Bölüm 5

Makine Öğrenmesi Yöntemleri ve Derin Sinir Ağlarıyla Trafik Kazalarında Hayatta Kalma Sonuçlarının Tahmin Edilmesi

Mustafa Bayram Gücen[1]

Özet

Bu çalışma, trafik kazalarında hayatta kalma olasılıklarını tahmin etmeye yönelik farklı makine öğrenimi algoritmalarının ve derin sinir ağlarının uygulanması ve metodolojik değerlendirmesine odaklanmaktadır. Çalışmada, parametrik yöntemler (Lojistik Regresyon) ile parametrik olmayan yöntemler (k-En Yakın Komşular, Karar Ağaçları, Rastgele Orman) ve derin öğrenme temelli Derin Sinir Ağları geniş veri setleri üzerinde sistematik bir şekilde uygulanmıştır. Veri ön işleme, model eğitimi, hiperparametre optimizasyonu ve çapraz doğrulama gibi titiz metodolojik adımlar, tahmin performansını ölçmede kullanılan doğruluk, kesinlik, duyarlılık, F1 skoru ve ROC eğrisi altında kalan alan gibi metriklerle desteklenmiştir.

Analizler, her bir modelin veri setindeki yapısal ve istatistiksel özellikleri yakalama kapasitesinin farklılık gösterdiğini ortaya koymaktadır. Bu bağlamda, trafik kazalarında hayatta kalma olasılıklarının tahminine yönelik stratejilerin geliştirilmesi, modelin hesaplama verimliliği, yorumlanabilirliği ve genelleme yeteneği gibi faktörlerin dikkate alınmasıyla optimize edilebileceği vurgulanmaktadır. Çalışmanın bulguları, trafik güvenliği alanında veri odaklı yaklaşımların metodolojik temellerini güçlendirmesi ve ileri araştırmalar için sağlam bir referans oluşturması açısından önem taşımaktadır.

1. Giriş

Karayollarında meydana gelen trafik kazaları, dünya genelinde ciddi bir halk sağlığı sorunu oluşturmakta olup, her yıl milyonlarca insanın yaşamını kaybetmesine veya kalıcı yaralanmalara yol açmaktadır. Dünya Sağlık Örgütü (DSÖ) tarafından yayımlanan 2004 tarihli Karayollarında Trafik

1 Arş. Gör. Dr., Yıldız Teknik Üniversitesi Fen Fakültesi Matematik Bölümü, mgucen@yildiz.edu.tr, 0000-0002-9920-1747

https://doi.org/10.58830/ozgur.pub900.c3726

Yaralanmaları Önleme Raporu, bu sorunun küresel çapta ne denli önemli olduğunu vurgulamış ve yol güvenliğinin insan yaşamı üzerindeki etkilerini detaylı bir şekilde ele almıştır. Trafik kazaları, özellikle insanların günlük hayatlarında en çok etkileşimde bulunduğu ve en karmaşık sistemlerden biri olan yol güvenliği ile doğrudan ilişkilidir. Dünya Sağlık Örgütü (2004) raporunda bu konuda yapılan küresel çabaların yetersiz kaldığını ve yol güvenliğinin artırılması için uluslararası düzeyde daha ciddi adımlar atılması gerektiğini belirtmiştir. Bu rapor, eğer mevcut politikalar güçlendirilmezse, trafik kazalarının önlenemeyeceği ve kazaların gelecekte daha da artacağı konusunda uyarılarda bulunmuştur. Bu bağlamda, trafik kazalarının önlenmesi amacıyla küresel bir çağrı yapılmış, dünya çapında ortak çözümler üretilmesi gerektiği ifade edilmiştir.

2023 yılı itibarıyla yayımlanan Dünya Sağlık Örgütü Küresel Yol Güvenliği Durum Raporu, trafik kazası ölümlerinde bir miktar azalma olduğunu gösterse de, hâlâ bu sorunun önüne geçilebilmiş değildir. Dünya Sağlık Örgütü (2023) raporunda 2023 yılı itibarıyla 1.19 milyon trafik kazası ölümünün kaydedilmesi, yol güvenliği konusunda yapılan iyileştirmelerin bir miktar etkili olduğunu göstermektedir. Ancak, bu rakamların hala yüksek olduğu ve trafik kazalarının hâlâ en önemli ölüm nedenlerinden biri olduğu gerçeği, çözümün ne kadar acil olduğunu ortaya koymaktadır. Özellikle düşük ve orta gelirli ülkelerde, trafik kazası ölümleri gençler ve çocuklar arasında önde gelen ölüm sebeplerinden biri olmaya devam etmektedir. Yaya, bisiklet sürücüsü ve motosiklet sürücüsü ölümleri, bu ülkelerde orantısız bir şekilde fazla olup, bu ölümlerin önlenmesi için özel stratejiler geliştirilmesi gerekmektedir. DSÖ, 2030 yılına kadar trafik kazası ölümlerini ve yaralanmalarını en az yarıya indirmeyi hedefleyen küresel bir hedef belirlemiştir. Bu hedefin gerçekleştirilmesi, yalnızca yerel değil, uluslararası iş birliğini ve küresel düzeyde koordinasyonun güçlendirilmesini gerektirmektedir. Bu amaçlarla çeşitli çalışmalar yapılmış, makine öğrenmesi ve derin öğrenme teknikleri kullanılmıştır. Behboudi vd. (2024), trafik kazası analizi ve tahmini konusunda makine öğrenimi tekniklerinin son yıllarda elde edilen önemli ilerlemelerini kapsamlı bir şekilde inceleyen bir derlemedir. Son beş yılda yapılan 191 çalışmayı ele alarak, kaza riskini, sıklığını, şiddetini, süresini tahmin etme ve kaza verilerinin genel istatistiksel analizine odaklanmıştır. Bu çalışmanın, kaza analizi ve tahminine ilişkin çok çeşitli alanlardaki son teknolojiye dair kapsamlı bir inceleme sunduğunu belirtmektedir. Derleme, farklı veri kaynaklarının ve gelişmiş makine öğrenmesi tekniklerinin entegrasyonunun tahmin doğruluğunu artırmada ve trafik verilerinin karmaşıklığını yönetmede etkinliğini vurgulamaktadır. Mevcut durumu haritalayarak literatürdeki boşlukları belirleyerek, bu

çalışma, Dünya Sağlık Örgütü'nün (DSÖ) hedeflerine paralel olarak, 2030 yılına kadar trafikle ilgili ölümleri ve yaralanmaları önemli ölçüde azaltmak amacıyla gelecekteki araştırmalara rehberlik etmeyi hedeflemektedir.

Makine öğrenimi ve Derin Sinir Ağları gibi ileri düzey teknolojiler, özellikle karmaşık verileri işleme ve verilerdeki gizli ilişkileri ortaya çıkarma konusunda önemli bir yere sahiptir. Bu teknolojiler, geleneksel yöntemlerin gözden kaçırabileceği önemli desenleri keşfetme yeteneğine sahip olup, büyük veri setlerini daha verimli bir şekilde analiz etme potansiyeline sahiptir. Theofilatos (2019), makine öğreniminin büyük veri setleri üzerinde uygulandığında, bu verilerdeki gizli ilişkileri ve karmaşık desenleri ortaya çıkarma gücüne sahip olduğunu göstermiştir. Benzer şekilde, Ballamudi (2019) ve Santos (2021) gibi araştırmacılar, derin sinir ağlarının özellikle yol güvenliği ve trafik kazalarının çözümünde önemli bir araç olduğunu vurgulamış, bu tekniklerin verilerdeki karmaşık ilişkileri anlamada üstün bir performans sergilediğini belirtmişlerdir. Kumar vd. (2021), araba kazalarının önlenmesi veya tespiti konusunda birçok çalışma ve araştırma yapıldığını belirtmişlerdir. Çoğu araştırma, kaza oluşturabilecek nesnelerin tespit edilmesine odaklanmaktadır. Önerilen sistemde, kazanın meydana geldiği durumları tespit etmek amacıyla, birbirine yakın hareket eden araçlardan uygun bilgilerin toplanarak makine öğrenimi araçlarıyla işlenmesi sağlanmıştır. Makine öğrenimi teknikleri, normal davranışlardan anormal davranışları ayırt etmede başarı sağlamıştır. Sistemin ana hedefi, trafik davranışlarını incelemek ve mevcut trafikten farklı hareket eden araçları yol kazası olarak değerlendirmektir.

Trafik kazalarının, emniyet kemeri kullanımı, çarpışma hızı ve yaş gibi faktörlerle ilişkili olduğu ve bu etmenlerin kazaların sonuçlarını önemli ölçüde etkilediği gerçeği, bu ileri düzey teknolojilerin trafik kazalarının anlaşılması ve çözülmesinde ne denli önemli bir rol oynayabileceğini göstermektedir.

Bu çalışmanın amacı, trafik kazalarında hayatta kalma sonuçlarını etkileyen faktörleri tahmin etmek ve analiz etmek için makine öğrenimi ve Derin Sinir Ağları tekniklerini kullanmaktır. Derin Sinir Ağları, verilerdeki karmaşık yapıları tanıma ve öğrenme yeteneği sayesinde, trafik kazalarının ve bu kazaların neden olduğu hayatta kalma sonuçlarının daha doğru bir şekilde analiz edilmesini sağlar. Bu yöntemler, kazaların oluşmasındaki etkenleri daha net bir biçimde ortaya koyabilmekte ve bu etkenler arasındaki ilişkilerin daha derinlemesine anlaşılmasını mümkün kılmaktadır. Özellikle, emniyet kemeri kullanımı, çarpışma hızı, yaş gibi faktörlerin kazaların sonuçlarını nasıl etkilediği üzerine yapılan analizler, kazaların önlenmesine yönelik stratejilerin geliştirilmesinde önemli bir katkı sağlayacaktır. Makine öğrenimi

ve Derin Sinir Ağları kullanılarak yapılan bu tür analizler, yol güvenliğinin artırılması adına daha etkili, hedeflenmiş ve bilimsel temellere dayalı stratejilerin oluşturulmasına olanak tanıyacaktır. Bu araştırma, trafik kazası ölümlerinin ve yaralanmalarının azaltılması, yol güvenliğinin artırılması ve hayatta kalma oranlarının iyileştirilmesi için daha doğru ve etkili çözümler geliştirmeyi hedeflemektedir. Sonuç olarak, bu çalışma, trafik kazalarının nedenlerini anlamada ve bu nedenlerin etkilerini analiz etmede, ileri düzey makine öğrenimi tekniklerinin kullanılmasının, güvenli yol sistemleri oluşturulmasına ve halk sağlığı sorunlarının çözülmesine katkıda bulunacak önemli bir adım olduğunu ortaya koymaktadır.

2. Literatür İncelemesi

Makine öğrenimi ve derin öğrenme, trafik kazalarıyla ilgili yaralanma şiddetini tahmin etmek ve trafik güvenliğini iyileştirmek için önemli araçlar olarak kullanılmaktadır. Literatürde, bu tekniklerin çeşitli yönleriyle kazaların şiddetini tahmin etme ve güvenlik önlemlerini geliştirme potansiyeli farklı araştırmacılar tarafından bir çok çalışmada araştırılmıştır.

Obasi ve Benson (2023) trafik kazalarının şiddetini tahmin etmek için Naive Bayes, Rastgele Orman, Lojistik Regresyon ve Yapay Sinir Ağları gibi makine öğrenimi modellerini kullanarak bir analiz çerçevesi önermiştir. Çalışmada, Birleşik Krallık'tan alınan 2005-2014 yıllarına ait trafik kazası verileri ile farklı makine öğrenimi modelleri karşılaştırılmış ve Random Forest ile Lojistik Regresyon modellerinin daha iyi sonuçlar verdiği gözlemlenmiştir. Bu çalışma, trafik kazası yönetimi ve kontrol önlemlerinin uygulanmasında makine öğrenimi tekniklerinin etkinliğini vurgulamaktadır.

Nandurge ve Dharwadkar (2017), yol trafik kazalarının temel faktörlerini belirlemenin, kaza verisi analizinin ana hedeflerinden biri olduğunu belirtmişlerdir. Yol kazası verilerinin heterojen yapısı nedeniyle analizlerin zorlayıcı olabileceği ifade edilmiştir. Bu heterojenliği aşmak için verilerin bölümlendirilmesi gerektiği vurgulanmıştır. Önerilen yöntem, yol kazası verilerinin segmentasyonunda ana görev olarak k-means kümeleme yöntemini kullanmaktadır. Ayrıca, k-means kümeleme algoritması ile tanımlanan kümelerin ve tüm veri setinin oluşumuyla ilgili durumları keşfetmek amacıyla ilişki kuralı madenciliği uygulanmıştır. K-means kümeleme ve ilişki kuralı madenciliğinin birleşik sonucu, önemli bilgilere ulaşılmasını sağlamaktadır. Patil vd. (2020) ise, Bengaluru'daki yol kazalarını analiz etmek için makine öğrenimi ve K-means algoritması kullanarak bir çalışmada kazaların yoğunlaştığı bölgeleri ve bu kazaların kök nedenlerini incelediler. Çalışma, trafik kazalarının gelişmiş ve gelişmekte olan ülkeler için büyük bir

tehdit olduğunu ve kazaların, hava koşulları, dik virajlar ve insan hataları gibi çeşitli faktörlerden kaynaklandığını vurgulamaktadır. Ayrıca, kazaların yol açtığı yaralanmaların bazen fark edilmeyen ancak sağlık üzerinde uzun vadeli etkiler yaratabilecek nitelikte olduğunu belirtmektedir. Zhengjing vd. (2021) ise, derin öğrenme tabanlı bir analitik çerçeve önererek, trafik kazası yaralanma şiddetini tahmin etmede kullanılan katkı sağlayan faktörlerin önemini analiz etmiştir. Çalışmada, CatBoost adı verilen makine öğrenimi yaklaşımı ile katkı sağlayan faktörlerin önemi ve bağımlılığı analiz edilmiş, K-means kümeleme kullanılarak veriler farklı sınıflara ayrılmıştır. Yüksek korelasyonlu faktörler kullanılarak derin öğrenme modeli ile yaralanma şiddeti tahmin edilmiştir. Sonuçlar, önerilen çerçevenin, trafik kazası yaralanma şiddetini tahmin etmede ve yayaların güvenliğini artırmada etkili olduğunu göstermektedir. Ahmed vd. (2021), trafik kazalarının ölüm ve yaralanmalara yol açmasının yanı sıra doğrudan ve dolaylı kayıplara neden olan küresel bir sorun olduğunu belirtmişlerdir. Çalışmada, büyük trafik verileri ve yapay zeka algoritmaları kullanılarak kaza şiddetinin tahmin edilmesi veya kazaların riskinin azaltılması için umut verici bir çözüm geliştirilmesi amaçlanmıştır. Mevcut çalışmalarda genellikle yol geometrisi, çevre ve hava koşulları gibi faktörler üzerinde durulmuş olsa da, alkol, ilaç kullanımı, yaş ve cinsiyet gibi insan faktörleri sıklıkla göz ardı edilmiştir. Çalışmada, kaza şiddetinin tahminine katkı sağlayan çeşitli faktörler incelenmiş ve tekli ve çoklu makine öğrenimi yöntemlerinin performansları, doğruluk, kesinlik, duyarlılık, F1 skoru ve ROC (receiver operator characteristic) eğrileri(AUC-ROC eğrisi altında kalan alan) gibi metriklerle karşılaştırılmıştır. Araştırma, kaza şiddetini sınıflandırma problemi olarak ele almış ve sonuçlar, Rastgele Orman(Random Forest) algoritmasının diğer yöntemleri geride bıraktığını ve her iki sınıflandırma türünde de en iyi sonuçları verdiğini göstermektedir. Araştırmada, Lojistik Regresyon ve kNN gibi tekli yöntemlerin benzer performanslar gösterdiği ancak ansambl yöntemlerinin, özellikle Rastgele Orman, XGBoost ve AdaBoost 'un daha yüksek doğrulukla kaza şiddetini tahmin edebildiği tespit edilmiştir.

Derin sinir ağları, görüntü sınıflandırma görevlerinde gösterdiği yüksek başarı sayesinde tıbbi görüntüleme, otonom sürüş, güvenlik sistemleri ve endüstriyel kalite kontrol gibi alanlarda yaygın olarak kullanılmaktadır. Röntgen ve MR gibi tıbbi görüntülerde hastalık belirtilerini tespit etmenin yanı sıra, otonom araçlarda trafik işaretlerini ve yayaları tanımlayarak güvenli sürüşe katkı sağlayan derin öğrenme modelleri, geniş veri kümeleri üzerinde gerçekleştirilen derin analizlerle model performansını optimize etmekte ve hata oranlarını azaltmaktadır. Ayrıca, açıklanabilir yapay zeka yöntemleri ile karar süreçlerinin şeffaflaştırılması, kazalara yol açabilecek

nedenlerin belirlenmesi ve risk analizi gibi kritik görevlerde etkin çözümler sunmaktadır. Derin sinir ağları aynı zamanda tahminlemede ve kazalara yol açabilecek nedenlerin incelenmesinde kullanılmaktadır. Singh vd. (2020), yol kazalarını tahmin etmek için derin sinir ağları modelinin kullanılmasını önermektedir. Derin sinir ağları, bir veya daha fazla gizli katman ve çok sayıda düğüm içerir. Çalışmada, alınan kaza verileri ve çok sayıda kaza kaydı bulunmaktadır ve elde edilen sonuçlar, Derin sinir ağlarının yol kazalarının tahmininde geliştirilmiş bir performans gösterdiğini ortaya koymaktadır. Lin vd. (2021), yüksek kaza riski tahmin modeli geliştirmek için derin öğrenme yöntemlerini kullanarak trafik kazası verilerini analiz etmiş ve iyileştirme için öncelikli kavşakları belirlemiştir. Trafik kazası verilerinin derlendiği bir veri tabanı oluşturulmuş ve bu veriler üzerinden kazaların olabileceği yüksek riskli kavşakları tahmin edebilmek amacıyla çeşitli makine öğrenimi yöntemleri kullanılarak bir kavşak kaza riski tahmin modeli geliştirilmiştir. Vicent vd. (2025), dünya genelinde her yıl binlerce kişinin trafik kazalarında hayatını kaybettiğine dikkat çekmektedir. Trafik kazalarında tıbbi yardıma ihtiyaç duyulup duyulmadığının tahmin edilmesi, bu konuda önemli bir sorundur. Bu amaçla, yazarlar, tıbbi yardıma ihtiyaç duyulup duyulmadığını ayırt etmek için geliştirilmiş yeni bir derin öğrenme modeli sunmaktadır.

Vasavi (2018), yol kazaları verilerinde gizli desenleri ortaya çıkarmak için makine öğrenimi tekniklerinin kullanımını incelemiştir. Çalışmasında, K-medoids ve beklenti-maksimizasyon algoritmaları ile kümeler oluşturulmuş ve a priori algoritması kullanılarak kazalar arasındaki ortak özellikler keşfedilmiştir. Sonuçlar, makine öğrenimi tekniklerinin yol kazalarındaki gizli örüntüleri başarılı bir şekilde çıkarabildiğini göstermiştir. Kodepogu vd. (2023), yol güvenliği yönetimi kapsamında, trafik kazalarının şiddetini tahmin eden modellerin geliştirilmesinin önemine odaklanmaktadır. Çalışma, kaza şiddetini etkileyen çok yönlü faktörleri inceleyerek, hız ve araç boyutu gibi taşıt özellikleri ile yol tasarımı ve trafik hacmi gibi yol karakteristiklerini ele almıştır. Ayrıca, sürücü demografisi, deneyimi ve hava koşulları gibi dış etkenler de analiz edilmiştir. Veri analizi ve makine öğrenimi tekniklerindeki son gelişmeler, trafik kazası şiddetinin tahmin edilmesi konusunda önemli bir rol oynamaktadır. Geleneksel istatistiksel yöntemlerin aksine, makine öğrenimi modelleri karmaşık değişken etkileşimlerini daha iyi yakalayarak daha yüksek tahmin doğruluğu sunmaktadır. Ancak, bu modellerin etkinliği, kullanılan verilerin kalitesine ve kapsamına bağlıdır. Çalışma, standart veri toplama ve raporlama yöntemlerinin önemini vurgulamaktadır. Mevcut literatürün titizlikle analiz edilmesi sonucunda, çalışmada yol kazası tahmin modellerinin temel kavramları, teorik çerçeveleri ve veri kaynakları ele alınmıştır. Bulgular, yol kazalarının sıklığını ve ciddiyetini azaltmayı

amaçlayan gelişmiş modelleme tekniklerinin sürekli olarak geliştirilmesi ve iyileştirilmesi gerektiğini göstermektedir. Bu tür çalışmalar, trafik kontrolü ve kamu güvenliği önlemlerinin iyileştirilmesine önemli katkılar sunmaktadır. Bai vd. (2025), trafik kazalarının bireylere, topluma ve altyapıya önemli maliyetler getirdiğini vurgulayarak, kazaların maliyetlerini tahmin etmek için geliştirilmiş makine öğrenimi tekniklerini entegre eden kapsamlı bir çerçeve önerdiler. Çalışmada model, tarihsel kaza verileri, hava koşulları, yol altyapısı ve demografik faktörler gibi bağlamsal bilgilerle eğitilmiştir. Çalışma, bu yaklaşımın geleneksel yöntemlere göre üstün tahmin doğruluğu sağladığını ve farklı faktörlerin maliyet tahminine etkilerini değerlendiren duyarlılık analizleri sunduğunu belirtmektedir. Önerilen çerçeve, trafik kazası maliyetlerini tahmin etmek için güçlü ve verimli bir çözüm sunmakta ve toplum ve altyapı üzerindeki olumsuz etkileri hafifletmeye yönelik proaktif önlemleri kolaylaştırmaktadır.

Bu çalışmalar, makine öğrenimi ve derin öğrenme tekniklerinin trafik kazalarının tahmin edilmesi, kazaların şiddetinin analiz edilmesi ve trafik güvenliğinin iyileştirilmesi konusunda etkili bir araç olduğunu göstermektedir. Ayrıca, bu tekniklerin trafik kazası verilerindeki gizli desenleri ortaya çıkarma, önemli faktörleri belirleme ve yol güvenliği önlemlerini geliştirme açısından büyük potansiyel sunduğu vurgulanmaktadır. Bu tür teknolojilerin, daha güvenli bir trafik ortamı yaratmak ve kazaların önlenmesine yönelik stratejiler geliştirmek için önemli bir rol oynayabileceği belirtilmiştir. Ayrıca, bu tekniklerin çeşitli trafik kazası veri setleriyle uyumlu olarak uygulanabilir olması, gelecekteki trafik kazası yönetimi ve kontrol sistemlerinin daha verimli hale gelmesine katkı sağlayacaktır.

3. Veri Kümesi ve Yöntem

3.1. Veri Kümesi

Bu çalışmada kullanılan veri kümesi, çeşitli veri kaynaklarından alınan trafik kazalarından alınan veriler baz alınarak elde edilmiş trafik kazalarına dair detaylı kayıtları içermektedir. Veri kümesi; yaş, cinsiyet, çarpışma hızı, kask kullanımı, emniyet kemeri kullanımı ve hayatta kalma durumu gibi faktörleri içeren demografik, davranışsal ve durumsal özellikleri kapsamaktadır. Bu özellikler, trafik kazalarında hayatta kalma olasılıklarını etkileyen çeşitli faktörleri anlamaya yönelik önemli bilgiler sunmaktadır. Veri kümesi, yol kazalarının daha iyi analiz edilmesi için toplanmış olan değişkenleri içermektedir ve bu faktörlerin hayatta kalma üzerinde nasıl bir etkisi olduğuna dair önemli ipuçları sağlayabileceği düşünülmektedir. Veriler, veri analizi öncesi makine öğrenimi algoritmalarına uygun olacak şekilde

ön işleme tabi tutulacaktır. Veri kümesi, https://www.kaggle.com/datasets/himelsarder/road-accident-survival-dataset web adresinden elde edilmiştir.

Veri kümesinde kullanılan değişkenlere ait bilgiler Tablo 1'de verilmiştir.

Tablo 1. Değişkenlere Ait Bilgiler

Değişken Adı	**Değişken Değerleri**
Yaş (Age)	18-69
Cinsiyet (Gender)	Erkek (Male), Kadın (Female)
Çarpışma Hızı (Speed_of_Impact)	20-119 km/s (20-119 km/h)
Kask Kullanıldı mı (Helmet_Used)	Evet (Yes), Hayır (No)
Emniyet Kemeri Kullanıldı mı? (Seatbelt_Used)	Evet (Yes), Hayır (No)
Hayatta Kalma Durumu (Survived)	Evet (1), Hayır (0)

3.2. Veri Önişleme Süreci

Veri önişleme, makine öğrenmesi projelerinde modelin başarısını doğrudan etkileyen zorunlu bir adımdır. Gerçek dünya verileri genellikle eksik, hatalı veya düzensiz olduğundan, ham verinin doğrudan kullanılması modelin performansını düşürebilir. Bu nedenle, eksik verilerin doldurulması, kategorik değişkenlerin sayısallaştırılması ve veri ölçeklendirme gibi işlemler uygulanarak verinin analize uygun hale getirilmesi sağlanır.

Eksik verilerin yönetimi, modelin güvenilirliğini artırmak için yaygın olarak ortalama, medyan veya mod gibi istatistiksel yöntemlerle gerçekleştirilir. Kategorik değişkenler, One-Hot Encoding gibi tekniklerle sayısal formata dönüştürülerek modelin bu verileri işlemesi sağlanır. Ayrıca, farklı ölçeklerdeki değişkenlerin dengelenmesi için Min-Max Scaling veya Standardizasyon yöntemleri kullanılarak verinin uygun bir dağılıma getirilmesi sağlanır. Bu adımlar, modelin daha doğru ve genelleştirilebilir sonuçlar üretmesine yardımcı olur. Aşağıda, veri önişleme sürecinin açıklamasını yaparak hangi yöntemlerin kullanıldığını detaylı şekilde ele alacağız.

3.2.1. Eksik Verilerin Doldurulması

Eksik verilerin doldurulması, modelin doğruluğunu artırmak ve analiz sürecini kesintisiz hale getirmek için yapılır. Eksik veriler şu şekilde doldurulmuştur:

Cinsiyet (Gender) Sütunu: Bu sütundaki eksik değerler, sütunun mod değeriyle doldurulmuştur. Mod, bir veri kümesindeki en sık tekrarlanan

değeri ifade eder. Yani, verilerin içinde en çok görülen kategori alınır ve eksik değerler bu kategoriyle doldurulur.

Çarpma Hızı (Speed_of_Impact) Sütunu: Bu sütundaki eksik veriler, sütunun ortalaması ile doldurulmuştur. Ortalama, verilerin toplamının, veri sayısına bölünmesiyle hesaplanır. Bu, eksik veriler için yaygın bir doldurma yöntemidir.

Kask Kullanımı (Helmet_Used) ve Emniyet Kemeri Kullanımı (Seatbelt_Used) Sütunları: Bu kategorik sütunlardaki eksik değerler de yine mod değeriyle doldurulmuştur. En sık görülen kategori seçilerek eksik değerler tamamlanır.

3.2.2. Kategorik Verilerin Sayısal Verilere Dönüştürülmesi

Makine öğrenmesi algoritmaları, kategorik verileri doğrudan işleyemez. Bu nedenle, One-Hot Encoding kullanılarak kategorik değişkenler sayısal verilere dönüştürülmüştür.

One-Hot Encoding yöntemiyle, her kategoriye karşılık gelen yeni bir sütun oluşturulur ve bu sütun, kategoriye ait her veri için 1 veya 0 değerleri alır. Örneğin, Cinsiyet sütununda Erkek ve Kadın gibi iki kategori varsa, bu iki kategori için ayrı sütunlar yaratılır. Verinin Erkek olması durumunda, Erkek sütunu 1, Kadın sütunu ise 0 olacaktır. Diğer kategori için de aynı işlem yapılır.

3.2.3. Veri Kümesinin Eğitim ve Test Olarak Ayrılması

Veri kümesindeki eğitim ve test kümeleri, modelin doğruluğunu değerlendirmek için ayrılır. Bu işlemin amacı, modelin eğitim verisiyle aşırı öğrenmesini (overfitting) engellemektir. Bu adımda stratifikasyon uygulanır:

Stratifikasyon (Stratification) veri kümesinin sınıflara göre oranlarını koruyarak eğitim ve test kümesine ayırma işlemidir. Yani, her iki küme de, bağımlı değişkenin (y) sınıflarına orantılı olarak ayrılır. Bu işlem, özellikle dengesiz sınıflara sahip veri setlerinde modelin daha dengeli bir şekilde eğitilmesini sağlar.

Stratifikasyon işlemi, her sınıfın veri kümesindeki oranını korur. Bu, eğitim ve test kümesinde de aynı sınıf oranlarının korunmasını sağlar.

3.2.4. Sayısal Verilerin Ölçeklendirilmesi

Sayısal veriler, farklı aralıklara ve büyüklüklere sahip olabilir. Bu, modelin öğrenme sürecini zorlaştırabilir. Bu nedenle, MinMaxScaler ve StandardScaler gibi ölçeklendirme yöntemleri kullanılmıştır. Her iki yöntem de verilerin belirli bir aralığa çekilmesini sağlar, ancak kullandıkları formüller farklıdır.

MinMaxScaler yöntemi, verileri belirli bir aralıkta (genellikle [0, 1]) yeniden ölçeklendirir. Verinin her bir değeri, minimum değeri çıkarılarak ve maksimum değere bölünerek normalize edilir. Bu işlem, tüm verilerin aynı aralığa çekilmesini sağlar.

MinMaxScaler, verileri [0,1] aralığına ölçeklendirmek için şu formülü kullanır:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Burada: X' ölçeklendirilmiş değer, X orijinal değer, X_{min} veri kümesindeki minimum değer ve X_{max} veri kümesindeki maksimum değer olarak alınmaktadır.

StandardScaler yöntemi, verilerin ortalama ve standart sapma kullanılarak yeniden ölçeklendirilmesini sağlar.

$$X' = \frac{X - \mu}{\sigma} \tag{2}$$

Burada: X' standardize edilmiş değer, X orijinal değer, μ veri kümesinin ortalaması ve σ veri kümesinin standart sapması olarak alınmaktadır.

Bu yöntemle her verinin, veri kümesinin ortalamasından çıkarılması ve standart sapmaya bölünmesiyle normalize edilmesi sağlanır. Bu, verilerin sıfır ortalamaya ve birim standart sapmaya sahip olmasını sağlar.

3.2.5. Korelasyon Analizi

Korelasyon, bağımsız değişkenlerle bağımlı değişken arasındaki ilişkiyi gözlemlememizi sağlar. Tablo 2 bu amaçla yapılan bir korelasyon analizinin sonuçlarını sunmaktadır. Bu analizde, parametrik korelasyon yöntemlerinden olan Pearson korelasyonun bir türü Point-Biserial korelasyon katsayısı (r_{pb}) ve non-parametrik korelasyon yöntemlerinden Spearman Korelasyon

Katsayısı(r_s , p_value) kullanılmıştır. Point-Biserial korelasyon katsayısı (r_{pb}), bir sürekli (nicel) değişken ile bir ikili (dichotomous) (örneğin, 0 ve 1 gibi) kategorik değişken arasındaki ilişkiyi ölçmek için kullanılır. Bu katsayı Pearson korelasyon katsayısının özel bir türüdür, yani hesaplama yöntemi Pearson korelasyon katsayısına dayanır. Formülü şu şekildedir:

$$r_{pb} = \frac{M_1 - M_0}{s} \times \sqrt{\frac{p_1 p_0}{n^2}} \tag{3}$$

Burada: $M_1, M_0,$ bağımsız değişkenin sürekli değişken için kategoriye göre ortalama değerleridir, s, sürekli değişkenin toplam standart sapmasıdır, $p_1, p_0,$ bağımsız değişkenin iki kategorisinin oranlarını ifade eder.

Nonparametrik korelasyon yöntemleri, verilerin normal dağılım göstermediği veya doğrusal olmayan ilişkilerin bulunduğu durumlarda, değişkenler arasındaki ilişkiyi değerlendirmek için tercih edilen güçlü araçlardır ve verilerin sıralanmasına dayalı olarak monotonik ilişkilerin kuvvetini ölçmede kullanılır Kendall (1948). Gibbons ve Chakraborti (2014) nonparametrik istatistiksel çıkarım yöntemlerini detaylı bir şekilde ele alarak, bu tekniklerin modern uygulamalardaki geçerliliğini ve pratik faydalarını ortaya koymaktadır. Bu yöntemler, özellikle parametrik testlerin varsayımlarının ihlal edildiği durumlarda, güvenilir sonuçlar üretebilmekte ve araştırmacılara verinin yapısal özelliklerini daha doğru bir şekilde yorumlama imkanı sağlamaktadır. Spearman Korelasyon Katsayısı (r_s), değişkenler arasındaki sıra korelasyonunu ölçen parametrik olmayan bir istatistiksel yöntemdir. Pearson korelasyon katsayısından farklı olarak, Spearman katsayısı değişkenlerin sıralarına dayalı olduğu için normal dağılım varsayımı gerektirmez ve değişkenler arasındaki doğrusal olmayan monoton ilişkileri de yakalayabilir Spearman (1904). Spearman Korelasyon Katsayısı(r_s) kullanılarak değişkenler arasındaki ilişki değerlendirilmiştir.

$$r_s = \frac{1 - 6\sum d_i^2}{n\left(n^2 - 1\right)} \tag{3}$$

p-değeri (p-value), hipotez testinde korelasyon katsayısının istatistiksel olarak anlamlı olup olmadığını belirlemek için kullanılır. Düşük bir p-değeri ($p<0,05$), iki değişken arasında istatistiksel olarak anlamlı bir ilişki olduğunu gösterirken, yüksek bir p-değeri ($p>0,05$) bu ilişkinin tesadüfi olabileceğini işaret eder. Tablo 2, cinsiyet, yaş, emniyet kemeri kullanımı, çarpışma

hızı ve kask kullanımı gibi özelliklerin bağımlı değişkenle olan ilişkisini göstermektedir.

Şekil 1'de verilen korelasyon matrisi, veri setindeki değişkenler arasındaki doğrusal ilişkileri gösteren simetrik bir matristir. Bu matris, değişken çiftleri arasındaki korelasyon katsayılarını hesaplayarak, güçlü veya zayıf ilişkilerin belirlenmesine olanak tanır. Pozitif veya negatif korelasyon değerleri, değişkenlerin birbirleriyle nasıl etkileşime girdiğini anlamada önemli bir istatistiksel araçtır.

Korelasyon katsayıları genellikle düşük seviyelerde olup, çoğu ilişkinin Spearman Korelasyon Katsayısı açısından istatistiksel olarak anlamlı olmadığı gözlemlenmektedir. Özellikle p-değerleri 0,05'ten büyük olan özelliklerde, bu değişkenlerin bağımlı değişkenle güçlü bir ilişki kurmadığı söylenebilir. Bu sonuç, analiz edilen değişkenler arasında zayıf ilişkilerin varlığını ve çoğunun etkisinin sonuçlar üzerinde çok fazla olmadığını ortaya koymaktadır.

Tablo 2. Bağımlı Değişken ile Bağımsız Değişkenler Arasındaki Korelasyon Değerleri

Değişken Adı	r_{pb}	r_s	**p-değeri (p-value)**
Cinsiyet (Gender)	0,121845	0,121845	0,085657
Yaş (Age)	0,110090	0,103877	0,120695
Emniyet Kemeri Kullanımı (Seatbelt_Used)	0,059263	0,059263	0,404518
Çarpışma Hızı (Speed_of_Impact)	0,042198	0,046602	0,552989
Kask Kullanımı (Helmet_Used)	-0,041353	-0,041353	0,560969

Korelasyon ile ilgili değerler daha yaygın bir görsel ifade biçimi olan olan Şekil 1. de gösterimi verilen spearman korelasyon matrisi üzerinde daha ayrıntılı ve simetrik olarak incelenebilir.

Şekil 1. Spearman Korelasyon matrisi(Yaş (Age), Cinsiyet (Gender), Çarpışma Hızı (Speed_of_Impact), Kask Kullanımı (Helmet_Used), Emniyet Kemeri Kullanımı (Seatbelt_Used), Hayatta Kalma Durumu (Survived) olarak verilmektedir.)

3.2.6. Veri Artırma Teknikleri

Veri analizi ve modelleme sürecinde kullanılan çeşitli özellik mühendislik teknikleri, modelin doğruluğunu ve genelleme yeteneğini artırmada önemli rol oynamaktadır. Bu çalışmada, bir yol kazası verisinde hayatta kalma tahminlerini geliştirmek amacıyla bazı gelişmiş özellik mühendislik adımları uygulanmıştır. Özellikle, birinci derece ve ikinci derece özellikler oluşturulmuş ve bu özelliklerin kazaların sonuçları üzerindeki etkisi araştırılmıştır.

İlk olarak, ortalama türev(Average Derivative (AD)) özellikleri hesaplanmıştır. Bu özellikler, kazaların şiddetini belirlemede önemli olabilecek hız değişimlerini temsil eder. Çarpma Hızı sütunu üzerindeki farklar kullanılarak, her iki pencere boyutu için (2 ve 3) ortalama türev özellikleri oluşturulmuştur. Bu, çarpma hızındaki değişimin zaman içindeki etkisini analiz etmek için bir yöntemdir. Bu tür türevsel özellikler, özellikle

zaman serisi verilerinde modelin daha doğru tahminler yapmasına olanak tanıyabilir.

Ardından, polinom özellikleri oluşturulmuştur. Bu adımda, ikinci dereceden polinom özellikleri hesaplanmış ve her bir sayısal özelliğin kendisiyle ve diğer sayısal özelliklerle olan etkileşimlerinin etkisi analiz edilmiştir. Polinom özellikleri sınıfı kullanılarak, sayısal verilerdeki etkileşimler, eğilimler ve doğrusal olmayan ilişkiler dikkate alınmıştır. Bu özellikler, modelin doğrusal olmayan ilişkileri öğrenmesini sağlayarak tahmin gücünü artırabilir.

Yapılan bu polinom dönüşümü ile elde edilen yeni özellikler, modelin daha zengin bir veri seti üzerinde çalışmasını sağlar. Bu özelliklerin ismi, orijinal veri setinde yer alan sayısal sütunlardan türetilmiştir ve oluşturulan özelliklerin adı, polinom dönüşümü işlemiyle birlikte, her bir özelliğin etkileşimini ve seviyelerini ifade etmektedir. Bu işlem, modelin karmaşık ilişkileri daha iyi öğrenmesine yardımcı olur.

Son olarak, cinsiyet ve hayatta kalma gibi orijinal kategorik değişkenler, polinom özelliklerle oluşturulan yeni veri setine eklenmiştir. Bu adım, kategorik özelliklerin etkilerini modelin öğrenme sürecine dahil ederken, aynı zamanda verilerin bütünlüğünü korumaktadır. cinsiyet ve hayatta kalma değerlerinin olduğu sütunlar, kazalarda hayatta kalma tahminlerini etkileyebilecek önemli faktörlerdir ve bu nedenle, polinom özelliklerle birlikte değerlendirilmiştir.

Sonuç olarak, yapılan özellik mühendisliği işlemleri, yol kazası verisinde hayatta kalma tahminlerinin doğruluğunu artırmak için önemli bir adımdır. Ortalama türev ve polinom dönüşümleri gibi teknikler, modelin karmaşık ilişkileri öğrenmesine ve daha hassas tahminlerde bulunmasına olanak tanımaktadır. Bu yaklaşım, veri analizi ve makine öğrenmesi uygulamalarında etkili bir özellik mühendisliği stratejisi olarak değerlendirilebilir.

3.3. Yöntem

Bu çalışmada kullanılan veri ön işleme ve özellik mühendislik tekniklerine dair yapılan adımlar detaylı bir şekilde açıklanmıştır. Verilerin işlenmesi ve modelin doğruluğunu artırmak amacıyla çeşitli yöntemler uygulanmıştır. Bu adımlar, verinin anlamlı özellikler haline getirilmesini ve modelin genel performansının iyileştirilmesini hedeflemektedir. İlk adımda, veri setinden kullanılacak sayısal sütunlar seçilmiştir. Bu sütunlar, Yaş (Age), çarpma hızı (Speed of Impact), kask kullanımı (Helmet Used) ve emniyet kemerinin kullanımı (Seatbelt Used) olarak belirlenmiştir. Bu sütunlar, kazaların sonuçlarını anlamada kritik faktörler olarak kabul edilmiştir. Veri setinde

ortalama türev yani Average Derivative (AD) özellikleri oluşturulmuştur. Bu işlemde, hızdaki değişimin zaman içindeki etkisi dikkate alınmıştır. Çarpma hızı sütunu üzerinden hesaplanan farklar, belirli bir pencere uzunluğuna göre alınarak hız değişiminin ortalama türevini temsil eden yeni özellikler eklenmiştir. Pencere uzunlukları olarak 2 ve 3 kullanılmıştır. Bu işlem, kazaların etkisinin zamanla nasıl değiştiğini daha iyi anlamaya olanak sağlar. Bir sonraki adımda, Polinom Özellikleri oluşturulmuştur. Bu işlemde, orijinal sayısal özelliklerin ikili etkileşimleri ve ikinci dereceden polinomları hesaplanmıştır. Polinom Özellikleri sınıfı kullanılarak, ikinci dereceden polinom özellikler oluşturulmuş ve bu özelliklerin sayısal veri setine eklenmesi sağlanmıştır. Bu teknik, modelin doğrusal olmayan ilişkileri öğrenmesini kolaylaştırarak tahmin doğruluğunu artırabilir. Son olarak, orijinal veri setinden Cinsiyet (Gender) ve Hayatta Kalma Durumu (Survived) gibi kategorik değişkenler, oluşturulan yeni özelliklerle birlikte eklenmiştir. Bu kategorik değişkenler, kazaların hayatta kalma durumu üzerinde önemli etkiler yaratabilir. Dolayısıyla, modelin bu faktörleri de dikkate alarak tahminlerde bulunması sağlanmıştır.

Tüm bu adımların sonunda, özellik mühendisliği ile zenginleştirilmiş bir veri seti elde edilmiştir. Bu veri seti, kazaların sonuçlarını etkileyebilecek daha fazla bilgiye sahip olarak, tahmin modelinin doğruluğunu ve genelleme yeteneğini artırma amacını taşımaktadır. Bu yöntem, özellikle karmaşık ilişkilerin ve doğrusal olmayan etkilerin göz önünde bulundurulması gereken makine öğrenmesi uygulamaları için etkili bir strateji olarak öne çıkmaktadır.

Model kurulumu yapıldıktan sonra, farklı makine öğrenmesi algoritmaları ve derin sinir ağları kullanılarak model performansları detaylı bir şekilde değerlendirilmiştir. Model performansını ölçmek için doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru (F1-score) ve ROC-AUC (Receiver Operating Characteristic - Area Under Curve) gibi temel sınıflandırma metrikleri kullanılmıştır. Bunlara ek olarak, algoritmaların hesaplama maliyetlerini değerlendirmek amacıyla işlem süresi (time taken) de analiz edilmiştir. Bu sayede, modellerin yalnızca doğruluğu değil, aynı zamanda verimliliği ve hesaplama süresi açısından da karşılaştırılması amaçlanmıştır.

Model eğitimi sırasında kullanılan veri setleri, önişleme adımları uygulanarak temizlenmiş ve eksik veriler uygun yöntemlerle tamamlanmıştır. Kategorik değişkenler One-Hot Encoding veya Etiket Kodlama (Label Encoding) ile sayısal formata dönüştürülmüş, sürekli değişkenler ise Min-Max Scaling veya Standardizasyon ile ölçeklendirilmiştir. Bu işlemler, modelin daha kararlı ve genelleştirilebilir sonuçlar üretmesini sağlamak

amacıyla gerçekleştirilmiştir. Ayrıca, eğitim ve test veri kümelerinin dengeli dağılımı sağlanarak modelin aşırı öğrenme (overfitting) riskinin azaltılması hedeflenmiştir.

Çalışmada kullanılan algoritmalar arasında lojistik regresyon (Logistic Regression), karar ağaçları (Decision Trees), rastgele ormanlar (Random Forest), ve destek vektör makineleri (Support Vector Machines - SVM) yer almaktadır. Derin öğrenme tarafında ise Derin sinir ağları (Deep Neural Networks - DNN) gibi modeller test edilmiştir. Bu modeller farklı hiperparametre ayarlarıyla eğitilmiş ve sonuçlar karşılaştırmalı olarak analiz edilmiştir.

Son olarak, elde edilen metrikler ve işlem süresi değerleri karşılaştırılarak her bir modelin avantajları ve dezavantajları belirlenmiştir. Yüksek doğruluk oranı sağlayan modellerin işlem süresi bakımından daha maliyetli olduğu, bazı algoritmaların ise daha hızlı ancak düşük doğruluklu sonuçlar ürettiği gözlemlenmiştir. Bu analizler doğrultusunda, belirli uygulama senaryolarına en uygun model seçimi konusunda öneriler geliştirilmiştir.

3.3.1. Lojistik Regresyon

Lojistik regresyon, sınıflandırma problemlerinde yaygın olarak kullanılan doğrusal bir modeldir. Bu çalışmada, veri kümesindeki ilişkileri incelemek amacıyla lojistik regresyon analizi kullanılmıştır. Bu model, bağımsız değişkenlerin bağımlı değişkenin olasılığını nasıl etkilediğini tahmin etmemize olanak tanır. Lojistik regresyon modeli, bağımlı değişkenin başarı olasılığını

$$P(y=1|X)=\frac{1}{\left(1+e^{-z}\right)} \tag{4}$$

şeklinde ifade eder. Bu dönüşüm, z 'yi 0 ile 1 arasında bir değere sıkıştırarak, gözlemin belirli bir sınıfa ait olma olasılığını verir. Burada doğrusal tahminci z, bağımsız değişkenlerin bir doğrusal kombinasyonudur ve $\beta 0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n$ şeklinde hesaplanır. Bu denklemde, β_0 sabit terim (intercept), $\beta_1,\beta_2,...,\beta_n$ modelin öğrenilen parametreleridir ve X kümesinden alınan yani veri kümesinden alınan $x_1,x_2,...,x_n$ bağımsız değişkenlerinin değerlerini ifade eder.

Lojistik regresyon modelinde, modelin tahmin performansı, doğruluk, duyarlılık, özgüllük, ROC eğrisi, AUC kullanılarak değerlendirilmiştir. Modelin katsayıları (β), bağımsız değişkenlerin bağımlı değişkenin olasılığı

üzerindeki etkisini gösterir. Pozitif katsayılar, bağımsız değişkenin değerindeki artışın bağımlı değişkenin başarı olasılığını artırdığını gösterirken, negatif katsayılar tersini gösterir. Katsayıların büyüklüğü, etkinin gücünü gösterir. Lojistik regresyon modelinin bazı varsayımları vardır: bağımlı değişkenin iki kategorili olması, bağımsız değişkenler ile bağımlı değişkenin logit dönüşümü arasında doğrusal bir ilişki olması, gözlemlerin birbirinden bağımsız olması ve çoklu doğrusallık olmaması (bağımsız değişkenler arasında yüksek korelasyon olmaması) (Hosmer vd., 2013).

Lojistik regresyon modelinde, optimal hiperparametre C 'nin belirlenmesi için bir arama işlemi gerçekleştirilmiştir. Bu işlemde, C değeri 0.001 ile 1000 arasında logaritmik bir şekilde değiştirilmiş ve model, eğitim verisi üzerinde 5 katlı K-Fold çapraz doğrulama (KFold(5)) kullanılarak değerlendirilmiştir. Her bir C değeri için çapraz doğrulama sonucu elde edilen F1 skoru hesaplanmış ve en yüksek ortalama F1 skoru ile en uygun C değeri belirlenmiştir. Sonuçlar, C değeri ile birlikte çapraz doğrulama skorları da sırasıyla yazdırılmıştır. En yüksek F1 skoruna sahip olan C değeri belirlenerek modelin performansı optimize edilmiştir (Pedregosa vd., 2011).

Son olarak, optimal C değeriyle Logistic Regression modeli yeniden eğitilmiş ve test verisi üzerinde modelin performansı, doğruluk, duyarlılık, özgüllük ve ROC eğrisi gibi metriklerle rapor edilmiştir. Zaman hesaplamaları da eklenerek modelin eğitilmesi ve değerlendirilmesi süresi kaydedilmiştir.

Lojistik regresyon modelinin bazı sınırlılıkları da bulunmaktadır. Öncelikle, model yalnızca iki kategorili bağımlı değişkenler için uygundur. Ayrıca, bağımsız değişkenlerle bağımlı değişken arasındaki ilişkinin doğrusal olduğu varsayılmaktadır. Bunun yanı sıra, lojistik regresyon modeli aykırı değerlere duyarlı olabilir, bu da modelin performansını olumsuz etkileyebilir.

Modelin sonuçları, bağımsız değişkenlerin bağımlı değişken üzerinde anlamlı bir etkiye sahip olduğunu göstermektedir. Bu bulgular, trafik kazalarında tahmin yaparken önemli çıkarımlar sunmakta ve söz konusu bağımsız değişkenlerin, kazaların oluşma olasılığı üzerindeki etkisini anlamamıza yardımcı olmaktadır.

3.3.2. k-En Yakın Komşu (kNN)

Bu çalışmada, veri kümesi üzerindeki sınıflandırma problemi için k-En Yakın Komşu (kNN) algoritması uygulanmıştır. kNN, veri noktalarını sınıflandırmak için en yakın komşuların sınıflarını kullanan basit ve etkili bir denetimli öğrenme yöntemidir. Algoritmanın başarısı, özellikle k parametresinin (komşu sayısı) doğru seçimine bağlıdır. Bu nedenle, modelin performansını en üst düzeye çıkarmak için, 1 ile 20 arasında değişen k

değerleri için 5 katlı çapraz doğrulama (KFold) kullanılarak en uygun k değeri belirlenmiştir. Çapraz doğrulama sürecinde, her bir k değeri için F1 skoru (sınıflandırma performansının bir ölçüsü) hesaplanmış ve en yüksek F1 skorunu veren k değeri optimal k olarak seçilmiştir. Bu optimizasyon süreci, modelin genelleme yeteneğini artırmayı ve aşırı uyum (overfitting) riskini azaltmayı amaçlamaktadır.

Optimizasyon sonucunda elde edilen optimal k değeri [optimal_k] olarak bulunmuştur. Bu değer kullanılarak, kNN modeli eğitim ve test verileri üzerinde tekrar eğitilmiştir. Modelin sınıflandırma performansı, sınıflandırma raporunu veren fonksiyon kullanılarak detaylı bir şekilde değerlendirilmiştir. Bu rapor, doğruluk(accuracy), kesinlik(precision), duyarlılık (recall) ve F1 skoru gibi metrikleri içermektedir. kNN algoritması, yeni bir veri noktasının sınıfını belirlemek için en yakın k komşusunun sınıflarını kullanır. Uzaklık hesaplaması genellikle Öklid uzaklığı ile yapılır:

$$d(x,y)=\sqrt{\sum(x_i-y_i)^2} \tag{5}$$

Burada $d(x,y), x$ ve y noktaları arasındaki uzaklığı, x_i ve y_i ise bu noktaların i. değerlerini temsil eder.

kNN algoritması trafik veri seti üzerindeki sınıflandırma problemi için başarılı bir şekilde uygulanmış ve optimal k parametresi kullanılarak modelin performansı artırılmıştır. Modelin sınıflandırma raporu, test verileri üzerinde tatmin edici sonuçlar verdiğini göstermektedir. Bu çalışma, kNN algoritmasının hiperparametre optimizasyonunun önemini ve modelin genel performansına olan etkisini vurgulamaktadır. Çapraz doğrulama gibi yöntemlerin kullanılması, modelin güvenilirliğini artırmakta ve gerçek dünya uygulamalarında daha iyi sonuçlar elde edilmesini sağlamaktadır (Cover ve Hart, 1967; Kuhn ve Johnson, 2013; Bishop, 2006; Hastie vd., 2009; Scikit-learn Documentation, 2025).

3.3.3. Karar Ağacı

Bu çalışmada, veri seti üzerinde sınıflandırma problemi için Karar Ağacı (Decision Tree) algoritması kullanılarak, veri odaklı karar verme süreçleri incelenmiştir. Karar ağaçları, özellikle karmaşık ve doğrusal olmayan ilişkileri modellemede etkili, denetimli öğrenme algoritmalarıdır (Breiman, 1986). Algoritma, veri setini özyinelemeli olarak alt kümelere ayırarak, her bir alt küme için homojenlik kriterlerini maksimize etmeyi amaçlar. Bu süreçte, bilgi kazancı (information gain) veya Gini safsızlığı (Gini impurity) gibi metrikler kullanılarak optimal özellik seçimi yapılır. Bilgi kazancı, bir özelliğin veri setini ne kadar iyi ayırdığını ölçer ve şu şekilde ifade edilir:

$$IG(X,A)=H(X)-\Sigma\left[\left(|X_v|/|X|\right)*H(X_v)\right] \quad (6)$$

Burada $IG(X,A)$, IG (Information Gain) - Bilgi Kazancı, X veri setinde A özelliğinin bilgi kazancını, *H(X)*, X veri setinin entropisini, X_v, A özelliğinin v değerine sahip alt kümesini ve |.| kümenin eleman sayısını temsil eder. Entropi, veri setindeki sınıf dağılımının homojenliğini ölçer ve şu şekilde hesaplanır:

$$H(X)=-\Sigma\left[p_i*log_2(p_i)\right] \quad (7)$$

Burada p_i, X veri setindeki i. sınıfının olasılığını temsil eder.

Modelin genelleme yeteneği, ağacın karmaşıklığını kontrol eden hiperparametrelerin doğru seçimine kritik derecede bağlıdır. Bu bağlamda, modelin aşırı uyumunu (overfitting) önlemek ve genelleme performansını artırmak amacıyla, k-katlı çapraz doğrulama (k-fold cross-validation) yöntemi kullanılarak hiperparametre optimizasyonu gerçekleştirilmiştir (Kohavi, 1995). Çapraz doğrulama, veri setini k eşit parçaya bölerek her bir parçayı test, diğerlerini eğitim verisi olarak kullanır ve bu işlemi k kez tekrarlar. Bu sayede, modelin farklı veri alt kümeleri üzerindeki performansı değerlendirilerek daha güvenilir sonuçlar elde edilir. Her bir hiperparametre kombinasyonu için sınıflandırma performansının bir ölçüsü olarak F1 skoru veya AUC-ROC hesaplanmış ve en iyi performansı veren kombinasyon optimal değer olarak belirlenmiştir.

Optimizasyon sonucunda elde edilen optimal hiperparametre değerleri kullanılarak, Karar Ağacı modeli eğitim ve test verileri üzerinde tekrar eğitilmiştir. Modelin sınıflandırma performansı, doğruluk(accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru ve AUC-ROC gibi metriklerle değerlendirilmiştir (Powers, 2011). Karar ağaçları, verileri özellik değerlerine göre dallara ayırarak karar verir. Ağacın derinliği arttıkça, model daha karmaşık hale gelir ve eğitim verilerine aşırı uyum sağlayabilir. Bu nedenle, modelin karmaşıklığını kontrol eden hiperparametrelerin doğru seçimi modelin performansı için kritik öneme sahiptir.

Sonuç olarak, Karar Ağacı algoritması, veri seti üzerindeki sınıflandırma problemi için başarılı bir şekilde uygulanmış ve hiperparametre optimizasyonu ile modelin performansı artırılmıştır. Modelin değerlendirme metrikleri, test verileri üzerinde tatmin edici sonuçlar verdiğini göstermektedir. Bu çalışma, Karar Ağacı algoritmasının hiperparametre optimizasyonunun önemini ve modelin genel performansına olan etkisini vurgulamaktadır. Çapraz doğrulama ve bilgi teorisi tabanlı metriklerin kullanılması, modelin

güvenilirliğini artırmakta ve gerçek dünya uygulamalarında daha iyi sonuçlar elde edilmesini sağlamaktadır.

3.3.4. Rastgele Orman

Birden fazla karar ağacının çıktılarının ortalamasını alarak tahmin yapan Rastgele Orman modeli, sınıflandırma problemi için uygulanmıştır. Rastgele Orman, birden fazla karar ağacının bir araya gelerek oluşturduğu, topluluk öğrenme yöntemlerinden biridir. Bu algoritma, karar ağaçlarının aşırı uyum sorununu azaltarak daha sağlam ve genellenebilir modeller oluşturmayı amaçlar. Algoritmanın performansı, modelin karmaşıklığını ve hesaplama maliyetini doğrudan etkileyen bir parametrenin doğru seçimine bağlıdır. Bu nedenle, modelin en iyi performansını elde etmek için, bu parametre farklı değerler için çapraz doğrulama yöntemiyle optimize edilmiştir. Çapraz doğrulama, veri setini farklı parçalara bölerek her bir parçayı test, diğerlerini eğitim verisi olarak kullanır ve bu işlemi tekrarlar. Bu sayede modelin genelleme yeteneği daha güvenilir bir şekilde değerlendirilir. Her bir parametre değeri için sınıflandırma performansının bir ölçüsü hesaplanmış ve en yüksek performansı veren parametre değeri optimal değer olarak belirlenmiştir (Breiman, 2001).

Optimizasyon sonucunda elde edilen optimal parametre değeri kullanılarak, Rastgele Orman modeli eğitim ve test verileri üzerinde tekrar eğitilmiştir. Modelin sınıflandırma performansı, detaylı değerlendirme fonksiyonları ile analiz edilmiştir. Rastgele Orman algoritması, her bir karar ağacını farklı bir rastgele alt küme üzerinde eğiterek ve her ağaçta rastgele seçilmiş özellikler kullanarak çeşitlilik oluşturur. Bu çeşitlilik, modelin genelleme yeteneğini artırır ve aşırı uyum riskini azaltır. Her bir ağaç, tahminini yapar ve sınıflandırma sonucu, ağaçların çoğunluğunun verdiği karara göre belirlenir. Bu topluluk yaklaşımı, tek bir karar ağacına göre daha sağlam ve güvenilir sonuçlar üretir (Liaw ve Wiener, 2002).

Rastgele Orman'da, her bir ağaç en iyi özelliği seçmek için bilgi kazancı (information gain) veya Gini safsızlığı (Gini impurity) gibi metrikleri kullanır. Bilgi kazancı, bir özelliğin veri setini ne kadar iyi ayırdığını ölçer ve entropi kavramı üzerinden hesaplanır. Entropi, veri setindeki sınıf dağılımının homojenliğini ifade eder ve bu metrikler, her bir düğümde en iyi kararın verilmesini sağlar (Quinlan, 1986).

Topluluk öğrenme yaklaşımı, her bir karar ağacının tahminlerini birleştirerek nihai tahmini yapar. Sınıflandırma problemlerinde, her ağaç bir sınıf tahmini yapar ve nihai tahmin, ağaçların çoğunluğunun verdiği karara göre belirlenir. Regresyon problemlerinde ise ağaçların tahminlerinin

ortalaması alınır. Bu oylama veya ortalama alma süreci, modelin daha kararlı ve güvenilir tahminler yapmasını sağlar (Breiman, 2001).

Rastgele Orman, her bir karar ağacını farklı bir rastgele alt küme üzerinde eğiterek çeşitlilik oluşturur. Bu sürece bootstrap aggregating (bagging) denir. Bagging, veri setinden rastgele örnekler seçerek farklı alt kümeler oluşturur. Ayrıca, her bir ağaçta rastgele seçilmiş özellikler kullanılır. Bu rastgelelik, ağaçlar arasındaki korelasyonu azaltır ve modelin genelleme yeteneğini artırır (Breiman, 1996).

Modelin performansı, çapraz doğrulama (cross-validation) yöntemiyle değerlendirilir. Sınıflandırma problemlerinde, doğruluk(accuracy), kesinlik(precision), duyarlılık (recall), F1 skoru ve ROC eğrileri, AUC hesabı gibi metrikler kullanılır. Bu değerlendirme metrikleri, modelin ne kadar başarılı olduğunu sayısal olarak ifade eder (Kohavi, 1995).

Sonuç olarak, Rastgele Orman algoritması başarılı bir şekilde uygulanmış ve optimal parametre değeri kullanılarak modelin performansı artırılmıştır. Modelin değerlendirme metrikleri, test verileri üzerinde tatmin edici sonuçlar verdiğini göstermektedir. Bu çalışma, Rastgele Orman algoritmasının parametre optimizasyonunun önemini ve modelin genel performansına olan etkisini vurgulamaktadır. Çapraz doğrulama gibi yöntemlerin kullanılması, modelin güvenilirliğini artırmakta ve gerçek dünya uygulamalarında daha iyi sonuçlar elde edilmesini sağlamaktadır (Liaw & Wiener, 2002).

3.3.5. Derin Sinir Ağları (Deep Neural Networks, DeepNN)

Bu çalışmada, ikili sınıflandırma problemi için Derin Sinir Ağları (Deep Neural Networks, DeepNN) kullanılarak veri odaklı modelleme gerçekleştirilmiştir. Derin Sinir Ağları, çok sayıda gizli katmana sahip yapay sinir ağı modelleri olup, karmaşık desenleri ve ilişkileri öğrenme yeteneğiyle öne çıkar. Algoritmada, özel bir sınıf oluşturularak, TensorFlow ve Keras kütüphaneleri kullanılarak bir derin sinir ağı modeli tanımlanmıştır. Bu model, girdi katmanından sonra 5 adet gizli katman içermekte ve her bir gizli katmanında ReLU aktivasyon fonksiyonu kullanmaktadır. Çıktı katmanında ise sigmoid aktivasyon fonksiyonu kullanılarak ikili sınıflandırma problemi çözülmüştür. Model, Adam optimizasyon algoritması ve binary cross-entropy hata fonksiyonu ile derlenmiştir (Kingma ve Ba, 2015).

Modelin eğitimi, fit metodu aracılığıyla gerçekleştirilmiştir. Bu metot, girdi verileri (X) ve hedef değişkenleri (y) kullanarak ağı eğitir Chollet (2017). Eğitim süreci, belirli sayıda epoch (tekrar sayısı) ve batch size (toplu işleme boyutu) parametreleri ile kontrol edilmiştir. Modelin performansı, eğitim ve test verileri üzerinde ayrı ayrı değerlendirilmiştir. Eğitim verileri

üzerinde modelin doğruluğu (accuracy) hesaplanırken, test verileri üzerinde de aynı şekilde doğruluk (accuracy), duyarlılık (precision), geri çağırma (recall), F1 skoru, AUC-ROC ve log-loss gibi metrikler hesaplanmıştır (Sokolova ve Lapalme, 2009).

Modelin çalışma yapısı, her bir nöronun çıktısının bir aktivasyon fonksiyonu ile hesaplanmasına dayanır. Bir nöronun çıktısı şu şekilde ifade edilebilir:

$$y = f\left(\sum\left(wi * xi\right) + b\right) \quad 8)$$

Burada y, nöronun çıktısını, f, aktivasyon fonksiyonunu; w_i, girdilerin ağırlıklarını, x_i, girdileri ve b, bias terimini temsil eder. Aktivasyon fonksiyonları, ağın doğrusal olmayan ilişkileri modellemesini sağlar. Ağırlıkların güncellenmesi ise gradyan inişi algoritması ile gerçekleştirilmiştir ve şu şekilde ifade edilir:

$$wi = wi - \alpha * \partial L / \partial wi \quad (9)$$

Burada α, öğrenme oranını; L, hata fonksiyonunu; ve $\partial L / \partial wi$, L'nin w_i'ye göre türevini temsil eder.

Derin Sinir Ağları, yapay zeka ve makine öğrenmesi alanlarında güçlü bir modelleme aracıdır ve özellikle karmaşık veri setlerinde yüksek doğruluk oranları sağlayabilmektedir. Bu çalışmada, derin sinir ağı modelinin performansı, Python programlama dili ve Keras kütüphanesi kullanılarak değerlendirilmiştir. Modelin temel yapısı, birden fazla gizli katman içeren bir yapıyı benimsemektedir. Derin Sinir Ağı modelinde kullanılan katmanlar, verinin daha derin ve soyut özelliklerini öğrenebilmek için çeşitli aktivasyon fonksiyonları (ReLU ve sigmoid) kullanarak veri setinin özelliklerini sırasıyla dönüştürür. Bu çalışmada, modelin giriş katmanına 256 nöron eklenmiş, ardından sırasıyla 128, 64, 32 ve 16 nöronlu katmanlar eklenmiştir. Modelin çıktısı, ikili sınıflandırma problemini çözebilmek amacıyla sigmoid aktivasyon fonksiyonu ile elde edilmiştir.

Modelin eğitilmesinde, eğitim verisi üzerinde 100 epoch boyunca eğitim yapılmıştır. Eğitim sırasında, her bir veri örneği, modelin parametrelerini optimize etmek amacıyla Adam optimizasyon algoritmasıyla güncellenmiştir. Eğitim süreci, her bir epoch sonunda modelin doğruluğunu ve kayıp fonksiyonunu izleyerek gerçekleştirilmiştir. Eğitimin ardından, modelin performansı test verisi üzerinde değerlendirilmiş ve test doğruluğu hesaplanmıştır.

Modelin başarı metrikleri arasında, doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F1 skoru ve ROC eğrisi-AUC (Area Under

the Curve) gibi önemli performans göstergeleri yer almaktadır (Hand ve Till, 2001). F1 skoru, modelin sınıflandırma başarımını tek bir metrikle değerlendirebilmek için kullanılırken, AUC değeri, modelin sınıflandırma doğruluğunu daha kapsamlı bir şekilde ölçmek için önemli bir araçtır. Bu çalışmada elde edilen AUC skoru, modelin ikili sınıflandırmada ne kadar etkili olduğunu gösterirken modelin tahmin doğruluğunu sayısal bir şekilde ifade etmektedir.

Sonuç olarak, derin sinir ağları, özellikle yüksek doğruluk gerektiren sınıflandırma problemlerinde başarılı bir çözüm sunmaktadır (Goodfellow vd., 2016). Bu model, karmaşık verilerdeki soyut ilişkileri öğrenme yeteneği ile geleneksel makine öğrenmesi yöntemlerine kıyasla daha üstün performans gösterebilmektedir. Ancak, modelin doğruluğu ve genel performansı, kullanılan verinin kalitesine ve modelin hiperparametrelerinin doğru bir şekilde optimize edilmesine bağlıdır. Bu çalışmada, hiperparametrelerin belirlenmesi için çeşitli optimizasyon teknikleri kullanılmakta olup, elde edilen sonuçlar, derin sinir ağlarının uygulanabilirliğini ve etkinliğini desteklemektedir.

Sonuç olarak, Derin Sinir Ağı modeli başarıyla eğitilmiş ve test verileri üzerinde tatmin edici sonuçlar elde edilmiştir. Modelin değerlendirme metrikleri, modelin genelleme yeteneğinin yüksek olduğunu göstermektedir. Bu çalışma, Derin Sinir Ağlarının karmaşık sınıflandırma problemlerinde kullanılabileceğini ve veri odaklı karar verme süreçlerinde önemli bir rol oynayabileceğini göstermektedir. Bu sonuçlar, farklı makine öğrenmesi algoritmalarının ve Derin Sinir Ağlarının veri setine göre farklı performans sergileyebileceğini ortaya koymaktadır. Model seçiminde yalnızca doğruluk değil, diğer metrikler ve işlem süresi gibi faktörlerin de göz önünde bulundurulması gerekmektedir.

4. Bulgular

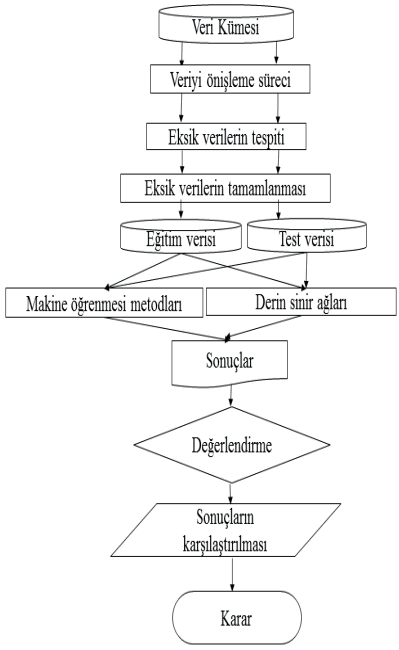
Bu çalışmada, trafik kazalarında hayatta kalma oranlarını tahmin etmek amacıyla farklı makine öğrenmesi modelleri ve Derin Sinir Ağları incelenmiş, modellerin performansları doğruluk(accuracy), kesinlik(precision), duyarlılık(recall) ve F1 skoru gibi metrikler açısından değerlendirilmiştir. Modellerin çalıştırılmasıyla elde edilen metrikler Tablo 2 de sunulmuştur. Modellerin ROC eğrileri-AUC grafiği Şekil 4 de verilmektedir. Elde edilen sonuçlar, kullanılan modeller arasında belirgin farklılıklar olduğunu göstermektedir. En yüksek doğruluk oranı %90,5 ile Rastgele Orman (Random Forest) modeli tarafından elde edilmiştir. Rastgele Orman

modeli, aynı zamanda %90,2 duyarlılık ve %90,99 F1 skoru ile en dengeli performansı sergileyen model olmuştur.

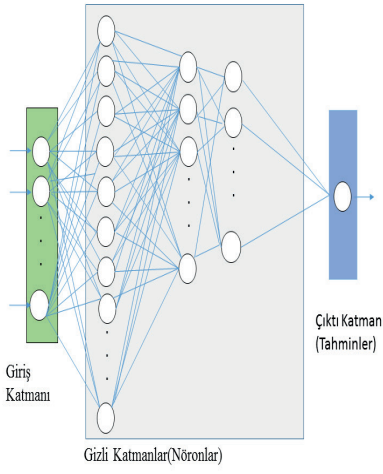
Karar Ağaçları (Decision Tree) modeli, %87,75 doğruluk oranı ile ikinci en iyi performansı göstermiştir. Bu modelin kesinlik ve duyarlılık değerleri de sırasıyla %89,2 ve %89,7 olarak hesaplanmıştır. Tek bir karar ağacı kullanan bu model, Rastgele Orman'a kıyasla biraz daha düşük performans göstermiş olsa da, hesaplama süresi açısından daha hızlı olduğu gözlemlenmiştir. k-En Yakın Komşular (KNN) algoritması ise %65,5 doğruluk oranı ile orta seviyede bir performans sergilemiştir. Özellikle yüksek değişkenlik gösteren veri setlerinde daha az başarılı olduğu ve işlem süresinin nispeten uzun olduğu belirlenmiştir.

Lojistik Regresyon modeli %54,75 doğruluk oranı ile en düşük performansı sergileyen model olmuştur. Özellikle doğrusal ayrılabilir olmayan veriler üzerinde düşük başarı göstermesi, bu modelin trafik kazalarında hayatta kalma tahminleri için yeterince uygun olmadığını düşündürmektedir. Derin Sinir Ağları (Deep Neural Networks) modeli ise %62 doğruluk oranı ile beklenenden düşük bir performans göstermiştir. Bu durum, modelin aşırı öğrenme (overfitting) veya yetersiz öğrenme (underfitting) problemleri yaşadığını düşündürmektedir. Modelin performansını artırmak için hiperparametre optimizasyonu ve daha büyük veri setleriyle eğitiminin yapılması gerekmektedir.

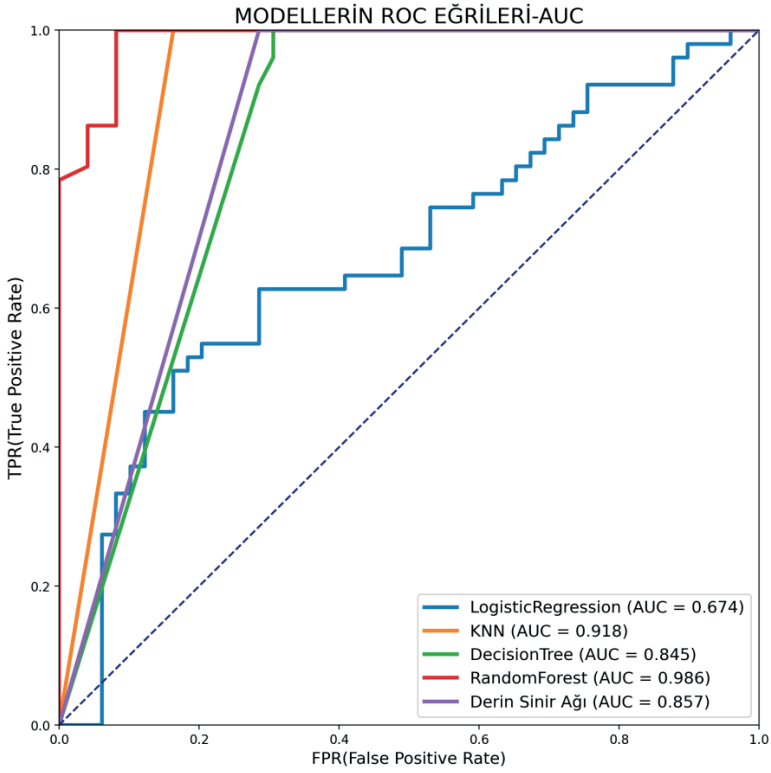
Genel olarak, ağaç tabanlı modellerin (Karar Ağaçları ve Rastgele Orman) diğer yöntemlere göre daha başarılı olduğu görülmüştür. Özellikle Rastgele Orman modeli, tüm performans metriklerinde en iyi sonuçları sunarak, trafik kazalarında hayatta kalma oranlarının tahmini için en uygun model olarak öne çıkmaktadır. Ancak, modelin işlem süresi ve karmaşıklığı dikkate alındığında, daha hızlı ancak nispeten düşük doğruluk oranına sahip olan Karar Ağaçları gibi modellerin de tercih edilebileceği sonucuna varılmıştır.



Şekil 2. Veri önişleme ve veri analizi süreçlerinin gösterimi



Şekil 3. Derin sinir ağları



Şekil 4. ROC eğrileri ve AUC değerleri

Tablo 2. Algoritmalardan elde edilen sonuçlar

Model	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru (F1 Score)	Geçen Süre (Time Taken)
Rastgele Orman	0,9050	0,897952	0,902073	0,909905	38,184784
Karar Ağaçları	0,8775	0,892689	0,897317	0,894149	0,632241
k-En Yakın Komşular	0,6550	0,672424	0,652073	0,660337	0,928479
Derin Sinir Ağı	0,6200	0,640050	0,681341	0,636005	165,2073
Lojistik Regresyon	0,5475	0,554645	0,587683	0,569715	15,173153

5. Sonuç

Bu Bu çalışma, trafik kazalarında hayatta kalma oranlarını tahmin etmek amacıyla trafik verileri üzerinden çeşitli makine öğrenmesi modellerinin performansını kapsamlı bir şekilde değerlendirmiştir. Elde edilen sonuçlar, modeller arasında belirgin farklılıklar olduğunu ortaya koymaktadır. Özellikle Rastgele Orman modeli, yüksek doğruluk oranı (%90,5) ve düşük

işlem süresiyle öne çıkarken, ROC eğrileri de bu modelin diğer yöntemlere göre daha üstün sınıflandırma performansı sunduğunu göstermiştir. Karar Ağaçları modeli ise nispeten daha düşük doğruluk oranına sahip olmasına rağmen, işlem süresi açısından daha hızlı olması nedeniyle avantaj sağlamaktadır.

Derin Sinir Ağları, görsel sınıflandırma alanında elde ettikleri üstün performanslarla öne çıkmaktadır. Bu modeller, görüntülerdeki karmaşık desenleri ve yapıları otomatik olarak öğrenme kabiliyetleriyle, görüntü sınıflandırmada geleneksel yöntemlere kıyasla daha doğru ve verimli sonuçlar sunmaktadır. Ancak, bu çalışmada kullanılan Derin Sinir Ağları modeli beklenen performansı sağlayamamış, hem doğruluk hem de işlem süresi açısından daha düşük sonuçlar üretmiştir. Bu durum, modelin aşırı öğrenme (overfitting) veya yetersiz öğrenme (underfitting) problemleri yaşadığını göstermektedir. Gelecekte, modelin hiperparametre optimizasyonu yapılarak ve daha büyük veri setleri ile eğitilerek performansının artırılması sağlanabilir.

Genel olarak, ağaç tabanlı modellerin (Karar Ağaçları ve Rastgele Orman) diğer yöntemlere göre daha başarılı olduğu görülmüştür. Özellikle Rastgele Orman modeli, tüm performans metriklerinde en iyi sonuçları sunarak, trafik kazalarında hayatta kalma oranlarının tahmini için en uygun model olarak öne çıkmaktadır. Bununla birlikte, modelin işlem süresi ve hesaplama karmaşıklığı dikkate alındığında, nispeten daha düşük doğruluk oranına sahip ancak daha hızlı olan Karar Ağaçları gibi modellerin de tercih edilebileceği gözlemlenmiştir.

Bu bağlamda, çalışmanın bulguları, trafik kazalarında hayatta kalma oranlarını tahmin etme sürecinde daha karmaşık modellerin her zaman daha iyi sonuç vermediğini, veri yapısı ve hesaplama verimliliği gibi faktörlerin de göz önünde bulundurulması gerektiğini göstermektedir. Özellikle büyük ölçekli veri setleri kullanıldığında, hesaplama maliyetleri ve modelin genelleme yeteneği gibi faktörlerin kritik olduğu görülmektedir. Daha karmaşık modellerin her zaman en iyi performansı sunmaması, veri özelliklerine uygun model seçiminin önemini vurgulamaktadır.

Gelecekteki çalışmalar, farklı veri setleri ve ek parametreler kullanarak modellerin genel geçerliğini artırmaya yönelik kapsamlı analizlere odaklanabilir. Özellikle, değişken seçim yöntemleri ve veri ön işleme tekniklerinin model başarısına etkisi araştırılabilir. Ayrıca, derin öğrenme modellerinin performansını artırmak amacıyla transfer öğrenme ve ileri seviye optimizasyon tekniklerinin uygulanması faydalı olabilir. Son olarak, gerçek

zamanlı tahmin sistemleri geliştirmek için donanım destekli hızlandırıcıların kullanılması, modellerin pratik kullanım alanlarını genişletebilir.

Veri Kümesi Kaynağı ve Kullanılabilirliği

Bu çalışma, Kaggle platformunda yer alan Road Accident Survival Dataset veri kümesi kullanılarak gerçekleştirilmiştir. Veri kümesi, trafik kazalarında hayatta kalma olasılıklarını etkileyen demografik, davranışsal ve durumsal faktörler gibi çeşitli etmenler hakkında kapsamlı bilgiler sunmakta ve bu etmenlerin kazalarda hayatta kalma şansına nasıl etki edebileceği konusunda derinlemesine analiz yapılmasına imkân tanımaktadır. Veri kümesi, CC0: Public Domain lisansı altında yayımlanmaktadır; bu, verilerin telif hakkı kısıtlamalarından arındırıldığı ve herkese ücretsiz olarak kullanım hakkı verildiği anlamına gelmektedir. Kişisel veri içermemekte olup, kullanımına yönelik herhangi bir kısıtlama bulunmamaktadır. Bu nedenle, veriler akademik çalışmalar, araştırmalar ve diğer bilimsel analizler için serbestçe kullanılabilir. Veri kümesine ulaşmak için Kaynakça da verilen Kaggle web sitesinde datanın yayınlandığı web adresi için bağlantısı üzerinden erişim sağlanabilir.

Kaynakça

Ahmed, S., Hossain, M. A., Bhuiyan, M. M. I., & Ray, S. K. (2021, December). A comparative study of machine learning algorithms to predict road accident severity. In 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS) (pp. 390-397).

Bai, M. L., Pamula, R., Subbarao, K., & Bharathi, S. (2025). Recommended System for Predicting Traffic Accident Costs using Enhanced Machine Learning Techniques. Journal of Electrical Engineering & Technology, 1-12.

Ballamudi, K. R. (2019). Road accident analysis and prediction using machine learning algorithmic approaches. Asian Journal of Humanity, Art and Literature, 6(2), 185-192.

Behboudi, N., Moosavi, S., & Ramnath, R. (2024). Recent advances in traffic accident analysis and prediction: a comprehensive review of machine learning techniques. arXiv preprint arXiv:2406.13968.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Breiman, L. (1986). Classification and regression trees. Wadsworth & Brooks/ Cole Advanced Books & Software.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Chollet, F. (2017). Deep learning with Python. Manning Publications.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.

Dünya Sağlık Örgütü (WHO). (2004). World report on road traffic injury prevention. https://www.who.int/publications/i/item/9241562609

Dünya Sağlık Örgütü (WHO). (2023). Global status report on road safety 2023. https://www.who.int/publications/i/item/9789240086517

Gibbons, J. D., & Chakraborti, S. (2014). *Nonparametric statistical inference: revised and expanded*. CRC press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 45(2), 171-186. https://doi.org/10.1023/A:1010920819831

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.

Kendall, M. G. (1948). Rank correlation methods. Oxford University Press.

Kingma, D. P. & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Kodepogu, K., Manjeti, V. B., & Siriki, A. B. (2023). Machine learning for road accident severity prediction. Mechatron. Intell. Transp. Syst, 2, 211-226.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1137–1143.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kumar, B., Basit, A., Kiruba, M. B., Giridharan, R., & Keerthana, S. M. (2021, July). Road accident detection using machine learning. In 2021 International Conference on System, Computation, Automation and Networking (ICSCAN) (pp. 1-5).

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Lin, D. J., Chen, M. Y., Chiang, H. S., & Sharma, P. K. (2021). Intelligent traffic accident prediction model for Internet of Vehicles with deep learning approach. IEEE transactions on intelligent transportation systems, 23(3), 2340-2349.

Ma, Z., Mei, G., & Cuomo, S. (2021). An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors. Accident Analysis & Prevention, 160, 106322. https://doi.org/10.1016/j.aap.2021.106322

Nandurge, P. A., & Dharwadkar, N. V. (2017, February). Analyzing road accident data using machine learning paradigms. In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 604-610).

Obasi, I. C., & Benson, C. (2023). Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. Heliyon, 9(8), e17875. https://doi.org/10.1016/j.heliyon.2023.e17875

Patil, J., Prabhu, M., Walavalkar, D., & Lobo, V. B. (2020, December). Road accident analysis using machine learning. In 2020 IEEE Pune Section International Conference (PuneCon) (pp. 108-112).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37–63.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.

Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine learning approaches to traffic accident analysis and hotspot prediction. Computers, 10(12), 157.

Sarder, H. (2025). Road accident survival dataset. Kaggle. https://www.kaggle.com/datasets/himelsarder/road-accident-survival-dataset

Singh, G., Pal, M., Yadav, Y., & Singla, T. (2020). Deep neural network-based predictive modeling of road accidents. Neural Computing and Applications, 32, 12417-12426.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72-101.

Theofilatos, A., Chen, C., & Antoniou, C. (2019). Comparing machine learning and deep learning methods for real-time crash prediction. Transportation research record, 2673(8), 169-178.

Vasavi, S. (2018). Extracting hidden patterns within road accident data using machine learning techniques. In Information and Communication Technology: Proceedings of ICICT 2016 (pp. 13-22). Springer Singapore.

Vicent, J. F., Curado, M., Oliver, J. L., & Pérez-Sala, L. (2025). A novel approach to predict the traffic accident assistance based on deep learning. Neural Computing and Applications, 37(7), 5343-5368.