POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Trafik kazalarının sınıflandırılmasında çok katmanlı algılayıcı, regresyon ve en yakın komşuluk algoritmalarının performans analizi

Performance analysis of multilayer perceptron, regression and nearest neighbor algorithms in classification of traffic accidents

***Yazar(lar) (Author(s))**: Emre KUŞKAPAN[1], Muhammed Yasin ÇODUR[2]*

ORCID[1]: 0000-0003-0711-5567

ORCID[2]: 0000-0001-7647-2424

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): Kuşkapan E. ve Çodur M. Y. "Trafik kazalarının sınıflandırılmasında çok katmanlı algılayıcı, regresyon ve en yakın komşuluk algoritmalarının performans analizi", ***Politeknik Dergisi***, 25(1): 373-380, (2022).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: **10.2339/politeknik.697530**

Trafik Kazalarının Sınıflandırılmasında Çok Katmanlı Algılayıcı, Regresyon ve En Yakın Komşuluk Algoritmalarının Performans Analizi

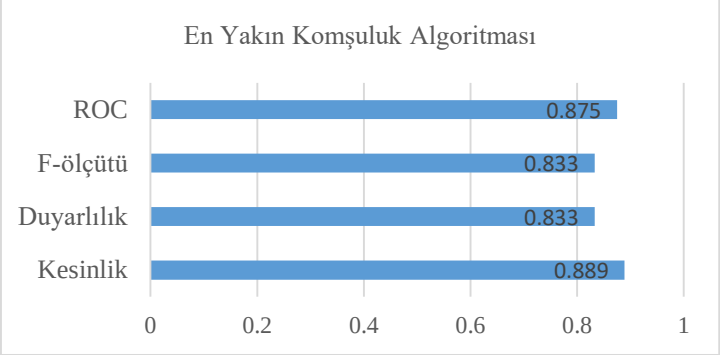
Performance Analysis of Multilayer Perceptron, Regression and Nearest Neighbor Algorithms in Classification of Traffic Accidents

Önemli noktalar (Highlights)

- *Makine öğrenmesi algoritmalarının trafik kazalarında uygulanması / Application of machine learning algorithms in traffic accidents*
- *En yakın komşuluk algoritması mevcut veri kümesi için en yüksek doğruluğa sahiptir. / The nearest neighbor algorithm has the highest accuracy for the current dataset.*
- *Trafik kazalarının doğru sınıflandırılması analiz işlemlerinin daha verimli şekilde yapılmasını sağlamaktadır. / Accurate classification of traffic accidents enables analysis processes to be carried out more efficiently.*

Grafik Özet (Graphical Abstract)

Trafik kazaları bir yapay zeka yöntemi olan makine öğrenmesi kullanılarak sınıflandırılmıştır. / Traffic accidents are classified using machine learning, which is an artificial intelligence method.



Şekil. En yakın komşuluk algoritmasının performans analizi **/ Figure.** Performance analysis of the nearest neighbor algorithm

Amaç (Aim)

Mevcut veri kümesi için makine öğrenmesi kullanılarak performans analizi ve hata ölçeklerinin tespit edilmesidir. / It is the performance analysis and detection of error scales using machine learning for the existing dataset.

Tasarım ve Yöntem (Design & Methodology)

Çalışmada makine öğrenmesi algoritmaları kullanılmıştır. / Machine learning algorithms were used in the study.

Özgünlük (Originality)

Trafik kazaları analizi için yüksek doğruluğa sahip algoritma tespit edilmiştir. / An algorithm with high accuracy has been determined for the analysis of traffic accidents.

Bulgular (Findings)

En yakın komşuluk algoritmasının daha iyi sonuçlar verdiği tespit edilmiştir. / It has been determined that the nearest neighbor algorithm gives better results.

Sonuç (Conclusion)

Trafik kazalarının incelenmesinde makine öğrenmesi algoritmalarının uygulanabilirliği ortaya konmuştur. / The applicability of machine learning algorithms in the examination of traffic accidents has been demonstrated.

Etik Standartların Beyanı (*Declaration of Ethical Standards*)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Trafik Kazalarının Sınıflandırılmasında Çok Katmanlı Algılayıcı, Regresyon ve En Yakın Komşuluk Algoritmalarının Performans Analizi

Araştırma Makalesi / Research Article

Emre KUŞKAPAN*, Muhammed Yasin ÇODUR

Mühendislik ve Mimarlık Fakültesi, İnşaat Mühendisliği Bölümü, Erzurum Teknik Üniversitesi, Türkiye

(Geliş/Received : 03.03.2020 ; Kabul/Accepted : 17.06.2021 ; Erken Görünüm/Early View : 28.06.2021)

ÖZ

Dünya genelinde artan nüfus ile birlikte taşıt sayısı da artış göstermektedir. Taşıt sayısının artışı ise birçok problemi beraberinde getirmektedir. Bu problemlerden en önemlisi ise trafik kazalarıdır. Trafik kazalarının maddi ve manevi önemli kayıplara sebep olabilme durumu bu alandaki çalışmaların gerekliliğini ortaya koymaktadır. Trafik kazalarının daha iyi analiz edilebilmesi ve kolay yorumlanabilmesi için sınıflandırma işlemine ihtiyaç duyulmaktadır. Bu kapsamda teknolojinin gelişmesi ve yapay zekâ teknolojilerinin insan hayatına girmesi ile çeşitli sınıflandırma yöntemleri ve bilgisayar programları geliştirilmektedir. Yapılan bu çalışmada; Ülkemizde yıllara göre meydana gelen trafik kaza verisi kullanılarak yıllar ölüm ve yaralanma durumlarına göre sınıflandırılmıştır. Daha sonra veri madenciliği algoritmaları olan çok katmanlı algılayıcı, regresyon ve en yakın komşuluk yöntemleri ile yılların trafik kaza sayılarına göre sınıflandırılma performansları ve hata ölçütleri WEKA analiz programı ile hesaplanmıştır. Her üç algoritmanın sınıflandırılma değerleri birbiri ile kıyaslandığında hem performans analizi hem de hata ölçütleri açısından birçok kriterde en yakın komşuluk algoritmasının daha iyi sonuçlar verdiği tespit edilmiştir. Yapılan bu çalışma sayesinde son yıllarda meydana gelen trafik kazalarında ölüm ve yaralanma oranının 2000'li yıllarının başında olduğu gibi tekrar yüksek risk seviyesine geldiği tespit edilmiştir. Bu durum karar vericilerin trafik kazalarını azaltmaya yönelik önlemlerini artırması adına önemlidir. Öte yandan yapılan sınıflandırma performanslarının incelenmesi sayesinde ise benzer özelliklere sahip veri kümesinin sınıflandırılması işleminde hangi algoritmanın tercih edilebileceği ortaya konmuştur.

Anahtar Kelimeler: Sınıflandırılma, veri madenciliği, algoritma performansları.

Performance Analysis of Multilayer Perceptron, Regression and Nearest Neighbor Algorithms in Classification of Traffic Accidents

ABSTRACT

With the increasing population worldwide, the number of vehicles also increases. The increase in the number of vehicles brings with it many problems. The most important of these problems is traffic accidents. The situation where traffic accidents can cause material and moral losses reveals the necessity of working in this field. Classification is needed for better analysis and easy interpretation of traffic accidents. In this context, various classification methods and computer programs are developed with the development of technology and the introduction of artificial intelligence technologies into human life. In this study; the years are classified according to death and injury situations by using traffic accident data occurring year by year in our country. Then, with the WEKA analysis program, multilayer perceptron, regression and classification performances and error criteria of the nearest neighbor methods were calculated. When the classification values of all three algorithms are compared with each other, it has been found that the nearest neighbor algorithm gives better results in many criteria in terms of both performance analysis and error criteria. Thanks to this study, it has been determined that the rate of death and injury in traffic accidents that have occurred in recent years has reached a high risk level again as it was in the early 2000s. This situation is important for decision makers to increase their measures to reduce traffic accidents. On the other hand, by examining the classification performances, it was revealed which algorithm can be preferred in the classification process of the data set with similar characteristics.

Keywords: Classification, data mining, performances of algorithm.

1. GİRİŞ (INTRODUCTION)

Dünya genelinde artan nüfus ile birlikte taşıt sayısı da her geçen gün artmaktadır. Artan taşıt sayısı da çeşitli problemlere sebep olmaktadır. Bu problemlerden en büyüğü ise trafik kazalarıdır. Trafik kazaları maddi kayıpların yanı sıra ölüm ve yaralanmalar gibi önemli manevi kayıplara neden olmaktadır. Trafik kazalarının önceden tespit edilebilmesi çok güç olmakla birlikte bu kazalar için çeşitli tahminler yürütülebilmektedir. Benzer şekilde trafik kazalarının sebep olduğu problemler tespit edilerek önlemler alınabilmektedir.

**Sorumlu Yazar (Corresponding Author)*
e-posta : emre.kuskapan@erzurum.edu.tr

Problemlerin tespit edilmesi için ise trafik kazaları sınıflandırılarak incelemelere daha kolay hale gelebilmektedir. Bu doğrultuda trafik kazalarının analizi, tahmini ve gruplandırılması için çeşitli yöntemler kullanılmaktadır. Kullanılmakta olan yöntemler ise veriyi farklı yönlerden inceledikleri için farklı sonuçlar verebilmektedirler. Yapılan çalışmalarda veriyi tek bir yöntemle incelemek yerine birden fazla yöntemle inceleyerek her bir yönteme ait sonuçların karşılaştırılması daha çok tercih edilmektedir.

Trafik kazalarının incelenmesi üzerine çok sayıda akademik çalışma mevcuttur. Akademik çalışmalar sayesinde trafik kazaları daha iyi şekilde yorumlanabilmektedir. Bu çalışmalarda ise çeşitli analiz yöntemleri kullanılabilmektedir. Murat ve Şekerler yaptıkları çalışmada, Denizli iline ait 2004, 2005 ve 2006 yıllarına ait trafik kaza sayılarını, bilgisayar programları kullanarak klasik ve bulanık kümelenme yöntemleriyle analiz etmişlerdir. Kümeleme yöntemi olarak K-ortalamalar yöntemi ve bulanık C-ortalamalar yöntemi kullanılmıştır. Elde edilen sonuçlar neticesinde yöntemlere doğrulama uygulanarak küme merkezlerine denk gelen kara noktalar belirlenmiştir [1]. Kashani ve Mohaymany 2011 yılında yaptıkları çalışmada; İran'daki çift şeritli ve çift yönlü kırsal yollarda trafik kazalarındaki yaralanmalarının şiddetini etkileyen faktörleri tanımlamışlardır. 2006-2008 yılları arasında bu bölgelerde meydana gelen trafik kaza istatistikleri veri madenciliğinin en yaygın yöntemlerinden biri olan sınıflandırma ve regresyon ağaçları kullanılarak analizler yapılmıştır [2]. Başka bir çalışmada ise Erzurum'un ilçelerinde meydana gelen trafik kazalarının Coğrafi Bilgi Sistemleri (CBS) ile değerlendirilmesi yapılmıştır. 2006-2008 yıllarında meydana gelen trafik kazaları esas alınarak ilçe haritası CBS ortamında sayısallaştırılmıştır. Böylece en çok trafik kazası, ölüm ve yaralanmanın meydana geldiği ilçeler görsel olarak harita üzerinde tespit edilmesi sağlanmıştır [3]. Çodur vd. yaptıkları çalışmada, trafik güvenliğini etkileyen ana faktörleri esas alıp 2005-2010 yılları arasında Erzurum Kuzey Çevre Yolunda meydana gelen trafik kaza verisi kullanarak kaza tahmin modeli oluşturmuşlardır. Tahmin modeli oluşturulurken genelleştirilmiş lineer regresyon modeli kullanmışlardır. Model sonucunda kazaya etki eden faktörlerdeki değişimler yorumlanmıştır [4]. Atalay vd. yaptıkları çalışmada, 1977-2006 yılları arasında meydana gelen aylık trafik kaza istatistikleri (şehir içi ve şehir dışı toplamı) kullanarak zaman serisi analiz yöntemi ile modelleme yapmışlardır. Yapılan analizler sonucunda çalışma döneminde kullanılan veriye göre en uygun modelin ARIMA(4,1,4) olduğu belirlenmiştir. Çalışmada aynı zamanda en uygun model kullanılarak aylık kaza tahminleri yapılmıştır [5]. Başka bir çalışmada ise Gupta vd. yaptıkları çalışmada, Hindistan'ın Mujjafarnagar bölgesinde polis kayıtlarından elde edilen trafik kazaları bilgileri kullanılarak trafik kazalarının şiddeti sınıflandırılmıştır. Sınıflandırma işlemi gerçekleştirilirken ortaklık kural madenciliği yöntemi kullanılarak üç veri kümesi (ölümcül, büyük yaralanmalı ve küçük yaralanmalı) elde edilmiştir [6]. Rovsek vd. ise yaptıkları çalışmada, trafik kazalarından kaynaklı olarak meydana gelen ölüm ve ciddi yaralanmaların giderek arttığını ve tedbirler alınması gerekliliğini vurgulamışlardır. Tedbirler alınmadan önce kazaların şiddetini etkileyen faktörlerin sınıflandırılmasına ihtiyaç duyulduğu belirtilmiştir. Bu kapsamda Slovenya'da 2005-2009 yılları arasında meydana gelen trafik kazaları esas alınarak trafik kaza faktörleri parametrik olmayan bir sınıflandırma ağacı kullanılarak tanımlanmıştır [7]. Syahputri vd. yaptıkları çalışmada, Endonezya'nın Medan kentinde 2018 yılında meydana gelen kazaları ciddiyetine göre gruplandırmışlardır. Daha sonra kentteki yollar Bulanık C-Ortalamalar Algoritması ile kümelenerek güvenli olup olmadıkları belirlenmiştir [8]. Murat vd. ise yaptıkları çalışmada, Denizli ilindeki trafik kazalarının mekânsal analizini sunmuşlardır. Analizde 2004, 2005 ve 2006 yıllarına ait trafik kaza bilgileri, K-ortalamalar yöntemi ve C-ortalamalar yöntemi kullanılarak analiz edilmiştir. Sonuçlar, 2004 ve 2005 yıllık veri kayıtlarına göre 2006 yılında trafik kazalarında artış olduğunu göstermiştir [9]. Atalay ve Tortum yaptıkları çalışmada, Türkiye'deki meydana gelen kaza bilgilerini kullanılarak her il için ölüm ve yaralanma oranları hesaplanmıştır. Buna göre, hem geleneksel k-ortalamalar hem de bulanık c-ortalamalar teknikleri yardımıyla kümeleme analizi yapılmıştır. Elde edilen sonuçlar neticesinde bulanık c-ortalamalar tekniğinin daha tutarlı sonuçlar verdiği tespit edilmiştir [10].

Özden ve Acı yaptıkları çalışmada, Adana ilinde 2005 ile 2014 yılları arasında meydana gelen yaralanmalı trafik kazalarına ait aylık bazdaki istatistikler ile yaralanmalı kaza sayısını tahmin edecek modeller geliştirmişlerdir. Tahmin modellerinde, İleri Beslemeli Çok Katmanlı Yapay Sinir Ağı (İBÇK-YSA), Fonksiyon Uydurma Yapay Sinir Ağı (FU-YSA), Genelleştirilmiş Regresyon Yapay Sinir Ağı (GR-YSA), Regresyon Ağacı (RA), Destek Vektör Makinesi (DVM) ve Çoklu Doğrusal Regresyon Analizi (ÇDR) yöntemleri kullanılmıştır. Çalışma sonucunda, DVM yönteminin her iki tahmin senaryosunda da en başarılı sonuçları verdiği görülmüştür [11]. Selvi ve Çağlar 2018 yılında yaptıkları çalışmada trafik kazalarının meydana gelmesini etkileyen faktörleri kullanarak trafik kazalarını sınıflandırmış ve haritalandırmışlardır. K-ortalamaları yöntemi, Kmedoids yöntemi ve kümeleme analizi yöntemleri arasında yer alan Aglomerasyonel ve Bölünmüş Hiyerarşik Kümeleme Yöntemi ile üretilen çok değişkenli haritaların gerçek değerlerinin karşılaştırılması yapılmıştır [12]. Mlouk ve Agouti yaptıkları çalışmada, veri madenciliği algoritmaları olan kural analizi, zaman serileri ve çok kriterli karar verme yöntemlerini kullanarak trafik kazalarını incelemişlerdir. Geliştirmiş oldukları sistemin; trafik kazalarını tahmin etmek ve yol güvenliğini artırabilmenin önemli bir çalışma olduğuna değinmişlerdir [13]. Das vd. yaptıkları çalışmada, Amerika Birleşik Devletlerinin bir eyaleti olan Louisiana'da 2010-2016 yılları arasında meydana gelen trafik kazalarını araç kusurları yönünden

incelemişlerdir. Çalışmada bir veri madenciliği yaklaşımı olan Bayes yöntemi kullanılarak araç özelliklerine göre (araç yaşı, lastiklerin durumu, fren yapısı vb.) trafik kazaları sınıflandırılmıştır [14].

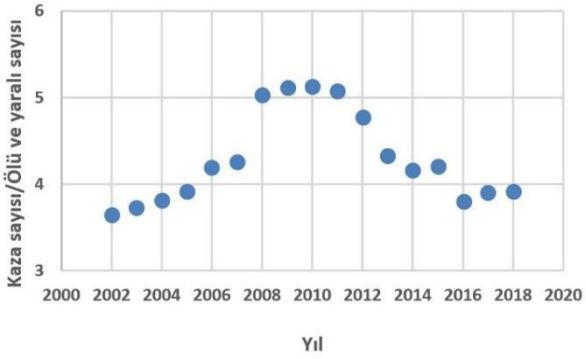
Yapılan bu çalışmada ise Ülkemizde 2002 ile 2018 arasındaki yıllar, trafik kaza sayısının ölüm ve yaralanma sayısına oranına göre sınıflandırılmıştır. Elde edilen üç değer aralığına göre belirtilen yıllardaki trafik kazaları risk durumlarına göre adlandırılmıştır. Daha sonra bu yıllardaki nüfus, toplam taşıt sayısı, toplam trafik kazası, kaza sonucu meydana gelen ölüm ve yaralanma sayıları kullanılarak veri madenciliği algoritmaları olan çok katmanlı algılayıcı, regresyon ve en yakın komşuluk yöntemlerinin performansları analiz edilmiştir. Son olarak trafik kaza veri setine göre her üç algoritmanın doğruluk oranları tespit edilip karşılaştırma yapılmıştır. Mevcut veri kümesine benzer çalışmalar için hangi algoritmanın daha iyi sınıflandırma yapacağı tespit edilmiştir.

2. MATERYAL VE METOD (MATERIAL and METHOD)

Ülkemizde artan nüfus ile birlikte taşıt sayısı da her geçen gün artmaktadır. Kentleşmenin artması ile yeni yaşam alanlarının meydana gelmesi sebebiyle karayolu yol ağı da her geçen gün genişlemektedir.Her iki durum bir arada değerlendirildiğinde ise günlük hayatta bir problem oluşmayacağı yanılgısı meydana gelebilmektedir. Fakat yol ağları artmış olmasına rağmen yüksek binalar sebebiyle kentlerde birim alanda yaşayan insan sayısı artmakta ve teknolojinin gelişmesi ile yüksek hızlı taşıtların sayısı da her geçen gün artmaktadır. Trafik kaza sayısının artmasında bu durumlara benzer olarak birçok sebep sayılabilmektedir. Bu nedenle trafik kazaları; nüfus ve taşıt sayısı gibi devamlı artış göstermiyor olsa da genel olarak artış eğilimindedir.

Çizelge 1'de 2002-2018 yılları arasında Ülkemizde nüfus, taşıt sayısı, trafik kaza sayısı ve bu kazalara bağlı olarak meydana gelen ölü ve yaralı sayısı gösterilmektedir. Kazalar beraberinde maddi ve manevi kayıplara sebebiyet verebilmektedir. Maddi kayıplar için telafiler sunulabilirken ölüm ve yaralanma ile sonuçlanabilen trafik kazaların telafisi oldukça zordur. Bu durumun önüne geçilebilmek için tek çözüm trafik kazalarının azaltılmasını sağlamaktır.

Çizelge incelendiğinde yıllara göre nüfus sayısı ve toplam taşıt sayısında devamlı artış olduğu görülmektedir. Trafik kaza sayısında ise genel olarak artış olmakla birlikte bazı yıllarda dalgalanmalar meydana geldiği görülmektedir. Benzer durum trafik kazaları sonucunda meydana gelen ölüm ve yaralanma durumunda da görülmektedir. Fakat trafik kaza sayısının artış göstermesi aynı şekilde ölüm ve yaralanma sayısının da artış göstereceği durumuna sebep olmadığı görülmektedir. Bu bağlamda yılların sınıflandırma işleminde meydana gelen trafik kaza sayısının bu kazalarda meydana gelen ölüm ve yaralanma sayısına oranı esas alınmıştır. Bu oranın düşük olması birim kazada meydana gelen ölüm ve yaralanma sayısının fazla olduğunu göstermektedir. Sınıflandırma işleminde ölüm ve yaralanma durumuna göre yıllar; yüksek riskli, orta riskli ve düşük riskli olarak ayrılmıştır. Bu durumda yıllara göre trafik kaza sayısının ölüm ve yaralanma sayısına oranını içeren grafik Şekil 1'de gösterilmektedir.



Şekil 1. Yıllara göre trafik kaza sayısının bu kazalardaki ölü ve yaralanma sayısına oranı (The ratio of the number of traffic accidents to the number of deaths and injuries in these accidents by years)

Çizelge 1. Yıllara göre Ülkemizde nüfus ve trafik kazalarında meydana gelen değişimler (Changes in population and traffic accidents in our country by years) [15]

Yıllar	Nüfus Sayısı	Toplam Taşıt Sayısı	Trafik Kazası Sayısı	Ölü ve Yaralı Sayısı
2002	65.022.300	8.655.170	439.777	120.505
2003	65.938.265	8.903.843	455.637	122.160
2004	66.845.635	10.236.357	537.352	140.864
2005	67.743.052	11.145.826	620.789	158.591
2006	68.626.337	12.227.393	728.755	173.713
2007	69.496.513	13.022.945	825.561	194.064
2008	70.363.511	13.765.395	950.120	188.704
2009	71.241.080	14.316.700	1.053.346	205.704
2010	72.137.546	15.095.603	1.106.201	215.541
2011	73.058.638	16.089.528	1.228.928	241.909
2012	73.997.128	17.033.413	1.296.634	271.829
2013	76.667.864	17.939.447	1.207.354	278.514
2014	77.695.904	18.828.721	1.199.010	288.583
2015	78.741.053	19.994.472	1.313.359	311.951
2016	79.814.871	21.090.424	1.182.491	311.112
2017	80.810.525	22.218.945	1.202.716	307.810
2018	82.003.882	22.865.921	1.229.364	313.746

Şekil 1 incelendiğinde ise trafik kaza sayısının, bu kazalarda meydana gelen ölüm ve yaralanma sayısına oranının 3 ila 6 arasında olduğu görülmektedir. Bu sebeple 3 ve 4 değerleri arasında bulunan yılların riski yüksek, 4 ve 5 değerleri arasında bulunan yılların riski orta, 5 ve 6 değerleri arasında bulunan yılların riski düşük olarak belirlenmiştir. Bu duruma göre aşağıda verilmiş olan Çizelge 2'de yıllara göre trafik kazalarından meydana gelen ölüm ve yaralanma risk durumları gösterilmektedir.

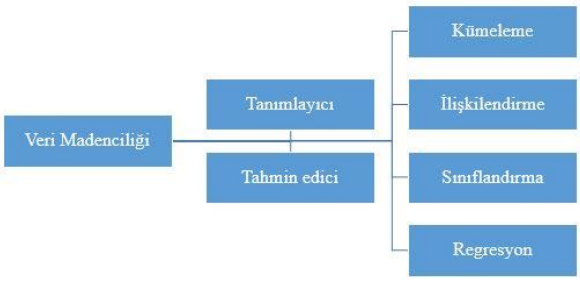
Çizelge 2. Yıllara göre Ülkemizde trafik kazalarının risk durumları (Risk situations of traffic accidents in our country by years)

Risk Durumu	Yıllar
Düşük	2008-2009-2010-2011
Orta	2006-2007-2012-2013-2014-2015
Yüksek	2002-2003-2004-2005-2016-2017-2018

Yılların risk durumları incelendiğinde devamlı artış veya azalış durumu söz konusu değildir. Bu sebeple risk durumu lineer değişim göstermemektedir. Lineer yapıya sahip olmayan veri sınıflandırılmasını doğru tahmin elde edebilen algoritmalar bulabilmek oldukça zordur. Sistem yapısına uygun olan algoritmalar tercih edilerek bu algoritmalar içerisinde veri kümesine doğruluğu yüksek olan algoritmalar belirlenmelidir. Algoritmaların tercih edilmesinde ise veri madenciliği kullanılmaktadır.

2.1 Veri Madenciliği (Data Minning)

Veri madenciliği, makine öğrenimi, istatistik ve veri tabanı sistemlerinin kesişimindeki yöntemleri içeren büyük veri kümelerinde kalıpları keşfetme işlemidir. Veri madenciliği, bir veri kümesinden bilgi (akıllı yöntemlerle) çıkarmak ve bilgiyi daha fazla kullanım için anlaşılabilir bir yapıya dönüştürmek için genel bir amacı olan bilgisayar bilimi ve istatistiklerinin disiplinler arası bir alt alanıdır. Ham analiz adımının yanı sıra, veri tabanı ve veri yönetimi yönleri, veri ön işleme, model ve çıkarımla ilgili hususlar, ilginçlik ölçütleri, karmaşıklık konuları, keşfedilen yapıların sonradan işlenmesi, görselleştirme ve çevrimiçi güncellemeyi de içermektedir.

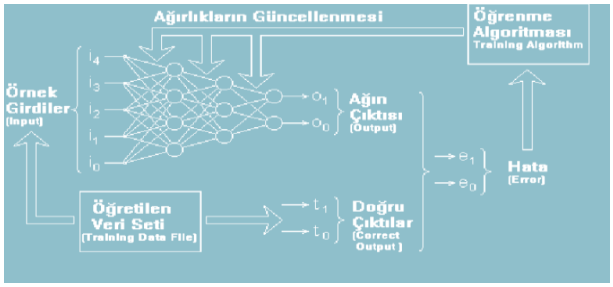


Şekil 2. Veri madenciliği teknikleri (Data mining techniques)

Şekil 2'de gösterildiği üzere veri madenciliği tanımlama ve tahmin etme gibi iki temel amacı içermektedir. Tanımlama işleminin gerçekleştirilmesi için kümeleme ve ilişkilendirme yöntemleriyle gerçekleştirilmektedir. Tahmin etme ise sınıflandırma ve regresyon yöntemleriyle gerçekleştirilmektedir [16]. Bu yöntemleri veri kümelerine uygulayabilmek için ise çeşitli algoritmalar üretilmiştir. Bu algoritmaların analizi için birçok bilgisayar programı kullanılmaktadır. Yapılan bu çalışmada bir makine öğrenmesi programı olan Weka (Waikato Environment for Knowledge Analysis) yazılımı kullanılmıştır. Bu yazılım içerisinde veri madenciliği algoritmalarını ve metotlarını bulundurmaktadır [17].

2.1.1 Çok katmanlı algılayıcı ile sınıflandırma (Classification via multilayer perceptron)

Yapay sinir ağlarının bir modelini oluşturan çok katmanlı algılayıcılar, denetimli bir öğrenme algoritmasıdır. Ağ bir giriş katmanı, bir gizli katman ve bir çıkış katmanı olmak üzere en az üç düğüm katmanından oluşur. Özellikle sınıflandırma ve genelleme yapma durumlarında etkin şekilde çalışmaktadır. Delta öğrenme kuralı olarak da adlandırılan bu ağın öğrenebilmesi için örnek giriş ve çıkışlardan oluşan eğitim seti şarttır [18]. Şekil 3'te çok katmanlı algılayıcının öğrenme yapısı aşamaları gösterilmektedir.



Şekil 3. Çok katmanlı algılayıcı için öğrenme yapısı (Learning structure for multilayer perceptron) [19]

Algılayıcı ile sınıflandırma genel olarak 1 numaralı denklemdeki formülasyon ile temsil edilmektedir. Burada ω ağırlıkların vektörünü, x girdilerin vektörünü, b sapmaları ve φ doğrusal olmayan aktivasyon fonksiyonunu ifade etmektedir [20].

$$y = \varphi(\sum_{i=1}^{n} \omega_i x_i + b) = \varphi(w^T x + b) \quad (1)$$

2.1.2 Regresyon ile sınıflandırma (Classification via regression)

Regresyon iki ya da daha çok değişken arasındaki ilişkiyi ölçmek için kullanılan istatistiksel bir yaklaşımdır. Eğer bir bağımlı ve bir bağımsız değişken varsa basit regresyon analizi, bir bağımlı değişken ve iki veya daha fazla bağımsız değişken varsa çoklu (multiple) regresyon analizi, hem bağımlı hem de bağımsız değişken sayısı iki veya daha fazla ise çok değişkenli (multivariate) regresyon analizi kullanılır [21]. Regresyon analizi ile değişkenler arasındaki ilişkinin varlığı, eğer ilişki var ise bunun gücü hakkında bilgi edinilebilir. Regresyon, iki (ya da daha çok) değişken arasındaki doğrusal ilişkinin fonksiyonel şeklini, biri bağımlı diğeri bağımsız değişken olarak bir doğru

denklemi olarak göstermekle kalmaz, değişkenlerden birinin değeri bilindiğinde diğeri hakkında kestirim yapılmasını sağlar. Genellikle bu iki (veya çok) değişkenlerin hepsinin niceliksel ölçekli olması zorunluluğu vardır. Regresyon modeli için hesaplanırken; 2 numaralı denklemde olduğu gibi y_i bağımlı değişken parametrelerin bir doğrusal birleşiği olmak koşulu ile x_i bağımsız değişkenler, β_0 ve β_1 parametreler ve ε_i hata terimi olması koşulu ile şu şekilde ifade edilmektedir [22].

$$y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \varepsilon_i \quad (2)$$

2.1.3 En yakın komşuluk ile sınıflandırma (Classification via nearest neighbour)

En yakın komşuluk algoritması örnek tabanlı öğrenme olarak da bilinmektedir. Bu algoritma yeni bir veri örneğinin bilinmeyen bir çıktı değerini tahmin etmek için geçmiş veri örneklerini, bilinen çıktı değerleriyle kullanılmasını sağlayan kullanışlı bir veri madenciliği tekniğidir [23]. Bu algoritma açık genelleme yapmak yerine yeni problem örneklerini eğitimde görülen ve bellekte depolanan örneklerle karşılaştırmaktadır. En yakın komşuluk algoritmasının en önemli avantajı ise modelini daha görülmemiş veri uyarlayabilmesidir. Hafızaya dayalı öğrenme olarak da bilinen EYK yeni bir örneği için bir değer veya sınıf tahmin ederken, bu örnek için daha önceki antrenman örnekleri ile arasındaki mesafeleri veya benzerlikleri hesaplamaktadır [24]. En yakın komşuluk algoritması; ana veri setindeki noktaların her birinin, esas değeri bilinmeyen test verisindeki bir noktaya olan uzaklıkların hesaplanması ile bulunmaktadır. Böylece en yakın uzaklığa sahip k sayıda gözlemin seçilmesi ile komşuluklar hesaplanmaktadır. Bu yöntem uzaklıkların hesaplanmasını yaparken, i ve j noktaları için 3 numaralı denklemde formülasyonu verilen Öklit uzaklığını kullanmaktadır [25].

$$d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (3)$$

2.2 Veri Madenciliği Performans ve Hata Ölçekleri Analizi (Data Mining Performance and Error Scales Analysis)

Veri madenciliğinde performans analizi yapılırken temel başarı ölçütü kavramları kullanılmaktadır. Bu kavramlar kesinlik, duyarlılık, F-ölçütü ve ROC kriterleridir. Bu kavramların değerleri hesaplanırken tahmin edilen ile eldeki veri kıyaslanma durumu hesaba katılmaktadır. Kıyaslama işleminde DP (doğru pozitif-doğruya doğru demek), DN (doğru negatif-doğruya yanlış demek), YP (yanlış pozitif-yanlışa doğru demek) ve YN (yanlış negatif-yanlışa yanlış demek) değerleri kullanılmaktadır. Şekil 4'te verilen karışıklık matrisi kullanılarak sınıflandırma algoritmalarının doğruluk değerleri hesaplanabilmektedir. Kesinlik ifadesi 4 numaralı denklemde de belirtildiği üzere sınıfı 1 olarak tahmin edilmiş doğru ve pozitif örnek sayısının, sınıfı 1 olarak tahmin edilmiş tüm örnek sayısına oranıdır [26].

		Öngörülen Sınıf	
		Sınıf=1	Sınıf=0
Gerçek Sınıf	Sınıf=1	DP	YN
	Sınıf=0	YP	DN

Şekil 4. Karışıklık matrisi (Confusion matrix)

Duyarlılık, 5 numaralı denklemde doğru sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranı olarak tanımlanmıştır. F-ölçütü ise 6 numaralı denklemde hem duyarlılık hem de kesinlik ifadelerini birlikte değerlendirmek amacıyla bu iki ifadenin harmonik ortalaması olarak belirtilmektedir [27]. ROC değeri ise model performansının genel olarak yorumlanabilmesi için oluşturulan eğri ile elde edilmektedir. Bu performans değerlerinin hepsi 0 ile 1 arasında değerler almaktadır.

$$\text{Kesinlik} = \frac{\text{DP}}{\text{DP+YP}} \quad (4)$$

$$\text{Duyarlılık} = \frac{\text{DP}}{\text{DP+YN}} \quad (5)$$

$$\text{F} - \text{ölçütü} = \frac{2\times\text{Duyarlılık}\times\text{Kesinlik}}{\text{Duyarlılık+Kesinlik}} \quad (6)$$

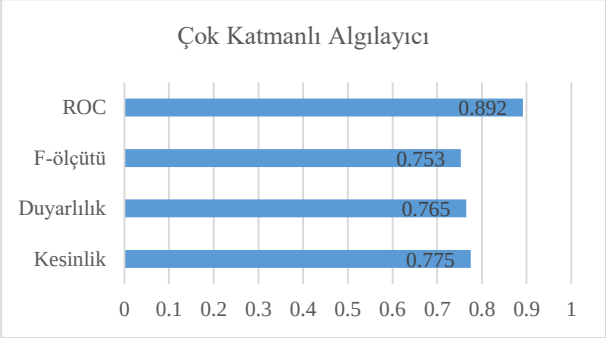
$$\text{Doğruluk} = \frac{\text{DP+DN}}{\text{DP+YP+YN+DN}} \quad (7)$$

Modelin hata ölçekleri ise doğruluk oranı, ortalama mutlak hata (MAE), kök hata kareler ortalaması (RMSE) ve Kappa istatistiği ile belirlenmektedir. 6 numaralı denklemde gösterilen doğruluk oranı algoritmanın başarısını gösteren en önemli kriter olmakla birlikte öngörülen değerin ölçüm değeri ile ne kadar uygun şekilde bulunduğunu ifade etmektedir. MAE tüm veri öngörülen değerler ile ölçüm değeri arasındaki farkın ortalaması olarak ifade edilmektedir. RMSE değeri model tarafından tahmin edilen değerler ve elde edilen ölçüm değeri arasındaki farkın ortalamasının karekökü alınarak hesaplanır [28]. Kappa değeri ise gözlemsel arasındaki uyuşmayı ölçmek için ifade edilen bir terimdir. Bu değer 1'e ne kadar yakınsa gözlemler arasında o kadar iyi uyuşma vardır demektir [29].

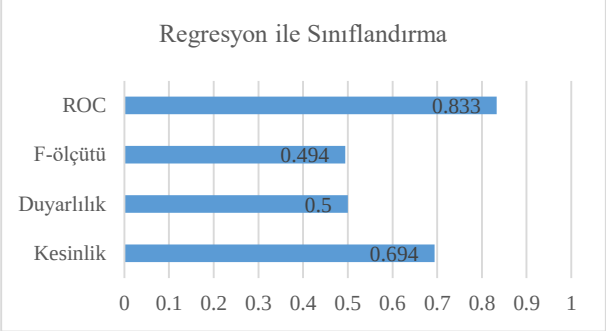
3. ARAŞTIRMA BULGULARI VE TARTIŞMA (FINDINGS and DISCUSSION)

Yılların sınıflandırılmasında veri madenciliği algoritmalarının performansı incelenirken veri kümesine ait sınıfların belirlenmiş olması gerekmektedir. Bu çalışmada veri kümesine göre trafik kazalarının ölüm ve yaralanmaya sebep olma durumlarına göre sınıflandırılması tercih edilmiştir. Sınıflandırmaya göre 2000'li yılların başında birim kazada meydana gelen ölüm ve yaralanma sayısının 2010 yılına kadar azalmakta

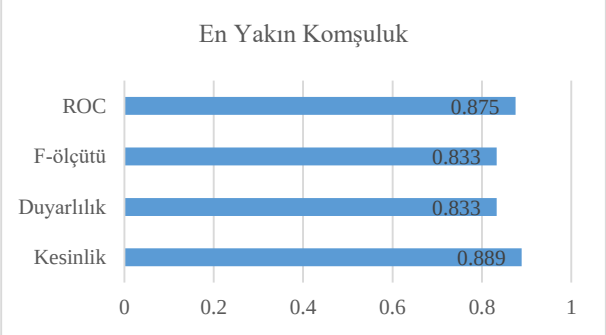
olduğu daha sonra tekrar artışa geçtiği görülmektedir. Yıllar içerisinde meydana gelen dalgalanmalar sınıflandırma algoritmalarının doğruluk oranlarını etkileyebilmektedir. Risk durumunun devamlı artış veya azalış durumunda olmaması ise sınıflandırma algoritma performanslarının kıyaslanabilmesi adına uygun bir yapı içermektedir.



Şekil 5. Çok katmanlı algılayıcı algoritması için performans değerleri (Performance values for multilayer perceptron algorithm)



Şekil 6. Regresyon algoritması için performans değerleri (Performance values for the regression algorithm)



Şekil 7. En yakın komşuluk algoritması için performans değerleri (Performance values for the nearest neighbour algorithm)

Algoritmaların performans değerleri gösterilirken veri kümesindeki tüm sınıflar ayrı ayrı gösterilmek yerine bu performans değerlerinin ağırlıklı ortalama değerleri tercih edilmiştir. Örneğin F-ölçütü her bir sınıf değeri için farklı değer almaktadır. Çalışmada F-ölçütü için her bir sınıfta elde edilen sonuçlar bulunarak bu değerlerin ağırlıklı ortalaması grafiklere işlenmiştir. Bu durumlara göre yılların sınıflandırmasında çok katmanlı algılayıcı algoritmasına ait performans değerleri Şekil 5'te gösterilmektedir. Grafiğe göre kesinlik ve duyarlılık ölçütlerinin 0,77 değeri civarında olduğu görülmektedir. Bununla birlikte F-ölçütü değeri kesinlik ve duyarlılık değerlerine bağlı bir ölçüt olmasına karşın iki kriterin de altında 0,75 değerinde yer almaktadır. Bu duruma sebep olarak da çok katmanlı algılayıcı algoritmasının genel olarak kesinlik ve duyarlılık performansının iyi, sınıf bazında ise dalgalanmaların meydana gelmiş olma durumu söylenebilir. Öte yandan bu durumu destekler nitelikte olarak ise ROC değerinin daha yüksek olması çok katmanlı algılayıcı algoritmasının ayrı ayrı sınıf performansından ise genel performansının daha iyi olduğu durumu ortaya çıkmaktadır. Şekil 6 incelendiğinde ise regresyon ile sınıflandırma algoritmasının performansının daha düşük değerlere sahip olduğu görülmektedir. Bunun yanı sıra ROC değerinin diğer kriterlere göre daha yüksek değerde olması çok katmanlı algılayıcı algoritmasında olduğu gibi ayrı ayrı sınıf performansının genel performansından daha iyi olduğunu göstermektedir.

Şekil 7'de ise yılların sınıflandırılmasında en yakın komşuluk algoritmasına ait performans değerleri gösterilmektedir. Grafikteki değerler incelendiğinde kesinlik değerinin oldukça yüksek olduğu görülmektedir. Aynı zamanda F-ölçütü değerlerinin duyarlılık değerine yakın olması sınıf bazındaki değerlerin de iyi olduğunu göstermektedir. Her iki algoritmayı karşılaştırdığımızda ise çok katmanlı algılayıcının ROC değeri daha yüksek olması itibariyle genel performans açısından iyi olduğu fikrini ortaya koyarken diğer üç kriterin en yakın komşuluk algoritmasında daha yüksek olması durumu da en yakın komşuluk algoritmasının sınıf bazında performansının iyi olduğu fikrini ortaya koymaktadır.

Çizelge 3. Çok katmanlı algılayıcı, regresyon ve en yakın komşuluk algoritmalarının hata ölçekleri (Multilayer perceptron, regression and error scales of the nearest neighbour algorithms)

Algoritma	Doğruluk (%)	MAE	RMSE	Kappa
Çok katmanlı algılayıcı	76.47	0.208	0.318	0.646
Regresyon	50	0.374	0.424	0.28
En yakın komşuluk	83.33	0.183	0.312	0.739

Her üç algoritma için de performans açısından öne çıktığı kısımlar olması itibariyle hangi algoritmanın bu veri kümesi için daha iyi sınıflandırma yaptığına karar verebilmek için hata ölçeklerine bakılması gerekmektedir. Çizelge 3'te her üç algoritma durumu için yılların sınıflandırılmasındaki doğruluk yüzdesi, hata değerleri ve Kappa istatistiği değerleri gösterilmektedir. Sonuçlara göre en yakın komşuluk algoritmasının doğruluk yüzdesi daha yüksek ve MAE, RMSE hata

ölçekleri daha düşüktür. Bununla birlikte ölçüm değeri ile tahmin edilen değer arasındaki uyuşmayı ifade eden Kappa istatistiği değerinin de en yakın komşuluk algoritmasında daha yüksek olduğu görülmektedir. Sonuçlar genel olarak değerlendirildiğinde yılların trafik kaza bilgilerine göre doğru sınıflandırılması için en yakın komşuluk algoritmasının çok katmanlı algoritma ve regresyon yöntemlerine göre daha iyi sonuçlar verdiği görülmektedir.

4. SONUÇLAR (CONCLUSIONS)

Ülkemizde karayolu ağ yapısının gelişmesi, yolların şeritlerinin artırılması ve araç teknolojilerinin ilerlemiş olması insanlarda doğal olarak karayolu yolculuğunun daha güvenli hale geldiği fikrine sebep olmaktadır. Bu durum sürücülerin daha dikkatsiz araç kullanmasına ve kurallara daha az uymasına sebep olmaktadır. Aynı zamanda karar vericiler de bu sebeplerle trafikteki bazı kısıtlamaları gevşetebilmektedir. Bu durumu incelemek amacıyla; Ülkemizde 2002-2018 yılları arasında meydana gelen ölümlü ve yaralanmalı trafik kaza sayısı ve bu kazalar sonucu oluşan ölüm ve yaralanma sayıları kullanılarak yıllar risk durumuna göre sınıflandırılmıştır. Bu kazalar ile meydana gelen ölüm ve yaralanma sayısına oranının 3 ila 6 arasında olduğu görülmüştür. Bu sebeple 3 ve 4 değerleri arasında bulunan yılların riski yüksek, 4 ve 5 değerleri arasında bulunan yılların riski orta, 5 ve 6 değerleri arasında bulunan yılların riski düşük olarak belirlenmiştir. Bu sınıflandırma doğrultusunda son yıllarda birim kazada meydana gelen ölüm ve yaralanma sayısının arttığı tespit edilmiştir. Bu durum karar vericiler için trafik kazalarının azaltılmasına yönelik tedbirleri artırması adına önemlidir. Çalışmanın bir diğer amacı ise belirlenen sınıflar doğrultusunda; veri madenciliği algoritmaları olan çok katmanlı algılayıcı, regresyon ve en yakın komşuluk yöntemleri uygulanarak bu algoritmaların veri kümesi için performans analizi ve hata ölçekleri hesaplanarak benzer veri kümelerine en uygun algoritmanın tespit edilmesidir. Her üç yönteme göre elde edilen performans değerleri ve hata ölçeği değerleri kıyaslandığında yılların doğru sınıflandırma işlemi için en yakın komşuluk algoritmasının daha iyi sonuçlar verdiği tespit edilmiştir. Bu doğrultuda trafik kazalarının ve benzer şekilde lineer değişim göstermeyen veri kümelerinin daha iyi incelenebilmesi için tercih edilen sınıflandırma işlemlerinde en yakın komşuluk algoritmasının çok katmanlı algılayıcı ve regresyon algoritmalarına göre daha tercih edilebilir bir yöntem olduğu fikri ortaya konmuştur.

ETİK STANDARTLARIN BEYANI (DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazarları çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

Emre KUŞKAPAN: Verilerin analizini yapmış ve sonuçları elde etmiştir.

Muhammed Yasin ÇODUR: Verileri temin etmiş ve sonuçları yorumlamıştır.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST)

Bu çalışmada herhangi bir çıkar çatışması yoktur.

KAYNAKLAR (REFERENCES)

[1] Murat Y.Ş. ve Şekerler A., "Trafik kaza verilerinin kümelenme analizi yöntemi ile modellenmesi". ***İmo Teknik Dergi***, 311, 4759-4777, (2009).

[2] Kashani A.T. and Mohaymany A.S., "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models". ***Safety Science***, 49, 1314–1320, (2011).

[3] Çodur M.Y., Tortum A. ve Çodur K.M., "Genelleştirilmiş lineer regresyon ile Erzurum kuzey çevre yolu kaza tahmin modeli". ***Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi***, 3 (1): 79-84, (2013).

[4] Tortum A., Çodur M.Y. and Kılınç B., "Modeling traffic accidents in Turkey using regression analysis". ***Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi***, 2 (3), 69-78, (2012).

[5] Atalay A., Tortum A. ve Gökdağ M., "Türkiye'de 1977-2006 yılları arasında meydana gelen aylık trafik kazalarının zamansal analizi". ***Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi***, 18(3), 221-229, (2012).

[6] Gupta M., Solanki V.K. and Singh V.K., "A novel framework to use association rule mining for classification of traffic accident severity". ***Ingeniería Solidaria***, 13(21), 37-44, (2017).

[7] Rovsek V., Batista M. and Bogunovic B., "Identifying the key risk factors of traffic accident injury severity on Slovenian roads using a non-parametrıc classification tree". ***Transport***, 32 (3), 1648-3480, (2017).

[8] Syahputri K., Sari R. M., Rizkya I., Tarigan U., Siregar I. and Farhan T. A., "Clustering the vulnerability of traffic accidents in Medan city with fuzzy c-means algorithm". ***Materials Science and Engineering***, 801(1), (2020).

[9] Murat Y. Ş., Kutluhan S. and Çakıcı Z., "Comparıson of fuzzy c-means and k-means clusterıng approaches ın spatıal analysıs of traffic accıdents." ***In Proceedıngs Book Of The Fourth Internatıonal Conference***, (2013).

[10] Atalay A. and Tortum A., "Türkiye'deki illerin 1997-2006 yılları arası trafik kazalarına göre kümeleme analizi". ***Pamukkale Üniversitesi Mühendislik Biilimleri Dergisi***, 16(3), (2010).

[11] Özden C. ve Acı Ç., "Makine öğrenmesi yöntemleri ile yaralanmalı trafik kazalarının analizi: Adana örneği". ***Pamukkale Üniversitesi Mühendislik Biilimleri Dergisi***, 24(2), 266-275, (2018).

[12] Selvi H.S., and Çağlar B., "Using cluster analysis methods for multivariate mapping of traffic accidents". ***Open Geosciences***, 10, 772-781, (2018).

[13] Mlouk A.A. and Agouti T., "DM-MCDA: A web-based platform for data mining and multiple criteria decision

analysis: A case study on road accident". ***SoftwareX***, 10, 100323, (2019).

[14] Das S., Dutta A. and Geedipally S.R., "Applying Bayesian data mining to measure the effect of vehicular defects on crash severity". ***Journal of Transportation Safety & Security***, DOI: 10.1080/19439962.2019.1658674, (2019).

[15] http://www.tuik.gov.tr/PreHaberBultenleri.do?id=27668, "Karayolu Trafik Kaza İstatistikleri". (2020).

[16] Alsagheer R.H.A., Alharan A.F.H. and Haboobi A.S.A., "Popular decision tree algorithms of data mining techniques: a review". ***International Journal of Computer Science and Mobile Computing***, 6 (6), 133-142, (2017).

[17] Kuşkapan, E. and Çodur, M.Y., "Examination of Aircraft Accidents That Occurred in the Last 20 Years in the World". ***Düzce University Journal of Science & Technology***, 9, 174-188, (2021).

[18] http://www.deeplearning.net/tutorial/mlp.html, "Multilayer Perceptron", (2019).

[19] https://medium.com/@isikhanelif/multi-layer-erceptron-mlp-nedir-4758285a7f15, "Multilayer Perceptron" (2019).

[20] https://pathmind.com/wiki/multilayer-perceptron, "Multilayer Perceptron" (2019).

[21] Mihăescu M.C., "Classification of learners using linear regression". ***Proceedings of the Federated Conference on Computer Science and Information Systems***, 717–721, (2011).

[22] https://tr.wikipedia.org/wiki/Regresyon_analizi, Lineer Regresyon. (2019).

[23] https://en.wikipedia.org/wiki/Instance-based_learning, Instance Based Learning. (2019).

[24] Kuşkapan, E., Çodur, M. Y., and Atalay, A., "Speed violation analysis of heavy vehicles on highways using spatial analysis and machine learning algorithms". ***Accident Analysis & Prevention***, *155*, (2021).

[25] Çavuşoğlu Ü. ve Kaçar S., "Anormal trafik tespiti için veri madenciliği algoritmalarının performans analizi". ***Academic Platform Journal of Engineering and Science***, 7 (2): 205-216, (2019).

[26] Coşkun C. ve Baykal A., "Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması". ***Akademik Bilişim'11 - XIII. Akademik Bilişim Konferansı Bildirileri***, İnönü Üniversitesi, Malatya, 51-58, (2011).

[27] https://www.statisticshowto.datasciencecentral.com/rmse, "Root Mean Square Error" (2019).

[28] Saxena A. and Jat M.K., "Analysing performance of SLEUTH model calibration using brute force and genetic algorithm–based methods". ***Geocarto International***, 35 (3): 256-279, (2020).

[29] Ardıl E., "Esnek Hesaplama Yaklaşımı ile Yazılım Hata Kestirimi". Yüksek Lisans Tezi, Trakya Üniversitesi, Fen Bilimleri Üniversitesi, Bilgisayar Mühendisliği Anabilim Dalı, (2009).