları temsil eder. Yapısal sıfırları, rastgele ortaya çıkan örnekleme sıfırlarından ayırt etmek psikososyal çalışmalarda önemli bir yere sahiptir (He, Tang, Wang ve Crits-Christoph, 2014).

ZIP için skor testi, p sıfır yığılmanın olasılığını göstermek üzere

$$H_0 : p = 0 \\ H_1 : p \neq 0 \tag{2.6}$$

hipotezleri altında gerçekleştirilmektedir. Skor istatistiği

$$\frac{\left(n_0 - ne^{\hat{\mu}}\right)^2}{ne^{\hat{\mu}}\left(1-e^{\hat{\mu}}\right) - n\bar{y}\left(e^{\hat{\mu}}\right)^2} \tag{2.7}$$

ile bulunmaktadır. Burada n gözlem sayısı, n_0 ise sıfır olan gözlem sayısıdır. ZIP için elde edilen skor istatistiği χ^2 dağılımına uymaktadır (Broek, 1995; Yang, Hardin ve Addy, 2010).

3. REGRESYON MODELLERİ VE KARAR AĞAÇLARI

Bu çalışmada, yürütülen simülasyon çalışmaları için sayma verilerine uygun olan regresyon ağaçları ve ilgili algoritmalar; uygulama bölümünde sayma verilerinde kullanılan regresyon modelleri ve karar ağacı algoritmaları kullanılmıştır.

3.1. Sayma Verisi İçin Regresyon Modelleri

Aktüerya, biyoistatistik, ekonomi, demografi ve sosyal bilimler gibi çeşitli alanlarda sayma verileri regresyon modelleri kullanılmaktadır.

Veriler, Poisson dağılımının varsayımları karşılanıyorsa standart Poisson modeli ile modellenebilir. Karşılanmıyorsa, ihlal edilen varsayım türüne göre seçilen alternatif bir sayım modeli kullanılmaktadır.

3.1.1. Poisson regresyon modeli

Sayma verileri için standart olarak doğrusal olmayan regresyon modeli olan Poisson regresyon modeli kullanılmaktadır ve standart olarak kesitsel verilere uygulandığı bilinmektedir. $\mathbf{X}$ değişkenleri verilmişken Poisson dağılımına uyan Y değişkeninin dağılımı

$$f\left(Y=y\right)=\frac{e^{-\mu}\mu^{y}}{y!}, \qquad y=0,1,2,... \tag{3.1}$$

ile gösterilmektedir.

Bağımlı değişken y, ilgilenilen olayın oluşum sayısını; $\mathbf{X}$, y'yi belirlediği düşünülen doğrusal bağımsız değişkenlerin vektörünü belirtmektedir. Log-lineer model için ortalama parametresi $\mu=\exp\left(\mathbf{X}'\boldsymbol{\beta}\right)$ üstel ortalama fonksiyonu olarak adlandırılmaktadır.

$y_1,...,y_n$ bağımsız rastgele değişkenler olmak üzere her y_i 'nin dağılımı Eş. 3.1'de verilen, ortalaması μ_i $(i=1,...,n)$ olan Poisson dağılımına sahip olsun. Bağımsız gözlemler verilmişken, parametre değerlerini elde etmek için kullanılan log- olabilirlik fonksiyonu

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \mathbf{X}_i' \boldsymbol{\beta} - \exp(\mathbf{X}_i' \boldsymbol{\beta}) - \ln y_i ! \right\} \tag{3.2}$$

ile gösterilmektedir. Log-olabilirlik fonksiyonunun katsayılara göre birinci dereceden kısmi türevi skor fonksiyonunu

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(y_i - \exp(\mathbf{X}_i' \boldsymbol{\beta}) \right) \mathbf{X}_i' \tag{3.3}$$

vermektedir (Cameron ve Trivedi, 1998: 61-62). Poisson maksimum olabilirlik tahmin edicisi $\hat{\boldsymbol{\beta}}$, skor fonksiyonunun sıfıra eşitlenip çözülmesi ile elde edilmektedir.

$$\sum_{i=1}^{n} \left(y_i - \exp(\mathbf{X}_i' \boldsymbol{\beta}) \right) \mathbf{X}_i' = 0 \tag{3.4}$$

Log-olabilirlik fonksiyonunun ikinci türevi Hessian matrisini

$$\mathbf{H} = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \left(\exp(\mathbf{X}_i' \boldsymbol{\beta}) \right) \mathbf{X}_i \mathbf{X}_i' \tag{3.5}$$

vermektedir. Modeldeki parametrelerin standart hataları, Hessian matrisi tersinin negatif işaretlisinin $(-\mathbf{H}^{-1})$ köşegenlerinin karekökü ile elde edilmektedir.

Sayma verisi modellerini tahmin etmek için Fisher puanlama yöntemine dayalı, yinelemeli olarak yeniden ağırlıklandırılmış en küçük kareler (Iteratively reweighted least squares (IRLS)- YAEKK) yöntemi ve Newton-Raphson tipi maksimum olabilirlik yöntemi olmak üzere iki yöntem kullanılmaktadır. YAEKK maksimum olabilirliğin bir alt türüdür. $\boldsymbol{\beta}$ katsayı tahminlerinin hesaplanması için standart metot Newton-Raphson yöntemidir.

$\mathrm{g} = \partial \ln \mathrm{L} / \partial \boldsymbol{\beta}$ olmak üzere $s+1.$ iterasyon

$$\hat{\boldsymbol{\beta}}_{s+1} = \hat{\boldsymbol{\beta}}_s - \mathbf{H}_s^{-1} \mathrm{g}_s \tag{3.6}$$

ile hesaplanmaktadır (Cameron ve Trivedi, 1998: 62, 94).

Newton-Raphson metodu ile maksimum olabilirlik tahmini algoritması

Poisson regresyon modelinin Newton-Raphson metoduna göre tahmin algoritmasının adımları aşağıda verilmiştir.

1. Tahmin için **β** katsayılarının başlangıç değerleri belirlenir. Başlangıç değerleri, lineer regresyon modeli uydurularak elde edilebilir.
2. Gradient, log-olabilirlik fonksiyonunun **β** katsayılarına göre birinci türevi, (Eş. 3.3) ve Hessian matrisi (Eş. 3.5) hesaplanır.
3. Eş. 3.6 ile yeni katsayılar elde edilir.
4. Elde edilen yeni katsayılar ile eski katsayılar arasındaki uzaklık ölçülür. Uzaklık önceden belirlenen bir tolerans değerinin üzerinde ya da iterasyon sayısı belirlenen bir değerin altında ise yeni katsayı değerleri ile algoritmanın 2. adımına dönülür.
5. Katsayılar arasındaki fark belirlenen değerden küçük olduğunda algoritma durur. Elde edilen son değerler modelin katsayı tahminini vermektedir.

Fisher puanlama Newton Raphson metodunun bir formudur. Log olabilirlik fonksiyonunun katsayılara göre ikinci türevinin gözlenen değeri yerine beklenen değeri olan Fisher bilgi matrisi kullanılmaktadır.

Newton-Raphson ve Fisher puanlama metoduna göre Poisson modeli maksimum olabilirlik tahmini için R programı (R Development Core Team, 2023) ile iki farklı fonksiyon oluşturulmuştur. Fonksiyonların kodları EK-1'de verilmiştir.

Yazılan kodlar ve Poisson regresyonu için kullanılan, R programında (R Development Core Team, 2023) bulunan "glm" fonksiyonu örnek bir veri setine uygulanmıştır. Örnek veri seti olarak uygulama bölümünde de kullanılan doktora makale sayısı verisi kullanılmıştır. Başlangıç **β** değerleri β_0 için bağımlı değişkenin ortalaması, $\beta_i \quad i = 1,...,5$ katsayıları için 0 olarak alınmıştır. Fonksiyonların çıktıları Çizelge 3.1, Çizelge 3.2 ve Çizelge 3.3'te verilmiştir.

Çizelge 3.1. Fonksiyonların katsayı tahminleri ve standart hataları

Katsayılar	glm fonksiyonu		Newton-Raphson fonksiyonu		Fisher puanlama fonksiyonu	
	Tahmin	Standart Hata	Tahmin	Standart Hata	Tahmin	Standart Hata
sabit terim	0,2656	0,0996	0,2656	0,0996	0,2656	0,0996
kadın	-0,2244	0,0546	-0,2244	0,0546	-0,2244	0,0546
evli	0,1573	0,0613	0,1573	0,0613	0,1573	0,0613
çocuk sayısı	-0,1849	0,0401	-0,1849	0,0401	-0,1849	0,0401
prestij	0,0254	0,0253	0,0254	0,0253	0,0254	0,0253
danışman yayın sayısı	0,0252	0,0020	0,0252	0,0020	0,0252	0,0020

Çizelge 3.2. Newton-Raphson algoritması iterasyonlarına göre **β** katsayıları

İterasyon	β_0	β_1	β_2	β_3	β_4	β_5
Başlangıç	0,5264408	0	0	0	0	0
1	0,2685318	-0,2243426	0,1579373	-0,1721126	0,0086444	0,0358844
2	0,2638107	-0,2244702	0,1583320	-0,1820726	0,0200037	0,0275850
3	0,2655905	-0,2244245	0,1573241	-0,1847977	0,0251234	0,0253357
4	0,2656246	-0,2244247	0,1573241	-0,1849141	0,0253771	0,0252309
5	0,2656247	-0,2244247	0,1573240	-0,1849143	0,0253776	0,0252307

Çizelge 3.3. Fisher puanlama algoritması iterasyonlarına göre **β** katsayıları

İterasyon	β_0	β_1	β_2	β_3	β_4	β_5
Başlangıç	0,5264408	0	0	0	0	0
1	0,3459045	-0,1570398	0,1105561	-0,1204789	0,0060511	0,0251191
2	0,2613806	-0,2280589	0,1600519	-0,1877043	0,0264372	0,0252268
3	0,2656339	-0,2244179	0,1573195	-0,1849090	0,0253757	0,0252307
4	0,2656247	-0,2244247	0,1573240	-0,1849143	0,0253776	0,0252307

Regresyon modelinin veriye uyumunu değerlendirmek için çeşitli testler kullanılmaktadır. Poisson regresyonu ile kullanılan uyum iyiliği testlerinden sapma uyum iyiliği testinin temelinde YAEKK yöntemi yakınsaması için kullanılan sapma istatistiği vardır. Sapma, log-olabilirlik fonksiyonunda μ yerine y değerinin konulmasıyla elde edilen doymuş log-olabilirlik ile tam model log-olabilirlik arasındaki fark

$$D = 2\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\mu_i}\right) - \left(y_i - \mu_i\right) \tag{3.7}$$

olarak tanımlanmaktadır (Hilbe, 2014: 75).

Sapma uyum iyiliği testi, sapma istatistiği ve gözlem sayısından parametre sayısının çıkarılmasıyla elde edilen artık serbestlik derecesi ile χ^2 dağılmaktadır. Hesaplanan χ^2 istatistiğinin *p* değeri 0,05'ten küçük ise modelin iyi uyduğu söylenebilir.

Sapma uyum iyiliği istatistiği yerine Pearson χ^2 istatistiği de kullanılmaktadır. Pearson uyum iyiliği testi, sapma uyum iyiliği testi gibi

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y_i - \mu_i\right)^2}{\mathrm{V}\left(\mu_i\right)} \tag{3.8}$$

biçimindedir (Hilbe, 2014: 78).

Modelin uygunluğu için kullanılan bir diğer ölçüt Akaike Bilgi Kriteri (AIC)'dir (Akaike, 1973).

$$AIC = -2L + 2k \tag{3.9}$$

L maksimum log-olabilirliği, *k* ise modeldeki parametre sayısını temsil etmektedir. Daha küçük AIC değerine sahip modelin daha uygun bir model olduğu söylenebilir.

Poisson dağılımının ihlal edilen varsayımına göre:

i. Veride sıfır olma olasılığı olmadığı durumlarda sıfır kesikli Poisson (Zero-Truncated Poisson- ZTP) modeli;
ii. Sıfır yığılma durumunda ZIP modeli;
iii. Belirli bir ortalama ile Poisson olasılık yoğunluk fonksiyonuna göre beklenenden daha çok veya daha az sıfır olması durumunda engel (hurdle) model kullanılır.

Sayma değerleri sayım aralığı içinde olması mümkün değilse kesikli Poisson modeli, herhangi bir nedenle sayım aralığı içinde değilse sansürlü Poisson modeli ya da önceki çözümlerin negatif binom versiyonları kullanılmaktadır (Hilbe, 2014: 40).

Sıfır yığılmalı Poisson modeli

Sıfır yığılmalı modellerde sayıların tahmin edilme mantığı, sıfırların olduğu durum ve sayıların rastgele olduğu durum olmak üzere iki ayrı durum gibi düşünmektir. (Hilbe, 2014: 198).

Beklenenden daha fazla sıfır sayısının gözlendiği sıfır yığılmalı sayma verilerini modellemek için Lambert (1992), baskılı devre kartlarındaki lehim kusurlarının sayısını modellediği çalışmasında ZIP modelini geliştirmiştir. İki bileşenden oluşan modelin ilk bileşenini kusurların nadir olduğu mükemmel bir durum (sıfır durumu), diğerini kusurların mümkün olduğu bir durum olarak tanımlamaktadır.

Lambert (1992), ZIP dağılımını

$$Y = \begin{cases} 0, & p \text{ olasılığı ile} \\ \sim Poisson(\mu), & (1-p) \text{ olasılığı ile} \end{cases}$$

olarak ifade etmektedir. Buradan, μ standart Poisson dağılımının ortalaması olmak üzere

$$\begin{aligned} \mathrm{P}(Y=0) &= p + (1-p)e^{-\mu} \\ \mathrm{P}(Y=y) &= \frac{(1-p)e^{-\mu}\mu^{y}}{y!}, \qquad y = 1,2,\ldots \end{aligned} \tag{3.10}$$

elde edilmektedir. Ortalaması ve varyansı

$$\begin{aligned} \mathrm{E}(Y) &= (1-p)\mu \\ \mathrm{V}(Y) &= (1-p)\mu(1+\mu p) \end{aligned} \tag{3.11}$$

olarak verilen ZIP regresyon modelinin log-olabilirlik fonksiyonu

$$L(\mu, p; y) = \begin{cases} \sum_{i=1}^{n} \log\left(p_i + (1-p_i)\exp(-\mu_i)\right), & y_i = 0 \\ \sum_{i=1}^{n} \left(\log(1-p_i) - \mu_i + y_i \log(\mu_i) - \log(y_i!)\right), & y_i > 0 \end{cases} \tag{3.12}$$

ile belirtilmektedir (Feng, 2021).

3.1.2. Negatif binom regresyon modeli

Negatif binom modeli, aşırı yayılmış verilerin modellenmesinde tek parametreli Poisson modelinden daha fazla esneklik sağlamaktadır. Negatif binom regresyon modeli, ortalama (μ) ve yayılım (α) parametrelerine sahiptir. Ortalama ve varyansı

$$\begin{aligned} &\mathrm{E}(Y) = \mu \\ &\mathrm{V}(Y) = \mu(1+\alpha\mu) \end{aligned} \tag{3.13}$$

ile tanımlanan negatif binom regresyon modelinin olasılık fonksiyonu

$$f(y; \mu, \alpha) = \begin{pmatrix} y + \frac{1}{\alpha} - 1 \\ \frac{1}{\alpha} - 1 \end{pmatrix} \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \tag{3.14}$$

ile belirtilmektedir ve log-olabilirlik fonksiyonu

$$\begin{aligned} L(y; \mu, \alpha) = \sum_{i=1}^{n} y_i \log\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\log(1+\alpha\mu_i) + \log\Gamma\left(y_i + \frac{1}{\alpha}\right) \\ - \log\Gamma(y_i+1) - \log\Gamma\left(\frac{1}{\alpha}\right) \end{aligned} \tag{3.15}$$

ile gösterilmektedir (Hilbe, 2014: 129, 130). α parametresi sıfıra yaklaştıkça model Poisson modeline yaklaşmaktadır.

Geleneksel negatif binom modeli, ikinci parametreye sahip olması haricinde Poisson dağılım varsayımlarına sahiptir. Negatif binom modeli, Poisson modelinden çok daha geniş

bir değişkenlik aralığının modellenmesine izin verir. Aşırı yayılmış Poisson verilerinin parametrelerini tahmin etmek için çoğunlukla negatif binom modeli kullanılmaktadır.

Sıfır yığılmalı negatif binom (ZNB) modeli

Ortalama μ olarak alındığında, tahmin edilen modelin varyansının negatif binom varyansını aştığı durumlar negatif binom aşırı yayılımı olarak tanımlanmaktadır (Hilbe, 2011: 180). Ortalaması ve varyansı

$$\begin{aligned} \mathrm{E}(Y) &= (1-p)\mu \\ \mathrm{V}(Y) &= \mu(1-p)(1+\mu(p+\alpha)) \end{aligned} \tag{3.16}$$

olan ZNB dağılımı

$$\mathrm{P}(Y=y) = \begin{cases} p+(1-p)\left[\left(\dfrac{1/\alpha}{\mu+1/\alpha}\right)^{1/\alpha}\right], & y=0 \\ (1-p)\dfrac{\Gamma(y+1/\alpha)}{\Gamma(1/\alpha)\,y!}\left(\dfrac{\mu}{\mu+1/\alpha}\right)^{y}\left(\dfrac{1/\alpha}{\mu+1/\alpha}\right)^{1/\alpha}, & y>0 \end{cases} \tag{3.17}$$

ile verilmiştir.

Negatif binom modelinde olduğu gibi ZNB modelinin yayılım parametresi α sıfıra yaklaştıkça model, ZIP modeline yakınsar. ZNB regresyon modelinin log olabilirlik fonksiyonu

$$L(\mu;y,\alpha) = \begin{cases} \displaystyle\sum_{i=1}^{n}\left[\ln(p_i)+(1-p_i)\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha}\right], & y_i=0 \\ \displaystyle\sum_{i=1}^{n}\left[\begin{array}{l}\ln(p_i)+\ln\Gamma\left(\dfrac{1}{\alpha}+y_i\right)-\ln\Gamma(y_i+1)- \\ \ln\Gamma\left(\dfrac{1}{\alpha}\right)+\left(\dfrac{1}{\alpha}\right)\ln\left(\dfrac{1}{1+\alpha\mu_i}\right)+y_i\ln\left(1-\dfrac{1}{1+\alpha\mu_i}\right)\end{array}\right], & y_i>0 \end{cases} \tag{3.18}$$

ile gösterilmektedir.

3.2. Karar Ağaçları

Karar ağaçları, bir dizi karar kuralı ve talimat ile verileri yinelemeli olarak bölümleyen hiyerarşik yapıdaki denetimli öğrenme modellerine verilen addır. Denetimli öğrenme hem bağımsız hem de bağımlı değişken değerlerini içeren örnek veriler üzerinden girdi ile çıktı arasındaki fonksiyonu öğrenen herhangi bir makine öğrenimi sürecini ifade eder (Sammut ve Webb, 2017).

Bir karar ağacı; veri araştırması için verinin temel özelliklerini koruyarak özet sağlama, anlamlı olarak yorumlanabilecek sınıflara ayırma ve bağımlı değişkenin gelecekteki değerlerini tahmin etme amacıyla genelleme yapma yollarından bir veya daha fazlasıyla kullanılabilir (Murthy, 1998).

Veri setinin tamamını model uydurmak için kullanmak aşırı uyuma neden olabilir bu da gelecek yeni veriler için kötü tahminler elde edilmesine sebebiyet verebilir. Veri madenciliği ve makine öğrenmesi modellerinin performanslarını ve seçilen modelin kalitesini değerlendirmek için yaygın olarak veri seti, eğitim ve test seti olmak üzere ikiye ayrılmaktadır. Model seçmek ve seçilen modelin tahmin hatasını elde etmek hedeflenmektedir. Eğitim seti, veriye uygun model kurma ve model parametrelerini tahmin etmek için kullanılmaktadır. Test seti ise seçilen modelin tahmin hatasının değerlendirilmesi amacıyla kullanılmaktadır (Hastie, Tibshirani ve Friedman, 2009: 219-223).

Eğitim ve test setlerinin nasıl seçileceği ve seçme oranlarına ilişkin genel bir kural olmamakla birlikte en kolay ve en yaygın kullanım yöntemi rastgele seçim yöntemidir. Rastgele seçim stratejisinin yanı sıra elde edilen setlerin veriyi daha iyi kapsamasını amaçlayan deterministik seçme stratejileri de bulunmaktadır (Joseph ve Vakayil, 2022). Eğitim seti tüm veri setinin %50-%80'i aralığında olacak şekilde değişen bir oran kullanılmaktadır. Sıklıkla eğitim seti veri setinin %80'i, test seti %20'si olacak şekilde seçim yapılmaktadır. Dobbin ve Simmon (2011) değişken sayısının gözlem sayısından büyük olduğu durumları ele aldığı çalışmasında veri setinin $2/3$'ünün eğitim, $1/3$'ünün test seti olmasını önerirken Joseph (2022) parametre sayısıyla ilişkili bir oran $(eğitim/test = \sqrt{p}/1)$ önermektedir.

Karar ağaçları hiyerarşik bir yapı oluşturmaktadır. Eğitim verisinin tamamını içeren ilk düğüme kök düğüm, kendisinden sonra bölünme oluşturmayan en alt sıradaki düğümlere uç veya yaprak düğüm, kök düğüm ile yaprak düğümler arasındaki düğümlere de ara düğüm adı verilmektedir.

Bir karar ağacı, bağımsız ve bağımlı değişkenleri içeren gözlemlerden oluşan eğitim setini kök düğümden itibaren istenilen kriterler sağlanıncaya kadar aynı ayırma prosedürünü uygulayarak bölmeye devam eder. Temelde, bütün uç düğümlerin saf olması (regresyon ağaçları için düğümdeki bütün bağımlı değişken değerlerinin aynı olması), belirlenen maksimum ağaç derinliğine ulaşılması, bir düğümde bulunması gereken minimum gözlem sayısına ulaşılması veya başka herhangi bir bölünmenin genel uyumu belirli bir faktör kadar iyileştirememesi gibi kriterlerden en az biri karşılanırsa bölünme durur (Greenwell, 2022: 69).

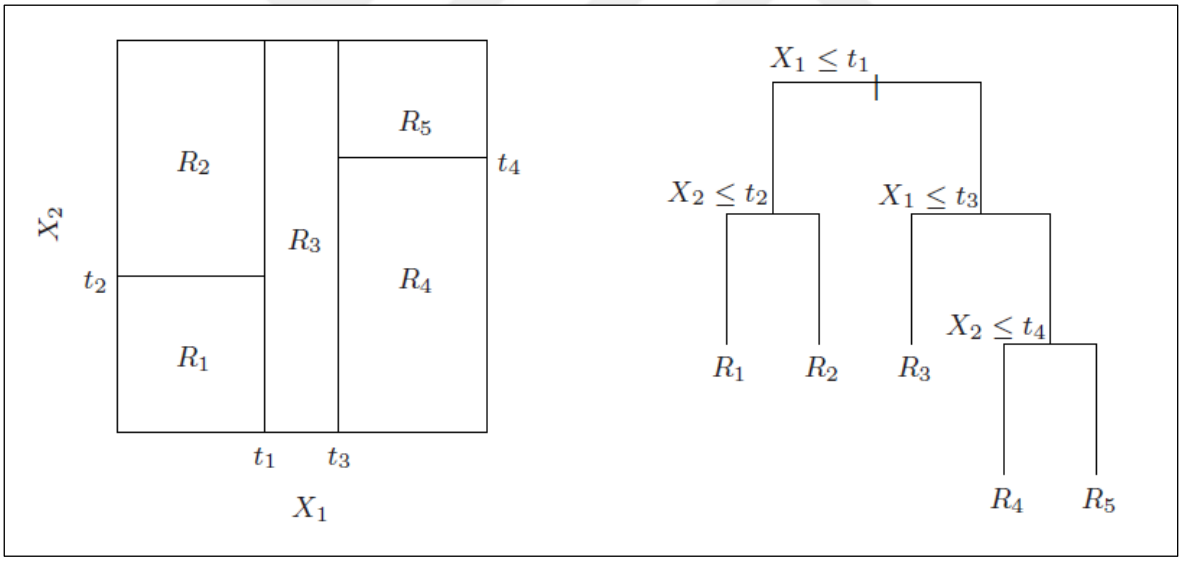
Çok katı durdurma kuralları küçük ve yetersiz ağaçlar oluştururken gevşek durdurma kuralları veriye fazla uyan çok büyük ağaçlar üretilmesiyle sonuçlanır. Uygun ağaç boyutunu belirlemek tahmin hatası performansı açısından önemlidir. Çok büyük ağaç aşırı uyuma, çok küçük ağaç eksik uyuma sebep olmaktadır. Bu durumda yeterince doğru sonuçları veren yeterli büyüklükteki ağaçları elde etmek için budama yapılmaktadır. İki tür budama stratejisi bulunmaktadır. Birincisi, eğitim setine tam uyumlu bir ağaç oluşturulup daha sonra uygun ağaç boyutuna küçültülmesiyle yapılan son budama yöntemi; ikincisi, ağaç oluşturulurken modele yeterli katkıyı sağlamayan dalların modelden çıkarılması ile yapılan ön budama yöntemidir. Karar ağaçlarının budanması için çeşitli yöntemler bulunmaktadır (Rokach ve Maimon, 2015: 69-75).

Ağaç tabanlı yöntemler, özellik uzayını bir dizi dikdörtgene ayıran ve her birine bir model ya da bir sabit yerleştiren basit ama güçlü yöntemlerdir. Her düğüm iki gruba ya da ikiden fazla gruba ayırılabilir. Ancak ikiden fazla düğüme ayırmak, verileri hızlı bir şekilde küçülterek alt düğümlere yetersiz veri bırakacağı için iyi bir strateji olarak görülmemektedir. Çoklu bölmeler ikili bölme ile de elde edilebileceğinden ikili ayırma stratejisi tercih edilmektedir (Hastie ve diğerleri, 2009: 311).

Karar ağaçları, ele alınan problem sınıflandırma ise sınıflandırma ağaçları, regresyon ise regresyon ağaçları olarak adlandırılmaktadır.

İlk ağaç tabanlı regresyon modeli, bağımlı değişkeni tahmin etmedeki hatayı mümkün olduğu kadar azaltacak en iyi alt grup kümesini tanımlamak ve ayırmak için oluşturulan AID algoritması ile tanıtılmıştır (Morgan ve Sonquist, 1963).

Geleneksel karar ağaçlarının basit bir gösterimi için X_1 ve X_2'nin bağımsız, Y'nin bağımlı değişken olduğu bir regresyon problemi ele alınırsa bağımsız değişkenlerin uzayının koordinat eksenlerine paralel çizgilerle bölümlenmesi Şekil 3.1'in solunda gösterilmektedir. Şekil 3.1'in sağında ise ikili yinelemeli bölümleme modeli grafik olarak gösterilmektedir. İlk olarak veri, ayırma değişkeni X_1 ile iki bölgeye ayrılır ve en az hatayı/ en iyi uyumu elde eden değer ayırma noktası olarak seçilir. Daha sonra ayrılan bu bölgeler de aynı prosedür ile bölünür ve bu işlem durdurma kuralı uygulanıncaya kadar devam eder. Bağımlı değişken, her yaprak düğümdeki ($R_1, R_2, ..., R_5$) bağımlı değişkenin ortalamasına göre ayarlanıp modellenir (Hastie ve diğerleri, 2009: 305).



Şekil 3.1. İkili yinelemeli bölümleme

İkiden daha fazla bağımsız değişken olduğunda da ağaçlar benzer şekilde oluşturulmaktadır. Bağımsız değişken uzayını farklı bölgelere ayırarak oluşturulan regresyon modellerinin parçalı sabit ve parçalı lineer olmak üzere iki türü vardır. İkisinin arasındaki temel fark bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiye yaklaşımdadır. Parçalı sabit modelde, her bölümde tahmin edilen çıktı değerinde değişim olmamakta aynı kalmaktadır. Parçalı lineer modelde ise her bölümde çıktı olarak sabit bir

değer kullanmak yerine tahmin edilen çıktı değeri ile bağımsız değişkenler arasında doğrusal bir ilişki takip edilmektedir.

Parçalı sabit modeller her bölüme sabit bir değer atadığı için genellikle daha az düzdür. Diğer taraftan parçalı lineer modeller her bölümde doğrusal bir ilişki ile değer belirlediği için bölümler arası daha pürüzsüz bir geçiş sağlamaktadır. Parçalı sabit modeller yorumlama açısından daha kolay ve sadedir ancak daha az esnektir, küçük farklılıkları yakalayamayabilirler. Parçalı lineer modeller ise daha karmaşıktır ancak girdi ile çıktı arasındaki doğrusal değişikliklere izin verdiği için daha esnektir (Yang, Liu, Tsoka ve Papageorgiou, 2015).

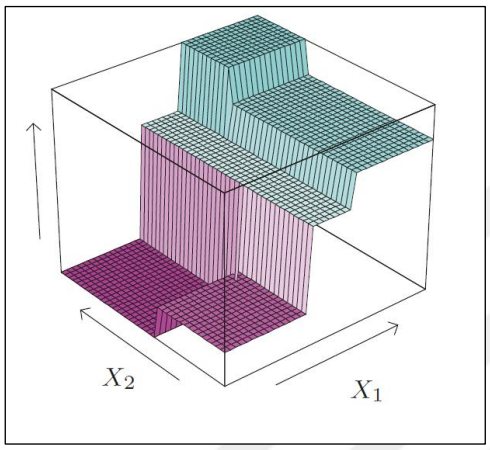
Karar ağaçları ve parçalı regresyonun ikisi de bağımsız değişken uzayını bölümlendirerek verideki doğrusal olmayan ilişkileri yakalamayı amaçlar. Karar ağaçları ağaç gibi bir yapı kullanırken parçalı regresyon, bağımsız değişken uzayını segmentlere ayırır. Karar ağacının her yaprak düğümü bir sabit ile bir segmenti temsil eder. Parçalı regresyon, parçalı lineer karar ağaçlarına benzer şekilde sürekli, parçalı doğrusal ilişkileri temsil etmek için kullanılabilir. Karar ağaçlarını parçalı regresyon teknikleriyle birleştirmek, verilerdeki hem parçalı sabit hem de parçalı doğrusal ilişkileri yakalayan daha güçlü modellere yol açabilir.

AID algoritmasının ortaya çıkışından bu yana sıralı en küçük kareler veya diğer kayıp fonksiyonları ile birçok parçalı sabit ya da parçalı lineer regresyon ağacı modeli geliştirilmiştir. Regresyon ağaçlarının en küçük kareler, Poisson, lojistik, kuantil ve sansürlü regresyon modelleri ile kullanıldığı bilinmektedir (Loh, 2014). Bu çalışmada regresyon ağacı algoritmalarından üçü; geleneksel olan ve hala popülerliğini yitirmeyen CART (Breiman ve diğerleri, 1984), ayırma değişken seçiminde yanlılığı en aza indirmeyi hedefleyen GUIDE (Loh, 2002) ve her düğüme parametrik model uydurarak oluşturulan MOB (Zeileis ve diğerleri, 2008a) algoritmaları incelenecektir.

CART algoritması parçalı sabit, GUIDE algoritması hem parçalı sabit hem parçalı lineer, MOB algoritması ise parçalı lineer regresyon ağaçları üretmektedir. Çalışmada GUIDE algoritması için yorumlaması daha kolay olan parçalı sabit regresyon ağacı seçilmiştir.

3.2.1. CART algoritması

Yaprak düğümlerdeki bağımlı değişkenin değerlerinin ortalamasını tahmin edilen değer olarak alan ve düğüm safsızlık fonksiyonu sapma kareler toplamı olan parçalı sabit regresyon ağaçları üretmektedir. Tahmin değerleri her yaprak düğümde sabit olduğundan ağaç, regresyon yüzeyinin bir histogram tahmini olarak düşünülebilir. (Bkz. Şekil 3.2; Hastie ve diğerleri, 2009)



Şekil 3.2. Parçalı sabit ikili regresyon ağacının üç boyutlu gösterimi

Bir ağaç tabanlı tahmin ediciyi belirlemek için üç element gereklidir:

1. Her ara düğümde bir ayırma noktası seçmenin bir yolu
2. Bir düğümün yaprak düğüm olduğunu belirlemek için bir kural
3. Her yaprak düğüme bir değer atamak için bir kural (Breiman ve diğerleri, 1984: 220)

Eğitim verisinin p bağımsız değişken ve bir bağımlı değişken ile N gözlem sayısına sahip olduğunu, algoritmanın M düğüme ayırdığını ve her düğüme sabit bir değer c_m atadığını varsayarsak model

$$f(X) = \sum_{m=1}^{M} c_m \, \mathrm{I}(x \in R_m) \tag{3.19}$$

olmaktadır.

En küçük kareler toplamı $\sum(y_i - f(x_i))^2$ kriter olarak alındığında, yaprak düğümün en iyi tahmini

$$c_m = \text{ort}\left(y_i \middle| x_i \in R_m\right) \tag{3.20}$$

olmaktadır.

Ayırma noktası için ayırma değişkeninin tüm farklı değerleri ele alınarak uygulanan tam kapsamlı arama yöntemi kullanılmaktadır. Ayırma değişkeni olarak *j*. ayırma değişkenini ve *s* ayırma noktasını düşünürsek alt düğümler

$$R_{sol}(j,s) = \left\{X \middle| X_j \le s\right\}$$
$$R_{sağ}(j,s) = \left\{X \middle| X_j > s\right\}$$

olarak tanımlanır. Bir sonraki adımda

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i \in R_{sol(j,s)}}\left(y_i - c_{sol}\right)^2 + \min_{c_2}\sum_{x_i \in R_{sağ(j,s)}}\left(y_i - c_{sağ}\right)^2\right] \tag{3.21}$$

çözümünü sağlayan *s* ayırma noktası aranır.

En iyi bölünme tespitinden sonra veriler iki bölgeye ayırılır ve ayırma işlemi ortaya çıkan tüm düğümlerde tekrarlanır. Çok büyük bir ağaç yetiştirilip maliyet karmaşıklığı budama stratejisi kullanılarak ağaç budanmaktadır. Sonrasında tahmin hatası tahminine göre budanmış son ağaç seçilir. Yetiştirilen büyük ağacın ara düğümlerinin daraltılmasıyla oluşan yeni ağaç *T*, bu ağacın yaprak düğümlerinin sayısı |*T*| ile gösterilsin ve yaprak düğümleri *m* ile indekslensin.

$$N_m = \#\left\{x_i \in R_m\right\}$$
$$c_m = \frac{1}{N_m}\sum_{x_i \in R_m} y_i \tag{3.22}$$

$$Q_m(T) = \sum_{m=1}^{|T|}\sum_{x_i \in R_m}\left(y_i - c_m\right)^2 \tag{3.23}$$

olmak üzere, maliyet karmaşıklığı

$$C_\alpha(T) = Q_m(T) + \alpha|T| \tag{3.24}$$

ile tanımlanmaktadır.

Burada her karmaşıklık parametresi α için Eş. 3.23'ü en aza indiren ağacı bulmak amaçlanmaktadır. Ayarlama parametresi ($\alpha \geq 0$), büyüdükçe daha küçük ağaç, küçüldükçe daha büyük ağaç üretilmektedir. $\alpha = 0$ olması durumunda budama olmayacak yani tam ağaç olacaktır. Tam ağaçtan kök düğüme kadar düğüm başına Eş. 3.23'teki en küçük artışı sağlayan ara düğüm budanarak en uygun ağaç bulunmaktadır. α'nın tahmini, çapraz doğrulama ile elde edilmekte ve son ağaç bulunmaktadır (Breiman ve diğerleri, 1984: 224-229; Hastie ve diğerleri, 2009: 307, 308).

CART algoritmasını Poisson modellere Ciampi (1991) uyarlamıştır. Alt düğümlerdeki sapmaların toplamını en çok azaltan düğümler seçilir ve AIC değerlerine göre budama yapılır.

CART algoritmasının R programında (R Development Core Team, 2023) uygulanması "rpart" paketi ile olmaktadır (Thernau ve Atkinson, 2023). Poisson regresyonu için ayırma yapılırken hata kareleri toplamı yerine düğüm içi sapma istatistiği

$$\begin{aligned} D &= \sum_{i=1}^{n}\left[c_i \log\left(\frac{c_i}{\hat{\mu}t_i}\right) - \left(c_i - \hat{\mu}t_i\right)\right] \\ \hat{\mu} &= \frac{\text{olay sayısı}}{\text{toplam zaman}} = \frac{\sum_{i=1}^{n} c_i}{\sum_{i=1}^{n} t_i} \end{aligned} \tag{3.25}$$

kullanılmaktadır. Burada c_i i. gözlem için gözlenen olay sayısını ve t_i i. gözlem süresini ifade etmektedir.

Alt düğümlerin sapmalarının toplamının en küçük olması,

$$\text{maks}\left[D_{ebeveyn} - \left(D_{sol} + D_{sağ}\right)\right] \tag{3.26}$$

çözümü, ayırma kriteri olarak alınmaktadır.

3.2.2. GUIDE algoritması

Ayırma değişkeninin seçiminde yanlılık ve yerel etkileşimlere duyarsızlık sorunlarını çözmeyi hedefleyen, parçalı lineer ve parçalı sabit regresyon modelleri oluşturan bir karar ağacı algoritmasıdır. Seçim yanlılığını çok küçük tutması, değişkenler arası etkileşimlere yer vermesi ve her türlü kategorik bağımsız değişkenleri kapsaması GUIDE algoritmasının özellikleri arasında yer almaktadır. Her bir bağımsız değişken ve değişkenlerin ikili kombinasyonu için regresyon modelinde belirlenen artıkların işaretlerine göre çapraz tablolar oluşturularak χ^2 ve p değerleri hesaplanır. En küçük p değerine sahip değişken ayırma değişkeni olarak seçilir. Ayırmadan sonra oluşan alt düğümlere uydurulan regresyon modellerinde toplam hata kareler toplamını veya artık işaretlerine göre gruplandırıldığında iki alt düğümdeki varyansların toplamını en aza indiren ayırma değişkeninin değeri veya değerler kümesi ayırma noktası olarak seçilir. Çok büyük bir ağaç yetiştirilir ve V-katlı çapraz doğrulama ile CART algoritmasında kullanılan maliyet karmaşıklığı budama yöntemi ile ağaç budanır (Loh, 2002).

Çapraz doğrulama, model tahmininin aşırı uyumunu önlemek için kullanılan bir yeniden örnekleme yöntemidir. V-katlı çapraz doğrulamada eğitim verisi, yaklaşık olarak eşit boyutlu, yerine koymadan rastgele seçilen V tane ayrık alt kümeye bölünür. $V-1$ alt küme ile model eğitilir. Doğrulama kümesi olarak adlandırılan kalan 1 kümeye model uygulanır ve performans ölçülür. Uygulama, V alt kümenin her biri doğrulama kümesi olana kadar tekrarlanır. V çapraz doğrulama kümesinden elde edilen ölçümlerin ortalaması çapraz doğrulanan performanstır (Berrar, 2018).

Loh (2006) GUIDE algoritmasını Poisson regresyonu için genişletmiştir. Poisson regresyon modelinde Anscombe artıkları kullanılmaktadır (Cao, 2019).

Loh (2002, 2008: 453-455) GUIDE algoritması parçalı sabit regresyon ağacının adımlarını açıklamaktadır:

1. Ele alınan düğümdeki her ayırma değişkeni için bir regresyon modeli uydurulur. En küçük artık hata kareler ortalamasını sağlayan değişken seçilir. Modelde $R^2 > 0,99$ ise

ya da gözlem sayısı önceden belirlenen, düğümde bulunması gereken minimum gözlem sayısının iki katından az ise algoritma durur. Aksi takdirde bölünme devam eder.

2. Modelin artıklarının işaretine göre gözlemler (pozitif ve pozitif olmayan olarak) iki gruba ayrılır.
3. Ayırma değişken seçimi iki ayrı teste göre değerlendirilir:

a. Eğrilik testi: Sayısal sürekli değişkenler için çeyreklikler sütunda, artık grupları satırda olmak üzere 2×4 boyutlu çapraz tablo oluşturulur. Kategorik değişkenler için kategori sayısı k sütunda olmak üzere $2\times k$ boyutlu çapraz tablo oluşturulur ve toplamı 0 olan sütunlar dahil edilmez. Her bir değişken için çapraz tablolar kullanılarak χ^2 istatistiği ve p değeri hesaplanır.

b. Etkileşim testi: Değişkenlerin ikili kombinasyonları ile çapraz tablolar oluşturulur. Sayısal değerli değişken çiftleri için değişkenler sütunda dört grup, satırda artık işaretleri 2 grup olacak şekilde 2×4 boyutlu çapraz tablolar oluşturulur. Kategorik değişken çiftleri için birinci değişkenin düzey sayısı k_1, ikinci değişkenin düzey sayısı k_2 olmak üzere değişkenlerin düzey sayısının çarpımı kadar sütuna ayrılarak $2\times k_1k_2$ boyutlu çapraz tablolar oluşturulur. Sayısal-kategorik değişken çiftleri için sayısal değişkenler ikiye, kategorik değişkenler düzey sayısına göre gruplara ayrılarak sütunda $2k$, satırda artık işaretlerinin 2 grubu olmak üzere $2\times2k$ boyutlu çapraz tablolar oluşturulur. Her bir değişken çifti için χ^2 istatistiği ve p değeri hesaplanır.

4. Çapraz tablolardan elde edilen en küçük p değeri eğrilik testinden geliyorsa o p değerine ait değişken ayırma değişkeni olarak seçilir. Aksi takdirde sayısal değişken çiftlerinden geliyorsa her değişkenin ortalamasına bölünerek elde edilen alt düğümlerden en küçük toplam hata kareler ortalamasını veren değişken, en az biri kategorik olan değişken çiftlerinden geliyorsa en küçük p değerine sahip olan değişken ayırma değişkeni olarak seçilir.
5. Seçilen ayırma değişkeni sıralı ise $X_j \le s$, kategorik ise $X_j \in A$ olan s veya A değeri ayırma noktası olarak belirlenir. A, ayırma değişkeni X_j'nin aldığı değerlerin bir alt kümesidir.

a. Ayırma değişkeni sıralı ise iki alt düğümde de belirlenen minimum gözlem sayısına sahip her s değeri için elde edilen alt düğümlere ayrı ayrı regresyon modeli uydurulur. Modellerin artık kareler toplamını en küçük yapan değer ayırma noktası olarak belirlenir.

b. Ayırma değişkeni kategorik ise sınıflandırma problemi gibi düşünülür. Alt düğümlerde minimum gözlem sayısını sağlayan her A kümesi için alt düğümlerin örneklem boyutuna göre ağırlıklandırılan toplam artık varyansını en küçük yapan küme ayırma noktası olarak belirlenir.

6. Her alt düğüme 1. adımdan başlanarak algoritma uygulanmaya devam eder.
7. Tam ağaç oluştuktan sonra CART algoritmasında uygulanan V-katlı çapraz doğrulama kullanılarak ağaç budanır. Tahmin hata kareler ortalamasının en küçük çapraz doğrulama tahmininin standart hatasının bir α katı dahilinde olan en küçük alt ağaç seçilir. Varsayılan değer olarak $V=10$, $\alpha=0,5$ alınmaktadır ve buna yarım SE kuralı denilmektedir.

3.2.3. MOB algoritması

Karar ağaçlarındaki yinelemeli bölümleme kavramını istatistiksel modellerle birleştiren modeller sunar. Temel fikir, her düğümün tek bir modelle ilişkili olmasıdır. Her bir yaprağın uygun bir modelle ilişkilendirildiği bir ağaç hesaplanarak, bir parçalı parametrik model uydurulur. Modelin amaç fonksiyonu, parametre tahmini ve bölümleme için kullanılır.

Gözlemleri bazı bağımsız değişkenlere göre bölümlendirerek her bölüme uygun bir model bulmak, tüm gözlemlere uydurulan tek bir modelden daha iyi uyum sağlamaktadır. Ayırma değişkenleri Z uzayının bir bölümüne ait parametre β ise bölümlendirilmiş model

$$M_R\left((X,Y),\{\beta_r\}\right) \tag{3.27}$$

ile gösterilmektedir.

Algoritma 3 adımdan oluşmaktadır:

1. $\hat{\beta}$'yı tahmin ederek geçerli düğümdeki gözlemlere modeli uydurma
2. Her ayırma değişkeni Z'ye göre parametre tahminlerinin kararlılığını değerlendirilerek en yüksek parametre kararsızlığıyla ilişkili Z_j değişkenini ayırma değişkeni olarak seçme

3. Amaç fonksiyonu Ψ'yi yerel olarak optimize eden ayırma noktasını bulma ve düğümü alt düğümlere ayırma

2. adımda istikrarsızlık yoksa bölümlenme durmaktadır. Aksi takdirde aynı prosedür alt düğümlerde de uygulanarak tekrar edilmektedir. Algoritmada ön budama yapılmaktadır.

Sıralı en küçük kareler tahmin edicisi yöntemi için amaç fonksiyonu Ψ, hata kareler toplamı iken maksimum olabilirlik tahmin edicisi için negatif log-olabilirlik olmaktadır.

N gözleme sahip β parametreli bir $M\left((X,Y),\beta\right)$ modelinde parametre tahmini, amaç fonksiyonunun en küçük değerini verecek şekilde hesaplanmaktadır.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \Psi\left((x,y)_i,\beta\right) \tag{3.28}$$

Eş. 3.28 ile tanımlanan $\hat{\beta}$ birinci dereceden koşullar ile hesaplanabilir.

$$\begin{aligned} &\sum_{i=1}^{n} \psi\left((x,y)_i,\hat{\beta}\right) = 0 \\ &\psi\left((x,y),\beta\right) = \frac{\partial \Psi\left((x,y),\beta\right)}{\partial \beta} \end{aligned} \tag{3.29}$$

Burada $\psi\left((\mathrm{x},\mathrm{y}),\beta\right)$, $\Psi\left((\mathrm{x},\mathrm{y}),\beta\right)$'ye karşılık gelen skor fonksiyonu veya tahmin fonksiyonudur.

Bölümleme işleminin ikinci adımında genelleştirilmiş bir M dalgalanma testi (Zeileis ve Hornik, 2007) kullanılarak seçilen ayırma değişkenleri Z_j üzerinden katsayı kararlılığı test edilir. Sayısal ve kategorik ayırma değişkenlerini değerlendirmek için iki farklı test istatistiği kullanılır.

Z_j değişkenine göre ayrılmış iki rakip bölüm, bölümlerdeki amaç fonksiyonları ile karşılaştırılabilir. Ayırma noktaları, maksimum olabilirlik fonksiyonunu yerel olarak optimize eden noktalar olarak hesaplanmaktadır (Zeileis ve diğerleri, 2008a).

4. UYGULAMA VE SİMÜLASYON

Gerçek veri setlerine regresyon ağacı algoritmaları ve regresyon modelleri uygulanmıştır. Ayrıca regresyon ağaçları ile simülasyon çalışmaları gerçekleştirilmiştir.

4.1. Gerçek Veri Setleri

İki farklı gerçek veri seti ile regresyon ağaçları oluşturulmuştur. Seçilen üç regresyon ağacı algoritması ve uygun regresyon modelleri doktora öğrencilerinin makale sayısı verisi ve kaza sayıları verisine uygulanmıştır.

4.1.1. Makale sayısı verisi

Uygulamada kullanılan veri biyokimya alanında doktora yapan öğrencilerin üretkenliğine ilişkin yapılan bir çalışmadan alınmıştır (Long, 1990). altı değişken 915 gözlemden oluşan veri seti R programı (R Development Core Team, 2023) "vcdExtra" paketinde bulunmaktadır (Friendly, 2023). Bağımlı değişken öğrencilerin doktora eğitiminin son üç yılında yayımladığı makale sayısı; bağımsız değişkenler ise cinsiyet, medeni hal, küçük çocuklarının sayısı, doktora bölümünün itibarı ve öğrenci danışmanlarının yayın sayısıdır. Değişkenlerin tanımları Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Makale sayısı veri setinin değişkenleri

Değişken	Açıklama	Değerleri
makale sayısı	Öğrencilerin makale sayısı	Min: 0 Maks: 19
kadın	Öğrencinin kadın olması	0: Hayır 1: Evet
evli	Öğrencinin evli olması	0: Hayır 1: Evet
çocuk sayısı	5 yaşından küçük çocuklarının sayısı	Min: 0 Maks: 3
prestij	Doktora bölümünün prestiji	Min: 1 Maks: 5
danışman yayın sayısı	Danışmanlarının yayın sayısı	Min: 0 Maks: 77

Veride kadın öğrencilerin sayısı 421 ile toplam öğrencilerin %46'sını, evli olan öğrencilerin sayısı 606 ile toplamın %66'sını oluşturmaktadır. Küçük çocuk sayısının ortalaması 0,5; standart sapması 0,76; bölümün prestij puan ortalaması 3,14; standart sapması 1,04; danışmanların ortalama yayın sayısının ortalaması 8,77; standart sapması 9,48'dir. Bağımlı değişken makale sayısının ortalaması 1,69; standart sapması 1,93; varyansı 3,71'dir.

Poisson regresyonu

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \quad (4.1)$$

ile model kurulmuştur. Poisson regresyon modelinin katsayı tahminleri Çizelge 4.2'de verilmiştir.

Çizelge 4.2. Makale sayısı Poisson modeli katsayıları

Katsayılar	Tahmin	$exp(\beta)$ değerleri	Standart hata	p değeri
Sabit	0,2656	1,3042	0,0996	0,0077
kadın	-0,2244	0,7990	0,0546	< 0,0001
evli	0,1573	1,1704	0,0613	0,0102
çocuk sayısı	-0,1849	0,8312	0,0401	< 0,0001
prestij	0,0254	1,0257	0,0253	0,3153
danışman yayın sayısı	0,0252	1,0256	0,0020	< 0,0001

Kurulan Poisson modeline göre diğer değişkenler sabitken kadın öğrencilerin makale sayısı erkek öğrencilerin makale sayısından %20 azdır. Evli olanların makale sayısı evli olmayanlarınkinden %17 daha fazladır. Küçük çocuk sayısındaki her bir artışın öğrencilerin makale sayısında yaklaşık %17'lik bir azalışa ve danışmanın yayın sayısındaki her bir artışın öğrenci makalelerinde %3'lük bir artışa sebep olduğu elde edilmiştir.

Makale sayısı ortalamasının varyansından küçük olduğu görülmektedir. Poisson varsayımlarından biri olan ortalama-varyans eşitliğinin sağlanamadığı söylenebilir. Aşırı yayılım test edildiğinde yayılım parametresi tahmininin 1'den büyük olduğu (1,826) gözlenmiştir. Aşırı yayılımın varlığı reddedilememiştir. Aşırı yayılım durumunda kullanılan modellerden negatif binom regresyon modeli R programında (R Development Core Team,

2023) “MASS” paketi (Ripley ve diğerleri, 2023; Venables ve Ripley, 2002) kullanılarak analiz sonuçları elde edilmiştir. Modelin katsayı tahminleri Çizelge 4.3’te verilmiştir.

Çizelge 4.3. Makale sayısı negatif binom modeli katsayıları

Katsayılar	Tahmin	$exp(\beta)$ değerleri	Standart hata	p değeri
Sabit	0,2129	1,2373	0,133	0,109
kadın	-0,2163	0,8055	0,073	0,003
evli	0,1528	1,1651	0,082	0,062
çocuk sayısı	-0,1763	0,8383	0,053	0,001
prestij	0,0293	1,0298	0,034	0,392
danışman yayın sayısı	0,0287	1,0291	0,003	< 0,001

Negatif binom modeline göre doktora öğrencileri tarafından yayınlanan makalelerin sayısı en çok danışmanlarının yayın sayısından etkilenmektedir. Danışman yayın sayısı arttıkça makale sayısı artmaktadır. Kadınların makale sayısı erkeklerin makale sayısından daha azdır. Küçük çocuklarının sayısı arttıkça makale sayısının azaldığı görülmektedir. Öğrencilerin medeni durumu ve doktora bölümünün prestijinin marjinal etkisi istatistiksel olarak anlamsız bulunmuştur.

Veri setinde makale sayısı değişkeninin 275 gözlemde 0 değerini aldığı, 0 değerine sahip gözlemlerin toplam gözlem sayısının %30’u olduğu görülmektedir. Sıfır yığılma durumunda kullanılan ZIP ile ZNB modelleri de R programı (R Development Core Team, 2023) “pscl” paketi (Jackman, 2023; Zeileis, Kleiber ve Jackman, 2008b) ile kurularak modellerin istatistikleri karşılaştırılmıştır. Modellerin karşılaştırılması Çizelge 4.4’te verilmiştir.

Çizelge 4.4. Makale sayısı verisi regresyon modellerinin karşılaştırması

Modeller	AIC	BIC	LR Ki-kare	Ki-kare p değeri
Poisson	3313,3	3342,3	3301,3	< 0,001
Negatif binom	3135,4	3169,1	3121,4	< 0,001
ZIP	3233,5	3291,3	3209,5	< 0,001
ZNB	3125,8	3188,4	3099,8	< 0,001

Kullanılan modeller karşılaştırıldığında AIC değerine göre en uygun modelin ZNB modeli olduğu görülmektedir. Onu negatif binom modeli izlemektedir. ZNB modelinin katsayıları Çizelge 4.5'te verilmiştir.

Çizelge 4.5. Makale sayısı verisi ZNB modeli katsayıları

Sayma modeli katsayıları (log link ile negatif binom)				
Katsayılar	Tahmin	$exp(\beta)$ değerleri	Standart hata	p değeri
Sabit	0,3655	1,4412	0,140	0,009
kadın	-0,1946	0,8232	0,076	0,010
evli	0,1015	1,1068	0,084	0,229
çocuk sayısı	-0,1516	0,8593	0,054	0,005
prestij	0,0156	1,0157	0,035	0,651
danışman yayın sayısı	0,0244	1,0247	0,003	< 0,001
Sıfır yığılma modeli katsayıları (logit link ile binom)				
Katsayılar	Tahmin	$exp(\beta)$ değerleri	Standart hata	p değeri
Sabit	-0,2943	0,7451	1,283	0,819
kadın	0,6580	1,9309	0,856	0,442
evli	-1,4741	0,2290	0,918	0,108
çocuk sayısı	0,6241	1,8667	0,441	0,157
prestij	-0,0113	0,9888	0,289	0,969
danışman yayın sayısı	-0,8765	0,4162	0,315	0,005

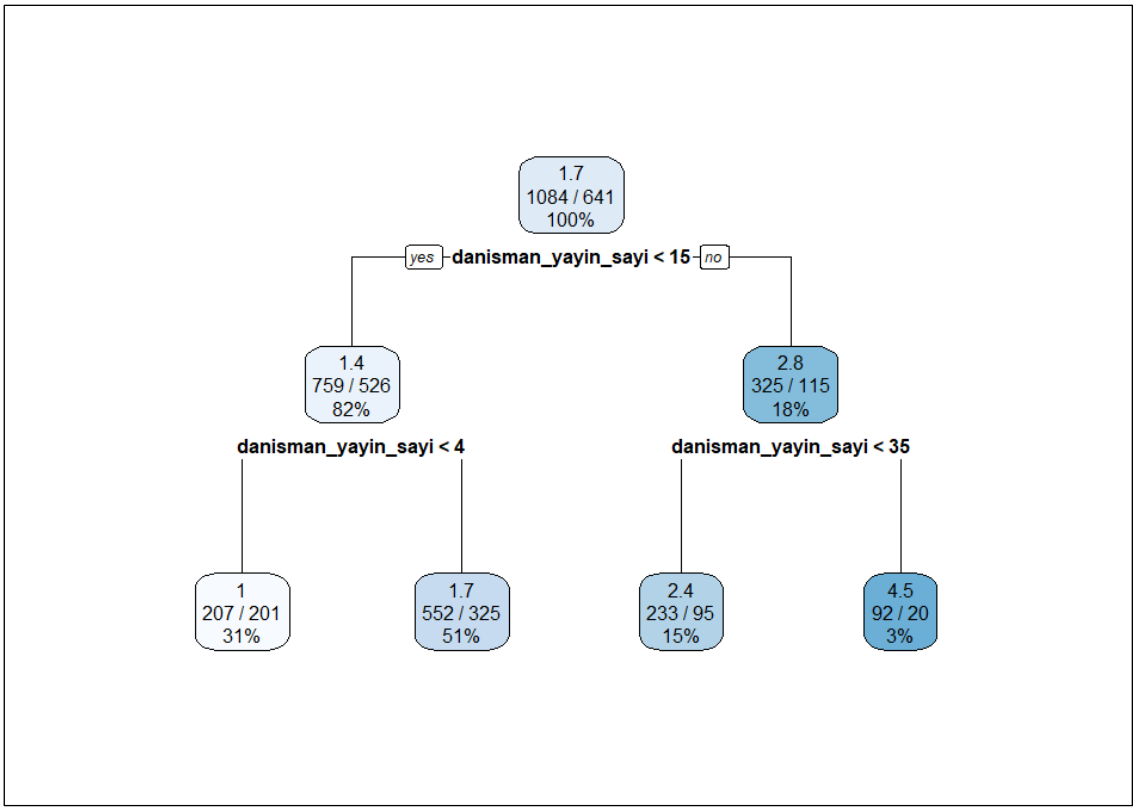
Negatif binom regresyon modeli kısmında danışman yayın sayısı, öğrencinin cinsiyeti ve çocuk sayısı değişkenlerinin istatistiksel olarak anlamlı olduğu görülmektedir. Aşırı sıfırları tahmin eden logit model kısmında danışman yayın sayısı değişkeni istatistiksel olarak anlamlı bulunmuştur. Modele göre diğer değişkenler sabit tutulduğunda kadın öğrencilerin ortalama makale sayısı erkek öğrencilerin ortalama makale sayısından %18 daha azdır. Öğrencilerin çocuk sayısındaki her bir birimlik artış makale sayısında ortalama %14 azalmaya neden olmaktadır. Danışman yayın sayısı bir birim arttığında öğrencilerin makale sayısında ortalama %2 artmaktadır. Theta ($\theta = 1/\alpha$) modelin yayılım parametresini göstermektedir. Theta değeri 2,6553 bulunmuştur. Sıfır yığılma modeline göre danışman yayın sayısındaki bir birimlik artış makale sayısının sıfır olma ihtimalini %58 azaltmaktadır.

Kullanılan modellerden çıkan sonuçlara göre öğrencilerin makale sayısını en çok etkileyen değişkenin danışmanlarının yayın sayısının olduğu görülmektedir.

Makale sayısı verisi ile regresyon ağaçları

Veri seti, 0,7'si eğitim seti, 0,3'ü test seti olarak ayrılmıştır. Eğitim setine CART, GUIDE ve MOB regresyon ağaçları algoritmaları uygulanmış, sonrasında test seti ile modellerin hata kareler ortalamaları incelenmiştir.

Poisson modeline bağlı olarak oluşturulan regresyon ağaçlarından elde edilen grafikler Şekil 4.1, Şekil 4.2 ve Şekil 4.3'te verilmiştir. Grafiklerde düğümlerin altındaki koşulu sağlayan gözlemler sola, sağlamayan gözlemler sağa ayrılmaktadır. CART ve GUIDE karar ağaçlarında yaprak düğümlerin değerlerinin genel ortalamanın üstünde olup olmamasına göre renkleri değişmektedir. Bu durumda, yaprak düğüm ortalaması 1,69'a eşit veya üstünde olan düğümler CART için koyu mavi, GUIDE için turuncu renk ile belirtilmektedir. CART ve GUIDE karar ağacına göre kök düğümü bölen en önemli değişken danışmanın yayın sayısı iken MOB karar ağacına göre öğrencinin cinsiyetidir.

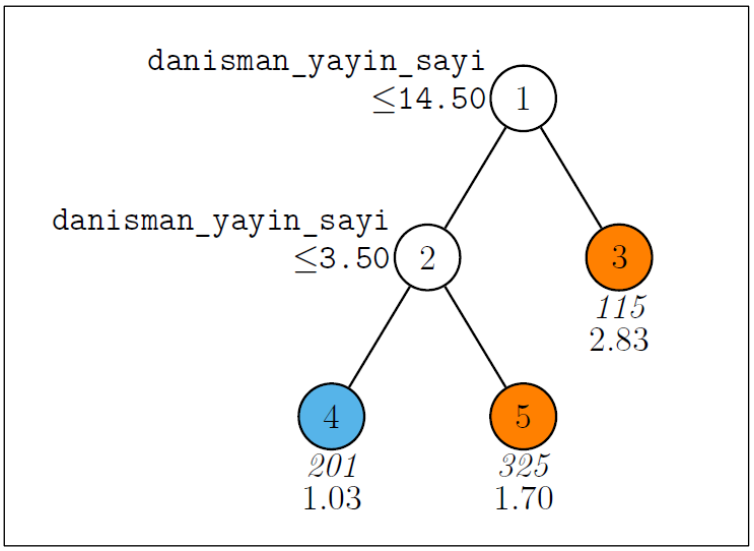


Şekil 4.1. CART algoritması ile makale sayısı regresyon ağacı

CART ile oluşturulan ağaçta, düğümün üstünde düğümün çıktı ortalaması, ortasında düğüme ait toplam makale sayısına gözlem sayısı oranı ve en altında düğümdeki gözlem sayısının toplam gözlem sayısına göre yüzdesi yazmaktadır.

CART algoritmasına göre kök düğümde makale sayısına etki eden en önemli değişkenler sırası ile danışman yayın sayısı ve prestij değişkenleridir. Şekil 4.1'de danışman yayın sayısı değişkeninin öğrencilerin makale sayısını etkileyen en önemli değişken olduğu görülmektedir. Danışman yayın sayısı 35 ve üzerinde olan alt grup, en fazla makale sayısı ortalamasına (4,5) sahiptir. En az makale sayısı ortalamasına sahip olan alt grup da yine danışmanın yayın sayısı ile belirlenmektedir ve danışmanının yayın sayısı 4'ün altında olan öğrenciler ortalama makale sayısı 1'dir.

CART algoritmasına göre kök düğümün ortalaması 1,69; ortalama sapması 1,9974'tür. 526 gözlemin bulunduğu ikinci ara düğümün ortalaması 1,44; ortalama sapması 1,7832'dir. 115 gözlemin bulunduğu ara düğümün ortalaması 2,82; ortalama sapması 2,1695'tir. İkinci ara düğümden çıkan danışman yayın sayısı değişkeninin 4'ten küçük değerlerini içeren, 201 gözleme sahip olan yaprak düğümün ortalaması 1,03; ortalama sapması 1,5621'dir. Danışman yayın sayısı değişkeninin 4'ten büyük değerlerini içeren, 325 gözleme sahip olan yaprak düğümün ortalaması 1,70; ortalama sapması 1,7959'dur. Danışman yayın sayısı değişkeninin 35'ten küçük değerlerini içeren, 95 gözleme sahip yaprak düğümün ortalaması 2,45; ortalama sapması 1,6445'tir. Danışman yayın sayısı değişkeninin 35'ten büyük değerlerini içeren, 20 gözleme sahip yaprak düğümün ise ortalaması 4,52; ortalama sapması 3,4849'dur.

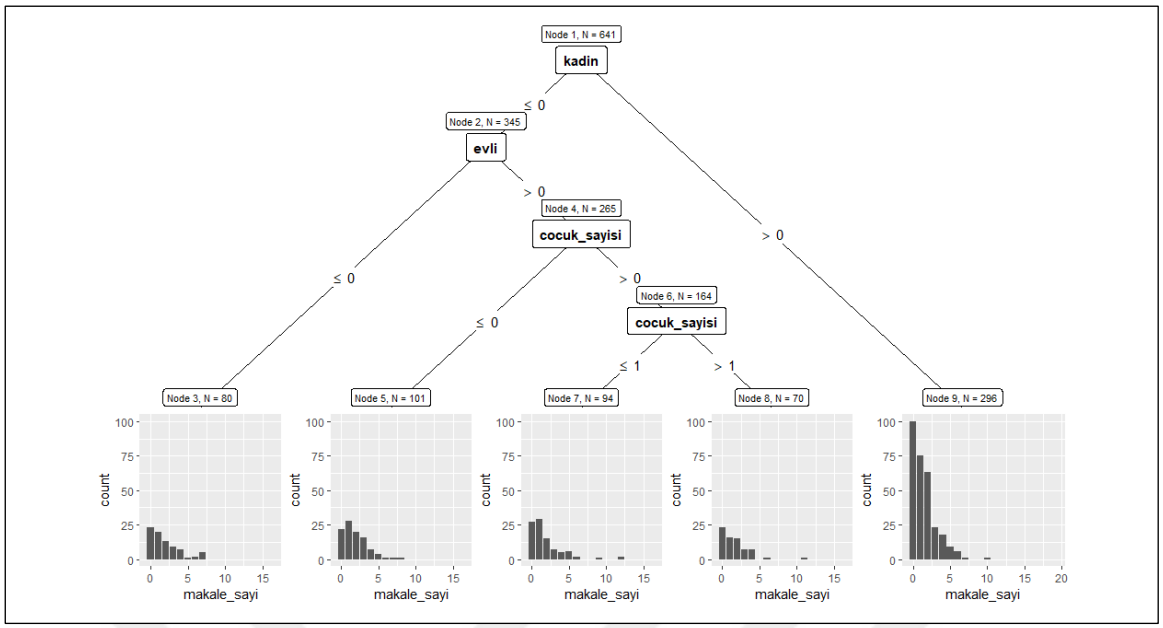


Şekil 4.2. GUIDE algoritması ile makale sayısı regresyon ağacı

GUIDE ile oluşturulan ağaçta yaprak düğümlerin altında düğümdeki gözlem sayısı ve düğümün çıktı ortalaması yer almaktadır. Yaprak düğümlerin turuncu olması düğüm ortalamasının genel ortalamadan daha büyük olduğu; mavi olması düğüm ortalamasının genel ortalamadan küçük olduğu anlamına gelmektedir. Düğümün yanında belirtilen koşulu sağlayan gözlemler sol alt düğüme, sağlamayanlar sağ alt düğüme ayrılmaktadır.

GUIDE algoritmasına göre kök düğümde bağımlı değişken makale sayısını etkileyen en önemli değişken danışman yayın sayısıdır. Kök düğümü ayıracak ikinci önemli değişken ise doktora bölümünün prestijini belirten prestij değişkenidir. Makale sayısına ait en yüksek ortalamanın, danışman yayın sayısının 14,5'ten daha büyük olduğu alt gruba düştüğü Şekil 4.2'de görülmektedir. Benzer şekilde en küçük makale sayısı ortalamasına sahip alt grup, danışman yayın sayısının 3,5'tan daha küçük olduğu gruptur. Danışman yayın sayısı 3,5 ile 14,5 arasında olan alt grubun ortalaması ise genel ortalama 1,69'a çok yakın elde edilmiştir.

GUIDE algoritmasına göre oluşturulan ağacın kök düğümünde 641 gözlem bulunmaktadır ve düğüm ortalaması 1,69; düğüm sapması 2,001'dir. 526 gözlemin bulunduğu 2 numaralı ara düğümde düğümün sapma değeri 1,787'dir. 115 gözlemi içeren 3 numaralı yaprak düğümün makale sayısı ortalaması 2,83; düğümün sapması 2,188'dir. 201 gözlemi içeren 4 numaralı yaprak düğümün makale sayısı ortalaması 1,03; düğümün sapması 1,570'tir. 325 gözlemi içeren 5 numaralı yaprak düğümün makale sayısı ortalaması 1,70; düğümün sapması 1,801'dir.



Şekil 4.3. MOB algoritması ile makale sayısı regresyon ağacı

MOB algoritması ile oluşturulan modelin grafiği R programı (R Development Core Team, 2023) "ggparty" paketi (Borkovec ve diğerleri, 2022) kullanılarak oluşturulmuştur. Grafiğin yapraklarında düğümlere ait makale sayısının çubuk grafiği çizdirilmiştir.

MOB regresyon ağacı algoritmasında yaprak düğümlerde bağımsız değişken için tek bir sabit değer bulunmamakta, bunun yerine her yaprak düğüme uydurulan bir regresyon modeli bulunmaktadır. Doktora makale sayısı verisi eğitim seti ile oluşturulan ağaç çıktısında evli olmayan erkekleri içeren gözlemlerin bulunduğu 3 numaralı yaprak düğümdeki modelde danışman yayın sayısı değişkeni istatistiksel olarak anlamlı bulunmuştur. Bu düğümde bulunan gözlemler için diğer değişkenler sabitken danışman yayın sayısındaki her bir birimlik artış öğrencilerin makale sayılarında ortalama %2,31'lik bir artışa neden olmaktadır. 5, 7, 8 ve 9 numaralı yaprak düğülerde bulunan modellere göre de danışman yayın sayısı değişkeni istatistiksel olarak anlamlı bulunmuştur ve bu düğümler için diğer değişkenler sabitken danışman yayın sayısındaki her bir birimlik artış öğrenci makale sayılarında sırasıyla %4,08; %2,06; %2,31 ve %2,61'lik bir artışa neden olmaktadır. Ayrıca kadın öğrencilerin olduğu 9 numaralı yaprak düğümde bulunan model için çocuk sayısı değişkeni de istatistiksel olarak anlamlı bulunmuştur. 9 numaralı yaprak düğümde bulunan gözlemler için öğrencilerin 5 yaşından küçük çocuklarının sayısındaki her bir artış makale sayısında %20'lik bir azalmaya sebep olmaktadır.

MOB algoritması ile oluşturulan ağacın öğrencilerin kadın olduğu 296 gözleme sahip Şekil 4.3'te en sağda bulunan 9 numaralı alt düğümün amaç fonksiyonunun değeri 492,45; artık sapma değeri 481,41'dir. Evli olmayan erkek öğrencileri içeren en solda bulunan 3 numaralı yaprak düğümün artık sapma değeri 149,98; amaç fonksiyonu 151,79'dur. Evli olan erkeklerden 5 yaşından küçük çocuğu bulunmayan 100 gözlemden oluşan 5 numaralı yaprak düğümün artık sapma değeri 172,03; amaç fonksiyonu 190,58'dir. Evli olan ve 5 yaşından küçük bir çocuğu bulunan erkeklerin oluşturduğu 98 gözlem bulunan 7 numaralı yaprak düğümün yaprak düğümün artık sapma değeri 199,84; amaç fonksiyonu 187,62'dir. Evli olan ve 5 yaşından küçük birden fazla çocuğu bulunan erkeklerin bulunduğu 67 gözlemden oluşan 8 numaralı yaprak düğümün amaç fonksiyonu 124,58; artık sapma değeri 126,24'tür. Ağacın kök düğümünün amaç fonksiyon değeri 1147,02'dir.

Modellerin performanslarını karşılaştırmak ve tahmin hatasını değerlendirmek için test verisindeki gerçek bağımlı değişkenin değerleri ile kurulan ağaç modelleri aracılığıyla tahmin edilen bağımlı değişkenin değerleri arasındaki farkın karelerinin ortalaması

$$MSE = \frac{\left(y_i - \hat{y}_i\right)^2}{274} \tag{4.2}$$

hesaplanmıştır. Eş. 4.2 ile hesaplanan değerler Çizelge 4.6'da verilmiştir.

Çizelge 4.6. Makale sayısı regresyon ağaçlarının hata kareler ortalaması

	CART	GUIDE	MOB
MSE	3,5194	3,8822	39,6423

Tahmin hatası en küçük olan CART algoritması iken en büyük olan MOB algoritması olduğu görülmüştür. GUIDE algoritması ile elde edilen tahmin hatası, CART algoritmasındakine yakın olarak elde edilmiştir.

Aynı eğitim ve test verisi ile Poisson, negatif binom, ZIP ve ZNB regresyonu da kurulmuş ve aynı formül ile hata kareler ortalamasına bakılmıştır. Test verisi ile hesaplanan hata kareler ortalaması en küçük ZIP regresyon modeli ile elde edilmiştir. Modellerin hata kareler ortalaması Çizelge 4.7'de verilmiştir.

Çizelge 4.7. Makale sayısı regresyon modellerinin hata kareler ortalaması

	Poisson	Negatif Binom	ZIP	ZNB
MSE	4,7607	4,7528	3,1567	3,1937

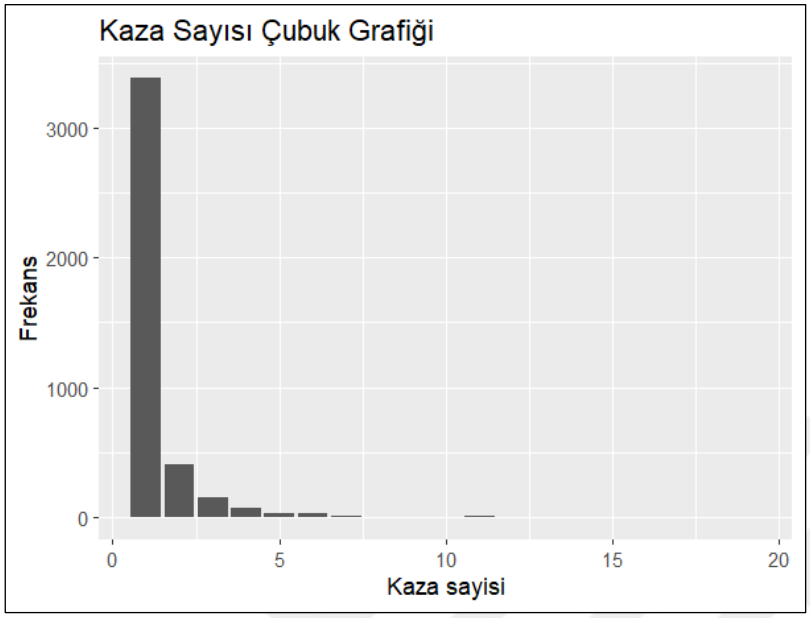
4.1.2. Kaza sayısı verisi

Kaza sayısı verisinde Ankara ilinde 2017-2022 yılları arasında gerçekleşen, ölüm ya da yaralanma ile sonuçlanan kaza bilgileri bulunmaktadır. Kullanılan veri Ankara Emniyet Müdürlüğü Trafik Şube Başkanlığı'ndan alınmıştır. Kazanın olduğu yolun fiziki özellikleri, çevre durumu ve hız limiti gibi kazaya sebep olabilecek unsurların yer aldığı veri setinde 4108 gözlem yer almaktadır. Bağımsız değişkenlere ait bilgiler Çizelge 4.8'de verilmiştir.

Çizelge 4.8. Kaza sayısı verisinin bağımsız değişkenleri

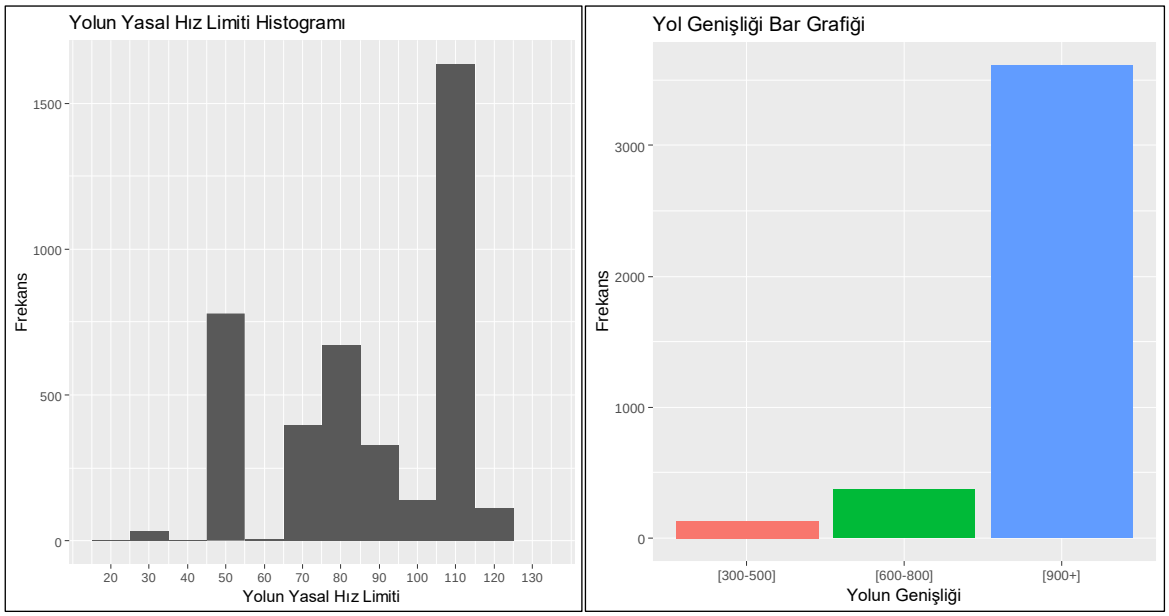
Değişken	Değişken tipi	Değerleri
Yolun tipi	Kategorik	1-Tek yönlü 2-Bölünmüş 3-İki yönlü 4-Diğer
Hız limiti	Sayısal	Min: 20 Maks: 130
Şerit sayısı	Sayısal	Min: 0 Maks: 8
Gün durumu	Kategorik	1-Gündüz 2-Alacakaranlık 3-Gece
Yolun yüzeyi	Kategorik	1-Kuru 4-Karlı 2-Buzlu 5-Sel /Su birikintili 3-Islak /Nemli 6-Diğer kaygan yüzey
Yol genişliği	Kategorik	1-[300-500] 2-[600-800] 3-[900 +]

Bağımlı değişken kaza sayısının minimum değeri 1, maksimum değeri 19'dur. Kaza sayısı değişkeninin çubuk grafiği Şekil 4.4'te verilmiştir.



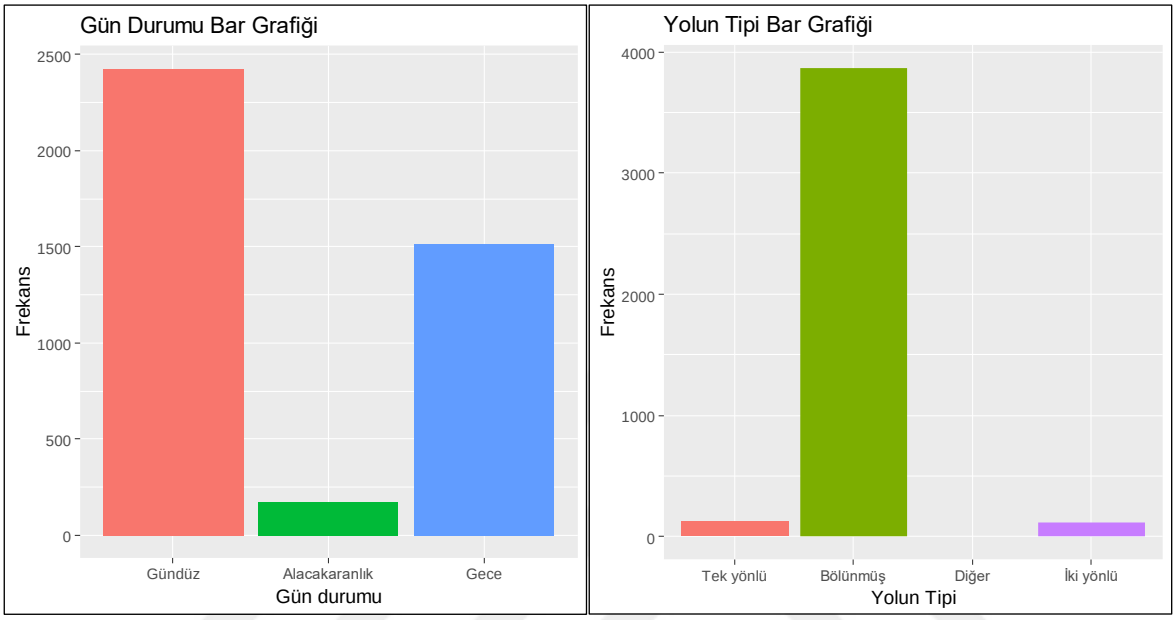
Şekil 4.4. Kaza sayısı değişkenine ait çubuk grafiği

Hız limiti değişkeninin minimum değeri 20, maksimum değeri 130'dur. Hız limiti değişkeninin histogramı Şekil 4.5'in solunda verilmiştir. Yol genişliği değişkeninin [300-500], [600-800] ve [900+] olmak üzere üç seviyesi bulunmaktadır. Yol genişliğine ait çubuk grafiği Şekil 4.5'in sağında verilmiştir.



Şekil 4.5. Hız limiti ve yol genişliği değişkenlerine ait frekans grafikleri

Gün durumu değişkeninin gündüz, alacakaranlık ve gece olmak üzere üç seviyesi bulunmaktadır. Gün durumu değişkeninin çubuk grafiği Şekil4.6'nın solunda verilmiştir. Yolun tipi değişkeninin tek yönlü, iki yönlü, bölünmüş ve diğer olmak üzere dört seviyesi bulunmaktadır. Yolun tipi değişkeninin çubuk grafiği Şekil 4.6'nın sağında verilmiştir.



Şekil 4.6. Gün durumu ve yol tipi değişkenlerinin frekans grafikleri

Bağımlı değişken kaza sayısının ortalaması 1,355; varyansı 1,127'dir. Bağımlı değişkenin varyans ve ortalaması birbirine yakın bir değerde olduğu için Poisson modeline uygun olduğu belirtilebilir. Yapılan aşırı yayılım testi ile de aşırı yayılımın olmadığı görülmüştür. Poisson regresyon modeli

$$\ln(\mu_i) = \beta_0 + \beta_{11}x_{i11} + \beta_{12}x_{i12} + \beta_{13}x_{i13} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{41}x_{i41} + \beta_{42}x_{i42} + \beta_{51}x_{i51} + \beta_{52}x_{i52} + \beta_{53}x_{i53} + \beta_{54}x_{i54} + \beta_{55}x_{i55} + \beta_{61}x_{i61} + \beta_{62}x_{i62} \quad (4.3)$$

uygulanmıştır.

Uygulanan Poisson regresyon modelinden elde edilen katsayılar Çizelge 4.9'da verilmiştir.

Çizelge 4.9. Kaza sayısı verisi Poisson regresyon modeli katsayıları

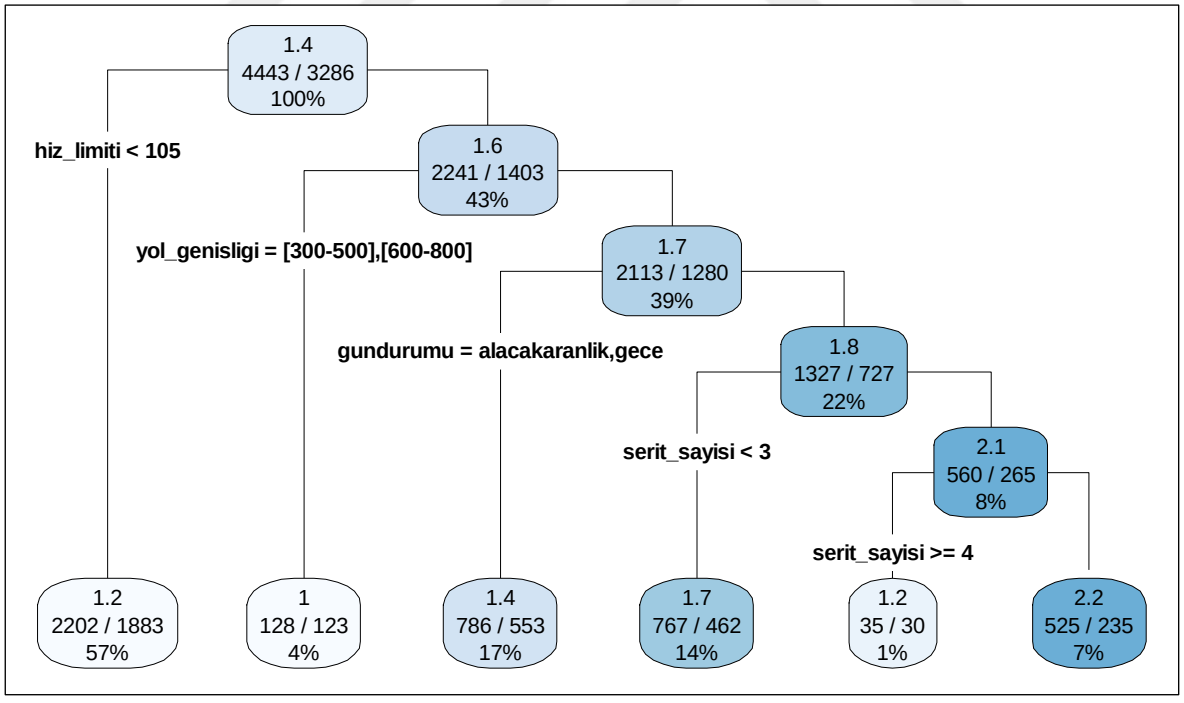
Katsayılar	Tahmin	$exp(\beta)$ değerleri	Standart hata	Z	p değeri
Sabit	-0,6062	0,5454	0,1422	-4,265	< 0,0001
Yol tipi - Bölünmüş	0,2872	1,3327	0,0889	3,233	0,0012
Yol tipi - İki yönlü	0,0736	1,0763	0,1283	0,573	0,5665
Yol tipi - Diğer	-0,0953	0,9091	0,4561	-0,209	0,8345
Hız limiti	0,0046	1,0046	0,0006	7,777	< 0,0001
Şerit sayısı	0,0142	1,0143	0,0168	0,847	0,3973
Gün durumu- Alacakaranlık	-0,3484	0,7058	0,0780	-4,466	< 0,0001
Gün durumu- Gece	-0,1368	0,8722	0,0285	-4,797	< 0,0001
Yol yüzeyi- Buzlu	-0,3380	0,7132	0,2302	-1,468	0,1421
Yol yüzeyi- Islak / Nemli	-0,1357	0,8731	0,0333	-4,072	< 0,0001
Yol yüzeyi- Karlı	-0,3076	0,7352	0,1223	-2,516	0,0119
Yol yüzeyi- Sel / Su birikintili	-0,4327	0,6488	0,2243	-1,929	0,0537
Yol yüzeyi- Diğer kaygan yüzey	-0,3161	0,7290	0,4088	-0,773	0,4394
Yol genişliği- [600-800]	0,1056	1,1114	0,1025	1,030	0,3030
Yol genişliği- [900 +]	0,3148	1,3700	0,0896	3,512	0,0004

Poisson regresyon modeline göre kaza sayısındaki değişimi açıklayan en önemli değişkenin en küçük p değerine sahip olan hız limiti değişkeni olduğu görülmektedir. Modelden elde edilen katsayılara göre hız limitindeki her bir birimlik artış kaza sayısında ortalama %0,46'lık bir artışa sebep olmaktadır. Yol tipi kategorik değişkeninin referans seviyesi tek yönlü olarak alınmıştır. Bölünmüş yolda gerçekleşen kaza sayısı, tek yönlü yoldaki kaza sayısından %33 daha fazladır. Gün durumu kategorik değişkeninin referans seviyesi gündüz olarak alınmıştır. Gündüze göre alacakaranlıkta %30, gece %13 daha az kaza gerçekleşmektedir. Yol yüzeyi kategorik değişkeninin referans seviyesi kuru yüzey olarak alınmıştır. Yol yüzeyi ıslak veya nemli olan yollarda kuru zemindekilere göre %13, karlı yollarda kuru zeminli yollara göre %17 daha az kaza meydana gelmektedir. Yol genişliği kategorik değişkeninin referans seviyesi [300-500] olarak alınmıştır. Yol genişliği 900'ün üzerinde olan yollarda gerçekleşen kaza sayısı [300-500] arası olan yollarda gerçekleşen kaza sayısından %37 daha fazladır.

Kaza sayısı verisi ile regresyon ağaçları

Kaza sayısı veri seti, 0,8'i eğitim seti, 0,2'si test seti olarak ayrılmıştır. 3286 gözlem bulunan eğitim setine CART, GUIDE ve MOB regresyon ağaçları algoritmaları uygulanmış sonrasında 822 gözlem bulunan test seti ile modellerin hata kareler ortalaması incelenmiştir.

Poisson modeli ile kurulan regresyon ağaçlarından elde edilen grafikler Şekil 4.7, Şekil 4.8 ve Şekil 4.9'da verilmiştir. Grafiklerde düğümlerin altındaki koşulu sağlayan gözlemler sola, sağlamayan gözlemler sağa ayrılmaktadır. CART ve GUIDE karar ağaçları için yaprak düğümlerin değerlerinin genel ortalamanın üstünde olup olmamasına göre renkleri değişmektedir. Bu durumda, yaprak düğüm ortalaması 1,35'e eşit veya üstünde olan düğümler CART için koyu mavi, GUIDE için turuncu renk ile belirtilmektedir. CART, GUIDE ve MOB karar ağaçlarına göre kök düğümü bölen en önemli değişken yolun yasal hız limitidir.

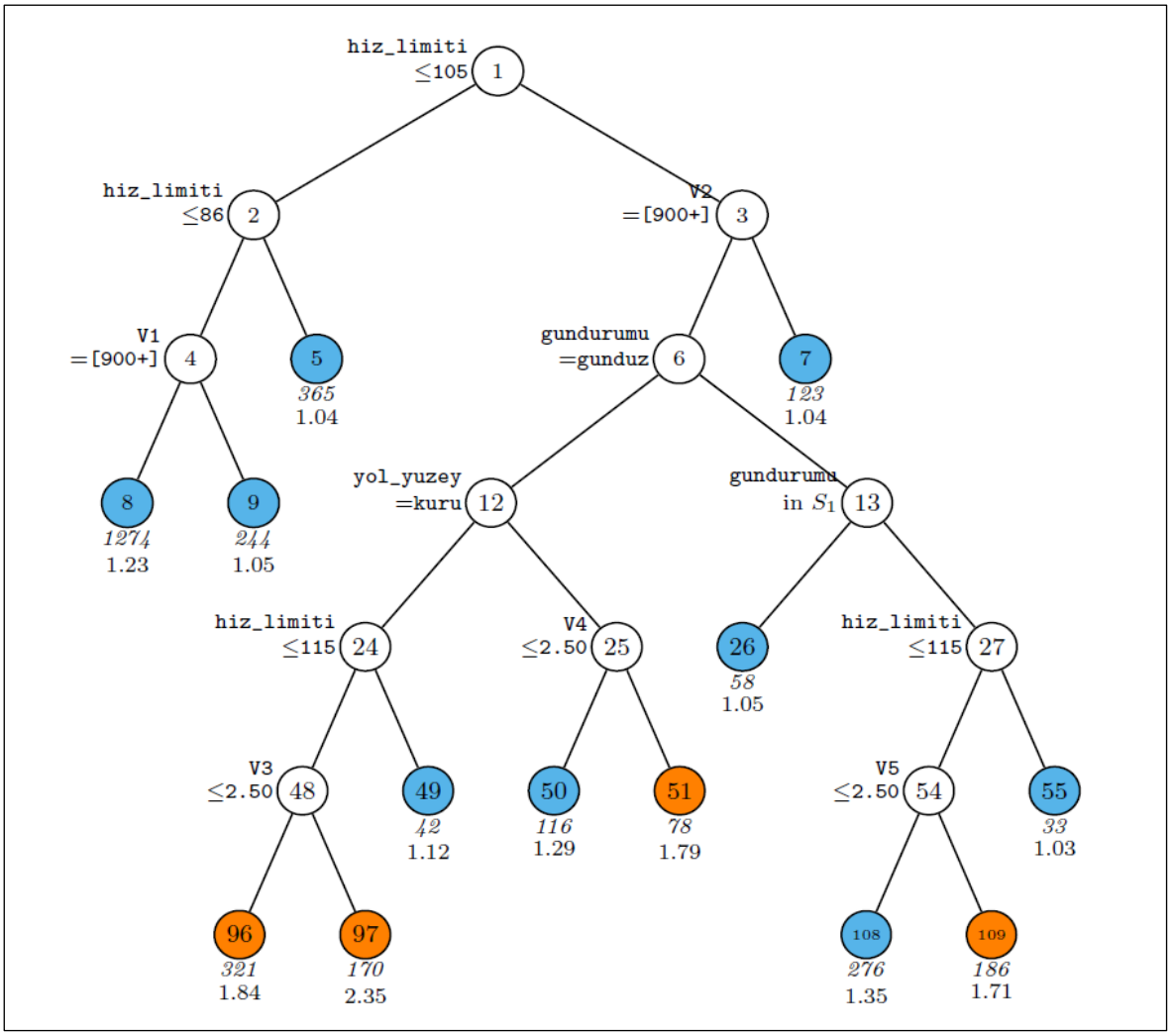


Şekil 4.7. CART algoritması ile kaza sayısı regresyon ağacı

CART ile oluşturulan ağaçta düğümün içerisinde en üstte düğümün çıktı ortalaması, ortasında düğüme ait toplam kaza sayısına gözlem sayısı oranı en altta düğümdeki gözlem

sayısının toplam gözlem sayısına göre yüzdesi yazmaktadır. Gözlemler düğümlerin sol alt dalında belirtilen koşulu sağlarsa sol alt düğüme, sağlanmazsa sağ alt düğüme ayrılır.

Eğitim veri setine uygulanan CART algoritmasının sonuçlarına göre, kaza sayısını etkileyen değişkenler önem sırasına göre hız limiti, şerit sayısı, gün durumu, yol genişliği ve yolun yüzeyidir. Kaza sayısını etkileyen en önemli değişken yolun yasal hız limitidir. Hız limiti 105'in altında olan yollarda gerçekleşen kaza sayıları içeren Şekil 4.7'de en solda 1883 gözlem bulunan yaprak düğümün kaza sayısı ortalaması 1,17; sapma ortalaması 0,1992'dir. Hız limiti 105'in üstünde ve yol genişliği 900'ün altında olan yollarda gerçekleşen ortalama kaza sayısı 1,04 ile en küçük ortalamaya sahip yaprak düğümü oluşturmaktadır. 123 gözlem içeren bu yaprak düğümün sapma ortalaması 0,0383'tür. Hız limiti 105'in, yol genişliği 900'ün üstünde olan yollarda alacakaranlık veya gece gerçekleşen ortalama kaza sayısı 1,42; düğümün sapma ortalaması 0,4674'tür. Hız limiti 105'in, yol genişliği 900'ün üstünde olan şerit sayısının 2,5'ten az olduğu yollarda gerçekleşen kazaların bulunduğu 462 gözlem içeren yaprak düğümün sapma ortalaması 0,6665; kaza sayısı ortalaması 1,66'dır ve genel ortalamadan daha fazladır. Hız limiti 105'in, yol genişliği 900'ün üstünde olan şerit sayısının 3,5'ten fazla olduğu yaprak düğümün ortalaması 1,17; ortalama sapması 0,1373'tür. Şerit sayısı 3 olan yolları içeren Şekil 4.7'nin en sağında bulunan alt düğümün sapma ortalaması 1,5806; kaza sayısı ortalaması ise 2,23'tür ve en büyük ortalamaya sahip düğümdür. Yolun yasal hız limiti, şerit sayısı ve yol genişliği arttığında kaza sayısı ortalamasının da arttığı görülmektedir.



Şekil 4.8. GUIDE algoritması ile kaza sayısı regresyon ağacı

Şekil 4.8'de V1 ve V2 yolun genişliğini, V3, V4 ve V5 şerit sayısını ve S1 gün durumu değişkeninin alacakaranlık seviyesini belirtmektedir. Düğümlerin altında italik olarak yazılan sayılar düğümün örneklem büyüklüğünü ve onun altında yazan sayılar düğümün ortalamasını ifade etmektedir. Kök düğümü bölebilecek ikinci en önemli değişken yol genişliği değişkenidir. GUIDE regresyon ağacında eğitim verisine ait kaza sayısının ortalaması 1,35'in üzerinde ortalamaya sahip yaprak düğümler turuncu renk ile 1,35'in altında ortalamaya sahip yaprak düğümler ise mavi renk ile temsil edilmektedir.

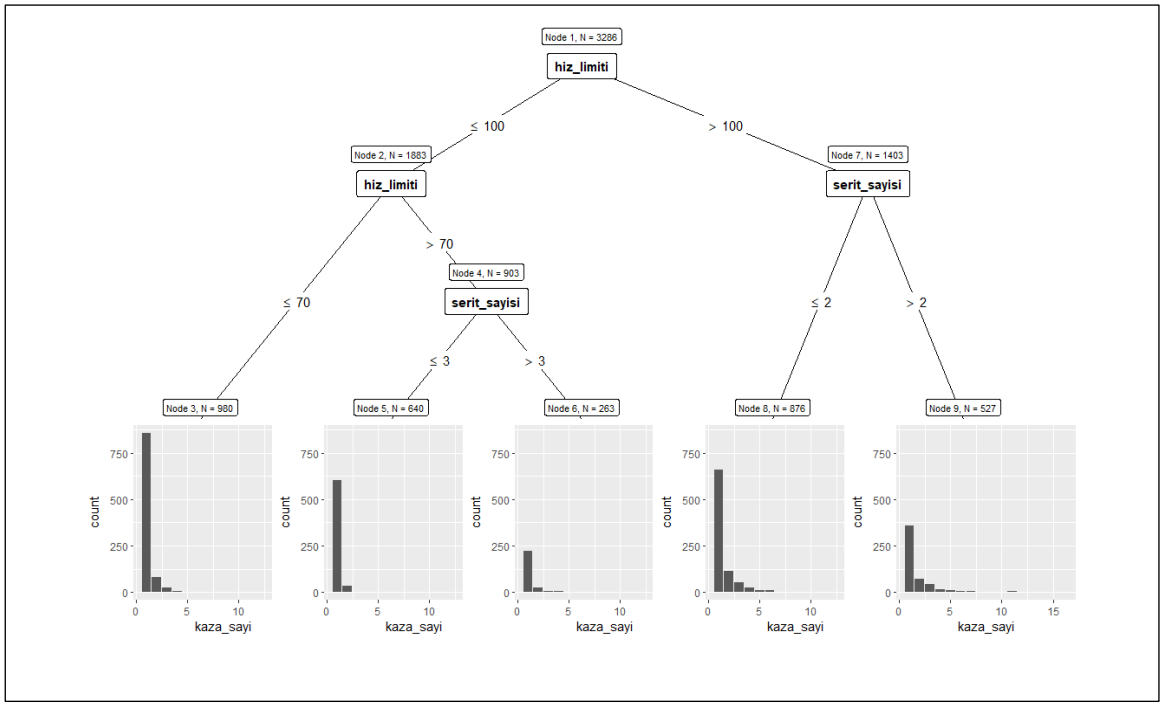
GUIDE algoritması ile oluşturulan ağacın düğümlerinin kaza sayı ortalaması ve düğüm içi ortalama artık sapması Çizelge 4.10'da verilmiştir.

Çizelge 4.10. Kaza sayısı verisi ile GUIDE algoritması ağacının düğümleri

Düğüm etiketi	Toplam gözlem	Düğüm ortalaması	Düğüm sapması
1	3286	1,352	0,4645
2	1883	1,169	0,1993
4	1518	1,202	0,2362
8T	1274	1,231	0,2671
9T	244	1,049	0,0515
5T	365	1,036	0,0263
3	1403	1,597	0,7439
6	1280	1,651	0,7887
12	727	1,825	0,9901
24	533	1,946	1,0330
48	491	2,016	1,0720
96T	321	1,841	0,7797
97T	170	2,347	1,5490
49T	42	1,119	0,1479
25	194	1,495	0,7902
50T	116	1,293	0,3014
51T	78	1,795	1,4300
13	553	1,421	0,4682
26T	58	1,052	0,0380
27	495	1,465	0,5051
54	462	1,496	0,5287
108T	276	1,351	0,3364
109T	186	1,710	0,7665
55T	33	1,030	0,0232
7T	123	1,041	0,0386

GUIDE algoritması ile oluşturulan regresyon ağacına göre hız limiti 86 ile 105 arasında olan yollarda gerçekleşen kaza sayısı ortalaması 1,04 ile genel kaza sayısı ortalamasından küçüktür. Yasal hız limiti 86'nın altında, yol genişliği 900'ün üzerinde olan yollarda kaza sayısı ortalaması 1,23 iken 900'ün altında yol genişliğine sahip yollarda 1,05'tir. Yasal hız limiti 105'in altında olan yollarda gerçekleşen kaza sayısı ortalaması genel kaza sayısı ortalamasından daha küçük olduğu görülmektedir. Yasal hız limiti 115'in üzerinde ve genişliği 900'ün üzerinde olan yollarda gece gerçekleşen ve hız limiti 105'in üzerinde olan yollarda gece gerçekleşen ortalama kaza sayıları da 1,03 ve 1,05 ile genel ortalamanın

altında yer almaktadır. Kaza sayı ortalaması 2,35 ile en büyük olan yaprak düğüm, hız limiti 115'in üzerinde, 900'den geniş, şerit sayısı 2,5'ten fazla olan kuru yollarda gündüz gerçekleşen kazaları içermektedir. Hız limiti ile birlikte yol genişliği ve şerit sayısının artması gerçekleşen ortalama kaza sayısını arttırdığı görülmektedir.



Şekil 4.9. MOB algoritması ile kaza sayısı regresyon ağacı

MOB regresyon ağacının yapraklarında düğümlere ait kaza sayısının çubuk grafiği çizdirilmiştir. MOB ağacında sadece iki değişkene, yolun yasal hız limiti ve şerit sayısı değişkenlerine göre alt düğümlere ayrıldığı görülmektedir.

MOB regresyon ağacı çıktısında parçalı lineer modeller üretmektedir. Her yaprak düğüme bir regresyon modeli uydurmaktadır. Kaza sayısı verisi eğitim seti ile elde edilen MOB regresyon ağacında 3, 5 ve 6 numaralı yaprak düğümlerde bulunan gözlemler için oluşturulan regresyon modellerine göre istatistiksel olarak anlamlı değişken bulunamamıştır. Yasal hız limiti 100'ün üzerinde ve 2 veya daha az şeridi olan yollara ait gözlemlerin bulunduğu 8 numaralı yaprak düğüm için oluşturulan modelde hız limiti, gün durumu- alacakaranlık, gün durumu- gece ve yol yüzeyi- ıslak değişkenleri anlamlı bulunmuştur. Yolun yasal hız limitindeki her bir birimlik artış diğer değişkenler sabitken meydana gelen kaza sayısında ortalama %4'lük bir azalmaya neden olmaktadır. Gün durumu alacakaranlık iken referans değer gündüze göre %34; gece iken gündüze göre %20

daha az kaza gerçekleşmektedir. Yol yüzeyi ıslak olması durumunda gerçekleşen kaza sayısı referans değer olan kuru yüzeyde gerçekleşen kaza sayısına göre %24 daha azdır. Hız limiti 100'ün üzerinde olan ve 2'den fazla şeridi bulunan yolları içeren 9 numaralı yaprak düğümdeki modele göre ise bölünmüş yollarda gerçekleşen kaza sayısının tek yönlü yollarda gerçekleşen kaza sayısına göre %93 daha fazla olduğu görülmektedir. Gündüz gerçekleşen kaza sayısına göre alacakaranlıkta gerçekleşen kaza sayısı %50, gece gerçekleşen kaza sayısı %22 daha azdır. Islak, karlı ve su birikintili yollarda referans düzey kuru yollara göre sırasıyla %22, %38 ve %56 daha az kaza gerçekleşmektedir. Yol genişliği 900'ün üzerinde olan yollarda gerçekleşen kaza sayısı, referans alınan [300-500] arasında olan yollarda gerçekleşen kaza sayısına göre %87 daha fazladır.

Eğitim verisine uygulanan MOB algoritmasına göre yolun yasal hız limiti 70'in altında olan yolları içeren, 980 gözlem bulunan yaprak düğümün artık sapması 226,92 ve amaç fonksiyon değeri 1143,657'dir. Hız limiti 70 ile 100 arasında, 3 veya daha az şeritli yollar için oluşturulan yaprak düğümde 640 gözlem bulunmakta ve artık sapması 48,458'dir. Hız limiti 70 ile 100 arasında, şerit sayısı 3'ün üstünde yollar için oluşturulan yaprak düğümde 263 gözlem bulunmakta ve artık sapması 68,980'dir. Hız limiti 100'ün üzerinde olan yollar ise şerit sayısına göre alt gruplara ayrılmaktadır. 2 veya daha az şeritli yollarla oluşturulan 876 gözlem bulunan alt grubun artık sapması 376,67'dir. 2'den fazla şeritli alt grupta 527 gözlem bulunmakta ve düğümün artık sapması 474,40'tır.

MOB algoritması ile oluşturulan regresyon ağacının düğümlerinin amaç fonksiyon değerleri 1 numaralı düğümden 9 numaralı düğüme kadar sırasıyla 4145,978; 2135,43; 1143,657; 991,773; 677,578; 314,195; 2010,548; 1160,759 ve 849,789'dur.

Modellerin performanslarını karşılaştırmak ve tahmin hatasını değerlendirmek için test verisinde bulunan gerçek y değerleri ile ağaç modelleri ile tahmin edilen $\hat{y}$ değerleri arasındaki farkın karelerinin ortalaması

$$MSE = \frac{\left(y_i - \hat{y}_i\right)^2}{822} \tag{4.4}$$

hesaplanmıştır. Eş. 4.4 ile hesaplanan değerler Çizelge 4.11'de verilmiştir.

Çizelge 4.11. Kaza sayısı regresyon ağaçlarının hata kareler ortalaması

	CART	GUIDE	MOB
MSE	1,2748	1,2434	26,3358

En küçük tahmin hatasına sahip olan GUIDE algoritması iken en büyük tahmin hatasına sahip olan MOB algoritmasıdır. Aynı eğitim ve test verisi ile Poisson regresyonu da kurulmuştur ve aynı formül ile hata kareler ortalamasına bakıldığında sonuç 2,4908 çıkmıştır.

4.2. Simülasyon

Simülasyon çalışması dört aşamada gerçekleştirilmiştir. İlk olarak, algoritmaların ayırma değişkenini seçmedeki yanlılığı incelenmiştir. İkinci olarak, algoritmaların değişken seçimindeki tip 1 hatası ve gücü değerlendirilmiştir. Gözlem ve tekrar sayısı 1000 olan simülasyon çalışmaları yürütülmüştür. Son olarak, uygulama bölümünde kullanılan veri setlerindeki değişkenler temel alınarak veri setleri üretilmiş modellerin performansları, algoritmaların kök düğümü bölmek için seçtikleri ayırma değişkenleri ve gücü incelenmiştir.

Simülasyon çalışmalarında veri üretimi ve ağaç oluşturma R programı (R Development Core Team, 2023) ile yapılmıştır. CART algoritması için “rpart” paketi (Therneau ve Atkinson, 2023), MOB algoritması için “partykit” paketi (Hothorn ve Zeileis, 2015) ve GUIDE algoritması için GUIDE programı 41.2 versiyonu (Loh, 2023) kullanılmıştır. GUIDE simülasyonlarını oluşturmak için GUIDE programı ile bağlantı kuran fonksiyonlar R programında (R Development Core Team, 2023) yazılmış ve sonuçlar elde edilmiştir. Kullanılan algoritmalarla regresyon ağaçları üretilirken Poisson dağılımı seçilmiş geri kalan ince ayarlar için algoritmaların varsayılan ayarları kullanılmıştır.

4.2.1. Değişken seçim yanlılığı

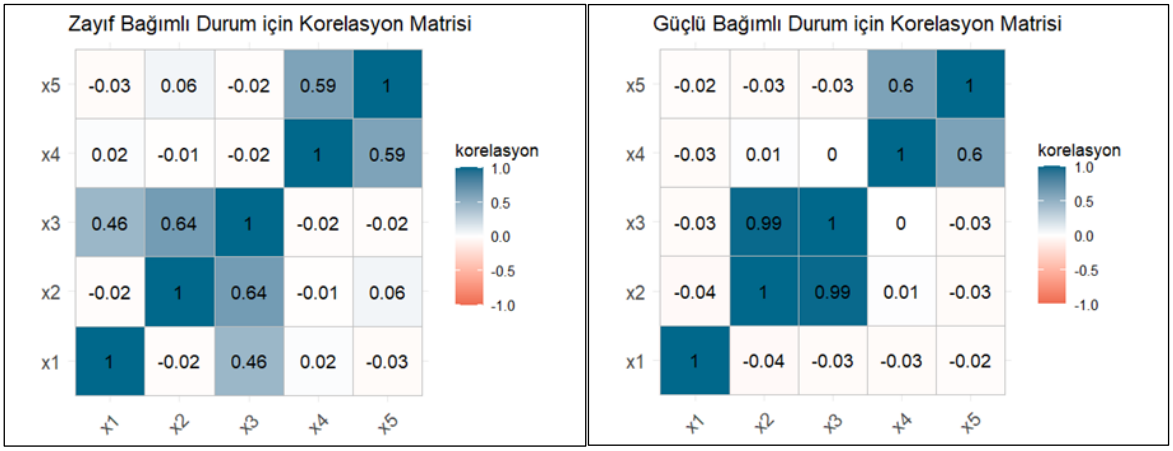
Bağımlı değişken Y, bağımsız değişken X’lerden bağımsız olarak üretilmiştir. X değişkenleri ve Y değişkeninin birbirinden bağımsız olması durumunda her bağımsız değişkenin seçilme olasılığının eşit olması beklenir. Bu durum değişken seçiminde yanlılık olmadığını göstermektedir.

Seçim yanlılığı simülasyonu için kullanılan modeller Loh (2002) çalışması örnek alınarak yapılmıştır. X değişkenlerinin birbirlerinden bağımsız, bazı değişkenlerin zayıf bağımlı ve güçlü bağımlı olduğu durum olmak üzere üç farklı bağımlılık durumu göz önüne alınarak çalışma yapılmıştır. Üçü sayısal ikisi kategorik beş bağımsız değişkenin marjinal dağılımları Çizelge 4.12'de verilmiştir. T, [1,3] aralığında tekdüze dağılımı; W, parametresi 1 olan üstel dağılımı; Z, standart normal dağılımı ve U, [0,1] aralığında tekdüze dağılımı göstermektedir. C_5, 1'den 5'e kadar eşit olasılıkla değer alabilen 5 seviyeli kategorik değişkeni belirtirken C_{10}, 1'den 10'a kadar eşit olasılıkla değer alabilen 10 seviyeli kategorik değişkeni belirtmektedir. [.] işlevi içindeki sayının en büyük tam sayı değerini vermektedir. Bütün değişkenler hem regresyon modeli kurmak için hem de düğümleri ayırmak için kullanılmıştır.

Çizelge 4.12. Bağımsız değişkenlerin dağılımı

	Bağımsız	Zayıf bağımlı	Güçlü bağımlı
X_1	T	T	T
X_2	W	W	W
X_3	Z	$T+W+Z$	$W+0{,}1Z$
X_4	C_5	$[UC_{10}/2]+1$	$[UC_{10}/2]+1$
X_5	C_{10}	C_{10}	C_{10}

Bağımsız değişkenlerin zayıf bağımlı ve güçlğ bağımlı olduğu durumlar için ele alınan bir veri setine ait bağımsız değişkenlerin korelasyon matrisleri Şekil 4.10'da verilmiştir.



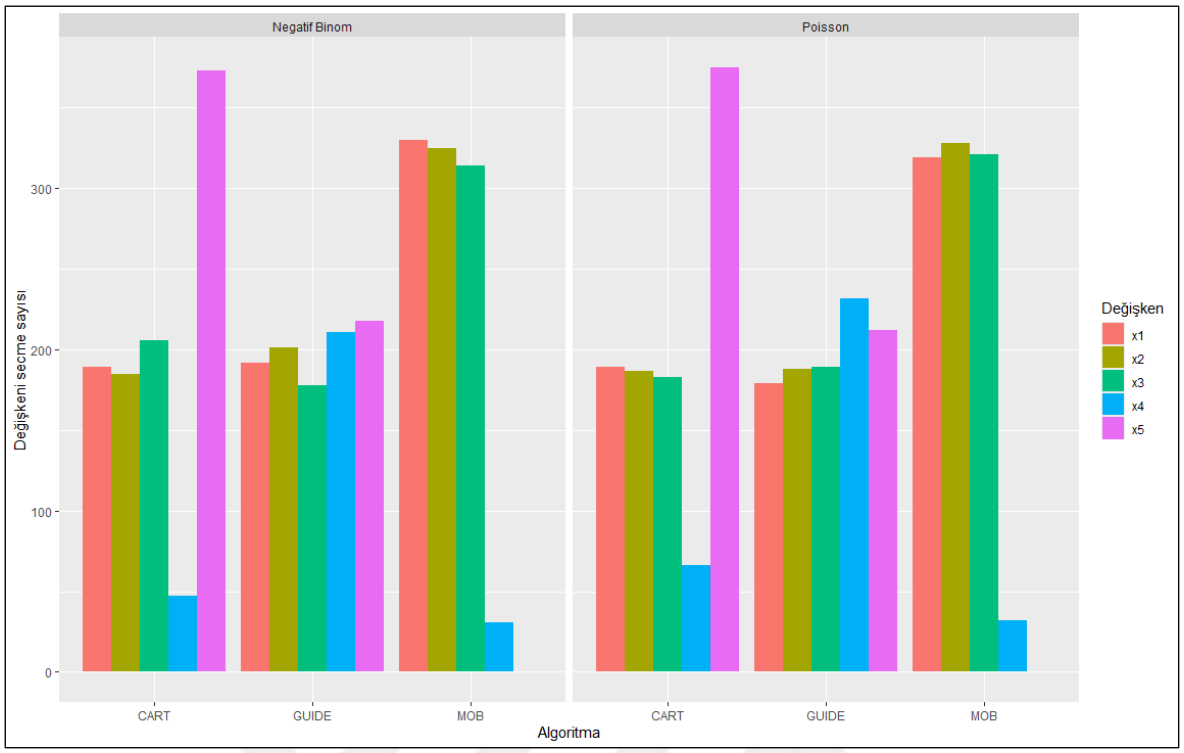
Şekil 4.10. Bağımsız değişkenlerin bağımlı oldukları durumlar için korelasyon matrisleri

Burada ağaç algoritmalarının her koşulda bölme yapması ya da kök düğümü bölecek en iyi ayırma değişkeninin bulunması istenmektedir. Bunun için CART algoritmasında karmaşıklık parametresi 0 olarak ayarlanmıştır. Böylece ağaç budanmamıştır. MOB algoritmasında ise parametre stabilite testi için anlamlılık düzeyi 1 olarak ayarlanmış böylece her durumda bölme yapması sağlanmış ve ağacın boyutunu mümkün olduğu kadar küçük tutmak için maksimum ağaç derinliği de 2 olarak ayarlanmıştır. GUIDE algoritmasının program çıktısında kök düğümü bölen en anlamlı ayırma değişkeni verilmektedir.

X'lerin birbirinden bağımsız olduğu ilk durum için bağımlı değişken *Y*'nin $\mu = 3$ ortalama parametresi ile Poisson dağılması durumunda algoritmaların ayırma değişkenlerini seçme sayısı Çizelge 4.13'ün sol tarafında, Y'nin $\mu = 3$, $\theta = 1{,}5$ yayılım parametresi ile negatif binom dağılması durumunda algoritmaların ayırma değişkenlerini seçme sayısı Çizelge 4.13'ün sağ tarafında verilmiştir. X değişkenlerinin birbirinden bağımsız olduğu durum için dağılımlara göre algoritmaların seçtiği değişkenlerin grafikleri Şekil 4.11'de gösterilmiştir.

Çizelge 4.13. Bağımsız değişkenlerin bağımsız olduğu durum için algoritmaların ayırma değişkenlerini seçme sayısı

$Y \sim Poisson(3)$				$Y \sim \text{NegatifBinom}(3;1,5)$			
	GUIDE	CART	MOB		GUIDE	CART	MOB
X_1	179	189	319	X_1	192	189	330
X_2	188	187	328	X_2	201	185	325
X_3	189	183	321	X_3	178	206	314
X_4	232	66	32	X_4	211	47	31
X_5	212	375	0	X_5	218	373	0

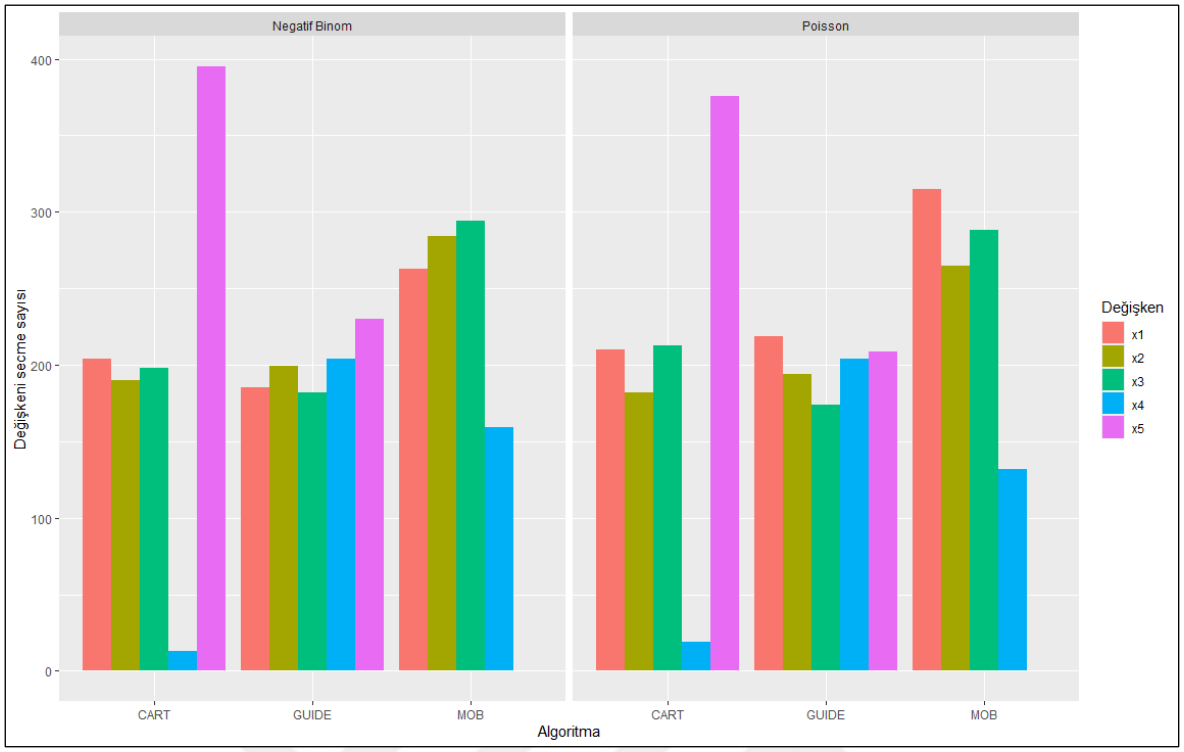


Şekil 4.11. Bağımsız durum için algoritmaların seçtiği değişkenlerin grafiği

Bazı X değişkenlerinin birbirleri ile bağımlı olduğu ikinci durum için bağımlı değişken Y $\mu=3$ ortalama parametresi ile Poisson dağıldığında algoritmaların ayırma değişkenlerini seçme sayısı Çizelge 4.14'ün sol tarafında; Y, $\mu=3$ ve $\theta=1,5$ yayılım parametresi ile negatif binom dağılımı ile üretildiğinde algoritmaların ayırma değişkenlerini seçme sayısı Çizelge 4.14'ün sağ tarafında verilmiştir. Bazı X değişkenlerinin zayıf bağımlı olduğu durum için dağılımlara göre algoritmaların seçtiği değişkenlerin grafikleri Şekil 4.12'de gösterilmiştir.

Çizelge 4.14. Bağımsız değişkenlerin zayıf bağımlı olduğu durum için algoritmaların ayırma değişkenlerini seçme sayısı

	$Y \sim Poisson(3)$				$Y \sim \text{NegatifBinom}(3;1,5)$		
	GUIDE	CART	MOB		GUIDE	CART	MOB
X_1	219	210	315	X_1	185	204	263
X_2	194	182	265	X_2	199	190	284
X_3	174	213	288	X_3	182	198	294
X_4	204	19	132	X_4	204	13	159
X_5	209	376	0	X_5	230	395	0

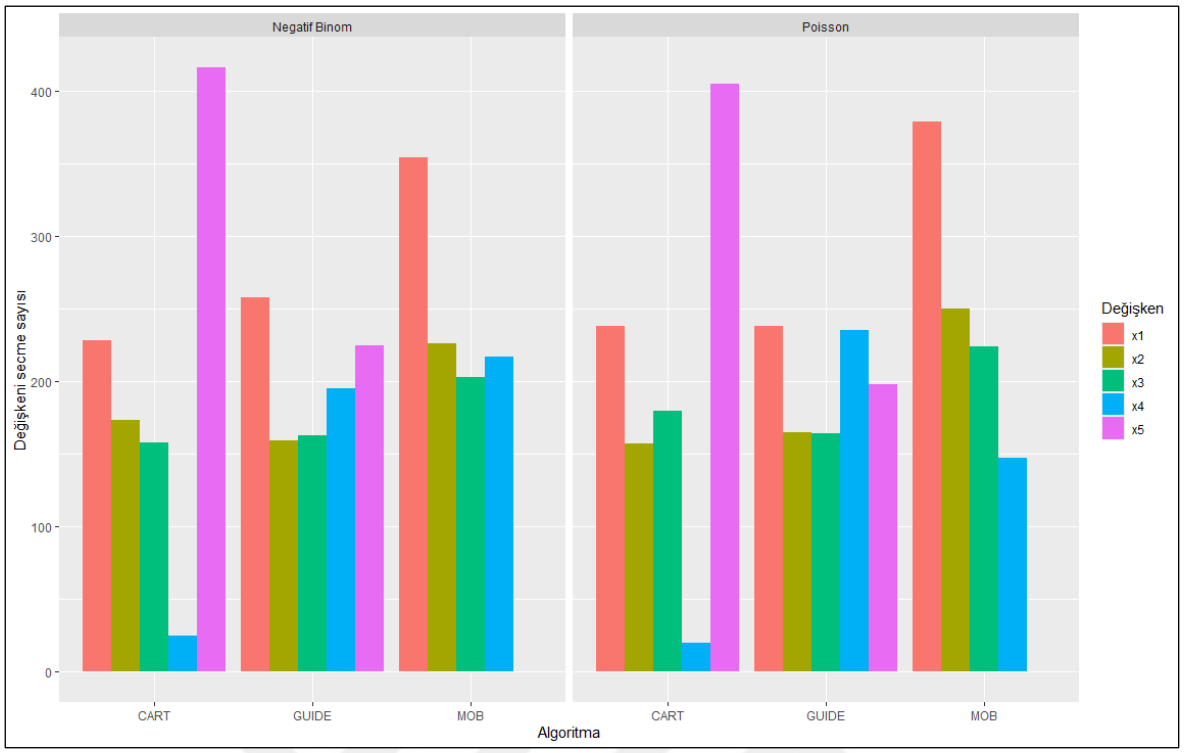


Şekil 4.12. Zayıf bağımlı durum için algoritmaların seçtiği değişkenlerin grafiği

Benzer şekilde bazı X değişkenlerinin güçlü bağımlı olduğu durum için elde edilen sonuçlar Çizelge 4.15 ve Şekil 4.13'te verilmiştir.

Çizelge 4.15. Bağımsız değişkenlerin güçlü bağımlı olduğu durum için algoritmaların ayırma değişkenlerini seçme sayısı

	$Y \sim Poisson(3)$				$Y \sim \text{NegatifBinom}(3;1,5)$		
	GUIDE	CART	MOB		GUIDE	CART	MOB
X_1	238	238	379	X_1	258	228	354
X_2	165	157	250	X_2	159	173	226
X_3	164	180	224	X_3	163	158	203
X_4	235	20	147	X_4	195	25	217
X_5	198	405	0	X_5	225	416	0



Şekil 4.13. Güçlü bağımlı durum için algoritmaların seçtiği değişkenlerin grafiği

Bağımlı değişkenin bağımsız değişkenlerden bağımsız olduğu, bağımsız değişkenlerin kendi aralarında bağımsız, zayıf bağımlı ve güçlü bağımlı olduğu üç durum için 1000 gözlem içeren 1000 ağaç üretilmiştir. Algoritmaların kök düğümü bölmeye en uygun olan ayırma değişkenleri kaydedilmiştir. Modellerde bağımlı değişken Y, bağımsız değişkenlerden bağımsız üretildiği için herhangi bir değişkeni eşit olasılıkla seçmesi algoritmaların ayırma değişkenini yansız olarak seçtiğini göstermektedir. Yansız değişken seçimi için beş bağımsız değişkenin bulunduğu 1000 tekrarlı simülasyon çalışmasında her değişkenin yaklaşık 200 kez seçilmesi beklenmektedir.

Kurulan modellere göre GUIDE algoritmasının değişkenleri yansız bir şekilde seçtiği, CART algoritmasının düzey sayısı az olan kategorik değişkenlerin aleyhine düzey sayısı fazla olan kategorik değişkenlerin lehine seçim yaptığı ve MOB algoritmasının düzey sayısı fazla olan kategorik değişkeni hiç seçmediği görülmektedir. MOB algoritması özellikle değişkenler birbirinden bağımsız olduğunda seçimini sayısal değişkenlerden yana yapmaktadır.

4.2.2. Tip 1 hata ve güç

Regresyon ağaçlarında bağımsız değişkenlerin bağımlı değişkenden bağımsız olması durumunda bölmenin gerçekleşmemesi, kök düğümün kalması beklenir. Ağacın bölünmemesi gerekirken bölünmesi tip 1 hata olarak değerlendirilebilir. Ağacın ayırma değişkeninin varlığında bölünmesi modelin gücü olarak tanımlanabilir. Tip 1 hata ve güç simülasyonları için kullanılan değişkenler Cao (2019) çalışmasında bulunan simülasyon örnek alınarak üretilmiştir. Bu çalışmada ikisi kategorik sekizi sürekli sayısal değişken olmak üzere on bağımsız değişken ve bir bağımlı değişken ile modeller kurulmuştur. X_3 değişkeni ile X_4 değişkeni kendi aralarında 0,5 korelasyon katsayısı ile bağımlı olup diğer değişkenler birbirinden bağımsız olarak üretilmiştir. Y değişkeni de X_i değişkenlerinden bağımsız olarak üretilmiştir. X_i değişkenlerinin marjinal dağılımları Çizelge 4.16'da verilmiştir.

Çizelge 4.16. Tip 1 hata ve güç için üretilen bağımsız değişkenlerin dağılımı

Bağımsız değişkenler	Dağılımları
X_1, X_2	$\sim Normal(0;1)$
X_3, X_4	$\sim Normal(0;1), cor(X_3, X_4) = 0,5$
X_5, X_6	$\sim$ Tekdüze$(0,1)$
X_7, X_8	$\sim$ Üstel(1)
X_9	$\sim$ Bernoulli$(0,5)$
X_{10}	$\sim$ Kategorik$\{1,2,3,4,5,6\}$

Algoritmalarda Poisson dağılımı seçilmiş ve varsayılan ayarlar ile ağaçlar oluşturulmuştur. Bağımlı değişken Y ortalaması 3 olan Poisson dağılımından ve ortalaması 3, yayılım parametresi 1,5 olan negatif binom dağılımından üretilmiştir. 1000 tekrar ile elde edilen tip 1 hata yüzdesi Çizelge 4.17'de verilmiştir.

Çizelge 4.17. Algoritmalara göre tip 1 hata oranları

	$Y \sim Poisson(3)$	$Y \sim \text{NegatifBinom}(3;1,5)$
GUIDE	0,031	0,029
CART	0,283	0,483
MOB	0,084	0,062

Üretilen 1000 veri seti için bölünme gerçekleşmemesi beklenmektedir. Bölünme gerçekleşen tekrar sayısının 1000'e bölünmesi ile elde edilen tip 1 hata oranlarına bakıldığında CART algoritmasının en yüksek orana, GUIDE algoritmasının en düşük orana sahip olduğu görülmektedir. Bölme olmaması gerekirken en çok CART algoritması veriyi bölmektedir.

Algoritmaların gücünü belirlemek için yürütülen simülasyonda bağımlı değişkenin ortalamasını en az bir bağımsız değişken ile ilişkilendirerek 5 model kurulmuştur. 1000 gözlem ve 1000 tekrar ile yürütülen çalışmada algoritmaların bölme yapıp yapmadığı incelenmiştir. Bağımsız değişkenler Çizelge 4.16'da belirtilen dağılımlardan üretilmiştir. Bağımlı değişkenin ortalama parametresi her model için Çizelge 4.18'de belirtildiği gibi elde edilerek Poisson dağılımı ve yayılım parametresi 1,5 olan negatif binom dağılımından üretilmiştir. Bütün değişkenler kullanılarak modeller kurulmuş algoritmalardan elde edilen sonuçlar incelenmiştir.

Çizelge 4.18. Modellere göre bağımlı değişkenlerin ortalama parametreleri

Modeller	Ortalama parametresi
M1	$\mu = \exp\left(3\left(1+X_5\right)+0.3\,\mathrm{I}\left(X_{10}=\{2,4,6\}\right)\right)$
M2	$\mu = \exp\left(3\left(1+X_5\right)+0.3\,\mathrm{I}\left(X_1 \le 0\right)\right)$
M3	$\mu = \exp\left(3\left(1+X_5\right)+0.3\,\mathrm{I}\left(\lvert X_1\rvert \le 0.7\right)\right)$
M4	$\mu = \exp\left(3\left(1+X_5\right)+0.3X_6\,\mathrm{I}\left(X_9=1\right)\right)$
M5	$\mu = \exp\left(3\left(1+X_5\right)+0.05\,\mathrm{I}\left(X_9=1\right)\left(1+X_5+X_6\right)\right)$

Algoritmaların bu modellerde bölme yapması beklenmektedir. Bölme yaptığı sayının tekrar sayısına bölünmesi ile elde edilen sonuçlar Çizelge 4.19'da verilmiştir.

Çizelge 4.19. Algoritmaların modellere göre gücü

	$Y \sim Poisson(\mu)$					$Y \sim \text{NegatifBinom}(\mu;1,5)$				
	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
GUIDE	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
CART	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
MOB	1,00	1,00	1,00	1,00	0,363	0,389	0,319	0,86	0,30	0,301

Yapılan simülasyona göre CART ve GUIDE algoritmaları bölünme gerçekleşmesi beklenirken veriyi kesin olarak bölmektedir. Bağımlı değişkenin Poisson dağılımına uygun üretildiği M5 modeli ve negatif binom dağılımına göre üretildiği M1, M2, M4, M5 modellerinde MOB algoritmasının gücünün oldukça düşük olduğu görülmektedir.

Tip 1 hata ve güç çalışmaları göz önüne alındığında MOB algoritmasının bölümleme yapmamaya, CART algoritmasının ise bölümleme yapmaya daha yatkın olduğu sonucu çıkarılmıştır.

4.2.3. Makale sayısı verisi ile simülasyon

Makale sayısı verisinin değişkenleri temel alınarak bir simülasyon modeli kurulmuştur. Verideki bağımsız değişkenlerin negatif binom modelindeki katsayı tahminleri kullanılarak benzetilen 1000 gözleme sahip 1000 tane veri seti üretilmiştir. Bağımsız değişkenler (X_i) orijinal verideki sıralama ile veri setine eklenmiştir. Bağımlı değişken, bağımsız değişkenler kullanılarak elde edilen ortalama ve 2,267 yayılım parametresi ile negatif binom dağılımından üretilmiştir. Üretilen veri setleri $70/30$ oranında eğitim ve test seti olarak ikiye ayrılmıştır. Eğitim setine uygulanan ağaç algoritmalarının sonuçları test seti ile tahmin hatasının belirlenmesi amaçlanmıştır. Simülasyon verileri ile oluşturulan regresyon ağaçlarının ilk bölmeyi gerçekleştirdiği bölme değişkenlerinin sayısı Çizelge 4.20'de verilmiştir.

Çizelge 4.20. Makale sayısı verisi simülasyonu ile elde edilen ilk ayırma değişkenlerinin frekansları

	CART	GUIDE	MOB
Bölme yok	0	53	880
X_1 (kadın)	6	129	16
X_2 (evli)	2	29	15
X_3 (çocuk sayısı)	10	98	15
X_4 (prestij)	1	2	26
X_5 (danışman yayın sayısı)	981	689	48

Çalışmanın temel alındığı doktora makale sayısı verisi ile yapılan uygulamada bağımlı değişkeni en çok etkileyen değişken simülasyon verisinde danışman yayın sayısı (X_5) değişkenine denk gelmektedir. Çıkan sonuçlara bakıldığında CART ve GUIDE algoritmaları değişken seçimini yüksek oranda X_5 değişkeninden yana kullandığı görülmektedir. Bölünme gerçekleşmesi beklenen simülasyon çalışmasına göre MOB algoritması oldukça düşük bir güce sahipken en yüksek güç tüm tekrarlarda veriyi bölümleyen CART algoritmasına aittir.

Üretilen 1000 veri seti için CART, GUIDE ve MOB algoritmalarının hata kareler ortalamasının (MSE) ortalaması, hata kareler ortalamasının karekökünün (RMSE) ortalaması, mutlak hata ortalamasının (MAE) ortalaması ve bunların standart hataları hesaplanmıştır. Sonuçlar Çizelge 4.21'de verilmiştir.

Çizelge 4.21. Makale sayısı simülasyonu ile elde edilen tahmin hataları ve güç

		CART	GUIDE	MOB
MSE	Ortalama	3,2077	3,2382	3,8355
	Standart hata	0,5115	0,5367	0,9200
RMSE	Ortalama	1,7856	1,7937	1,9476
	Standart hata	0,1389	0,1443	0,2062
MAE	Ortalama	1,3336	1,3432	1,3127
	Standart hata	0,0665	0,0670	0,1975
GÜÇ		1,00	0,947	0,120

Tahmin için girilen yeni verilere en iyi uyumu sağlayan CART algoritmasıdır. MOB algoritması çıktı olarak sabit bir değer yerine bir regresyon modeli vermektedir. Çoğunlukla bölünme gerçekleştirmemesi tahmin için çoğunlukla tüm veriye uydurduğu regresyon modelini kullandığını göstermektedir.

4.2.4. Kaza sayısı verisi ile simülasyon

Kaza sayısı verisinin değişkenleri temel alınarak bir simülasyon modeli kurulmuştur. Verideki bağımsız değişkenlerin değerleri ve Poisson modelinin katsayı tahminleri kullanılarak benzetilen 4000 gözleme sahip 1000 tane veri seti üretilmiştir. Bağımsız değişkenler (X_k) orijinal verideki sıralama ile veri setine eklenmiştir. Bağımlı değişken, bağımsız değişkenler kullanılarak elde edilen ortalama ile Poisson dağılımından üretilmiştir. Üretilen veri setleri $80/20$ oranında eğitim ve test seti olarak ikiye ayrılmış. Test seti ile eğitim setine uygulanan ağaç algoritmalarının tahmin hatasının belirlenmesi amaçlanmıştır. Simülasyon verileri ile oluşturulan regresyon ağaçlarının ilk bölmeyi gerçekleştirdiği bölme değişkenlerinin sayısı Çizelge 4.22'de verilmiştir.

Çizelge 4.22. Kaza sayısı simülasyonu ile elde edilen ilk ayırma değişkenlerinin frekansları

	CART	GUIDE	MOB
Bölme yok	219	0	855
X_1 (yolun tipi)	2	7	77
X_2 (hız limiti)	678	667	8
X_3 (şerit sayısı)	0	0	7
X_4 (gün durumu)	57	218	3
X_5 (yolun yüzeyi)	14	23	45
X_6 (yol genişliği)	30	85	5

Simülasyonda seçilen kök düğümü bölen en önemli değişken incelendiğinde CART ve GUIDE algoritmalarının en çok orijinal veride yolun hız limitine denk gelen X_2 değişkenini seçtiği görülmektedir. MOB algoritması ise çoğunlukla bölme yapmamıştır. Orijinal veride ilk ayırma değişkeni olarak seçtiği X_2 değişkenini yalnızca 8 kez seçmiştir. GUIDE algoritmasının 1000 veri setinin hepsinde bölme gerçekleştirdiği görülmektedir. Kaza sayısı verisi temel alınarak yapılan simülasyon çalışması için CART algoritmasının

gücü 0,781; GUIDE algoritmasının gücü 1; MOB algoritmasının gücü ise 0,145 olarak elde edilmiştir.

Simülasyonda üretilen test verileri ile elde edilen tahmin sonuçlarına göre MSE ortalaması, RMSE ortalaması, MAE ortalaması ve bunların standart hataları hesaplanmıştır. Sonuçlar Çizelge 4.23'te verilmiştir.

Çizelge 4.23. Kaza sayısı simülasyonu ile elde edilen tahmin hataları ve güç

		CART	GUIDE	MOB
MSE	Ortalama	1,3963	1,3825	1,7453
	Standart hata	0,0687	0,0683	0,9250
RMSE	Ortalama	1,1813	1,1754	1,3009
	Standart hata	0,0291	0,0290	0,2300
MAE	Ortalama	0,9529	0,9421	0,9604
	Standart hata	0,0216	0,0218	0,2486
GÜÇ		0,781	1,00	0,145

Kaza sayısı verisi temel alınarak yapılan simülasyon çalışmasında en iyi performansı GUIDE algoritması en kötü performansı MOB algoritması göstermiştir.

5. SONUÇ

Karar ağaçları ile ilgili geniş bir literatür bulunmaktadır. Bütün algoritmaları karşılaştırmak oldukça güç olduğu için bu çalışma tek ağaç üreten üç algoritma ile yapılmıştır. Özyinelemeli bölümleme yöntemlerinden sayma verilerine uygun olduğu düşünülen 3 algoritma ele alınmıştır. Literatürde sayma verileri için bu algoritmaları karşılaştıran simülasyon çalışmasına rastlanmamıştır ve bu çalışmanın ilk olduğu düşünülmektedir.

Ağaç tabanlı yöntemler, parametrik bir forma ihtiyaç duymaması, kolay yorumlanması, işlem hızı, tahmin doğruluğu, uygulama kolaylığı ve kullanılan yazılımlara erişilebilirlik açısından tercih edilmektedir. Kullanılan üç algoritmaya da ticari olmayan yazılımlar üzerinden erişilebilmektedir.

CART algoritması bilinen en eski regresyon ağacı algoritmalarından biridir. Anlaşılması ve yorumlaması kolay grafik çıktısı veren algoritma oldukça kullanışlıdır. Değişken seçim yanlılığı göz önüne alındığında sayısal tipteki değişkenleri benzer oranlarda seçtiği, düzey sayısı az olan kategorik değişkenlerin aleyhine düzey sayısı fazla olan kategorik değişkenlerin lehine seçim yaptığı görülmektedir. Tahmin performansı ve gücü yüksek, tip 1 hatası karşılaştırılan diğer iki algoritmaya göre yüksek elde edilmiştir.

MOB algoritmasının parçalı lineer ağaç modeli kullanması tahmin yeteneğini artırırken yorumlamayı zorlaştırmaktadır. Grafik çıktısının anlaşılması ve yorumlanması kullanılan diğer iki modele göre daha zordur. İstenildiği takdirde ayırma değişkenleri ve regresyon değişkenleri ayrı ayrı girilebilmektedir. Yapısal kırılma testleri ile ayırma değişkeni seçerek yansız değişken seçimine ulaştığı ancak bazı X değişkenleri hem ayırma hem de model uydurma için kullanıldığında algoritmanın yansız olmadığı belirtilmektedir (Loh, 2014). Yapılan simülasyonda ayırma değişkeni seçimini sayısal değişkenlerin lehine kategorik değişkenlerin aleyhine yönelik yaptığı görülmüştür. Düğümlere uydurulabilen regresyon modellerinin çeşitli olması bir avantaj gibi gözükmekte ancak pratiklik açısından engel teşkil etmektedir. Algoritmanın tip 1 hatasının iyi olduğu fakat gücünün çok düşük olduğu görülmektedir. Bölümleme yapması gereken veriyi bölmekte iyi olmadığı anlaşılmaktadır.

GUIDE algoritmasının uygulanması kendi programı aracılığıyla yapılmaktadır. Diğer iki algoritmadan daha uğraştırıcı gözükse de işlem hızı oldukça yüksektir. Tekli regresyon

ağaçları için lineer ve Poisson regresyonun da içinde bulunduğu yedi farklı regresyon modeli uygulanabilmektedir. Tip 1 hata, güç ve tahmin açısından oldukça iyi performans göstermektedir. Poisson ve negatif binom modelleri için ayırma değişkenlerini diğer iki algoritmanın aksine yansız bir şekilde seçtiği görülmüştür. Anlaşılması ve yorumlanması kolay grafik çıktıları vermektedir.

Genel olarak bakıldığında değişkenlerin yansız olarak seçme, tip 1 hata ve güç bakımından karşılaştırılan üç algoritma arasında en iyi sonuç GUIDE algoritması ile elde edilmiştir. Gelecek verileri tahmin etme açısından özellikle aşırı yayılımın varlığında en iyi sonucu CART algoritması vermiştir. Gerçek veriler üzerinden yapılan uygulamalarda CART ve GUIDE algoritmaları ile birbirine yakın değerde tahmin hataları edilmiştir. Belirtilen iki algoritmadan test seti ile elde edilen tahmin hatası Poisson ve negatif binom regresyon modellerinden daha iyi sonuç vermiştir. Sıfır yığılmalı durumda sıfır yığılmalı modeller, regresyon ağaçlarına çok yakın olmakla birlikte onlardan daha küçük tahmin hatası vermiştir. MOB algoritmasının gerçek verilerdeki test seti ile oluşturulan tahmin hatası regresyon modellerinden ve karşılaştırılan iki algoritmadan daha yüksek elde edilmiştir.

Sayma verilerinin analizi için standart metot Poisson veya negatif binom regresyon modeli olsa da ağaç tabanlı yöntemlerin de yorumlama ve tahmin için kullanılabileceği açıkça görülmektedir. Uygulamada ve seçim yanlılığı dışındaki simülasyonlarda model olarak Poisson seçilmiş ve ince ayar yapılmamış algoritmaların yazılımda bulunan varsayılan ayarlarıyla çalışma gerçekleştirilmiştir. Ele alınan veriye ve probleme göre GUIDE ve MOB algoritmalarında ayırma değişkeni belirlenmesi veya MOB algoritması için düğüm modelinin değiştirilmesi ile farklı sonuçların elde edilebileceği göz önüne alınmalıdır.

KAYNAKLAR

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B.N., and Caski, F. (Eds.), *Second International Symposium on InformationTheory*. Budapest: Akademia Kiado, pp. 267-281

Berrar, B. (2018). Cross-validation. In Cannataro, M., Gaeta, B., and Khan, M.A. (Eds.), *Encyclopedia of Bioinformatics and Computational Biology Volume 1*. United States: Elsevier, pp. 542-545.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees*. (First Edition). New York: CRC Press.

Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, 51(2), 738-743.

Cao, L. (2019). *CounTree: Regression tree for count data*, Doktora Tezi, University of Wisconsin-Madison, Wisconsin.

Cameron, A.C., and Trivedi, P.K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics,* 46(1990), 347-364.

Cameron, A.C., and Trivedi, P.K. (1998). *Regression Analysis of Count Data.* (First Edition). New York: Cambridge University Press.

Chaudhuri, P., Huang, M.C., Loh, W.Y., and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica,* 4(1), 143-167.

Chaudhuri, P., Lo, W.D., Loh, W.Y., and Yang C.C. (1995). Generalized Regression Trees. *Statistica Sinica,* 5(2), 641-666.

Choi, Y., Ahn, H., and Chen, J.J. (2005). Regression trees for analysis of count data with extra Poisson variation. *Computational Statistics and Data Analysis,* 49(2005), 893-915.

Ciampi, A. (1991). Generalized regression trees. *Computational Statistics and Data Analysis,* 12, 57-78.

Costa, V.G., and Pedreira, C.E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review,* 56, 4765-4800.

Dobbin, K. K., and Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, 4(31), 1-8.

Feng, C.X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. *Journal of Statistical Distributions and Applications,* 8(1), 8.

Greenwell, B.M. (2022). *Tree-based methods for statistical learning in R* (First Edition). Boca Raton: CRC Press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning* (Second Edition). New York: Springer.

He, H., Tang, W., Wang, W., and Crits-Christoph, P. (2014). Structural zeroes and zero-inflated models. *Shanghai Archives of Psychiatry,* 26(4), 236-242.

Hilbe, J.M. (2011). *Negative Binomial Regression.* (Second Edition). Cambridge: Cambridge University Press.

Hilbe, J.M. (2014). *Modelling Count Data.* (First Edition). New York: Cambridge University Press.

Hothorn, T., and Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, 16(1), 3905-3909.

İnternet: Borkovec, M., Madin, N., Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., and Yutani, H. (October 13, 2022). *Package "ggparty"*. URL: https://cran.r-project.org/web/packages/ggparty/ggparty.pdf

İnternet: Erhardt, E.B. *Logistic regression and Newton-Raphson.* URL: https://statacumen.com/teach/SC1/SC1_11_LogisticRegression.pdf Son Erişim Tarihi: 15.09.2023

İnternet: Friendly, M. (August 22, 2023). *Package "vcdExtra".* URL: https://cran.r-project.org/web/packages/vcdExtra/vcdExtra.pdf

İnternet: Jackman, S. (May 10, 2023). *Package "pscl".* URL: https://cran.r-project.org/web/packages/pscl/pscl.pdf

İnternet: Kleiber, C., and Zeileis, A. (October 12, 2022). *Package "AER".* URL: https://cran.r-project.org/web/packages/AER/AER.pdf

İnternet: Loh, W.-Y. (2023). *User Manual for GUIDE ver. 41.2*. URL: https://pages.stat.wisc.edu/~loh/treeprogs/guide/guideman.pdf

İnternet: R Development Core Team (2023). R: A language and environment for statistical computing. URL: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf

İnternet: Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., and Firth, D. (May 4, 2023). *Package 'MASS'*. URL: https://cran.r-project.org/web/packages/MASS/MASS.pdf

İnternet: Therneau, T.M., and Atkinson, E.J. (October 9, 2023). An introduction to recursive partitioning using the rpart routines. *Mayo Foundation.* URL: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining*, 15(4), 531-538.

Joseph, V.R., and Vakayil, A. (2022). SPlit: An optimal method for data splitting. *Technometrics,* 64(2), 166-176.

Kleiber, C., and Zeileis, A. (2008). *Applied Econometrics with R.* New York: Springer-Verlag.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics,* 34(1), 1-14.

Lee, S.-K., and Jin, S. (2006). Decision tree approaches for zero-inflated count data. *Journal of Applied Statistics,* 33(8), 853-865.

Liu, N.-T., Lin, F.-C., and Shih, Y.-S. (2020). Count regression trees. *Advances in Data Analysis and Classification,* 14, 5-27.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica,* 12(2002), 361-386.

Loh, W.-Y. (2006). Regression tree models for designed experiments. In *Second E. L. Lehmann Symposium, IMS Lecture Notes-Monograph Series Vol. 49*. Bethesda, MD: Institute of Mathematical Statistics, pp. 210-228.

Loh, W.-Y. (2008). Regression by parts. In Chen, C., Hardle, W., and Unwin, A. (Eds.), *Handbook of data visualization*. Berlin: Springer, pp. 447-469.

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review,* 82(3), 239-348.

Long, J.S. (1990). The origins of sex differences in science. *Social Forces,* 68(4), 1297-1316.

Morgan, J.N., and Sonquist, J.A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association,* 58(302), 415-434.

Murthy, S.K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery,* 2, 345-389.

Rokach, L., and Maimon, O. (2015). *Data Mining with Decision Trees Theory and Applications.* (Second Edition). Singapore: World Scientific Publishing, 69-75.

Rusch, T., and Zeileis, A. (2013). Gaining insight with recursive partitioning of generilazed linear models. *Journal of Statistical Computation and Simulation,* 83(7), 1301-1315.

Sammut, C., and Webb, G.I. (2017). Supervised Learning. In *Encyclopedia of Machine Learning and Data Mining*. Boston: Springer, pp. 1213-1214.

Venables, W.N., and Ripley, B.D. (2002). *Modern Applied Statistics with S.* (Fourth Edition). New York: Springer.

Yang, C.-C. (1993). *Tree structured Poisson regression*. Doktora Tezi, University of Wisconsin-Madison, Wisconsin.

Yang, Z., Hardin, J.W., and Addy, C.L. (2010). Score tests for zero-inflation in overdispersed count data. *Communication in Statistics – Theory and Methods,* 39(11), 2008-2030.

Yang, L., Liu, S., Tsoka, S., and Papageorgiou, L.G. (2015). Mathematical programming for piecewise linear regression analysis. *Expert Systems with Applications,* 44, 156-167.

Zeileis, A., and Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, 61(4), 488-508.

Zeileis, A., Hothorn, T., and Hornik, K. (2008a). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2), 492–514.

Zeileis, A., Kleiber, C., and Jackman, S. (2008b). Regression models for count data in R. *Journal of statistical software*, 27(8), 1-25.

EKLER

EK-1. Maksimum olabilirlik tahmin fonksiyonları

```
loglike_func <- function(y, X, beta){
   # Poisson log-olabilirlik fonksiyonu
   y <- as.matrix(y)
   beta <- as.matrix(beta)
   X <- as.matrix(X)
   xbeta <- X %*% beta
   l <- y * xbeta - exp(xbeta) - log(factorial(y))
   return(sum(l))
}

score_func <- function(y, X, beta){
   # Poisson regresyonu icin gradient / score fonksiyonu
   y <- as.matrix(y)
   beta <- as.matrix(beta)
   X <- as.matrix(X)
   xbeta <- X %*% beta
   s <- t(y - exp(xbeta)) %*% X
   return(s)
}

hessian_func <- function(X, beta){
   # Poisson regresyonu icin Hessian matrisi fonksiyonu
   beta <- as.matrix(beta)
   X <- as.matrix(X)
   h <- matrix(0, nrow = length(beta), ncol = length(beta))
   for (i in 1:length(X[,1])){
      h <- h - t(exp(X[i,] %*% beta) %*% X[i,]) %*% X[i,]
   }
   return(h)
}

newton_raphson_func <- function(y, X, beta1, maxiter = 50, tol = 1e-6){
   # Poisson regresyonunun beta katsayilarini belirlemek icin
   # Newton Raphson metodu fonksiyonu
   beta0 <- rep(-Inf, length(beta1))
   # beta değerlerinin arasindaki uzaklik
   diff_beta <- sqrt(sum((beta1 - beta0)^2))

   i <- 1
   while ((i < maxiter) & (diff_beta > tol)){
      beta0 <- beta1
      # Hessian matrisi
      h <- hessian_func(X = X, beta = beta0)
      # skor fonksiyonu
      s <- score_func(y, X, beta0)
      beta1 <- beta0 - solve(t(h), t(s))
      # beta değerlerinin arasindaki uzaklik
      diff_beta <- sqrt(sum((beta1 - beta0)^2))
```

EK-1. (devam) Maksimum olabilirlik tahmin fonksiyonları

```
      beta_hist <- rbind(beta_hist, matrix(beta1, nrow = 1))
      i <- i + 1
   }
   out <- list()
   out$beta_hist <- beta_hist
   out$beta_coef <- beta1
   out$std_error <- sqrt(-diag(solve(hessian_func(X, beta1))))
   out$iteration <- i-1
   return(out)
```

"fs_func" fonksiyonu, lojistik regresyon için oluşturulmuş Neton-Raphson algoritmasını kullanan koddan Poisson modeline uyarlanmıştır (Erhardt, bt).

```
fs_func <- function(X, y, beta1, tol = 1e-6, maxiter = 50){
   # Fisher's scoring for estimation of
   # Poisson regression model (with line search)
   beta1 <- as.matrix(beta1)
   X <- as.matrix(X)
   # initial beta2
   beta2 <- as.matrix(rep(-Inf, length(beta1)))
   # Euclidean distance between beta1 and beta2
   diff_beta <- sqrt(sum((beta1 - beta2)^2))

   # log-likelihood
   llike_1 <- loglike_func(y, X, beta1)
   llike_2 <- loglike_func(y, X, beta2)
   # log-likelihood difference
   diff_like <- abs(llike_1 - llike_2)
   if (is.nan(diff_like)) {diff_like <- 1e9}

   # initial iteration value
   i <- 1
   # line search step without zero
   alpha_step <- seq(-1, 2, by = 0.1)[-11]

   # iteration history
   NR_history <- data.frame(i, diff_beta, diff_like, llike_1, step_size = 1)
   # beta history
   beta_history <- matrix(beta1, nrow = 1)
   while ((i <= maxiter) & (diff_beta > tol) & (diff_like > tol)) {
      i <- i+1
      # update beta values
      beta2 <- beta1
      # variance matrix
      v2 <- diag(as.vector(exp(X %*% beta2)))
      # score function
      score <- score_func(y, X, beta2)
```

EK-1. (devam) Maksimum olabilirlik tahmin fonksiyonları

```
      # increment
      increm <- solve(t(X) %*% v2 %*% X, t(score))

      # line search
      llike_alpha_step <- rep(NA, length(alpha_step))
      for (i_step in 1:length(alpha_step)) {
         llike_alpha_step[i_step] <-
            loglike_func(y, X, beta2 + alpha_step[i_step]*increm)
      }
      # step size index for maximum log-likelihood
      ind_max_alpha_step <- which(llike_alpha_step == max(llike_alpha_step))[1]
      # update beta values
      beta1 <- beta2 + alpha_step[ind_max_alpha_step] * increm
      # Euclidean distance between beta1 and beta2
      diff_beta <- sqrt(sum((beta1 - beta2)^2))

      # update log-likelihood
      llike_2 <- llike_1
      llike_1 <- loglike_func(y, X, beta1)
      diff_like <- abs(llike_1 - llike_2)

      NR_history <- rbind(NR_history, c(i, diff_beta, diff_like, llike_1,
                                        alpha_step[ind_max_alpha_step]))
      beta_history <- rbind(beta_history, matrix(beta1, nrow = 1))
   }
   # output
   out <- list()
   out$beta_MLE <- beta1
   out$iteration <- i -1
   out$NR_history <- NR_history
   out$beta_history <- beta_history
      v1 <- diag(as.vector(exp(X %*% beta1)))
      linv1 <- solve(t(X) %*% v1 %*% X)
   out$beta_cov <- linv1
   out$std_error <- sqrt(diag(linv1))

   return(out)
}
```

Gazili olmak ayrıcalıktır