Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2021) 13(1), 66-73.

10.29137/umagd.705156

Cilt/Volume:13 Sayı/Issue:1 Ocak/January 2021

Araştırma Makalesi / Research Article

Trafik Kazalarının Makine Öğrenmesi Yöntemleri Kullanılarak Değerlendirilmesi

Evaluation of Traffic Accidents Using Machine Learning Methods

Arzu Altın Yavuz[1], Barış Ergül[1], Ebru Gündoğan Aşık[*2]

[1]Eskişehir Osmangazi Üniversitesi Fen Fakültesi İstatistik Bölümü, 26010 Eskişehir, TÜRKİYE

[2]Karadeniz Teknik Üniversitesi Fen Fakültesi İstatistik ve Bilgisayar Bilimleri Bölümü, 61080 Trabzon, TÜRKİYE

***Başvuru/Received**: 17/03/2020* ***Kabul / Accepted**: 20/12/2020* ***Çevrimiçi Basım / Published Online:** 18/01/2021*

***Son Versiyon/Final Version**: 18/01/2021*

Öz

Türkiye'de meydana gelen trafik kazaları, sebep oldukları maddi/manevi kayıplar sebebiyle gündemin ilk sırasında olma durumunu korumaktadır. Trafik kazaları, insan, yol, araç, iklim, çevre koşulları gibi birçok etkenin bir araya gelmesi sonucu oluşmaktadır. Trafik kazaları sonucu, telafi edilebilen kazalar olabileceği gibi, telafisinin olanaksız olduğu kazalar da olabilmektedir. Trafik kazalarının sayısını ve etkilerini en aza indirebilmek için genel olarak kazaya sebep olan etkenlerin tespit edilip ortadan kaldırılması gerekmektedir. Trafik kazalarına neden olan etkenlerin belirlenebilmesi için geçmiş kaza verilerinden yararlanılmaktadır. Kaza analizinde önemli olan var olan durumun model yardımıyla doğru bir şekilde sınıflandırılmasıdır. Trafik kazaları için literatür çalışmaları incelendiğinde, genel olarak diskriminant analizi, lojistik regresyon analizi ve logaritmik doğrusal modellerin kullanıldığı görülmektedir. Bu çalışmada 2012 ile 2016 yılları arasında Antalya ili ve ilçelerinde sonucu ölümlü, yaralanmalı olarak gerçekleşen 3181 trafik kazası ele alınmıştır ve makine öğrenme yöntemleri kullanılarak trafik kazalarının sınıflandırılması yapılmıştır. Ele alınan makine öğrenme yöntemlerinin performansları çeşitli ölçütlere göre karşılaştırılması sonucunda kaza verilerini en yüksek doğrulukla sınıflandıran yöntemin Naive Bayes olduğu tespit edilmiştir.

Anahtar Kelimeler

"Trafik, kaza analizi, sınıflandırma, karar ağaçları, WEKA, makine öğrenmesi"

Abstract

Traffic accidents occurred in Turkey caused their financial/moral losses to continue to maintain its status by virtue of being first during the agenda. Traffic accidents occur as a result of the combination of many factors such as people, roads, vehicles, climate and environmental conditions. As a result of traffic accidents, there may be accidents that can be recovered or accidents that cannot be compensated. In order to minimize the number and effects of traffic accidents, the factors causing the accident in general should be identified and eliminated. In order to identify the factors that cause traffic accidents accident history data are utilized. The important thing in accident analysis is the correct classification of the existing situation with the help of the model. When literature studies for traffic accidents are examined, it is seen that discriminant analysis, logistic regression analysis and logarithmic linear models are generally used. In this study, 3181 traffic accidents that occurred as a result of death or injury in Antalya province and its districts between 2012 and 2016 were considered and the classification of traffic accidents was made using machine learning methods. As a result of comparing the performances of the machine learning methods discussed according to various criteria, it was determined that the method that classifies the accident data with the highest accuracy is Naive Bayes.

Key Words

"Traffic, accident analysis, classification, decision tree, WEKA, machine learning"

***Sorumlu Yazar:** ebrugundogan@ktu.edu.tr*

1. Giriş

Karayolu ile ulaşım, bütün dünya için diğer ulaşım türlerine oranla daha çok tercih edilmektedir. Türkiye için, elde edilen veriler ışığında yaklaşık %95'lik bir oranla en çok kullanılan ulaşım türü, karayolu ulaşımıdır (Bolakar, 2014). 2018 yılında Türkiye'de 428.074 adet trafik kazası meydana gelmiş, bunlardan 2.712 adedi ölüm ve 183.710 adedi yaralanma ile sonuçlanmıştır. Bu kazalarda 3.373 kişi hayatını kaybetmiş, 310.109 kişi yaralanmıştır (EGM Trafik İstatistik Bülteni, 2018). Yoğun bir kullanıma sahip olan karayollarında meydana gelen trafik kazaları neden oldukları maddi/manevi kayıplar sebebi ile ulaşım alanında önemli konulardan biridir. Trafik kazalarının oluşmasında birçok değişken rol almaktadır ve bu durum trafik kazalarının oluşmasını belirleme sürecini karmaşık hale getirmektedir. Bu nedenle trafik güvenliğini sağlama ve sürdürülebilir kılmak oldukça karmaşık ve zor bir hal almaktadır (Özden ve Acı, 2018). Trafik kazalarının başlıca oluşma sebepleri, insan, yol, araç, iklim, çevre koşulları olarak sıralanabilmektedir. Trafik kazaları sonucu, telafi edilebilen kazalar olabileceği gibi, telafisinin mümkün olmadığı kazalar da olabilmektedir. Trafik kazalarının sayısını ve etkilerini en aza indirebilmek için ülkeler, bu konularda çeşitli stratejiler ve uygulamalar geliştirmektedirler. Trafik kazalarının sayısını ve maddi/manevi kayıpları azaltabilmek için kazaya sebep olan etkenlerin tespit edilip ortadan kaldırılması benimsenen bir yoldur.

Birçok sebep ile meydana gelen trafik kazalarının önlenmesinde mevcut eksikliklerin olduğu görülmektedir. Sistemin sağlıklı işlemesi açısından çeşitli yaptırımlar ile birlikte bazı kuralların da uygulamaya yardımcı olması önemlidir. Trafik güvenliğindeki eksikliklerden kaynaklanan sorunların çözümüne yönelik değerlendirme yapılırken, öncelikle var olan yapıdaki sorunların tespit edilmesi ve sonrasında çözüm odaklı güncellemelerin geliştirilmesi uygun bir yaklaşım olacaktır (Aron vd., 2015; Tolunay ve Gökdeniz, 2002).

Literatürde farklı yöntemler kullanılarak yapılandırılan birçok trafik kaza tahmin modeli vardır. Bu çalışmalar, istatistiksel yöntemler kullanılarak gerçekleştirilmektedir. İstatistik alanındaki gelişmeler ve son yıllardaki makine öğrenmesi yaklaşımları ile kaza nedenlerinin tespiti önemli bir çalışma alanıdır. Bilgisayar programlarının geliştirilmesi de bu çalışmaların hızını arttırmıştır. Dolayısı ile artık trafik kazalarının nedenlerinin tespiti için uzun süreli gözlemlerin ve analizlerin yapılması beklenmeksizin sürecin değerlendirilmesi yapılabilmektedir. Özgan vd., (2004) yılındaki çalışmalarında, 1999-2002 yılları arasında Sivas İli 'nde meydana gelen trafik kazalarına ait tutanak ve raporlardaki verileri kullanarak meydana gelebilecek kazaların sayısını tahmin etmeye çalışmışlardır. Chong vd., (2005) yılında yaptıkları çalışmada, trafik kazaları sonucu oluşan yaralanma derecesinin kestirimini yapmak için Yapay Sinir Ağları (YSA) ve Regresyon Ağaçları (RA) yöntemlerini karma şekilde kullanan yeni bir model önerisi yapmışlardır. Chang ve Wang, (2006) yılındaki çalışmalarında, Karar Ağacı (KA) yöntemi ile araç tipinin kaza şiddetini etkileyen en önemli değişken olduğu sonucuna ulaşmışlardır. Shon ve Shin, (2010) yılında yaptıkları çalışmalarında, YSA, Lojistik Regresyon (LRA) ve KA yöntemlerini kullanarak Kore'de olan trafik kazalarının şiddetini etkileyen değişkenleri belirlemişlerdir. Kwon vd., (2014) yılında yaptıkları çalışmalarında Naive Bayes (NB) ile KA yöntemlerini kullanarak trafik kazalarına sebep olan değişkenleri nispi önemlerine göre sıralamışlardır. Muhammed vd., (2017) yılında yaptıkları çalışmada, karayollarında oluşan trafik kazalarını karar ağaçları algoritmalarını kullanarak trafik kazalarının tahminini yapmaya çalışmışlardır. Jiahia vd., (2019) yılında yapılan çalışmalarında, WEKA 'daki C4.5 algoritmalarını kullanarak trafik kazalarını analiz etmişlerdir.

Trafik kaza analizleri için literatür genel olarak değerlendirildiğinde çalışmaların büyük kısmında diskriminant analizi, lojistik regresyon analizi, logaritmik modeller, yapay sinir ağları (YSA) ve fuzzy yaklaşımların kullanıldığı görülmektedir (Bektaş, 2012; Ahmed, 2017; Delen vd., 2006). Genel olarak bu yöntemler, veri setinin ölümlü/yaralamalı şeklinde sınıflandırılmasını sağlayan yöntemlerdir. Bir sınıflandırma sonucu tespit edilen sonuçlar değerlendirilirken, ele alınan sınıflandırıcının doğru sınıflandırma oranı göz önüne alınır. Doğru sınıflandırma oranı düşük olan bir yöntemden elde edilen sonuçların geçerli ve güvenilir olması beklenemez.

Bu çalışmada, farklı istatistiksel sınıflandırma algoritmaları ve karar ağaçları ele alınmıştır. Öncelikle bu yöntemler tanıtılacaktır. Daha sonra ise, 2012 ile 2016 yılları arasında Antalya ili ve ilçelerinde gerçekleşen trafik kaza tutanaklarından elde edilen veriler yardımıyla trafik kazalarının doğru sınıflandırılmasını en iyi sağlayan model belirlenecektir.

2. Materyal ve Metot

Çok değişkenli istatistiksel yöntemler, birden çok değişkenin oluşturduğu değişken kümesinin yapısını belirleyerek, ele alınan problemin yapısına uygun çözümler için gerekli olan betimleyici bilgiyi sağlar. Aynı zamanda, değişken grupları arasındaki karşılıklı ilişkiyi ölçme olanağı sağlayan istatistik tekniklerini oluşturur.

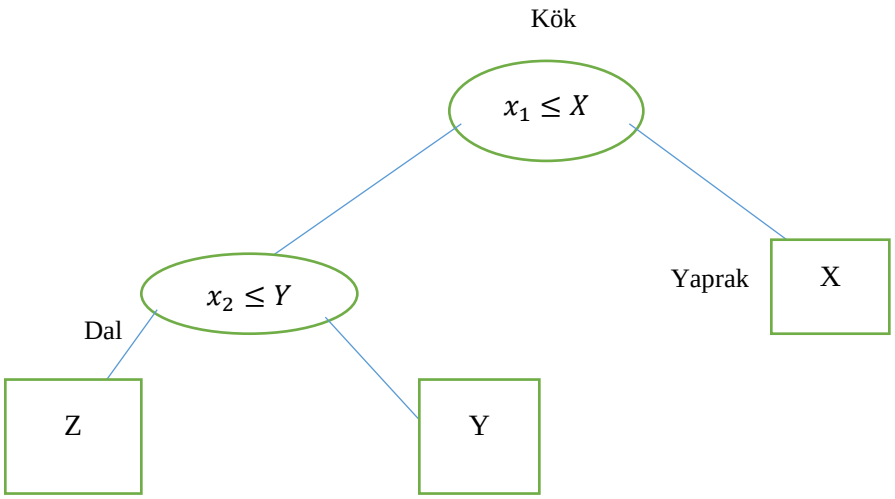
Sınıflandırma problemi istatistiksel karar verme süreci olarak görülebilir. Bu düzeyde, araştırmacı için iki çeşit karar verme süreci bulunmaktadır. İlki, grubun ayırt edici özelliklerini araştırmak ve bu özellikleri bir fonksiyon haline getirmek; diğeri, bu ayırt edici fonksiyonlar yardımıyla gözlemleri gruplara atamaktır. Sınıflandırma, veri madenciliğinde eğiticisi olan ve olmayan olarak ikiye ayrılır. Eğiticisi olan öğrenmede, grup sayısı ve hangi gruba ait olduğu önceden bilinen verilerin, deneme süreci ile hangi gruba ait olduğu bulunmaya çalışılır. Karar ağaçları eğiticisi olan öğrenme sınıflandırmasında yaygın olarak kullanılmaktadır. Ayrıca karar

ağaçları, eğer verinin ait olduğu grup biliniyorsa, sınıflandırma ağaçları yöntemi ile modellenmelidir (Akçetin ve Çelik, 2014; Breiman vd., 1984).

Veri madenciliği yaklaşımı; karar ağaçları, sınıflandırma ve tahmin için sıklıkla kullanılmaktadır. Kolay yorumlanması ve anlaşılması açısından karar vericiler için bir avantaj sağlamaktadır. Bu çalışmada karar ağaçları, lojistik regresyon ağacı, J48, Basit CART (Classification and Regression Trees), rassal ağaç ve rassal orman karar ağaçları yöntemleri ele alınmıştır.

2.1. Karar Ağaçları

Karar ağaçları, denetimli makine öğrenmesi sınıfındandır. Karar ağaçları kolay uygulanabilmesinin yanı sıra, kolay yorumlanabilmesi; nitel, nicel, sürekli ve kesikli değişkenlere uygulanabilmesi ve güvenilir sonuçlar vermesi sebebiyle sıklıkla kullanılmaktadır (Akşehirli, 2012). Karar ağaçları, tek kökten başlamaktadır. Karar düğümlerine doğru ilerleyen, etiketlenmiş yapraklarda son bulan bir sınıflandırma ağacıdır. Basit bir karar ağacının yapısı Şekil 1 'de gösterilmiştir. Şekil 1'den görüleceği üzere karar ağaçları, kök, dal ve yapraklardan oluşmaktadır.



Şekil 1. Karar Ağacı Yapısı

Karar ağaçlarında kullanılacak veri seti için eğitim ve test verisi, veri setindeki sınıf oranları dikkate alınarak tüm veri setinin 2/3'si eğitim (in Bag) ve 1/3'i test verisi (Out of Bag, OOB) olarak kullanılmaktadır. Karar ağacı oluşturmak için, eğitim veri seti içinde veriyi en iyi tanımlayan veri seçilir. Bu veri ile ağacın dalı ve yaprakları olarak bilenen ayrıştırma işlemi yapılır ve yeni bir veri seti oluşturulur. Ayrıştırılan dal üzerinde bulunan örneklerden yeni bir belirleyici veri bulunur ve yeni dallar oluşturulur. Veri setini ayrıştıracak başka veri kalmamışsa ve kalan verilerin değerini taşıyan başka veri yoksa işlem sonlandırılır. Aksi durumda alt veri setini ayrıştırmak için yeniden belirleyici bir veri bulunur (Albayrak, 2015).

Karar ağaçlarında, bir veri için sınıflandırma, üzerinden geçilen her dal ve bütün yaprakların doğru olduğu yollar üzerinden takip edilerek yapılır (Freund ve Mason, 1999).

2.2. J48 Karar Ağacı

Veri madenciliği alanında yaygın bir kullanıma sahip olan J48 algoritması bir karar ağacı sınıflandırıcısıdır. J48 sınıflandırıcısı C4.5 karar ağacı olarak da bilinmektedir. Bu algoritma, veriyi yukarıdan aşağıya doğru bir dağılım ile sınıflandırmaktadır. En yüksek bilgi kazancına sahip nitelikten başlayarak verilerin bölünmesi ile nihai karar ağacına ulaşılır (Quinlan, 1993; Öztürk ve Mesut, 2016).

Karar ağacı yapısı, daha küçük bölümlere neden olan her düğümde bölümlenen bir veri seti (eğitim seti) ile başlamaktadır. Bu sayede, özyinelemeli bir bölünme stratejisi izlenmektedir. Bir veri kümesine ek olarak, bir dizi nitelik de iletilir. Nesneler bir olay, bir aktivite veya nitelikler o nesneyle ilgili bilgiler olabilmektedir. Veri kümesindeki her demet için, bir nesnenin belirli bir sınıfa ait olup olmadığını belirleyen bir sınıf etiketi ile ilişkilendirilme yapılmaktadır. Her bir düğümdeki entropi değerleri kullanılarak en yüksek bilgi kazancına ulaşılmaya çalışılmaktadır. En yüksek bilgi kazancına sahip veri üzerinden verilerin bölünmesiyle oluşturulan dallardan hareket ederek sonuca ulaşmaktadır (Quinlan, 1993).

J48 algoritmasında ilk adım bilgi kazanımını hesaplamaktır. Bilgi kazanımı hesaplanırken kullanılan entropi formülü Eşitlik (1) de verildiği gibidir (Şeker, 2012).

$$Entropi(D) = -\sum_{i=1}^{m} p_i \, log_2(p_i) \tag{1}$$

Daha sonra her bir özellik için ayrı ayrı bilgi hesaplaması yapılır.

$$Entropio_A(D) = \sum_{i=1}^{v} \frac{|D_j|}{|D|} \times Entropi(D_j) \tag{2}$$

İlk aşamada hesaplanan bilgi değerinden, belirli bir özellik için elde edilen bilgi değeri çıkarılarak ilgili özelliğin bilgi kazanımı belirlenir. Tüm özellikler için elde edilen bilgi kazanımları içerisinde en yüksek bilgi kazanımını sağlayan özellik karar ağacında ilk basamakta yer alır. Daha sonra süreç benzer şekilde devam eder ve karar ağacını şekillenir. Son olarak geriye doğru budama yapılarak karar ağacının nihai şekli elde edilir.

J48 karar ağacı, veri yapısını anlamlandırmada güçlü ve hızlı yollar sunar. Veri hazırlamak için az zaman gerektirmesi, yorumlama kolaylığı sağlaması ve doğrusal olmayan ilişkilerden etkilenmemesi bakımından tercih edilir. Dezavantajları ise, çok aşırı dal içere karar ağacı çizimleri karmaşık bir yapı sunmakta ve zaman alıcı olmaktadır. Büyük ağaçların sunumu ve anlaşılması zordur.

J48 Con. Karar Ağacı ise, tahmin hatasının azaltılması amacı ile mevcut karar ağacına yeni dallar ekler. Bu teknik, ağacı karmaşık yapıdan kurtarmaktadır ve ağacın oluşma zamanını azaltmaktadır.

2.3. Basit CART (Classification and regression trees)

CART algoritması, eğitim veri setini her bir yaprak olmayan düğümlerde iki parçaya bölmektedir. Bu bölme işlemi, eğitim veri setini tamamen ayrışmaz duruma gelene dek bölmektedir. Bölme işlemi, ayrışmayıncaya dek sürmektedir ve sonrasında durmaktadır (Ma, 2013).

Breiman vd., (1984), tarafından önerilen bu algoritma, sayısal ve isimsel veri setlerinden oluşan değişkenleri kullanarak analiz yapmaktadır. CART karar ağacı, ikili olarak özyinelemeli biçimde bölünen bir yapıya sahip olması nedeniyle artık yeni bir bölünmenin gerçekleşmeyeceği duruma kadar devam eder ve sonraki aşamada uçtan başlayarak köke doğru budama işlemi başlar. Her budama işlemi sonrası test verisi kullanılmak suretiyle, olası en başarılı karar ağacı tespit edilmeye çalışılır. Her bir düğümdeki binom dağılımının genellemesi ile elde edilen Gini indeksi değerleri kullanılarak en yüksek bilgi kazancına ulaşılmaya çalışılmaktadır. Birbiri ile ilişkili değişkenler olması durumunda, iyi performans vermediği söylenmektedir.

Gini indeksi Eşitlik 3 de verildiği gibidir.

$$g(t) = 1 - \sum_j p^2(j/t) \tag{3}$$

Burada, t yaprağındaki j sınıfına ait koşullu olasılık $p(j/t)$ ile gösterilmiştir.

2.4. Rassal Ağaç

Rassal ağaç, her bir düğümde daha önce belirlenen sayıda rastgele özellik seçilmesiyle başlamaktadır. Bu algoritmada ağacın dallarında budama yapılmamaktadır. Karar ağaçlarında, her bir düğümde en çok bilgi veren özellik seçilirken; rastgele ağaç yönteminde bu seçim tesadüfi olarak gerçekleşmemektedir (Witten vd., 2016). Rassal Ağaç, ağaç kümesinde her ağacın örnekleme alınma şansının eşit olduğu veya ağaçların "uniform" dağılıma sahip olduğu bir ağaçtır (Breiman, 2001). Bununla birlikte, her düğümde K rassal öneme sahip bir dizi olası ağaçtan rassal olarak oluşturulmuş bir ağaç olduğu düşünülür. Ayrıca kategorik değişkenler için olasılıkların tahminine izin vermektedir (Zhao ve Zhang, 2008; Başar ve Akan, 2018).

2.5. Rassal Orman

Rastgele orman, her biri birbirinden bağımsız ve aynı dağılım kullanılarak elde edilen eğitim veri setinden rassal olarak elde edilmiş bir örnekleme dayanan karar ağacıdır. Rassal Orman algoritması, tek bir karar ağacı oluşturmak yerine çok sayıda karar ağacının kararlarının birleştirilmesine dayanmaktadır. Bu karar ağaçlarının her biri, birbirinden bağımsızdır ve eğitim kümesinden bootstrap tekniği ile seçilen farklı örneklerden oluşturulur. Farklı eğitim kümeleri için, aynı dağılımdan gelen rassal özellik seçimi kullanılmaktadır. Karar ağaçlarını oluştururken bütün ağaçlarda ilgili nitelik taraması yapılır. Sonrasında, diğer ağaçlardaki nitelikler birleştirilir. Bu şekilde en çok kullanılan öznitelik seçilir. Seçilen öznitelik karar ağacına dahil edildikten sonra diğer aşamalarda da aynı işlemler tekrarlanmaktadır. RO, karar ağacı oluşturmak için CART algoritmasını kullanmaktadır (Daş ve Türkoğlu, 2014). Ağaç yapısının oluşturulması için, her bir dalda kullanılacak örneklem sayısı ve oluşacak ağaç sayısının belirlenmesi önem taşımaktadır (Pal, 2005). Dallara ayırıcı özellikteki değişkenin seçimi "Bagging" yaklaşımından farklıdır. Burada, tüm değişkenler içerisinden rassal olarak m=$\sqrt{p}$ değişken seçilmektedir. Ayrıca her aşamadaki m sabittir (Breiman, 2001). Regresyon ağaçlarında m=p/3 olarak alınmaktadır (Yılmaz, 2014).

Eğitim sırasında birçok karar ağacının oluşturulması, sınıflandırmada başarı oranının yüksek olmasını sağlamaktadır. Rastgele ormanın en önemli avantajı, diğer karar ağacı yöntemlerinden daha hızlı çalışması olarak gösterilmiştir. Bu algoritmada budama işlemi yoktur (Breiman, 2001). Budamanın olmaması RO algoritmasını diğer karar ağacı yöntemlerinden daha avantajlı hale getirmektedir. Birbiri ile ilişkili değişkenler olması durumunda, diğer karar ağaçlarına göre performansı daha yüksektir.

2.6. Lojistik Regresyon Ağacı

Lojistik Regresyon Ağacı, lojistik regresyon ve karar ağacının birleştirilmesi ile meydana gelmiş bir sınıflandırma modelidir. Lojistik Regresyon ağacı, regresyon analizi yapısına sahip bir karar ağacıdır. Bu ağaç yapısında, ağacın her dalı için lojistik regresyon analizi yapılmaktadır, daha sonra C4.5 karar ağacı kullanılarak dallar ayrılmaktadır. Son aşama, ağacın budanması aşamasıdır (Long, 1993; Landwehr vd., 2005).

Lojistik regresyon ağacı, ortalama olarak hem karar ağaçları hem de lojistik regresyondan daha iyi performans göstermektedir. Ayrıca daha iyi yorumlanabilen bir model sunarken, güçlendirilmiş karar ağacı topluluklarıyla rekabet içinde performans sergilemektedir.

Eğitim veri setine bağlı olarak, karmaşıklığını kolayca ve otomatik olarak ayarlayabilen bir modeldir. Lojistik regresyon ağacı büyütülürken; her bölünmüş düğüm, aday bir yaprak düğümü olarak kabul edilir. Bu nedenle ağaçtaki her düğümle bir lojistik regresyon modeli ilişkilendirilmektedir. Çapraz doğrulama (Cross Validation), uygun sayıda yinelemeyi belirlemek için kullanılır. C4.5 algoritması, lojistik regresyon ağacı düğümlere sığmadan önce temel ağaç yapısını oluşturmak için kullanılmaktadır. Ağaç büyütüldükten sonra hem eğitim hatasını hem de ağacın karmaşıklığını göz önünde bulunduran düşük maliyetli budama kullanılarak budama yapılmaktadır (Breiman vd.,1984).

2.7. Naive Bayes

Naive Bayes Classifier, bayes teoremine dayanan en popüler makine öğrenimi sınıflandırma tekniklerinden biridir. Naive Bayes sınıflandırması birçok farklı alanda da kullanılmaktadır (Cihan vd., 2018). Mevcut ve geçmiş frekans oluşumlarını hesaplamak için kullanılan Naive Bayes Sınıflandırıcısının hesaplanışı Eşitlik (4) de verildiği gibidir (Aydındağ ve Kırcı, 2019).

$$P(A \setminus B) = P(B \setminus A)*P(A)/ P(B) \quad (4)$$

Burada;

P(A): A'nın prior olasılığı. Sadece A olayının sayısını içerir.

P(A \ B): B verildiğinde A'nın posterior olasılığı.

P(B \ A): A verildiğinde B'nin posterior olasılığı.

P(B): B'nin prior olasılığını ifade etmektedir.

Tablo 1. Trafik Kaza Verileri İçin Kullanılan Makine Öğrenme Yöntemleri

Yaklaşım	**Kullanılan Teknik**
Bayes	Naive Bayes
Regresyon	Lojistik Regresyon
Karar Ağacı	Karar Ağacı, Basit CART, J48, Düzeltilmiş J48, Rassal Ağaç, Rassal Orman

2.8. Karşılaştırma Ölçütleri

Sınıflandırma problemlerinde modelin başarısını değerlendirebilmek için sık kullanılan temel kavramlar kontenjans tablosundan hesaplanan değerlerdir. Doğruluk ölçütünün hesaplanmasında, doğru sınıflandırma tablosu kullanılmaktadır. Doğru sınıflandırma tablosunun köşegen elemanları TP (True Pozitif) ve TN (True Negatif) değerleridir. Bu değerler doğru sınıflandırmayı temsil eder. FP (False Pozitif) ve FN (False Negatif) değerleri ise köşegen dışı elemanlardır ve yanlış sınıflandırma durumlarını ifade eder. Doğruluk değerinin hesaplanması eşitlik (5)'te verildiği gibidir. Eşitlik (5) de T_p doğru pozitif, T_n doğru negatif, F_p yanlış pozitif ve F_n yanlış negatif ifadelerine karşılık gelmektedir (Kaur ve Chhabra, 2014).

$$Doğruluk = \frac{T_p+T_n}{T_p+F_p+T_n+F_n} \quad (5)$$

Burada;

- True Pozitif - TP: Yaralanmalı olarak sonuçlanan bir kazanın yaralanmalı olarak sınıflandırılması
- False Pozitif - FP: Ölümle sonuçlanan kazaların yaralanmalı olarak sınıflandırılması
- True Negatif - TN: Ölümle sonuçlanan kazaların ölümlü olarak sınıflandırılması
- False Negatif - FN: Yaralanmalı olarak sonuçlanan kazaların ölümlü olarak sınıflandırılmasını

ifade etmektedir. Doğruluk değeri (DSO), yöntemlerin değerlendirilmesinde kullanılan model doğruluğunun bir ölçüsüdür. DSO ne kadar yüksek ise, o yöntem daha iyi demektir.

3. Uygulama

2012 ile 2016 yılları arasında Antalya ili ve ilçelerinde toplam 30232 adet trafik kazası olmuştur. Bu çalışmada, meydana gelen trafik kazaların sonucu, ölümlü ve yaralamalı olarak gerçekleşen 3181 trafik kazası veri seti olarak kullanılmıştır. Analize alınan bağımsız değişkenler; kaza yeri, yol tipi, yolun kaplama cinsi, yolun sınıfı, hava durumu, yol yüzeyi, trafik lambasının durumu, aydınlatma, trafik görevlisi durumu, emniyet şeridi durumu ve bağımlı değişken kaza türü değişkenidir.

Bu çalışmada, makine öğrenmesi modelinde yapılan testin hatasını tahmin edebilmek için model seçiminde sıklıkla kullanılan k-kat çapraz doğrulama test yöntemi kullanılmıştır. Bir test ve eğitim seti olarak verilen verileri bir k sayısına göre ayırdıktan sonra, sınıflandırıcı, modelin güvenilirliğini doğrulamak için değerlendirilmektedir.

Verilerin analizi için kullanılan WEKA yazılımı; Waikato Üniversitesinde geliştirilmiş, GNU lisansı ile çalışan açık kaynak kodlu bir yazılımdır (Witten ve Frank, 2005). Makine öğrenmesi algoritmalarını içeren WEKA, temel olarak sınıflandırma, kümeleme, demetleme, birliktelik analizi, veri ön işleme gibi temel veri madenciliği işlemlerini yapabilmesinin yanı sıra, eğitimli ve eğitimsiz öğrenme yöntemlerini de içeren bir yazılımdır. Makine öğrenmesi paketleri bakımından güçlü bir yazılım olan WEKA programının seçilmesinin nedeni, sınıflandırma tekniklerini gerçek veriler üzerinde kolay ve anlaşılır şekilde karşılaştırabilme olanağı da sağlamasıdır. Algoritmalar, veri kümesine doğrudan veya Java kodundan çağrılarak uygulanabilen WEKA, aynı zamanda yeni makine öğrenme algoritmaları geliştirmek için uygundur (Patterson vd., 2008).

Bu çalışmada, açık kaynak kodlu yazılım olan WEKA yardımıyla Karar Ağaçları, Lojistik Regresyon Ağaçları, J48 Ağaçları, Basit CART, Rassal Ağaç, Rassal Orman ve Naive Bayes yöntemleri uygulanmıştır.

WEKA kullanılarak analiz edilen trafik kazaları veri seti için 3 farklı karşılaştırma ölçütü kullanılmıştır. Sonuçlar Tablo 2'de verildiği gibidir.

Tablo 2. Antalya İli Trafik Kazaları için Sınıflandırma Analizi Sonuçları

Ölçüt	L.R.A.	J48	J48 Con.	K.A.	Basit CART	R.A.	R.O.	Naive Bayes
DSO	98,679	97,679	98,4981	88,679	87,679	89,616	91,679	99,012
TP	0,987	0,977	0,985	0,887	0,877	0,896	0,917	0,990
FP	0,013	0,023	0,015	0,113	0,123	0,104	0,083	0,010

DSO ölçütüne göre Naive Bayes algoritması trafik kazalarını %99.012 doğrulukla sınıflandırmaktadır. Bu yöntemden sonra en yüksek doğru sınıflandırma lojistik regresyon ağacı kullanılarak elde edilmiştir. Sıralama düzeltilmiş J48, J48, rassal orman, rassal ağaç, karar ağacı ve basit CART şeklinde devam etmektedir. Yaralanmalı kazaların yaralanmalı olarak sınıflandırılmasını temsil eden TP değeri benzer şekilde bir sıralama sunmaktadır. Ölümle sonuçlanan bir kazanın yaralanmalı olarak tespitini ifade eden FP değerinin oldukça küçük olması istenir. FP değerleri incelendiğinde en düşük FP değerini veren yöntemin Naive Bayes algoritması olduğu, bunu lojistik regresyon ağacının takip ettiği görülmektedir. Tüm ölçütler göz önünde bulundurulduğunda trafik kazalarının sınıflandırılması için en etkin sonucu veren yöntemin Naive Bayes olduğu görülmektedir. Naive Bayes yöntemi sonucunda, Antalya İli trafik kazaları için en etkili değişkenlerin yolun kaplama cinsi ve yolun sınıfı olduğu belirlenmiştir.

4. Sonuç ve Öneriler

Makine öğrenme algoritmaları pek çok alanda olduğu gibi trafik çalışmalarında son yıllarda sıklıkla kullanılmaktadır. Günümüzde makine öğrenmesi ve veri madenciliği yöntemleriyle doğru kararlar alınmasına olanak sağlayan ve trafik kazalarının önlenmesinde etkili olan pek çok yöntem geliştirilmektedir. Bu yöntemler sayesinde var olan durumun tespiti doğru bir şekilde analiz edilerek gerek önlemlerin alınması gerek yeni düzenlemeler yapılmaktadır. Bu çalışmada, 2012 ile 2016 yılları arasında Antalya ili ve ilçelerinde ölümlü, yaralanmalı ve maddi hasarlı olmak üzere toplam 30232 adet trafik kazası değerlendirmeye alınmıştır. Bunlar içerisinden sonucu, ölümlü ve yaralamalı olarak gerçekleşen 3181 trafik kazası veri seti olarak belirlenmiştir. Ölümlü ve yaralanmalı kazaların sonuçlarının telafisi mümkün olmadığından, kazaya neden olan unsurların belirlenmesi önemlidir. Ancak bunun yapılabilmesi için kaza analizinde kullanılacak sınıflandırma yönteminin etkinliğinin yüksek olması gerekmektedir. Bu çalışmada WEKA yazılımı yardımıyla Karar Ağaçları, Lojistik Regresyon Ağaçları, J48 Ağaçları, Basit CART, Rassal Ağaç, Rassal Orman ve Naive Bayes yöntemleri kullanılmıştır. Doğru sınıflandırma ölçütü bakımından incelendiğinde en etkili sonucu Naive Bayes yönteminin verdiği belirlenmiştir. Daha sonra ise lojistik regresyon karar ağaçlarının etkili sonuçlar verdiği görülmüştür. Naive Bayes yöntemine göre, yolun kaplama cinsi ve yolun sınıfı değişkenlerinin kaza türünü etkileyen etkili değişkenler olduğu tespit edilmiştir. Singh ve Kaur (2014) çalışmalarında, yol karakteristikleri ile kaza türü arasında ilişki olduğundan bahsetmişlerdir. Bu çalışmada, bu sonuç doğrulanmıştır. Sonuç olarak, kaza analizinde Bayes ve regresyon yaklaşımını kullanan yöntemlerin performansının, diğer karar ağaçlarından daha yüksek olduğu tespit edilmiştir. Trafik kazaları ile ilgili politika üretenler için yol gösterici nitelikte olan bu çalışmada, yolun kaplama cinsi ve yol sınıfı değişkenlerinin trafik kazalarını etkilediği görülmüştür. Yol kaplama türü ve yol sınıfının değiştirilmesi ile ilgili yapılacak çalışmalar sayesinde, trafik kazalarının azalacağı ön görülmektedir.

Referanslar

Ahmed, L.A. (2017). Using logistic regression in determining the effective variables in traffic accidents. Applied Mathematic Science, 11(42), 2047-2058.doi:10.12988/ams.2017.75179

Akçetin, E., & Çelik, U. (2010). İstenmeyen elektronik posta (spam) tespitinde karar ağacı algoritmalarının performans kıyaslaması. İnternet Uygulamaları ve Yönetimi, 5(2), 43-56.doi:10.5505/iuyd.2014.43531

Akşehirli, Ö. (2012). Tıbbi Araştırmalarda Destek Vektör Makinelerinin Kullanımı. Yüksek Lisans Tezi. Düzce Üniversitesi, Düzce.

Albayrak, S. (2015). CE 4850 data mining sınıflama ve kümeleme yöntemleri. Ders Notları, Bilgisayar Mühendisliği, Yıldız Teknik Üniversitesi, İstanbul.

Aron, M., Billot, R., ElFaouzi, N., & Seidowsky, R. (2015). Traffic ındicators, accidents and rain: some relationships calibrated on a french urban motorway network. Transportation Research Procedia, 10, 31-40. doi: 10.1016/j.trpro.2015.09.053

Aydindag Bayrak, E., & Kirci, P. (2019) Intelligent big data analytics in health. In Early Detection of Neurological Disorders Using Machine Learning Systems, IGI Global 252-291. doi:10.4018/978-1-5225-8567-1.ch014

Doğruyol Başar, M., & Akan, A. (2018). Chronic kidney disease prediction with reduced individual classifiers, Electrica, 18(2), 249-255. doi: 10.26650/electrica.2018.99255

Bektaş, S. (2012). Çok şeritli bölünmüş karayollarında kaza tahmin modeli. İleri Teknoloji Bilimler Dergisi, 1 (1), 27-24.

Bolakar, H. (2014). Yapay Sinir Ağları ile Trafik Kazalarının Modellenmesi: Erzurum İli Örneği. Yüksek Lisans Tezi. Atatürk Üniversitesi, Erzurum.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. doi :10.1023/A:1010933404324

Chang, L., & Wang, H. (2006). Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention, 38(5), 1019-1027. doi: 10.1016/j.aap.2006.04.009

Chong, M., Abraham, A., & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. Informatica, 29(1), 89-98.

Cihan, Ş., Karabulut, B., Arslan, G., & Cihan, G. (2017). Koroner arter hastalığı riskinin veri madenciliği yöntemleri ile incelenmesi. International Journal of Engineering Research and Development, 10(1), 85-93. doi: 10.29137/umagd.419663

Daş, B., & Türkoğlu, İ. (2014). DNA dizilimlerinin sınıflandırılmasında karar ağacı algoritmalarının karşılaştırılması, Elektrik-Elektronik-Bilgisayar-Biyomedikal Mühendisliği Sempozyumu, 27 – 29.

Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis & Prevention, 38(3), 434-444. doi: 10.1016/j.aap.2005.06.024

EGM Trafik İstatistik Bülteni, 2018.

Freund, Y., & Mason, L., (1999). The alternating decision tree learning algorithm. Paper presented at the Proceedings of the Sixteenth International Conference on Machine Learning, 1-10.

Jiajia, L., Jie, H., Ziyang, L., Hao, Z., & Chen, Z. (2019). Traffic accident analysis based on C4.5 algorithm in WEKA. MATEC Web of Conferences, 272(10),1-8. doi: 10.1051/matecconf/201927201035

Kaur, G., & Chhabra, A. (2014). Improved J48 classification algorithm for the prediction of diabetes. International Journal of Computer Applications, 98(22), 13-17. doi:10.5120/17314-7433

Kwon, O.H., Rhee, W., & Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. Journal of Accident Analysis and Prevention, 75, 1-15. doi: 10.1016/j.aap.2014.11.005

Landwehr, N., Hall, M., & Frank, E. (2005). Logistic model trees. Machine Learning, 59(1-2), 161-205. doi:10.1007/s10994-005-0466-3

Long, W. J., Griffith, J. L., Selker, H. P., & D'Agostino, R. B. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. Computers in Biomedical Research, 26(1), 74-97. doi: 10.1006/cbmr.1993.1005

Ma, Y. (2013). The research of stock predictive model based on the combination of cart and DBSCAN. Ninth International Conference on Computational Intelligence and Security, 159-164.

Muhammad, L.J., Salisu, S., Yakubu, A., Malgwi, Y.M., Abdullahi, E.T., Mohammed, I.A., & Muhammad, N.A. (2017). Using decision tree data mining algorithm to predict causes of road traffic accidents, its prone locations and time along kano–wudil highway. International Journal of Database Theory and Application, 10(1), 197-206. doi: 10.14257/ijdta.2017.10.1.18

Özden, C., & Acı, Ç. (2018). Makine öğrenmesi yöntemleri ile yaralamalı trafik kazalarının analizi: Adana örneği. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 24(2), 266-275. doi: 10.5505/pajes.2016.87847

Özgan, E., Ulusu, H., & Yıldız, K. (2004). Trafik kaza verilerinin analizi ve kaza tahmin modeli. SAU Fen Bilimleri Enstitüsü Dergisi, 8(1), 160-166. doi:10.16984/saufbed.47078

Öztürk, E., & Mesut, A. (2016). Makine öğrenmesi kullanılarak jpeg xr standardında dosya boyutu belirleme işlemi, 24th Signal Processing and Communication Application Conference (SIU),1-4.

Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217–222. doi:10.1080/01431160412331269698

Patterson, D., Liu, F., Turner, D., Concepcion, A., & Lynch, R. (2008). Performance comparison of the data reduction system, Proceedings of the SPIE Symposium on Defense and Security, 1-6.

Quinlan, J. R. (1993). C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc.

Singh, M., & Kaur, A. (2014). A Review on Road Accident in Traffic System using Data Mining Techniques. International Journal of Science and Research, 5(1), 1530-1535.

Sohn, S.Y., & Shin, H. (2010). Pattern recognition for road traffic accident severity in Korea. Ergonomics, 44(1), 107-117. doi:10.1080/00140130120928

Şeker, S. (2012). Weka, http://bilgisayarkavramlari.sadievrenseker.com/2009/06/01/weka/.

Tolunay, M. K., & Gökdeniz, İ. (2002). Trafik bilincinin oluşması ve kurallara uyumu sağlamada kampanyaların yeri ve önemi. Uluslararası Trafik ve Yol Güvenliği Kongresi, 1-11.

Witten, I.H., & Frank, E. (2005). Data mining practical machine learning tools and techniques 2rd edition. San Fransisco, Morgan Kaufmann Publications.

Witten, I. H., Frank, E., & Hall, M. (2016). Data mining: Practical machine learning tools and techniques 3rd edition. USA: Morgan Kaufmann Publications.

Yılmaz, H. (2014). Random Forests Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama. Yüksek Lisans Tezi. Eskişehir Osmangazi Üniversitesi, Eskişehir.

Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. Advances in Space Research, 41(12), 1955-1959. doi: 10.1016/j.asr.2007.07.020.