|  |
| --- |
| Makale / ResearchPaper |

**Trafik Kazalarının Şiddetinin Tahmini ve Sosyolojik Etkenleri Üzerine Analiz**

Kerem Okur, Nursena Taşköprü

1Necmettin Erbakan Üniversitesi,Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü. Konya/TÜRKİYE

[mkre2808@gmail.com](mailto:mkre2808@gmail.com), [nursenataskopru@gmail.com](mailto:nursenataskopru@gmail.com)

|  |
| --- |
| Öz: Trafik kazaları; sadece fiziksel ve teknik etmenlerle sınırlı kalmayıp, psikolojik veya sosyal faktörlerin etkisiyle de şekillenebilmektedir. Bu çalışma, trafik kazalarının şiddetini tahmin etmeye yönelik yapay zeka modelleri geliştirmeyi ve kazaların oluşumunda etkili olan değişkenleri analiz etmeyi amaçlamaktadır. Literatürdeki çeşitli makine öğrenmesi algoritmaları karşılaştırılarak en iyi performans gösteren modeller incelenmiş, bunun yanı sıra gün ,zaman , yaş, cinsiyet, eğitim seviyesi, gelir düzeyi, iş stresi gibi faktörlerin kazaların ciddiyetine olan etkileri sorgulanmıştır. Çalışma, trafik güvenliğine yönelik daha etkili ve bütüncül stratejiler geliştirilmesine katkı sunmayı hedeflemektedir.Anahtar Kelimeler: Trafik kazaları, Makine öğrenmesi, Yapay zeka, Trafik güvenliği, Kaza ciddiyeti |

**1. Giriş**

Trafik kazaları, bireylerin ve toplumların güvenlik algısını sarsan, ekonomik ve sosyal sonuçlar doğuran önemli olaylardır. Dünya Sağlık Örgütü’nün (WHO) raporuna göre, her yıl trafik kazaları nedeniyle yaklaşık 1,3 milyon insan hayatını kaybetmekte, 50 milyondan fazla kişi yaralanmaktadır [1]. Türkiye’de Emniyet Genel Müdürlüğü (EGM) ve Karayolları Genel Müdürlüğü (KGM) verilerine göre, 2022 yılında toplam 1.232.957 trafik kazası meydana gelmiştir. Bu kazaların 197.261'i ölümlü ve yaralanmalı, 1.035.696'sı ise maddi hasarlı olarak kaydedilmiştir. Kaza yerinde hayatını kaybeden kişi sayısı 2.282 olup, kazadan sonra otuz gün içinde ölenlerin sayısı 2.947 olarak belirlenmiştir. Böylece toplam ölüm sayısı 5.229’a ulaşırken, yaralanan kişi sayısı 288.696 olarak bildirilmiştir [2].

Trafik kazalarının yarattığı kayıplar yalnızca can kayıplarıyla sınırlı kalmamakta, aynı zamanda ekonomik ve sosyal etkileriyle de toplumları derinden etkilemektedir. Kazalar nedeniyle ortaya çıkan hastane ve rehabilitasyon masrafları, iş gücü kayıpları ve sigorta maliyetleri, ülkelerin ekonomik yükünü artırmaktadır. Dünya Sağlık Örgütü verilerine göre trafik kazaları nedeniyle oluşan mali kayıplar, çoğu ülkenin gayri safi yurt içi hasılasının %3’üne mal olmaktadır[3]. Bu durum, trafik güvenliği politikalarının geliştirilmesini ve trafik kazalarına neden olan faktörlerin bilimsel analizlerle ortaya konmasını zorunlu hale getirmektedir.

Trafik kazalarının incelenmesi, bu kazaların nedenlerini anlamak ve önleyici tedbirler geliştirmek açısından büyük önem taşımaktadır. Kazalara yol açan faktörler arasında sürücü hataları, yol ve çevre koşulları, hava durumu ve araç teknik arızaları gibi unsurlar bulunmaktadır. Ancak, yalnızca fiziksel ve teknik etkenleri incelemek yeterli değildir. Sosyolojik faktörler de kazaların oluşumunda önemli bir rol oynamaktadır. Sürücülerin risk alma eğilimleri, trafik kurallarına uyma alışkanlıkları, toplumsal normlar ve kültürel etkiler, bireylerin sürüş davranışlarını doğrudan etkilemektedir. Örneğin, Dünya Sağlık Örgütü verilerine göre trafik kazası sebepli 100 Bin insan başına düşen ölüm sayısı Japonya’da 2.1 iken Türkiye’de 6.7 ve Venezuella’da 39’dur[4].

Trafik kazalarının analiz edilmesi, yalnızca mevcut durumun anlaşılmasını değil, aynı zamanda trafik güvenliğini artırmaya yönelik stratejilerin belirlenmesini de sağlayacaktır. Büyük veri setleri üzerinde makine öğrenmesi gibi analiz yöntemleri kullanılarak, kazaların hangi faktörlerden etkilendiği, hangi bölgelerde daha sık meydana geldiği ve en riskli sürücü gruplarının kimler olduğu gibi kritik sorulara yanıt bulunabilir. Böylece, daha bilinçli trafik politikaları oluşturularak, kazaların azaltılması ve can kayıplarının en aza indirilmesi mümkün olacaktır.

Literatürde trafik kazaları veri setleri üzerinde yapılmış çalışmalar mevcuttur. Bunların bir kısmı aşağıda özetlenmiştir.

Ennaji Fatima Zohra ve ark.[5] yaptıkları çalışma da ABD Kazaları (2016-2023) veri seti kullanılmıştır. Araştırmacılar veri boyutunu küçültmek ve bir eyalete odaklanmak için veri setindeki en çok trafik kazasının gerçekleştiği eyalet olan Kaliforniya üzerinde çalıştılar. Kazanın ciddiyetini tahmin eden bir model oluşturmak için ANN (Artifical Neural Network), Random Forest ve Logistic Regression algoritmalarını kullanarak başarılarını kıyasladılar. En yüksek doğruluk oranına %90,7 ile Random Forest algoritması ile ulaştılar.

Koteswararao Kodepogu ve ark.[6] yaptıkları çalışmada Addis Ababa şehrinin tarfik kazalarını içeren bir veri setininin on özellikten oluşan küçültülmüş bir halini kullandılar. Kazanın ciddiyetini tahmin etmek için beş farklı makine öğrenmesi algoritmasını kıyasladılar. En yüksek doğruluk oranını %85.3 ile Naive Bayes ve Adaptive Boost Classifier algoritmalarından elde ettiler. Flask Framework ile modeli bir web uygulamasına entegre ettiler.

Mübariz Manzoor ve ark.[7] yaptıkları çalıma da ABD’de yaşanan 4.2 milyon tarfik kazasını içeren bir veri seti üzerinde çalıştılar. Çalışma trafik kazasının ciddiyetini tahmin etmek için Random Forest ve CNN tabanlı bir topluluk modeli geliştirerek modellerini altı diğer algoritma ile kıyasladılar. İlk aşama da veri setindeki 48 özelliğin tamamını eğitime dahil ettiler. Önerdikleri topluluk modeli %81.2 doğruluk oranı ile birinci oldu. İkinci aşamada trafik kazasını en çok etkilediğini tepit ettikleri 20 özellik modelleri eğittiler. Önerdikleri model %99.1 ile birinci oldu.

Bharti Sharma ve ark.[8] yaptıkları çalışma da 300 veriden oluşan bir veri kümesini analiz ederek MLP (Çok Katmanlı Algılayıcılar) ve SVM’i kıyasladılar. Kazaya en çok sebep olan iki faktörün hız ve alkol olduğunu tespit ederek bu iki özellik üzerine yoğunlaştılar. SVM test veri kümesinde %94 başarı gösterirken MLP %60 doğruluk oranı gösterdi.

Rabia Emhamed AlMamlook ve ark.[9] yaptıkları çalışma da 2010-2016 yılları arasında Michigan eyaletinde meydana gelen 270.000’den fazla trafik kazası kaydının içeren bir veri seti kullandılar. Kazanın ciddiyetini tahmin etmek için Naive Bayes, Random Forest, Logistic Regression ve AdaBosst algoritmalarını karşılaştırdılar. En iyi sonucu %75.5 doğruluk oranı ile Random Forest gösterdi.

Daniel Santos ve ark.[10] yaptıkları çalışma da 2016-2019 yılları arasında Portekiz’de meydana gelen yaklaşık 28.000 trafik kazası kaydını içeren bir veri kümesi üzerinde çalıştılar. Kazanın ciddiyetini tahmin etmek için Random Forest, Navie Bayes ve Logistic Regression algoritmalarını kıyasladılar. En iyi sonucu %73 doğruluk oranı ile Random Forest verdi.

[Sümbal Malik](https://ieeexplore.ieee.org/author/37087844420) ve ark.[11] yaptıkları çalışma da İngiltere Ulaştırma Bakanlığı tarafından yayınlanan ve yaklaşık 122.000 örnek içeren bir veri seti kullandılar. Naive Bayes, Random Forest, Logistic Regression, Decision Tree, Bagging ve AdaBoost algoritmalarını birbirleriye karşılaştırdılar. En iyi sonucu %98.80 doğruluk oranı ile Random Forest gösterdi.

[Şeref Elkheder](https://onlinelibrary.wiley.com/authored-by/Alkheder/Sharaf) ve ark.[12] yaptıkları çalışma da 2008-2013 yılları arasında Abu Dabi’de meydana gelen kazaları içeren 5976 örnekten ouşan bir veri seti kullandılar. Veri ön işleme sırasında 48 nitelikten oluşan veri setini 16 niteliğe düşürdüler. WEKA veri madenciliği yazılımı ile yapay sinir ağı oluşturarak test ettiler. Modelleri %74,6 test doğruluk oranı gösterdi.

Kahramanmaraş İli’ne ait trafik kazaları verileri üzerinde yapılan bir çalışmada, J48 algoritması kullanılarak trafik kazalarının ciddiyeti tahmin edilmiş ve bu algoritmanın en yüksek doğruluk oranına sahip olduğu belirlenmiştir [13]. Araştırmacılar, kaza verilerini analiz etmek için Naive Bayes, Destek Vektör Makineleri (SVM), k-En Yakın Komşu (k-NN), PART, AdaboostM1, J48 ve Rastgele Orman (Random Forest) gibi algoritmaları karşılaştırmış ve J48 algoritmasının %87,5 doğruluk oranı ile en iyi performansı gösterdiğini ortaya koymuşlardır.

Trafik kazalarına neden olan faktörleri analiz etmek için makine öğrenmesi yöntemine ek olarak aynı zamanda yapısal eşitlik modelleri (SEM) gibi istatistiksel yöntemler de yaygın olarak kullanılmaktadır. Ankara’da gerçekleştirilen bir çalışmada, 2010-2015 yılları arasında meydana gelen trafik kazalarının verileri yapısal eşitlik modeli ile analiz edilmiş ve kazaların oluşumunda çevresel ve bireysel faktörlerin önemli bir rol oynadığı tespit edilmiştir [14]. Araştırmacılar; kazanın yeri, hava şartları, yol durumu ve sürücü davranışlarının kaza ciddiyeti üzerinde doğrudan etkili olduğunu belirlemiş ve bu değişkenlerin modellenmesiyle tahmin doğruluğunun artırılabileceğini göstermiştir.

Literatürdeki çalışmalar incelendiğinde araştırmacıların hedeflerinin kazanın teknik sebeplerini analiz etmek ve ciddiyetini en iyi tahmin eden algoritmayı bulmak olduğu söylenebilir. Bu çalışma da kazanın ciddiyetini tahmin eden yapay zeka modelleri geliştirmenin yanında yaş, cinsiyet, eğitim seviyesi, gelir düzeyi, iş stresi, madde kullanımı gibi demografik, sosyoekonomik ve psikososyal faktörlerin trafik kazalarındaki etkisi araştırılacaktır.

Çalışmada cevabı araştırılacak sorulardan bazıları şunlardır:

Eğitim seviyesi ile trafik kazasının ciddiyeti arasında bir korelasyon var mı? Bu soru, akademik eğitim almış insanların daha az eğitim almış veya eğitimsiz insanlara göre daha güvenli araç kullanıp kullanmadığı, kurallara uyma oranında ciddi bir farklılık olıp olmadığını bulmayı hedefler.

Hangi mesleğe sahip insanların trafik kazası riski daha yüksektir? Bu soru, mesleklerin gelir düzeyinin, çalışma şartlarının, iş stresinin ve yoğunluğunun trafik kazaları üzerinde etkisini bulmayı amaçlar. İş gücü gerektiren mi yoksa zihinsel çaba gerektiren mi mesleklerin daha çok risk altında olduğu da bu soru kapsamında araştırılacaktır.

Matematik eğitimi almış olmak trafik kazası riskini azaltır mı? Bu soru, matematik eğitiminin sürücülerin analitik düşünme becerilerini geliştirerek trafikte daha bilinçli ve öngörülü hareket etmelerine katkı sağlayıp sağlamadığını araştırmayı amaçlar. Matematik, bireylerin problem çözme yeteneklerini, dikkat süresini, hızlı karar alma becerilerini ve mantıksal düşünme kapasitelerini artırabilir. Bu beceriler, trafikte karmaşık durumları analiz etme, hız-mesafe hesaplamaları yapma ve olası riskleri daha iyi tahmin etme açısından önemli olabilir.

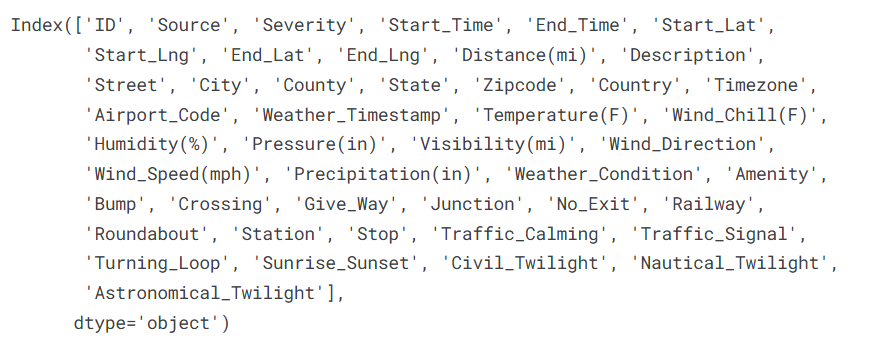
Hangi yaş aralığı ve cinsiyet daha fazla risk altındadır? Bu soru, sürücülerin yaş ve cinsiyet faktörlerinin trafik kazalarına etkisini anlamayı amaçlamaktadır. Genç sürücülerin sürüş deneyiminin daha az olduğu bilinirken yaşlı sürücülerin de yaşlanmadan kaynaklı dikkat ve reflekslerde azalma olduğu bilinmektedir. Bu soru kaza riski en az ve en fazla olan yaş aralıklarını bulmayı hedefler. Bununla beraber hangi cinsiyetin daha ciddi kazalara karıştığı araştırılacaktır.

Eyaletlerin gelir düzeylerinin, yaşam şartlarının, genel suç oranlarının, ceza kanunlarının trafik kazaları üzerindeki etkisi ne seviyededir? Bu soru, sosyoekonomik faktörlerin trafik kazaları üzerindeki etkisini incelemeyi amaçlamaktadır. Eyaletlerin gelir düzeyleri ve yaşam standartları, sürücülerin psikolojisinden sahip olduğu araçların kalitesine ve yol altyapısına kadar birçok faktörü etkileyebilir. Genel suç oranları da trafik kazalarıyla doğrudan veya dolaylı olarak ilişkili olabilir. Yüksek suç oranına sahip bölgelerde dikkatsizlik, agresif sürüş veya tehlikeli araç kullanımı daha sık görülebilir. Cezaları daha ağır olan eyaletlerde sürücüler kuralların dışına çıkmaktan kaçınabilir ve kaza riski azalabilir.

**2. Materyal ve Metot**

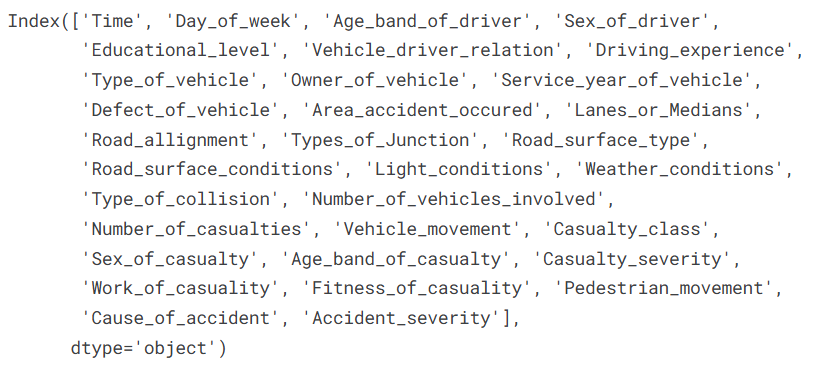
**2.1. Materyal**

Çalışmada iki farklı veri setinin kullanılması planlanmıştır. Birinci veri seti, 2016 yılından 2023 yılına kadar ABD’nin 49 eyaletinde meydana gelen yaklaşık 7.7 milyon trafik kazasının kaydını içerir. Veriler ABD ve eyalet ulaştırma bakanlıkları, kolluk kuvvetleri ve çeşitli kuruluşlar tarafından toplanmıştır. Veri setindeki sütun bilgileri aşağıdaki görselde verilmiştir.



**Görsel 1: ABD Kazaları Veri Seti Özellikleri**

İkinci veri seti Etiyopya’nın başkenti Addis Ababa şehrinde gerçekleşmiş 12.316 tarfik kazasının kaydını içerir. Veriler şehrin polis teşkilatları tarafından akademik çalışmalarda kullanılması amacıyla toplanmıştır. Veri setindeki sütun bilgileri aşağıdaki görselde verilmiştir.



**Görsel 2: Addis Ababa Kazaları Veri Seti Özellikleri**

**2.2. Metot**

Bu bölümde çalışmada kullanılan ön işleme teknikleri ve öğrenme algoritmalarından bahsedilecektir.

**2.2.1 ABD Kazaları Veri Seti**

Bilgisayarlarımız orijinal veri seti üzerinde çalışmaya imkan vermediği için 60.000 veriden oluşan bir örneklem üzerinde çalışılmıştır.

1. **Eksik Veriler:** Örneklem veri setinde eksik veriler sadece kazanın trafikteki etkilerinin bittiği enlem ve boylam kordinatlarını içeren iki özellikte toplanmıştır. Bu özelliklere ait verilerin %49.98’inin eksik veri olması ve ilgili özelliklerin çok önemli olmadığı düşünüldüğü için iki özellik silinmiştir.
2. **Aykırı Veriler:** Veri setimizde nümerik değer alan “Sıcaklık”, “Hissedilen Sıcaklık”, “Nem” , “Basınç”, “Yağış Miktarı”, “Rüzgar Hızı” , “Görüş Mesafesi” olmak üzere 7 tane özellik vardır. Bu özellikler için kutu grafiği çizdirildiğinde grafiğin dışında kalan binlerce veri olduğu görülmüştür. Sayının bu denli çok olması, aykırı verilerin yanlış ölçümler nedeniyle değil farklı iklimler yaşanan eyaletlerden gelen verilerin sayısının dengesiz olması sebebiyle olduğunu düşündürmüştür. Texsas, New Mexico gibi sıcak ve kurak eyaletlerin nüfusunun, Alaska gibi soğuk eyaletlerin nüfusundan çok daha fazla olması dolayısıyla kaza kayıtlarının çoğu yüksek nüfuslu eyaletlerden gelmiş ve iklim farkı, az nüfuslu eyaletlerin verilerinin ortalamadan uzak olmasına yol açmıştır. İklimin kaza sonuçarını sınıflandırmada önemli bir etken olabileceği düşünülerek kutu grafiğinin dışındaki veriler silinmemiş ve bu aşamada müdahale edilmemiş ancak sonraki aşamada kategorize edilmiştir.
3. **Sürekli Verilerin Ayrıklaştırılması:** Yukarıda bahsedilen 7 özellik ve onların dışında “Kazanın gerçekleştiği saat”, “Kazanın tarfikteki etkilerinin bittiği saat”, “Kazanın gerçekleştiği yerin enlem ve boylam kordinatları”, “Kaza bölgesinin uzunluğu” özellikleri sürekli değerler aldığı için Kmeans kümeleme algoritması ile ayrıklaştırılarak kategorilere bölünmüştür.
4. **Özellik Çıkarımı:** “Kazanın gerçekleştiği saat” sütunundaki saat verisinden “Kazanın tarfikteki etkilerinin bittiği saat” sütunundaki saat verisi çıkartılarak “Kaza Etki Süresi” ismini verdiğimiz yeni bir özellik elde edilmiştir.
5. **Yeni Özellikler Ekleme:** Her bir eyaletin nüfus ve gayrisafi yurt içi hasıla (GSYH) bilgileri araştırılarak iki yeni özellik olarak veri setine eklenmiştir. Ardından bu iki sütundaki verilere de c maddesindeki işlemler uygulanmıştır.
6. **Özellik Seçimi:** Tüm veriler kategorik hale getirildikten sonra öncelikle Cramers V Korelasyon matrisi çizdirilerek inputlar ile hedef değişkenin ilişkisi incelenmiştir. Matrise göre sonuçla en çok ilişkili olan ilk 5 , ilk 9 ve ilk 16 özellik seçilmiştir. Ayrıca chi-square testi ile 8 ve 16 özellik seçilmiştir. Tüm özelliklerin de dahil edildiği kümeyle beraber toplamda 6 tane küme oluşmuş ve her küme üzerinde Naive Bayes, Karar Ağacı, Rastgele Orman ve Yapay Sinir Ağları varsayılan parametrelerle çalıştırılmıştır. Tüm algoritmalar için en iyi sonuçlar korelasyon matrisindeki sonucu en çok etkileyen 16 özelliğin seçildiği kümede elde edilmiştir. Sonuçlar aşağıdaki tablolarda verilmiştir.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tüm Özellikler** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.5560 | 0.5684 | 0.5560 | 0.5205 |
| Karar Ağacı | 0.5929 | 0.5922 | 0.5929 | 0.5924 |
| Random Forest | 0.6736 | 0.6677 | 0.6736 | 0.6665 |
| YSA | 0.6828 | 0.6804 | 0.6928 | 0.6745 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **5 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.5817 | 0.5575 | 0.5817 | 0.5316 |
| Karar Ağacı | 0.5979 | 0.5945 | 0.5979 | 0.5919 |
| Random Forest | 0.5977 | 0.5943 | 0.5977 | 0.5917 |
| YSA | 0.5967 | 0.5954 | 0.5967 | 0.5920 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **9 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.5780 | 0.5469 | 0.5780 | 0.5377 |
| Karar Ağacı | 0.6042 | 0.6005 | 0.6042 | 0.5915 |
| Random Forest | 0.6044 | 0.6007 | 0.6044 | 0.5917 |
| YSA | 0.6047 | 0.6013 | 0.6047 | 0.5915 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **16 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.5948 | 0.5772 | 0.5948 | 0.5734 |
| Karar Ağacı | 0.6538 | 0.6484 | 0.6538 | 0.6489 |
| Random Forest | 0.6707 | 0.6653 | 0.6707 | 0.6646 |
| YSA | 0.6871 | 0.6866 | 0.6871 | 0.6770 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Chi2 8 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.5746 | 0.5478 | 0.5746 | 0.5348 |
| Karar Ağacı | 0.6009 | 0.5973 | 0.6009 | 0.5871 |
| Random Forest | 0.6014 | 0.5974 | 0.6014 | 0.5878 |
| YSA | 0.6004 | 0.5972 | 0.6004 | 0.5909 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Chi2 16 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.5919 | 0.5831 | 0.5919 | 0.5651 |
| Karar Ağacı | 0.6298 | 0.6289 | 0.6298 | 0.6254 |
| Random Forest | 0.6425 | 0.6397 | 0.6425 | 0.6361 |
| YSA | 0.6512 | 0.6487 | 0.6512 | 0.6436 |

1. **Model Eğitimi ve Değerlendirmesi:** Algoritmalar 16 özellikten oluşan kümede daha başarılı oldukları için çalışmanın kalanına bu küme üzerinden devam edilmiştir. Ayrıca Naive Bayes her durumda en kötü sonucu gösterdiği için elenmiş ve çalışmaya kalan üç algoritma ile devam edilmiştir. GridSearcCV fonksiyonu ile parametre optimizasyonu yapılmış ve K-Fold Cross Validation tekniği ile her algoritma farklı parametrelerle 5 Fold çalıştırılmıştır. Ardından algoritmalar üçlü olarak Friedman testi ile, aralarında ikili olarak Wilcoxon Signed-Rank testi ile kıyaslanmıştır.

**2.2.2 Addis Ababa Kazaları Veri Seti**

Veri seti 12.316 veriden oluşmaktadır. Veri setinin orijinal ve ön işlenmiş hali internette yayınlanmıştır. Projede ön işleme kısmını kendimiz yapmak istediğimiz için orijinal veri kümesi üzerinde çalışılmıştır.

1. **Eksik Veriler:** Veri setinde hemen hemen her özellikte az veya çok eksik veriler bulunmaktadır. Kendisine ait eksik veri oranı %30’un üzerinde olan ve sonuçla çok alakalı olmayan özellikler silinmiştir. Bunun dışında kalan eksik veriler nearest neighbor hot deck yöntemi ile doldurulmuştur.
2. **Aykırı Veriler:** Veri setinde aykırı veri bulunmamıştır.
3. **Sayısal Verilerin Ayrıklaştırılması:** “Saat”, “Kaza yapan araç sayısı”, “Yaralı Sayısı” gibi nümerik özellikler Kmeans kümeleme algoritması ile ayrıklaştırılarak kategorilere bölünmüştür.
4. **Özellik Seçimi:** Tüm veriler kategorik hale getirildikten sonra öncelikle Cramers V Korelasyon matrisi çizdirilerek inputlar ile hedef değişkenin ilişkisi incelenmiştir. Matrise göre sonuçla en çok ilişkili olan ilk 2 , ilk 6, ilk 8 ve ilk 12 özellik seçilmiştir. Ayrıca chi-square testi ile 6 ve 10 özellik seçilmiştir. Tüm özelliklerin de dahil edildiği kümeyle beraber toplamda 7 tane küme oluşmuş ve her küme üzerinde Naive Bayes, Karar Ağacı, Rastgele Orman ve Yapay Sinir Ağları varsayılan parametrelerle çalıştırılmıştır. Veri seti dengesiz olduğu için kriterimiz f1 skorudur. Genel olarak en iyi sonuçlar korelasyon matrisindeki sonucu en çok etkileyen 8 özelliğin seçildiği kümede elde edilmiştir. Sonuçlar aşağıdaki tablolarda verilmiştir.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tüm Özellikler** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.8158 | 0.7423 | 0.8158 | 0.7690 |
| Karar Ağacı | 0.7343 | 0.7463 | 0.7343 | 0.7401 |
| Random Forest | 0.8457 | 0.7864 | 0.8457 | 0.7763 |
| YSA | 0.7967 | 0.7561 | 0.7967 | 0.7738 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.8457 | 0.7153 | 0.8457 | 0.7750 |
| Karar Ağacı | 0.8457 | 0.7153 | 0.8457 | 0.7750 |
| Random Forest | 0.8457 | 0.7153 | 0.8457 | 0.7750 |
| YSA | 0.8457 | 0.7153 | 0.8457 | 0.7750 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **6 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.8457 | 0.7153 | 0.8457 | 0.7750 |
| Karar Ağacı | 0.8292 | 0.7821 | 0.8292 | 0.7968 |
| Random Forest | 0.8418 | 0.7934 | 0.8418 | 0.8003 |
| YSA | 0.8477 | 0.8583 | 0.8477 | 0.7797 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **8 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.8392 | 0.7146 | 0.8392 | 0.7719 |
| Karar Ağacı | 0.8275 | 0.8010 | 0.8275 | 0.8109 |
| Random Forest | 0.8500 | 0.8179 | 0.8500 | 0.8200 |
| YSA | 0.8496 | 0.8399 | 0.8496 | 0.7845 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **12 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.8409 | 0.7149 | 0.8409 | 0.7728 |
| Karar Ağacı | 0.7252 | 0.7515 | 0.7252 | 0.7374 |
| Random Forest | 0.8448 | 0.7805 | 0.8448 | 0.7816 |
| YSA | 0.8448 | 0.7708 | 0.8448 | 0.7777 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Chi2 6 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.8451 | 0.7181 | 0.8451 | 0.7754 |
| Karar Ağacı | 0.8207 | 0.7371 | 0.8207 | 0.7706 |
| Random Forest | 0.8357 | 0.7468 | 0.8357 | 0.7763 |
| YSA | 0.8448 | 0.7151 | 0.8404 | 0.7746 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Chi2 10 Özellik** | **Doğruluk** | **Keskinlik** | **Duyarlılık** | **F1 Score** |
| Naive Bayes | 0.8314 | 0.7404 | 0.8314 | 0.7724 |
| Karar Ağacı | 0.7483 | 0.7378 | 0.7483 | 0.7430 |
| Random Forest | 0.8175 | 0.7435 | 0.8175 | 0.7729 |
| YSA | 0.8431 | 0.7410 | 0.8431 | 0.7750 |

1. **Model Eğitimi ve Değerlendirmesi:** Algoritmalar 8 özellikten oluşan kümede daha başarılı oldukları için çalışmanın kalanına bu küme üzerinden devam edilmiştir. Ayrıca özellik seçiminde kullanılmayan KNN de bu aşamda kullanılmış ve 5 algoritma ile çalışılmıştır. GridSearcCV fonksiyonu ile parametre optimizasyonu yapılmış ve K-Fold Cross Validation tekniği ile her algoritma farklı parametrelerle 5 Fold çalıştırılmıştır. Ardından algoritmalar toplu olarak Friedman testi ile, aralarında ikili olarak Wilcoxon Signed-Rank testi ile kıyaslanmıştır.

**3. Bulgular ve Tartışma**

**3.1 ABD Kazaları**

ABD kazaları veri setinden alınan örneklem veri setinde, seçilen 16 özellikten oluşan alt kümede algoritmalar farklı parametrelerle 5 Fold çalıştırılmıştır. En iyi sonucu veren parametrelerle, en iyi fold test sonuçları genel olarak ve sınıf bazlı olarak aşağıda verilmiştir.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Doğruluk | Keskinlik | Duyarlılık | F1 Score |
| Karar Ağacı | 0.6678 | 0.6714 | 0.6678 | 0.6579 |
| Random Forest | 0.6929 | 0.6922 | 0.6929 | 0.6840 |
| YSA | 0.6857 | 0.6866 | 0.6857 | 0.5753 |

**Tablo 1:** Genel Sonuçlar

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıflar / Metrikler | Keskinlik | Duyarlılık | F1-Skor | Veri Sayısı |
| Hafif | 0.70 | 0.77 | 0.73 | 3750 |
| Orta | 0.70 | 0.39 | 0.50 | 3179 |
| Ağır | 0.65 | 0.74 | 0.69 | 3750 |
| Ölümlü | 0.64 | 0.74 | 0.69 | 3236 |

**Tablo 2:** Karar Ağacı Sınıf Bazlı Tablo

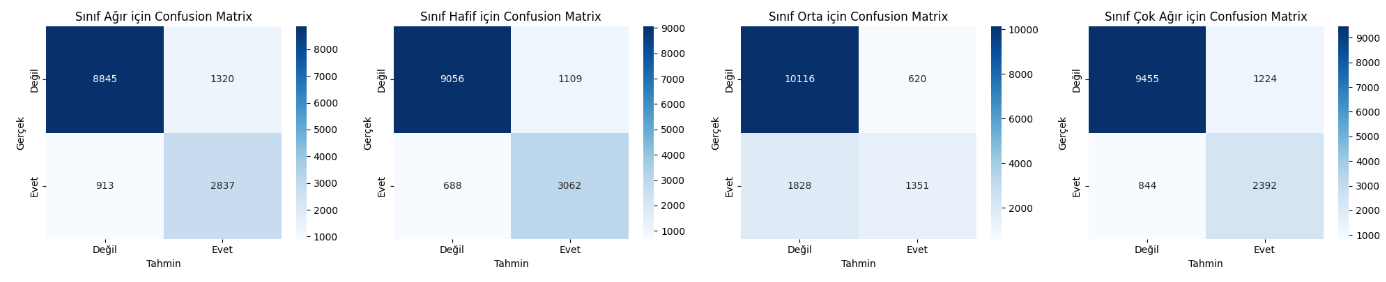
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıflar / Metrikler | Keskinlik | Duyarlılık | F1-Skor | Veri Sayısı |
| Hafif | 0.73 | 0.82 | 0.77 | 3750 |
| Orta | 0.69 | 0.42 | 0.52 | 3179 |
| Ağır | 0.68 | 0.76 | 0.72 | 3750 |
| Ölümlü | 0.66 | 0.74 | 0.70 | 3236 |

**Tablo 3:** Rastgele Orman Sınıf Bazlı Tablo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıflar / Metrikler | Keskinlik | Duyarlılık | F1-Skor | Veri Sayısı |
| Hafif | 0.70 | 0.84 | 0.77 | 3750 |
| Orta | 0.69 | 0.41 | 0.51 | 3179 |
| Ağır | 0.70 | 0.71 | 0.70 | 3750 |
| Ölümlü | 0.65 | 0.75 | 0.70 | 3236 |

**Tablo 4:** Yapay Sinir Ağları Sınıf Bazlı Tablo

Üç algoritmanın kıyaslandığı Friedman testinde %95 güven düzeyinde algoritmalar arasında anlmalı bir fark olduğu hipotezi doğru bulunmuştur. Sonrasında Wilcoxon Signed-Rank testi ile ikili kıyaslamalar yapıldığında %95 düzey güzeyinde algoritmalar arası anlamlı farklar olduğu hipotezi doğru bulunmuştur. Sonuçlara ve testlere göre Rastgele Orman algoritmasının diğerlerinden biraz daha iyi performans gösterdiği söylenebilir.



**Görsel 1:** Rastgele Orman Sınıf Bazlı Karmaşıklık Matrisleri

**3.2 Addis Ababa Kazaları**

Addis Ababa kazaları veri setinde, seçilen 16 özellikten oluşan alt kümede algoritmalar farklı parametrelerle 5 Fold çalıştırılmıştır. En iyi sonucu veren parametrelerle, en iyi fold test sonuçları genel olarak ve sınıf bazlı olarak aşağıda verilmiştir.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Doğruluk | Keskinlik | Duyarlılık | F1 Score |
| KNN | 0.8308 | 0.8012 | 0.8308 | 0.8113 |
| Naive Bayes | 0.8392 | 0.7146 | 0.8392 | 0.7719 |
| Karar Ağacı | 0.8282 | 0.8022 | 0.8282 | 0.8119 |
| Random Forest | 0.8470 | 0.8166 | 0.8470 | 0.8222 |
| YSA | 0.8496 | 0.8399 | 0.8496 | 0.7845 |

**Tablo 1:** Genel Sonuçlar

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıflar / Metrikler | Keskinlik | Duyarlılık | F1-Skor | Veri Sayısı |
| Hafif | 0.87 | 0.94 | 0.90 | 2604 |
| Orta/Ağır | 0.41 | 0.24 | 0.31 | 436 |
| Ölümlü | 0.44 | 0.21 | 0.28 | 39 |

**Tablo 2:** KNN Sınıf Bazlı Tablo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıflar / Metrikler | Keskinlik | Duyarlılık | F1-Skor | Veri Sayısı |
| Hafif | 0.84 | 0.99 | 0.91 | 2604 |
| Orta/Ağır | 0.00 | 0.00 | 0.00 | 436 |
| Ölümlü | 0.00 | 0.00 | 0.00 | 39 |

**Tablo 3:** Naive Bayes Sınıf Bazlı Tablo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıflar / Metrikler | Keskinlik | Duyarlılık | F1-Skor | Veri Sayısı |
| Hafif | 0.87 | 0.93 | 0.90 | 2604 |
| Orta/Ağır | 0.41 | 0.27 | 0.32 | 436 |
| Ölümlü | 0.38 | 0.21 | 0.27 | 39 |

**Tablo 4:** Karar Ağacı Sınıf Bazlı Tablo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıflar / Metrikler | Keskinlik | Duyarlılık | F1-Skor | Veri Sayısı |
| Hafif | 0.87 | 0.96 | 0.91 | 2604 |
| Orta/Ağır | 0.51 | 0.24 | 0.33 | 436 |
| Ölümlü | 0.47 | 0.18 | 0.26 | 39 |

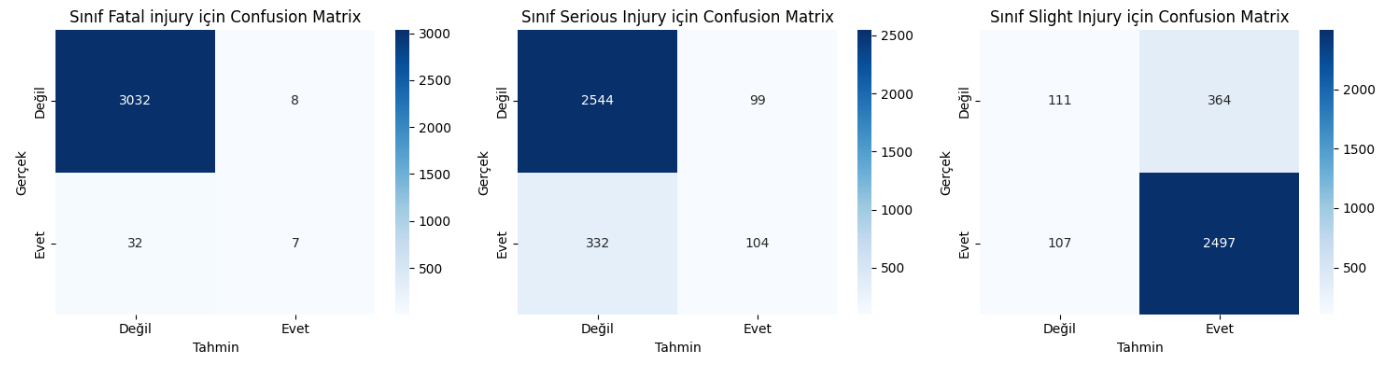
**Tablo 5:** Rastgele Orman Sınıf Bazlı Tablo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sınıflar / Metrikler | Keskinlik | Duyarlılık | F1-Skor | Veri Sayısı |
| Hafif | 0.85 | 1.00 | 0.92 | 2604 |
| Orta/Ağır | 0.86 | 0.03 | 0.05 | 436 |
| Ölümlü | 0.00 | 0.00 | 0.00 | 39 |

**Tablo 6:** Yapay Sinir Ağları Sınıf Bazlı Tablo

Sonuçlara bakıldığında yapay sinir ağları ve naive bayes belirgin şekilde daha kötü performans göstermiştir. Kalan 3 algoritma Friedman testinde kıyaslanmış ve %95 güven düzeyinde algoritmalar arasında anlamlı bir fark olduğu sonucu doğru bulunmuştur. Sonrasında algoritmaların arasında ikili olarak yapılan Wilcoxon Signed-Rank testinde Rastgele Orman ve KNN karşılaştırılmış, anlamlı bir fark bulunmamıştır. Sonra Karar Ağacı ve KNN karşılaştırılmış ve anlamlı bir fark bulunmamıştır. Rastgele Orman ve Karar Ağacı karşılaştırıldığında anlamlı fark bulunmuştur.

Sonuçlara ve testlere göre Rastgele Orman algoritmasının diğerlerinden biraz daha iyi performans gösterdiği söylenebilir. Aşağıda Rastgele Orman algoritmasına ait sınıf bazlı karmaşıklık matrisleri verilmiştir.



**Görsel 2:** Rastgele Orman Sınıf Bazlı Karmaşıklık Matrisleri

Rastgele Orman algoritması her iki veri setinde de diğer algoritmalara göre daha iyi performans göstersede, sonuçlar istenen düzeyde olmamıştır. Gelecekte tek bir algoritmaya bağlı kalmayıp topluluk modellerinin denenmesinin sonuçların başarısı artırabileceğini düşünüyoruz.

**4. Sonuç ve Öneriler**

Özellikler ve hedef değişken arasındaki ilişkiler incelendiğinde bazı bulgular elde edilmiştir.

Hava sıcaklığı ile kazanın şiddeti arasında bir ilişki bulunmuştur. Daha soğuk havalarda gerçekleşen kazaların sonuçları daha ciddi olmaktadır. Sıcaklığın 0 santigrat derecenin altında olduğu havalarda gerçekleşen kazaların %73 ağır veya ölümlü iken, sıcaklığın 26 santigrat derecenin üzerinde olduğu havalarda gerçekleşen kazaların %41’i ağır veya ölümlüdür.

Kar yağışlı havalarda gerçekleşen kazaların %76’sı ağır veya ölümlü iken, açık/parçalı bulutlu havalarda gerçekleşen kazaların %51’i ağır veya ölümlüdür.

Yaya geçidinin yakınlarında gerçekleşen kazaların yalnızca %18’i ağır veya ölümlü iken yaya geçidinin olmadığı ortamlarda gerçekleşen kazalarda bu oran %52’dir.

Kavşaklarda gerçekleşen kazaların %64’ü ağır veya ölümlü iken %36’sı daha hafiftir.

Trafik Işığının olduğu bölgelerde gerçekleşen kazaların %23’ü ağır veya ölümlü iken trafik ışığının olmadığı bölgelerde bu oran %55’tir.

Günün en ağır kazaları saat olarak 19.00 ve 00.00 arasında gerçekleşmektedir. Bu saat aralığında gerçekleşen kazaların %60’ı ağır iken, 05.34 ve 10.16 arasında gerçekleşen kazalar günün en az tehlikeli kazalarını oluştrumakta ve %39’ ağır olmaktadır.

Sürüş deneyimi 2-5 yıl arasında olan insanların yaptığı kazaların daha ağır olduğu bulunurken, deneyimi 2 yılın altında ve 5 yılın üstünde olan insanların yaptığı kazalar daha hafiftir.

Kazalarda ölüm oranı en yüksek aracın motosiklet, en düşük aracın bisiklet olduğu görülmüştür.

Araçların hangi durumda kaza yaptığı incelendiğinde en ağır sonuçların sollama sonucu oluşan kazalarda ortaya çıktığı görülmüştür.

Gün olarak pazartesi günü gerçekleşen kazaların daha hafif, cumartesi günü gerçekleşen kazaların daha ağır olduğu görülmüştür.

Görüş mesafesi ve kaza şiddeti arasında parabolik bir ilişki testpit edilmiştir. Düşük görüş mesafesinde kazalar ağır olurken, görüş mesafesi arttıkça kazanın ağır sonuçları azalmakta ancak görüş mesafesi 38 km’nin üzerine çıktığında kazanın sonuçları yine ciddileşmektedir.

Sürücünün cinsiyeti, mesleği, yaşı, eğitim düzeyi gibi faktörlerle kazanın ciddiyeti arasında kayda değer bir ilişki bulunmamıştır. Aynı şekilde şehrin nüfusu ve gelir düzeyi ile kazaların ciddiyeti arasında kayda değer bir ilişki bulunmamıştır.

**Teşekkür**

Projenin pratik uygulama çalışmalarının gerçekleştirilebilmesi için vermiş oldukları katkılar adına ………………………………. teşekkür ederiz. MF-11111 kodlu bu proje, … Üniversitesi Bilimsel Araştırma Projeleri Birimi tarafından desteklenmiştir. Bu çalışmanın ortaya çıkmasında verdiği destekten ötürü Bilimsel Araştırma Projeleri Birimine teşekkür ederiz.

**Kaynaklar**

[1] Dünya Sağlık Örgütü (WHO), “Global Status Report on Road Safety,” 2023. [Çevrimiçi]. Erişim: <https://www.who.int/publications/road-safety-report>

[2] Emniyet Genel Müdürlüğü ve Karayolları Genel Müdürlüğü, "Trafik Kazaları İstatistikleri 2022,"2023.[Çevrimiçi].Erişim:https://www.kgm.gov.tr/SiteCollectionDocuments/KGMdocuments/Trafik/TrafikKazalariOzeti2022.pdf

[3] World Health Organization, "Road traffic injuries," WHO, [Çevrimiçi]. Available: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>. [Erişim: 08-Mar-2025].

[4] World Health Organization, “Road traffic deaths,” Global Health Observatory (GHO) data, [Çevrimiçi]. Available: <https://apps.who.int/gho/data/node.main.A997>. [Erişim: 08-Mar-2025].

[5] E. F. Zohra, K. Maryam ve E. E. Hasna, "AccidentSeverityPredictionusing Machine Learning: A casestudy on the US AccidentsDataset," 2023 17th International Conference on Signal-Image Technology& Internet-BasedSystems (SITIS), IEEE, pp. 242-246, 2023.

[6] K. Kodepogu, V. B. Manjeti, and A. B. Siriki, "Road accident severity prediction using machine learning," Mechatronics and Intelligent Transportation Systems, vol. 2, pp. 211-226, 2023.

[7] M. Manzoor, M. Umer, S. Sadiq, A. Ishaq, S. Ullah, H. A. Madni, and C. Bisogni, "RFCNN: Traffic accident severity prediction based on decision level fusion of machine and deep learning model," IEEE Access, vol. 9, pp. 128359-128371, 2021.

[8] B. Sharma, V. K. Katiyar ve K. Kumar, "Trafficaccidentprediction model usingsupportvectormachineswithGaussiankernel," Proceedings of Fifth International Conference on Soft Computing for Problem Solving: SocProS 2015, cilt 2, ss. 1-10, SpringerSingapore, 2016.

[9] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh ve A. A. Frefer, "Comparison of machinelearningalgorithmsforpredictingtrafficaccidentseverity," 2019 IEEE Jordan International Joint Conference on ElectricalEngineeringand Information Technology (JEEIT), ss. 272-276, IEEE, 2019.

[10] D. Santos, J. Saias, P. Quaresma, and V. B. Nogueira, "Machine learning approaches to traffic accident analysis and hotspot prediction," Computers, vol. 10, no. 12, p. 157, 2021.

[11] S. Malik, H. El Sayed, M. A. Khan ve M. J. Khan, "Road accidentseverityprediction—a comparativeanalysis of machinelearningalgorithms," 2021 IEEE Global Conference on ArtificialIntelligenceand Internet of Things (GCAIoT), ss. 69-74, IEEE, 2021.

[12] S. Alkheder, M. Taamneh, and S. Taamneh, "Severity prediction of traffic accident using an artificial neural network," J. Forecasting, vol. 36, no. 1, pp. 100-108, 2017.

[13] A. Demir, "Trafik Kazalarının Makine Öğrenmesi Yöntemleriyle Analizi ve Tahmini: Kahramanmaraş İçin Örnek Bir Çalışma," Journal of Intelligent Transportation Systems, cilt 9, no. 4, ss. 211-226, 2023. [Çevrimiçi]. Erişim: <https://dergipark.org.tr/tr/download/article-file/2600712>.

[14] B. Çelik, "Trafik Kazalarını Etkileyen Faktörlerin Yapısal Eşitlik Modeli ile İncelenmesi," Journal of TransportationSafety, cilt 6, no. 2, ss. 272-290, 2021. [Çevrimiçi]. Erişim: https://dergipark.org.tr/tr/download/article-file/455302.