ISSN 2757-5543 GÜFFD 5. Cilt (2): 149-160 (2024)

Gazi Üniversitesi Fen Fakültesi Dergisi
https://dergipark.org.tr/tr/pub/guffd

GÜFFD GAZİ ÜNİVERSİTESİ FEN FAKÜLTESİ DERGİSİ

Trafik Kaza Sayılarının Regresyon ve Karar Ağacı Yöntemleri ile Modellenmesi: Ankara Devlet Yolları Örneği

Mine Fulya Gürsel[1], Hatice Tül Kübra Akdur[*2]

[1] Gazi Üniversitesi, Fen Bilimleri Enstitüsü, 06500, Ankara, Türkiye
[2]Gazi Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 06500, Ankara, Türkiye

Öne Çıkanlar

- *Trafik kaza sayılarının Poisson regresyonu ile modellenmesi.*
- *Sayım verileri için regresyon ağaçlarının oluşturulması.*
- *GUIDE, MOB ve CART algoritmalarının trafik kazası verilerinde kullanımı.*

Makale Bilgileri

Geliş: 29/03/2024
Kabul: 04/08/2024

Anahtar Kelimeler

Poisson regresyon, Karar ağacı algoritmaları, Trafik kaza sayısı

Öz

Karayolu trafik sistemleri insanların her gün karşılaşmak zorunda olduğu tehlikeli sistemlerdendir. Güvenli olmayan karayolu trafik sistemlerinin halk sağlığına ve kalkınmaya zarar verdiği bilinmektedir. Dünya Sağlık Örgütü verileri her yıl yaklaşık 1,3 milyon insanın trafik kazası sonucu yaşamını yitirdiğini göstermektedir. Türkiye'de 2022 yılında 197 bin insan trafik kazaları sonucunda yaralanmış veya hayatını kaybetmiştir. Trafik kazalarının oluşumunda çeşitli faktörler etkili olabilir. Bu faktörlerin incelenmesi, trafik kazalarının analiz ve tahmini için istatistiksel yöntemlerin yanı sıra makine öğrenmesi yöntemleri de kullanılmaktadır. Bu çalışmada Ankara ili ve ilçeleri devlet yollarında 2017-2020 yıllarında meydana gelen trafik kazaları, sayma verisine uygun regresyon modeli ve karar ağacı yöntemleri ile modellenmiştir. Analiz bulguları sonucunda, çeşitli faktörlerin trafik kazalarını nasıl etkilediğine dair kritik bilgileri ortaya koymuştur. CART algoritması yasal hız sınırlarını en önemli belirleyici olarak belirlemiştir. MOB ve GUIDE algoritmaları, belirli koşulların kaza oranlarını nasıl etkilediğine ilişkin ek incelikli bilgiler sağlamıştır. Bulgular, trafik güvenliğini artırmak ve politika kararlarını bilgilendirmek için birden fazla faktörün dikkate alınmasının önemini vurgulamıştır. Karşılaştırmalı performans değerlendirmesine göre, CART algoritmasının en düşük tahmin hatalarına sahip olduğu, onu yakından GUIDE algoritmasının takip ettiği, MOB algoritmasının ise daha yüksek tahmin hatasına sahip olduğu görülmüştür.

Modeling the Number of Traffic Accidents with Regression and Decision Tree Methods: State Highways of Ankara Example

Highlights

- Modeling the number of traffic accidents with Poisson regression.
- Creating regression trees for count data.
- Utilizing of GUIDE, MOB and CART algorithms in traffic accident data.

Article Info

Received: 29/03/2024
Accepted: 04/08/2024

Keywords

Poisson regression, Decision tree algorithms, Number of traffic accidents

Abstract

Road traffic systems are among the dangerous systems that people have to encounter every day. Unsafe road traffic systems are known to harm public health and development. World Health Organization data show that approximately 1.3 million people die as a result of traffic accidents every year. In Turkey, 197 thousand people will be injured or die as a result of traffic accidents in 2022. Various factors may be effective in the occurrence of traffic accidents. In addition to statistical methods, machine learning methods are also used to examine these factors and analyze and predict traffic accidents. In this study, traffic accidents that occurred on the state roads of Ankara province and its districts between 2017 and 2020 were modeled with regression models and decision tree methods suitable for counting data. As a result of the analysis findings, critical information was revealed about how various factors affect traffic accidents. The CART algorithm has identified legal speed limits as the most important determinant. The MOB and GUIDE algorithms provided additional nuanced insights into how specific conditions affect accident rates. The findings highlighted the importance of considering multiple factors to improve traffic safety and inform policy decisions. According to the comparative performance evaluation, it was found that the CART algorithm had the lowest prediction errors, followed closely by the GUIDE algorithm, while the MOB algorithm had higher prediction errors.

Araştırma Makalesi

* Sorumlu Yazar/Corresponding Author: Hatice Tül Kübra Akdur, haticesenol@gazi.edu.tr

1. GİRİŞ

Karayolu taşıma sistemlerinin kullanımının amacı insanları, eşyaları, hayvanları, gıda maddelerini verimli, sağlıklı ve güvenilir olarak bir yerden bir yere taşımaktır. Trafik kazaları ve yolda meydana gelen çalışmalar olumsuzluklar karayolu sisteminin işleyişini bozmaktadır. Dünya Sağlık Örgütü (DSÖ) verilerine göre her yıl trafik kazaları sonucu yaklaşık 1,3 milyon insan hayatını kaybetmekte ve 50 milyona yakın insan da yaralanmaktadır [1]. 2022 yılında Türkiye'de 197 bin 261'i ölümlü ya da yaralanmalı olmak üzere 1 milyon 232 bin 957 trafik kazası meydana gelmiştir [2]. Trafik kazası, bir karayolu taşıtının diğer bir taşıta, yayaya, hayvana, ağaca veya herhangi başka bir nesneye çarpması olarak tanımlanabilir. Trafik kazaları genellikle yaralanma, maddi zarar ve ölümle sonuçlanır. Dünya Sağlık Örgütü (DSÖ) ve Dünya Bankasınca ortak olarak hazırlanan trafik kazalarının önlenmesine ilişkin dünya raporunda, karayolu trafik kazaları büyük bir halk sağlığı ve gelişim sorunu olduğu belirtilmiştir [3]. Trafik kazalarının oluşmasında hava şartları, sürücünün sosyoekonomik durumu ve eğitim düzeyi, yolun tasarımı/ çevresi, sürüş hızı gibi birçok risk faktörü etkili olabilir.

Trafik kazalarının analiz ve tahmini için geleneksel olarak uygulanan regresyon modellerinin yanı sıra son zamanlarda karar ağaçları, yapay sinir ağları, destek vektör makineleri, kümeleme algoritmaları gibi çeşitli makine öğrenmesi modelleri de kullanılmaktadır. Kaza tahmini için makine öğrenmesi ile negatif binom regresyonu modellerinin karşılaştırıldığı çalışmalarda makine öğrenmesi tekniklerinin çoğunlukla regresyon modellerinden daha iyi performans gösterdiği belirtilmiştir [4].

Literatürde, makine öğrenmesi yöntemleriyle yapılan trafik kazaları analizinde sıklıkla sınıflandırma algoritmalarını kullanan çalışmalar olduğu; regresyon ağacı algoritmalarını kullanan çalışmaların az olduğu görülmektedir. 2017-2020 yılları arasında Kahramanmaraş ilinde gerçekleşen trafik kazaları sınıflandırma, birliktelik kuralı yöntemleri ile analiz edilmiştir [5]. Antalya ilinde 2012-2016 yılları arasında gerçekleşen ölümlü, yaralanmalı trafik kazaları ile yapılan çalışmada sınıflandırma algoritmalarının performansları karşılaştırılmıştır [6]. Adana ili 2005-2014 yıllarındaki trafik kazaları verisiyle yapılan çalışmada, yaralı sayısı ve yaralanmalı kaza sayısının bağımlı değişken olarak alındığı çoklu doğrusal regresyon analizinin yanı sıra regresyon ağacı, yapay sinir ağları ve destek vektör makineleri yöntemleri karşılaştırılmıştır [7]. Al-Asadi vd. (2022) trafik kazalarının önceden tahmin edilmesi, ulaşımın ve kamu güvenliğinin iyileştirilmesi için Türkiye'de 2029 yılına kadar trafik kazalarındaki ölü veya yaralı sayısını tahmin etmek için karar ağaçlarını da içeren üç makine öğrenme tekniğini uygulamıştır [8].

Türkiye'de yapılan literatürden araştırıldığı kadarıyla bu istatistiksel teknikleri kullanarak Ankara ili ve ilçelerinde trafik kaza sayılarını ve ilişkili risk faktörlerini belirleyen bir bilimsel çalışmaya rastlanmamıştır. Bu çalışma Ankara ili ve ilçeleri devlet yolları özelinde trafik kaza sayılarına odaklanarak kaza sayılarına etki eden risk faktörlerini belirlemeyi amaçlamıştır.

1. MATERYAL VE YÖNTEM

Bu çalışmada kullanılan veriler Ankara Emniyet Genel Müdürlüğü Trafik Eğitim ve Araştırma Dairesi Başkanlığı'ndan temin edilmiştir. Türkiye'nin başkenti olması sebebiyle Ankara ilinde yapılan iyileştirmeler tüm Türkiye'ye örnek olacağından bu çalışmada Ankara'ya ait kaza sayıları ele alınmıştır. 2017-2020 yılları arasında Ankara ili devlet yolunda polis sorumluluk bölgesinde meydana gelen kazalara ait bilgiler Emniyet Genel Müdürlüğü'nden resmi olarak gerekli izinler alınarak temin edilmiştir.

2995 gözlem bulunan bu veri seti, bir trafik kazasının bazı temel özelliklerini içermektedir, bu özellikler arasında kaza yol adı, yol şerit sayısı, yasal hız sınırı, yol genişliği, yolun tipi, yüzeyi ve gündüz/gece durumu yer almaktadır. Veri setindeki bu bilgiler, trafik kazalarının ne zaman, nerede ve nasıl meydana geldiğine dair önemli ipuçları sağlayarak trafik kazalarının önüne geçmek için alınabilecek önlemleri belirlemede kullanılabilir. Veri setinde yer alan bağımsız değişkenler ve açıklamaları Çizelge 1'de verilmiştir.

Çizelge 1. *Bağımsız değişkenler ve açıklamaları*

Değişkenin Adı	**Açıklama/ Tipi**	**Düzeyleri**
Yol şerit sayısı	Sayısal/ Kesikli	
Yolun yasal hız sınırı	Sayısal/ Sürekli	
Yol genişliği (cm)	Sıralı/ Kategorik	[300-500] [600-800] [900 +]
Yolun tipi	Kategorik	1. Tek yönlü 2. Bölünmüş 3. İki yönlü 4. Diğer
Yolun yüzeyi	Kategorik	1. Kuru 2. Buzlu 3. Diğer kaygan 4. Islak/ nemli 5. Karlı 6. Su birikintili
Gün durumu	Kategorik	1. Gündüz 2. Gece 3. Alacakaranlık

2.1. Poisson Regresyon Modeli Analizi

Çalışmada kullanılan veri setinin modellenmesinde sayma verileri için geliştirilmiş Poisson regresyon modeli ve ilgili karar ağacı teknikleri kullanılmıştır. Poisson regresyon modelinde aşırı yayılımı test etmek için AER R paketi kullanılmıştır [9]. MOB karar ağacı için partykit ve CART karar ağacı için rpart R paketleri kullanılmıştır [10,11]. GUIDE karar ağacını elde etmek için Loh (2023) tarafından geliştirilen GUIDE ver. 41.1 programı kullanılmıştır [12].

Regresyon analizi bir bağımlı değişken ile bir veya daha fazla bağımsız değişkenler arasındaki neden-sonuç ilişkisini belirler. Klasik regresyonda bağımlı değişken normal dağılıma sahipse uygulanabilir. Trafik kaza sayısı sayım yoluyla elde edilmiş kesikli tam sayı türünde bir veridir. Sayım verileri sağlık, mühendislik, sigortacılık, psikoloji, eğitim gibi birçok farklı alanda sıklıkla karşımıza çıkmaktadır. Klasik doğrusal regresyon varsayımları sayım verileri için sağlanmaz ve varsayımların sağlanamadığı durumlar sonuçların hatalı olmasına neden olur. Bağımlı değişken kesikli (sayma verisi) olduğu durumlarda Poisson, negatif binom, com-poisson vb. modeller uygulanır. Bu modeller ayrıca bağımlı değişkenin normal dağılım göstermediği durumlar için veri dönüşümüne alternatif olarak kullanılabilmektedir. Poisson regresyonu ve negatif binom regresyon modelleri sayım verilerinde en çok kullanılan modellerdendir [13]. $Y_1,...,Y_n$ bağımsız rasgele değişkenler ve $Y_i \sim Poisson(\mu_i)$ olmak üzere

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \quad (1)$$

Eş.1'de verilen model log doğrusal modeli olarak adlandırılır. Eşitlikte ortalama bağımsız değişkenlerin üstel bir fonksiyonu olmaktadır:

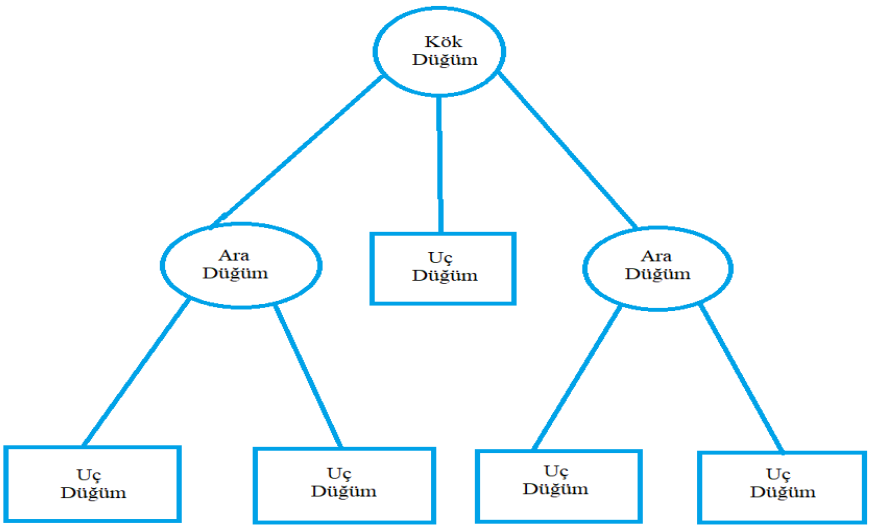
$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}) \quad (2)$$

Poisson regresyon modelinde $E[Y_i \mid X_i] = Var[Y_i \mid X_i] = \mu_i ; \mu_i = \exp[X_i^T \beta]$ olmak üzere ortalama ve varyansın eşitliğine eşit yayılım denilmektedir. Gözlemlenen veriler bu varsayımın izin verdiğinden daha fazla değişkenlik gösteriyorsa, bu aşırı dağılıma işaret eder. Bu çalışmada ele alınan veri setinde aşırı yayılım problemi görülmemiştir. Bu nedenle Poisson regresyon modeli yeterli olduğu için negatif binom regresyon modeli veri setine uygulanmamıştır.

2.2. Karar Ağaçları

Bir karar ağacı modeli, büyük bir gözlem koleksiyonunu belirli bir hedef değişkene göre daha küçük homojen gruba bölmek için uygulanan bir dizi kuraldan oluşur. Hedef değişken genellikle kategoriktir ve karar ağacı modeli, belirli bir kaydın hedef kategorilerden her birine ait olma olasılığını hesaplamak veya kaydı en olası kategoriye atayarak sınıflandırmak için kullanılır. Bir hedef değişken ve bir dizi açıklayıcı değişken verildiğinde, karar algoritmaları hangi değişkenlerin en önemli olduğunu otomatik olarak belirler ve ardından gözlemleri doğru çıktı kategorisine göre sıralar [14]. Karar ağaçları parametrik olmayan bir veri analizi yöntemi olması sebebiyle parametrik yöntemlerin sahip olması gereken varsayımlara gerektirmediği için sıklıkla tercih edilirler. Karar ağaçları genelden özele ve aşağıya doğru eğitilmiş verilerden oluşmuştur. Karar ağaçlarının yapısına bakıldığı zaman kök, dallar ve yapraklardan oluşmaktadır. Karar ağacının ilk bölümü kök düğümdür. Karar ağaçlarında kök düğüm bağımlı değişken üzerinde en çok etkiye sahip bağımsız değişkene, bir başka deyişle özniteliklerden birine karşılık gelmektedir. Kök düğümle başlayan ağacın aşağıya doğru inildikçe veri kümelerini daha küçük gruplara ayırdığını görebiliriz. Bu bölümlere de dal adı verilir. Bu kökten başlayıp dallara doğru uzanan ağaç yapısında ki her boğuma düğüm denir. Karar ağaçlarının yapısı Şekil 1'de gösterilmiştir. Bağımlı değişken kategorik ise sınıflama, sürekli ise regresyon ağacı oluşturmaktadır.

Veri madenciliği yazılımlarında yaygın olarak kullanılan karar ağacı algoritması Breiman vd. (1984) tarafından geliştirilen sınıflama ve regresyon ağaçları (Classification and Regression Trees: CART) algoritmasıdır [15]. Bu çalışmada, bağımlı değişkeninin sayma verisi olması durumunda kullanılan veri madenciliği algoritmalarından CART, Genelleştirilmiş, yansız, etkileşim tespiti ve tahmini (Generalized, Unbiased, Interaction Detection and Estimation: GUIDE) ve Model Tabanlı Özyinelemeli Bölümleme (Model-based Recursive Partitioning: MOB) algoritmaları kullanılacaktır. Bu algoritmalar, ağaç benzeri yapı diyagramını ve önemli bilgilerin çıkarılabileceği karar kurallarını üretmektedir.



Şekil 1. *Karar ağacı gösterimi*

2.3. CART Algoritması

İkili ağaçlar oluşturan CART algoritmasında bağımlı değişken kategorik ya da sürekli veri tipinde olabilir. Sınıflandırma durumlarında, gini ya da twoing kriteri, regresyon durumlarında ise en küçük kareler (EKK) yöntemi kullanılmaktadır. CART algoritması her adımda ilgili grubun, kendinden daha homojen olan iki alt gruba ayrılmasını sağlamaktadır. Yani her dal ikili alt gruplara ayrılarak büyümektedir [15]. Sayım verileri için Ciampi (1991), CART'ı düğüm modeli olarak Poisson regresyonuna uyacak şekilde sapma (deviance) fonksiyonu kullanarak genişletmiştir [16]. Ağaçları budamak için önem testi veya Akaike bilgi kriterini kullanmıştır [15]. Poisson regresyonu için CART algoritması, geleneksel CART algoritmasına benzer bir süreç izler, ancak ayırma ve değerlendirme kriterleri, özellikle Poisson dağılan bağımlı değişkenlerine uyarlanır. Ortaya çıkan ağaç, girdi özelliklerine dayalı olarak bağımlı değişkenin oranlarını tahmin eden bir tahmin modeli sağlar. CART algoritması, tüm olası özellik ayrımlarını inceleyerek başlar ve Poisson olabilirliğine veya sapmasına dayalı olarak her bir bölümün verileri ne kadar iyi ayırdığını değerlendirir. Sapmayı en aza indiren veya olabilirliği en üst düzeye çıkaran ayrımı seçer. Poisson regresyonunda, ayırma kriterleri sapma veya olabilirlik oranı testine dayanmaktadır. Sapma, gözlemlenen ve tahmin edilen sayılar arasındaki farkı temsil eder ve olasılık, model verilen verileri gözlemleme olasılığını ölçer. Bir ayırma değişkeni seçildiğinde, veri kümesi seçilen özellik ve ayırma noktasına göre iki alt kümeye bölünür. İşlem daha sonra bir durdurma kriteri karşılanana kadar her bir alt kümede yinelemeli olarak tekrarlanır. Bu kriter maksimum ağaç derinliği, yaprak başına minimum örnek sayısı veya kullanıcı tarafından tanımlanan diğer koşullar olabilir. Her yinelemeli adımda, algoritma durdurma kriterine ulaşana kadar verileri bölmeye devam eder. Durdurma kriteri karşılandığında, verilerin o alt kümesi için nihai bir tahmini temsil eden bir yaprak düğüm oluşturulur. Poisson regresyonunda, yaprak düğüm, o alt küme için bağımlı değişkeninin ortalama veya tahmin edilen oranını tutar.

R paketi rpart ile Poisson regresyon ağacı, CART ağacı yöntemi ile kullanılabilmektedir [11].

2.4. GUIDE Algoritması

Bu algoritma bağımlı değişken kategorik olduğunda bir sınıflama ağacı bağımlı değişken sayısal olduğunda bir regresyon ağacı oluşturulmasını sağlayan değişken seçimi yanlılığını ortadan kaldırmak için önerilmiş çok amaçlı bir makine öğrenme algoritmasıdır [17]. GUIDE algoritması, sayım verisi modellerinde regresyon ağacı oluşturmak için adapte edilmiştir [18]. Her bir düğümde bağımsız değişkenlerin istatistiksel olarak önemini ki-kare testine dayalı artıkları kullanan bir yöntemle (eğrilik testi) belirlemeyi amaçlar. Etkileşimlerin tespiti için hem sayısal bağımsız değişken çiftleri hem de kategorik bağımsız değişken çiftleri için artıkların işaretlerine dayalı bir ki-kare testi uygulanır. Ayrıca, sayısal ve kategorik bağımsız değişken çiftleri için benzer şekilde bir etkileşim testi yapılır. Algoritma, her düğümde, bir ayırma değişkenine karar vermek için her bağımsız değişken ve etkileşimleri için χ^2 ki-kare istatistiği ve p değerini hesaplar ve en küçük p değerine ait değişkeni ayırma değişkeni olarak seçer [17]. En küçük p değeri bir eğrilik testinden geliyorsa, düğümü bölmek için ilgili X değişkenini seçmek doğaldır. En küçük p değeri bir etkileşim testinden geliyorsa, etkileşimli iki değişkenden birini seçmek gerekir. İki değişkenin eğrilik p değerlerine dayanarak seçim yapılabilir, ancak amaç her düğüme sabit bir model sığdırmak olduğundan, seçim en küçük hata kareler toplamındaki azalmaya dayalı yapılır. Etkileşim testindeki iki değişken de sayısal ise, düğüm her değişkenin örnek ortalamalarına göre bölünür; her bölünme için, her bir alt modele ait hata kareler toplamı elde edilir, daha küçük hata kareler toplamını veren değişken seçilir. Aksi halde en az bir değişken kategorik ise eğrilik p değeri daha küçük olan seçilir. GUIDE yarı-parametrik bir karar ağacı algoritmasıdır. Bu yöntem, her düğümdeki verilerin istatistiksel özelliklerine dayalı şekilde verileri bölerek karar ağacı oluşturur. Yöntem, değişkenlerin doğrusal veya doğrusal olmayan etkileşimlerini hesaba katar ve verileri bölerek karar ağacı oluşturur. GUIDE algoritması, çeşitli endüstriyel ve bilimsel uygulamalarda kullanılmıştır. Örneğin, biyomedikal araştırmalarda akciğer kanseri ile ilişkili ölüm oranlarının modellenmesi, mühendislik çalışmalarında baskılı devre kartlarının elektronik bileşenlerin dalga lehimlenmesine ilişkin lehim atlama sayılarının modellenmesi gibi farklı bilim alanlarından ortaya çıkan sayma verilerinde GUIDE algoritmasının kullanımı görülmüştür [18,19].

GUIDE algoritmasının öne çıkan özellikleri arasında göz ardı edilebilir seçim yanlılığı, yüksek doğruluk oranları, verilerdeki etkileşimlerin keşfi ve modelin yorumlanabilirliği yer alır. Sonuç olarak, GUIDE, yarı-parametrik bir yaklaşım kullanır. Yöntem, verilerin daha iyi anlaşılmasına yardımcı olur ve birçok uygulama alanında kullanılabilir.

2.5. MOB Algoritması

MOB yöntemi, mevcut düğümdeki örneğe parametrik bir model uydurur ve her bir ayırma değişkeni boyunca parametrelerin kararlılığını değerlendirir. Parametre kararsızlığı testi, her bir ayırma değişkenine karşılık gelen ampirik dalgalanma süreci dikkate alınarak gerçekleştirilmektedir [20]. Parametre kararlılığının sıfır hipotezi altında, süreç bir Brownian köprüsüne yakınsar. Sıfırdan sapmaları yakalayan sürece bir skaler fonksiyonel uygulanarak bir test istatistiği elde edilir. MOB, ayırma değişken seçimini ayırma küme seçiminden ayırır. Ayrım değişkeni seçildikten sonra, tüm olası ikili ayrım kümeleri göz önünde bulundurularak ayrım kümesi seçilir. Ayırma olasılığını maksimize eden küme seçilir. Prosedür, düğüm boyutu çok küçük olana veya ilgili kararsızlık testleri anlamlı olmayana kadar yinelemeli olarak uygulanır. Her bir uç düğümde, gelecekteki gözlemleri tahmin etmek için tanımlanmış parametrik regresyon fonksiyonu kullanılır. MOB yönteminin ayrıntılı açıklaması Zeileis vd. (2008) ve Kleiber ve Zeileis (2008)'de bulunabilir [20,21].

MOB algoritmasında hedef, her düğümün bir Poisson modeliyle ilişkilendirildiği bir ağaç oluşturmaktır. Algoritma, bir düğümün bölünmesinin gerekli olmadığını belirlemek için parametre kararsızlığına yönelik dalgalanma testlerini kullanır. İlk olarak tüm gözlemler geçerli düğümdeki Poisson modeline uydurulur. Her ayırma değişkeni ilişkin parametre tahminlerinin kararlı olup olmadığı değerlendirilir. Genel bir kararsızlık varsa, en yüksek parametre kararsızlığıyla ilişkili değişken ayırma değişkeni olarak seçilir aksi halde algoritma durdurulur. Hedef fonksiyonu yerel olarak optimize eden bölünme noktaları (düğümler) hesaplanır. Düğümler gerekirse alt düğümlere bölünür ve prosedür her alt düğümde tekrarlanır. Parametre kararsızlığının test edilmesi için genelleştirilmiş M-dalgalanma testleri kullanılır (sayısal değişkenler için supLM istatistiği, kategorik değişkenler için χ^2 istatistiği). Minimum p değerinin anlamlılık düzeyinin altına düşüp düşmediğini kontrol edilir (varsayılan $\alpha = 0{,}05$; çoklu testler için Bonferroni ayarı yapılır). Bölünme noktasını bulmak için kapsamlı bir arama prosedürü benimsenmiştir. Özyinelemeli bölümleme algoritmasının bir yinelemesi, parametre kararsızlığı testinde önemli bir kararsızlık tespit edilmediğinde sona erer. Algoritma model tabanlıdır ve olabilirlik fonksiyonuna dayalı ölçümler kullanır. Ayırma değişkeninin seçimi, en yüksek parametre kararsızlığına dayanmaktadır. Özetle MOB, düğümlerin ne zaman bölüneceğine karar vermek için dalgalanma testleri kullanarak model tabanlı regresyonu özyinelemeli bölümlemeyle birleştirir. Özellikle potansiyel kararsız parametrelere sahip regresyon problemlerine uygun, esnek bir yaklaşım sağlar.

3. BULGULAR VE TARTIŞMA

Poisson regresyon modeline X_1 = Yolun tipi, X_2 = Yolun yasal hız sınırı, X_3 = Yol şerit sayısı, X_4 = Gündurumu, X_5 = Yolun yüzeyi, X_6 = Yol genişliği bağımsız değişkenleri sırasıyla dâhil edilerek analiz gerçekleştirilmiştir. Veri setine uydurulan Poisson log-lineer modeli aşağıda verilmiştir:

$$\begin{aligned} log\,\mu_i = {} & \beta_0 + \beta_{11}x_{i11} + \beta_{12}x_{i12} + \beta_{13}x_{i13} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{41}x_{i41} + \beta_{42}x_{i42} + \beta_{51}x_{i51} \\ & + \beta_{52}x_{i52} + \beta_{53}x_{i53} + \beta_{54}x_{i54} + \beta_{55}x_{i55} + \beta_{61}x_{i61} + \beta_{62}x_{i62} \end{aligned}.$$

Poisson regresyon modelinin parametre tahminleri ile ilgili sonuçlar Çizelge 2'de verilmiştir.

Çizelge 2. *Poisson regresyon modeli sonuçları*

Katsayılar	Katsayı tahmini	Standart hata	*z* değeri	*p* değeri
Sabit	-0,3672	0,1948	-1,885	0,0594
Yolun tipi-Bölünmüş	0,2951	0,1011	2,918	0,0035 **
Yolun tipi-İki yönlü	0,0868	0,1552	0,559	0,5761
Yolun tipi-Diğer	-0,1241	0,4587	-0,270	0,7868
Yolun yasal hız sınırı	0,0037	0,0007	5,311	< 0,0001 ***
Yol şerit sayısı	0,0349	0,0197	-1,772	0,0764
Gün durumu-Alacakaranlık	-0,3404	0,0925	-3,680	0,0002 ***
Gün durumu-Gece	-0,1398	0,0332	-4,211	< 0,0001 ***
Yolun yüzeyi-Buzlu	-0,3320	0,2509	-1,323	0,1858
Yolun yüzeyi-Diğer kaygan	-0,2952	0,4090	-0,722	0,4705
Yolun yüzeyi-Islak/ Nemli	-0,1514	0,0396	-3,826	0,0001 ***
Yolun yüzeyi-Karlı	-0,2587	0,1471	-1,759	0,0786
Yolun yüzeyi-Su birikintili	-0,3813	0,2591	-1,472	0,1411
Yol genişliği- [600-800]	0,0445	0,1658	0,268	0,7885
Yol genişliği- [900 +]	0,2722	0,1560	1,745	0,0810

Yolun tipi kategorik değişkeninde tek yönlü yol referans kategori olarak alınmıştır. Bölünmüş yoldaki ortalama tahmini kaza sayısı tek yönlü yoldaki kaza sayısından yüzde 34 daha fazladır. Yolun yasal hız sınırındaki 1 birimlik (diğer değişkenleri sabit tutulmak kaydıyla) artış tahmini ortalama kaza sayısında binde 4'lük bir artışa sahiptir. Gün durumu kategorik değişkeninde gündüz referans olarak alınmıştır. Alacakaranlık ve gece gerçekleşen ortalama kaza sayısı gündüz gerçekleşen ortalama kaza sayısından sırasıyla yüzde 29 ve yüzde 13 daha azdır. Yolun yüzeyi değişkeninde kuru yol referans olarak alınmıştır. Islak/nemli yollardaki ortalama kaza sayıları kuru yoldaki ortalama kaza sayısından yüzde 14 daha azdır.

Çizelge 3. *Algoritmaların tahmin hatalarının karşılaştırılması*

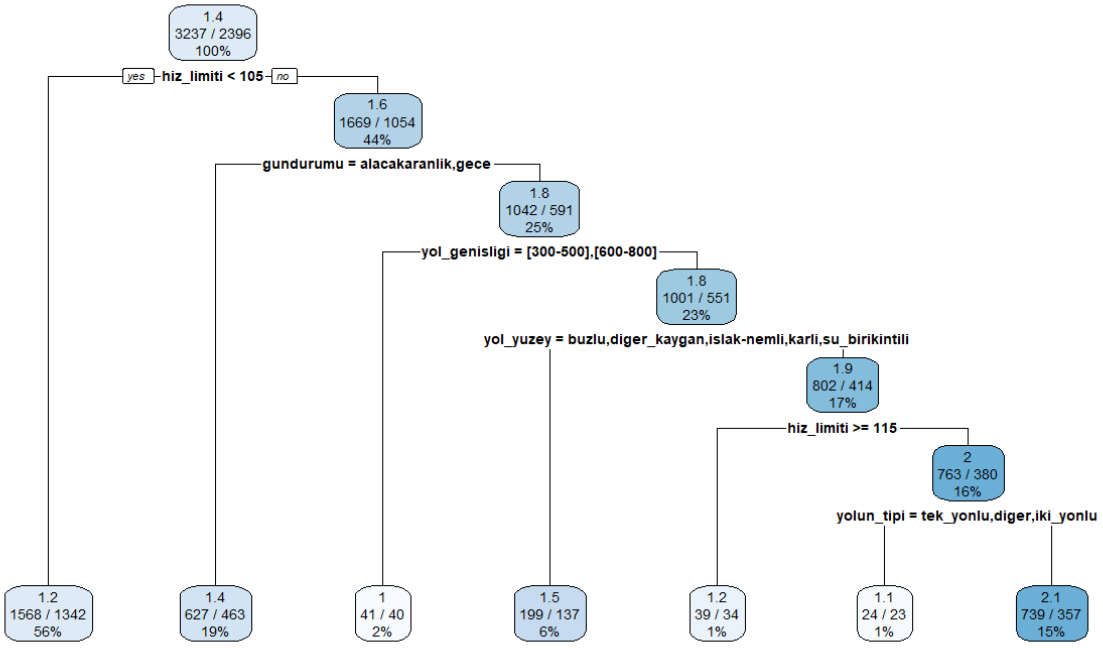
Algoritma	Hata kareler ortalaması	Mutlak hata ortalaması
CART	0,8766	0,5084
GUIDE	0,8881	0,5207
MOB	9,5042	2,9299

Veri seti, modellerin tahmin performanslarını değerlendirmek için 0,8'i eğitim 0,2'si test verisi olacak şekilde ikiye ayrılmıştır. Eğitim seti ile karar ağaçları oluşturulmuş, test seti ile gerçek değerler ve tahmin değerleri arasındaki hata kareler ortalaması ve mutlak hata ortalamaları hesaplanmıştır. Sonuçlar Çizelge 3'te verilmiştir. CART algoritmasından elde edilen karar ağacında her düğümün altında belirtilen koşul sağlanırsa düğüm sola, sağlanamazsa düğüm sağa ayrılmaktadır. CART algoritması Poisson modelinin sapma (deviance) ve olabilirlik değerlerini kullanarak bölünmeleri gerçekleştirir. CART algoritmasının R programı rpart paketi aracılığıyla elde edilen karar ağacı Şekil 2'de verilmiştir. Buna göre, kaza sayılarının belirlenmesinde verilen değişkenlerden en önemlisinin yasal hız sınırı olduğu görülmektedir. Yasal hız sınırı 105 km/saat' in altında olan yollarda meydana gelen kaza sayılarının ortalaması 1,2'dir ve bu yollarda meydana gelen kazalar toplam kaza sayısı verisinin yüzde 56'sını oluşturmaktadır. Yasal hız sınırı 105'e eşit veya daha büyük olan yollarda ortalama 1,6 kaza meydana gelmiştir. Hız sınırı >= 105 olan yollarda alacakaranlık ve gece şartlarında ortalama 1,4 kaza yaşandığı görülmektedir. Hız sınırı >= 105 olan yollarda gündüz şartlarında ve yol genişliği 900 cm'nin altında olanlarda ortalama 1 kaza yaşandığı görülmektedir. Hız sınırı >= 105, yol genişliği 900 cm'nin üstünde olan ve gündüz vaktinde kuru olamayan yollarda ortalama 1,5 kaza meydana gelmiştir. Hız sınırı 105 ile 115 km/saat arasında olan, yol genişliği 900 cm'nin üstünde, kuru yollarda ve gün ışığında daha fazla kaza meydana gelmiştir (ortalama = 2,1) bu da genel verinin ortalamasından daha yüksek bir ortalamaya sahip olduğunu göstermektedir.

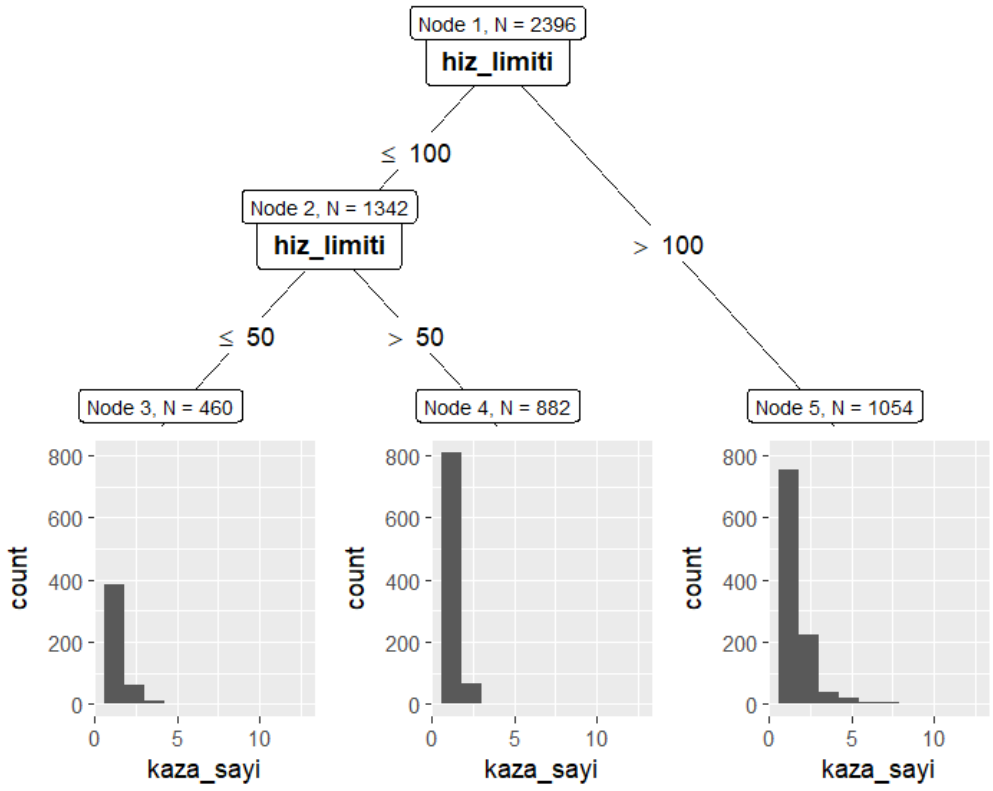
MOB algoritmasının R programı partykit paketi aracılığıyla elde edilen karar ağacı Şekil 3'te verilmiştir. MOB regresyon ağacının yapraklarında düğümlere ait kaza sayısının histogram grafiği çizdirilmiştir. MOB algoritmasında ağaç diyagramının yanı sıra her düğümdeki parametre tahminleri, standart hata ve p değerlerinin olduğu iki farklı tablo daha sağlanmaktadır.

Parametre tahminlerinin her bir düğümdeki p değerleri incelenerek anlamlı olan tahmin değerlerinin ortalama kaza sayısı üzerindeki etkileri aşağıdaki gibi özetlenecektir. 3. ve 4. düğümde istatistiksel olarak anlamlı değişken bulunamamıştır. Hız limiti 100'ün üzerinde olan yolları içeren 5. düğümde bölünmüş yollarda gerçekleşen ortalama kaza sayısı tek yönlü yollara göre yüzde 60 daha fazladır. Hız limitindeki her bir birimlik artış kaza sayısında yüzde 4'lük azalmaya neden olmaktadır. Alacakaranlık ve gece gerçekleşen kaza sayıları gündüz gerçekleşen kaza sayısından sırasıyla yüzde 43 ve yüzde 22 daha az olduğu görülmektedir. Islak/ nemli yollarda gerçekleşen kaza sayısı kuru yollara göre yüzde 22 daha azdır. Bu sonuçlar, düğüm 5'te belirli yol ve çevre koşullarının kaza sayısını önemli ölçüde etkilediğini göstermektedir. Özellikle, bölünmüş yolların kaza oranını artırdığı, hız limitinin artmasının kaza oranını azalttığı, alacakaranlık ve gece vakti ile ıslak-nemli yol yüzeylerinin kaza oranlarını azalttığı görülmektedir. Bu bulgular, trafik güvenliğini artırmak için bu faktörlere odaklanmanın önemini vurgular.

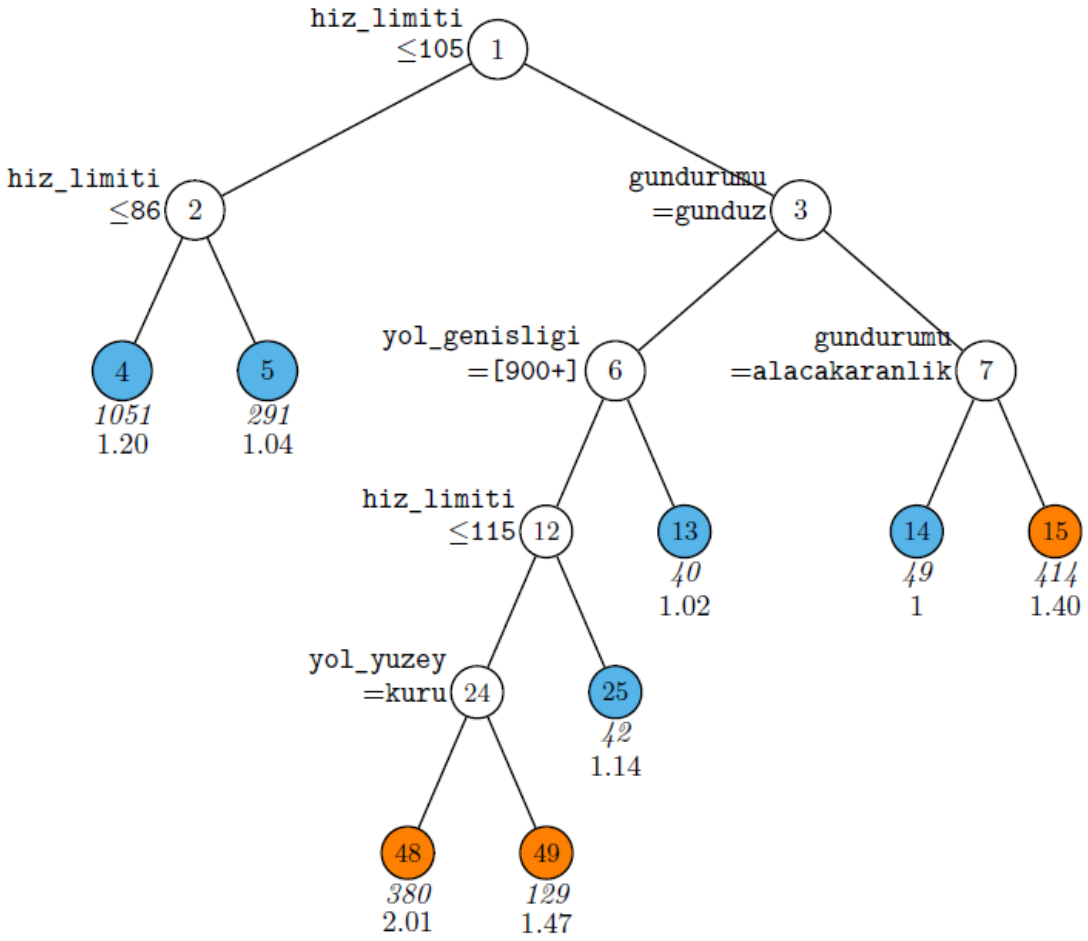
GUIDE algoritmasında parçalı sabit Poisson regresyon modeli kullanılmıştır. GUIDE algoritmasından elde edilen karar ağacı Şekil 4'te verilmiştir. Her bölünmede, sadece koşul sağlandığında sol dala gidilir. Düğümlerin altında italik olarak belirtilen sayılar örneklem büyüklüğünü ve altında yazan sayılar kaza sayısının ortalamasını vermektedir. Kök düğümün kaza sayısı ortalaması olan 1,35 değerinin altında ve üstünde ortalamalara sahip uç düğümler sırasıyla gök mavisi ve turuncu renklidir. Kök düğümünde ikinci en iyi ayırma değişkeni yol genişliğidir. 86 ve altındaki hız sınırları ortalama 1,20 kaza sayısıyla; 86 ile 105 arasındaki hız sınırları ortalama 1,04 kaza sayısı ile ilişkilidir. 105'in üzerinde hız sınırı olan 900 cm'nin altındaki yollarda gündüz gerçekleşen kaza sayıları ortalaması 1,02 ile genel ortalamanın altındadır. 105 ile 115 arasında hız sınırı olan 900 cm'nin üstündeki yollarda, gündüz kuru zeminli yollarda gerçekleşen kaza sayısı ortalama 2,01 ile en yüksek ortalamaya sahip yaprak düğümü oluşturmaktadır. Yasal hız sınırı 105'in üzerinde olan yollarda, gece gerçekleşen kaza sayısı ortalaması 1,40'tır. Yasal hız sınırı 105'in üzerinde olan yollarda, alacakaranlıkta gerçekleşen kaza sayısı ortalaması 1 ile en küçük ortalamaya sahip yaprak düğümü oluşturmaktadır.



Şekil 2. *CART algoritmasına ait karar ağacı*



Şekil 3. *MOB algoritmasına ait karar ağacı*



Şekil 4. *GUIDE algoritmasına ait karar ağacı.*

4. SONUÇ

Bu çalışmada, Ankara ili devlet yollarında 2017-2020 yılları arasında gerçekleşen trafik kazaları verileri kullanılarak yol özelliklerinin, yasal hız sınırının ve gün durumunun kaza sayısı üzerindeki etkisi incelenmiştir. Analizler, yolun yüzeyi, yolun yasal hız sınırı, yol şerit sayısı, gün durumu, yol genişliği ve yolun tipi gibi faktörlerin kaza sayısında belirleyici olduğunu göstermektedir. Poisson regresyon analizinden elde edilen sonuçlara göre; çok şeritli yollarda tahmini ortalama kaza sayısının, tek yönlü yollara göre daha fazla olduğu görülmüştür. Alacakaranlıkta ve gece meydana gelen ortalama kaza sayısının, gündüz meydana gelen ortalama kaza sayısından daha az olduğu görülmüştür. Bununla birlikte, ıslak veya nemli yollarda ortalama kaza sayısının kuru yollara göre daha az olduğu tespit edilmiştir. Bu çalışmada trafik kazalarını modellemek için üç farklı algoritma kullanılmıştır: CART, GUIDE ve MOB. Her algoritma, çeşitli yol ve çevre koşullarının kaza sayıları üzerindeki etkilerini farklı açılardan değerlendirmiştir. CART algoritması, yasal hız sınırını en önemli değişken olarak belirlemiştir. Hız sınırı 105 km/saat'in altında olan yollarda ortalama kaza sayısı 1,2 olup, toplam kazaların %56'sını oluşturmaktadır. Hız sınırı 105 km/saat veya daha yüksek olan yollarda ise ortalama kaza sayısı 1,6'dır. Bu algoritma, kaza sayısının yol genişliği ve günün saati gibi diğer faktörlere bağlı olarak da değiştiğini göstermiştir. GUIDE algoritması, parçalı sabit Poisson regresyon modeli kullanarak detaylı bir ayrım yapmıştır. Hız sınırı 86 km/saat ve altındaki yollarda ortalama kaza sayısı 1,20 iken, 86-105 km/saat arası yollarda bu sayı 1,04'e düşmektedir.

Hız sınırı 105 km/saat'in üzerinde olan yollarda ve gündüz kuru zeminli koşullarda ortalama kaza sayısı 2,01 ile en yüksek değeri göstermiştir. Gece ve alacakaranlık gibi koşullarda ise kaza sayıları genel ortalamanın altında kalmıştır.

MOB algoritması, her düğümde parametre tahminleri, standart hata ve p değerleri ile birlikte kaza sayısının histogramını sunmaktadır. Hız limiti 100'ün üzerinde olan yollar için bölünmüş yollarda kaza sayısı tek yönlü yollara göre %60 daha fazla bulunmuştur ve hız limitindeki her bir birimlik artış kaza sayısında %4'lük bir azalmaya neden olmaktadır.

Yasal hız sınırlarının uygulanması ve bunlara saygı gösterilmesinin rolü vurgulanarak, hız sınırlarına titizlikle uyulduğunda kazalar gözle görülür biçimde daha az olacaktır. Gün ışığı güvenliği garanti etmemektedir. Hız sınırları, özellikle gündüz yoğun olan yollarda, dikkatli inceleme gerektirir. Hız düzenlemeleri ile yol koşulları arasında doğru dengeyi yakalamak zorunludur. Tek veya çift şeritli yollar trafik kazaları açısından daha düşük risk taşır ancak çok şeritli yollarda şerit sayısının kaza oranlarını önemli ölçüde etkilediği görülmüştür. Daha geniş yolların daha fazla dikkat gerektirdiği görülmüştür. Yol şerit sayısının planlanması ve uygun düzenlemelerin yapılması, trafik güvenliğini artırmak için önemli olacaktır. Yol şerit sayısının yanı sıra diğer faktörlerin de kazalara etkisi olabileceği unutulmamalıdır. Yolun yüzeyine göre kaza sayısı analizinde, farklı yol yüzeylerinin kaza sayısında önemli farklılıklar olduğu görülmüştür. Örneğin yol yüzeyi buzlu olduğunda yoldaki ortalama tahmini kaza sayısı, kuru yoldaki kaza sayısından daha azdır. Buna göre buzlu yolda sürücüler daha temkinli davrandığı için böyle bir sonuç çıktığı düşünülebilir. Gün durumu için kaza sayısını incelediğimizde gündüz gerçekleşen kaza sayısının gece ve alacakaranlığa göre daha fazla olduğu görülmektedir. Gündüz kaza sayılarının fazla olmasını etkileyen faktörler fazla olabilir (trafik yoğunluğunun fazla olması, insanların bir yere yetişmesi için fazla hız yapması gibi). Bu bilgiler, trafik yönetimi yetkililerine trafik kazalarını azaltmak ve yol güvenliğini artırmak için hedefe yönelik müdahaleler geliştirme konusunda rehberlik edebilir.

TEŞEKKÜR

Bu çalışma Gazi Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, Veri Bilimi Yüksek Lisans programında "Sayma Verilerinde Karar Ağaçları" isimli yüksek lisans tezinden türetilmiştir. Tez çalışması için Ankara Emniyet Genel Müdürlüğü Trafik Eğitim ve Araştırma Dairesi Başkanlığı'ndan gerekli izinler alınarak veri temin edilmiştir. Tezde ve makalede kullanılmak üzere veri setini sağladıkları için Ankara Emniyet Genel Müdürlüğü Trafik Eğitim ve Araştırma Dairesi Başkanlığı'na teşekkür ederiz.

ÇIKAR ÇATIŞMASI/ÇAKIŞMASI BİLDİRİMİ

Yazarlar arasında çıkar çatışması/çakışması bulunmamaktadır.

YAZAR KATKI ORANLARI

Mine Fulya Gürsel: Araştırma, Yazılım, İçerik analizi, Makalenin yazımı- İnceleme ve Düzenleme. **Hatice Tül Kübra Akdur:** Metodoloji, Materyal temini, Yazılım, Makalenin yazımı- İnceleme ve Düzenleme, Danışman/Kontrolörlük

KAYNAKLAR

[1] World Health Organization. (2015). Global status report on road safety 2015. *World Health Organization*.
[2] TİK Karayolu Trafik Kaza İstatistikleri URL: https://data.tuik.gov.tr/Bulten/Index?p=Karayolu-Trafik-Kaza-Istatistikleri-2022-49513, Son Erişim Tarihi: 21.08.2023.
[3] World Health Organization. (2004). World report on road traffic injury prevention: summary. *In World Report on Road Traffic Injury Prevention: Summary*, IX-52.
[4] Silva, P.B., Andrade, M. & Ferreire, S. (2020). Machine learning applied to road safety modeling: A systematic literatüre review. *Journal of Traffic and Transformation Engineering*, 7(6), 775-790.

[5] Bolat, H., Yücesan, M. ve Utku A. (2022). Trafik kazalarının makine öğrenmesi yöntemleriyle analizi ve tahmini: Kahramanmaraş için örnek bir çalışma. *International Journal of Pure and Applied Sciences*, 8(2), 490-506.
[6] Yavuz, A.A., Ergül, B. ve Aşık, E.G. (2021). Trafik kazalarının makine öğrenmesi yöntemleri kullanılarak değerlendirilmesi. *Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi*, 13(1), 66-73.
[7] Özden, C. ve Acı, Ç. (2018). Makine öğrenmesi yöntemleri ile yaralanmalı trafik kazalarının analizi: Adana örneği. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 24(2), 266-275.
[8] Al-Asadi, M., TaşdemiR, Ş., & ÖRNEK, H. K. (2022). Predict the number of traffic accidents in Turkey by using machine learning techniques and python tools. *Artificial Intelligence Studies*, 5(2), 35-46.
[9] Kleiber, C. & Zeileis, A. (2022). Package "AER". URL: https://cran.r-project.org/web/packages/AER/AER.pdf, Son Erişim Tarihi: 19 Aralık 2023.
[10] Hothorn, T. & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, 16(1), 3905-3909.
[11] Atkinson, E.J. & Therneau, T.M. (2000). An introduction to recursive partitioning using the RPART routines. *Rochester: Mayo Foundation*, 2000.
[12]Loh, W.Y.(2023).User Manual for GUIDE ver. 41.1. URL:
https://pages.stat.wisc.edu/~loh/treeprogs/guide/guideman.pdf, Son Erişim Tarihi: 13.11.2023.
[13] Cameron, A.C. & Trivedi, P.K. (2013). Regression analysis of count data (2nd edition*). Cambridge: Cambridge University Press*.
[14] Olson, D.L. & Shi, Y. (2007). Introduction to business data mining (1st edition). *Boston: McGraw-Hill/Irwin*.
[15] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). Classification and regression trees. *CRC Press*.
[16] Ciampi, A. (1991). Generalized regression trees. *Computational Statistics & Data Analysis*, 12(1), 57-78.
[17] Loh, W.Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12(2002), 361 – 386.
[18] Choi, Y., Ahn, H., & Chen, J. J. (2005). Regression trees for analysis of count data with extra Poisson variation. *Computational Statistics & Data Analysis*, 49(3), 893-915.
[19] Loh, W.Y. (2006). Regression tree models for designed experiments. *IMS Lecture Notes-Monograph Series*, 49: 210-228.
[20] Zeileis, A., Hothorn, T. & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*,, 17(2), 492–514.
[21] Kleiber, C. & Zeileis, A. (2008). Applied Econometrics with R. New York. *Springer*.