



ABD TRAFİK KAZALARINDA HAVA DURUMU, YOL TİPİ VE SAAT BİLGİSİNE GÖRE KAZA ŞİDDETİNİN TAHMİNİ

BİTİRME PROJESİ RAPORU

Makine Öğrenmesi ile Çok Sınıflı Sınıflandırma

Dilan Sazan

Proje Deposu: github.com/Dilan-Sazan/abd-kaza-tahmini

Ağustos, 2026

İÇİNDEKİLER

1. GİRİŞ

- 1.1. Projenin Amacı
- 1.2. Proje Kısıtları

2. MATERYAL VE YÖNTEM

- 2.1. Kullanılan Veri Setleri
- 2.2. Geliştirme Ortamı ve Kütüphaneler
- 2.3. Hedef Değişken: Severity
- 2.4. Başarım Ölçütünün Belirlenmesi
- 2.5. Yöntem Akışı

3. UYGULAMA

- 3.1. İlçe Referans Tablosunun Oluşturulması
 - 3.1.1. Veri Kaynaklarının Birleştirilmesi
 - 3.1.2. Bağımsız Şehir Çakışmalarının Çözülmesi
- 3.2. Veri Hazırlama
 - 3.2.1. Parçalı Okuma ve Eyalet Süzme
 - 3.2.2. Hedef Sızıntısı Taşıyan Sütunların Çıkarılması
 - 3.2.3. Eksik Değer Yönetimi
 - 3.2.4. Referans Tablosuyla Birleştirme
 - 3.2.5. Türetilen Değişkenler
- 3.3. Keşifsel Veri Analizi
 - 3.3.1. Aykırı Değer Tespiti ve Fiziksel Sınır Kırpması
 - 3.3.2. Hedef Değişken Dağılımı
 - 3.3.3. Saat Kırılımı
 - 3.3.4. Yol Tipi Kırılımı
 - 3.3.5. Hava Durumu Kırılımı
 - 3.3.6. Kentsel/Kırsal Kırılımı
 - 3.3.7. Korelasyon Analizi
- 3.4. Modelleme
 - 3.4.1. Özellik Seçimi
 - 3.4.2. Kategorik Değişkenlerin Kodlanması
 - 3.4.3. Eğitim/Test Ayrımı
 - 3.4.4. Model 1 — Baseline (DummyClassifier)
 - 3.4.5. Model 2 — Lojistik Regresyon
 - 3.4.6. Model 3 — Random Forest
 - 3.4.7. Model Karşılaştırması

4. SONUÇ

5. KAYNAKÇA

1. GİRİŞ

1.1. Projenin Amacı

Trafik kazalarının şiddeti, olay anındaki çevresel koşullarla ilişkilidir. Hava durumu, yolun tipi, kazanın gerçekleştiği saat ve bölgenin kentsel yapısı gibi etkenlerin kaza sonucuna nasıl yansıdığının bilinmesi; trafik yönetimi, acil müdahale planlaması ve yol güvenliği çalışmaları açısından değer taşımaktadır.

Bu çalışmanın amacı, bir trafik kazası meydana geldiği anda bilinebilen bilgilerden yola çıkarak kazanın trafiğe etki şiddetini (Severity) tahmin eden bir makine öğrenmesi modeli geliştirmektir. Modele girdi olarak yalnızca kaza anında gözlemlenebilir değişkenler verilmiş; kaza sonrasında ölçülen bilgiler bilinçli olarak dışarıda bırakılmıştır.

Çalışma kapsamında üç temel soruya yanıt aranmıştır:

- Çevresel koşullar (hava durumu, yol tipi, saat, kentsel/kırsal yapı) ile kaza şiddeti arasında ölçülebilir bir ilişki var mıdır?
- Bu ilişki doğrusal bir modelle temsil edilebilir mi, yoksa doğrusal olmayan bir yaklaşım mı gerekmektedir?
- Sınıf dağılımının ileri derecede dengesiz olduğu bir veri setinde model başarımı hangi ölçütlerle değerlendirilmelidir?

1.2. Proje Kısıtları

Çalışma sırasında karşılaşılan ve yöntemsel kararları doğrudan etkileyen kısıtlar aşağıda sıralanmıştır.

Veri boyutu. Ham veri seti 7,7 milyon satır ve yaklaşık 3 GB büyüklüğündedir. Dosyanın tamamının belleğe alınması mümkün olmadığından, veri 500.000 satırlık parçalar hâlinde okunmuş ve süzme işlemi okuma sırasında uygulanmıştır.

Coğrafi kapsam. Hesaplama maliyetini makul düzeyde tutmak ve iklim/yol yapısı bakımından birbirinden farklı bölgeleri karşılaştırabilmek amacıyla çalışma üç eyaletle sınırlandırılmıştır: Florida (subtropikal iklim, düz arazi), New York (yoğun kentsel yapı) ve Minnesota (sert kış koşulları). Süzme sonrası veri seti 1.278.270 satıra inmiştir.

Sınıf dengesizliği. Hedef değişkenin %81,69'u tek bir sınıfta (Severity 2) toplanmıştır. En seyrek sınıf olan Severity 1 ise verinin yalnızca %0,66'sını oluşturmaktadır. Bu dengesizlik hem başarımların ölçütü seçimini hem de model parametrelerini belirlemiştir.

Hedef sızıntısı riski. Ham veri setinde bulunan bazı sütunlar kaza sonrasında ölçülmektedir ve hedef değişkeni doğrudan ele vermektedir. Bu sütunlar model kurulmadan önce veri setinden çıkarılmıştır.

Ölçüm tutarsızlıkları. Veri seti farklı trafik API'lerinden derlendiği için sensör kaynaklı fiziksel olarak imkânsız değerler ve yıllar içinde değişen terminoloji farklılıkları içermektedir. Bu durumlar keşifsel analiz aşamasında tespit edilerek ele alınmıştır.

2. MATERYAL VE YÖNTEM

2.1. Kullanılan Veri Setleri

Çalışmada biri ana, üçü destekleyici olmak üzere dört veri kaynağı kullanılmıştır.

Veri Kaynağı	İçerik	Kullanım Amacı
US_Accidents_March23.csv (Kaggle)	ABD genelinde 7,7 milyon trafik kaydı, 47 sütun	Ana veri seti; hedef değişken ve çevresel değişkenler
co-est2024-alldata.csv (US Census Bureau)	İlçe bazında nüfus tahminleri	nufus_2022 değişkeninin türetilmesi
2024_Gaz_counties_national (US Census Gazetteer)	İlçe bazında kara alanı (km ²)	nufus_yogunluğu değişkeninin hesaplanması
NCHSurb-rural-codes.csv (CDC / NCHS)	İlçe bazında kentsel-kırsal sınıflandırma kodları	kentsel_kırsal değişkeninin türetilmesi

Tablo 2.1.1. Çalışmada kullanılan veri kaynakları

Destekleyici üç kaynak, ilçe bazında birleştirilerek tek bir referans tablosuna dönüştürülmüş; ana veri setine bu tablo üzerinden bağlanmıştır. Böylece ham veride bulunmayan nüfus, nüfus yoğunluğu ve kentsel/kırsal sınıf bilgileri modele kazandırılmıştır.

2.2. Geliştirme Ortamı ve Kütüphaneler

Bileşen	Sürüm / Açıklama
İşletim sistemi	Windows
Geliştirme ortamı	Visual Studio Code — Jupyter Notebook eklentisi
Programlama dili	Python 3.12
Veri işleme	pandas, numpy
Görselleştirme	matplotlib, seaborn
Makine öğrenmesi	scikit-learn
Ara veri formatı	Apache Parquet
Sürüm kontrolü	Git / GitHub

Tablo 2.2.1. Geliştirme ortamı ve kullanılan kütüphaneler

Temizlenmiş veri, ara çıktı olarak CSV yerine Parquet biçiminde saklanmıştır. Parquet sütun tabanlı bir formattır; veri tiplerini koruduğu ve sıkıştırma sağladığı için tekrar okuma süresi CSV'ye kıyasla belirgin biçimde kısadır.

2.3. Hedef Değişken: Severity

Hedef değişken Severity, 1 ile 4 arasında değer alan sıralı (ordinal) bir etikettir. Veri setinin dokümantasyonuna göre bu değer yaralanma şiddetini değil, kazanın trafik akışına etkisinin derecesini ifade etmektedir. 1 en düşük, 4 en yüksek etkiyi göstermektedir.

Bu ayrım projenin yorumlanması açısından belirleyicidir: model, bir kazanın ne kadar ağır yaralanmaya yol açacağını değil, trafiği ne ölçüde etkileyeceğini tahmin etmektedir.

Severity	Kayıt Sayısı	Oran
2	1.044.196	%81,69

Severity	Kayıt Sayısı	Oran
3	201.712	%15,78
4	23.905	%1,87
1	8.457	%0,66
Toplam	1.278.270	%100,00

Tablo 2.3.1. Hedef değişkenin sınıf dağılımı

2.4. Başarım Ölçütünün Belirlenmesi

Sınıf dağılımının ileri derecede dengesiz olması nedeniyle doğruluk (accuracy) ölçütü bu çalışmada yanıltıcıdır. Hiçbir öğrenme gerçekleştirmeyen ve her kayda çoğunluk sınıfını atayan bir model dahi yaklaşık %82 doğruluk elde edecektir. Bu değer modelin başarısını değil, verinin dengesizliğini yansıtır.

Bu nedenle başarımlı ölçütü olarak makro ortalama F1 skoru (macro-F1) seçilmiştir. Macro-F1, her sınıf için F1 skorunu ayrı ayrı hesaplayıp bunların ağırlıksız ortalamasını alır. Böylece seyrek sınıflar (Severity 1 ve 4) çoğunluk sınıfıyla eşit ağırlık taşır; yalnızca çoğunluk sınıfını tahmin eden bir model düşük macro-F1 değeri üretir.

Ayrıca tüm modellerde `class_weight="balanced"` parametresi kullanılmıştır. Bu parametre, eğitim sırasında seyrek sınıflara ait örneklerle daha yüksek ağırlık atayarak modelin bu sınıfları tamamen göz ardı etmesini engeller.

Regresyon problemlerinde kullanılan R^2 ve RMSE gibi ölçütler bu çalışmada uygulanabilir değildir. Severity sıralı bir etikettir; sınıflar arasındaki mesafe tanımlı olmadığından "ne kadar yanlışlık" sorusu anlamlı bir karşılık bulmaz. Değerlendirme, tahmin edilen sınıfın doğru olup olmadığı üzerinden yapılmıştır.

2.5. Yöntem Akışı

Çalışma, her biri ayrı bir Jupyter Notebook dosyası olarak yürütülen dört aşamada gerçekleştirilmiştir.

Aşama	Dosya	İçerik
1	00_referans_olustur.ipynb	Census, Gazetteer ve CDC verilerinin ilçe bazında birleştirilmesi
2	01_veri_hazirlama.ipynb	Parçalı okuma, süzme, temizleme, birleştirme, değişken türetme
3	02_eda.ipynb	Keşifsel veri analizi, aykırı değer kontrolü, korelasyon analizi
4	03_model.ipynb	Özellik seçimi, kodlama, model kurulumu ve karşılaştırma

Tablo 2.5.1. Çalışma akışı ve notebook yapısı

Aşamaların ayrı dosyalarda tutulması, her adımın bağımsız olarak çalıştırılabilmesini ve ara çıktıların yeniden üretilmesine gerek kalmadan kullanılabilmesini sağlamıştır.

3. UYGULAMA

3.1. İlçe Referans Tablosunun Oluşturulması

3.1.1. Veri Kaynaklarının Birleştirilmesi

Ham kaza verisi ilçe adı bilgisi içermekte, ancak o ilçenin nüfusu, yüzölçümü veya kentsel/kırsal niteliği hakkında bilgi taşımamaktadır. Bu bilgiler üç ayrı resmî kaynaktan alınmış ve tek bir referans tablosunda birleştirilmiştir.

Birleştirme işleminde anahtar olarak ilçe adı yerine GEOID (FIPS kodu) kullanılmıştır. İlçe adları eyaletler arasında tekrar edebildiği ve yazım farklılıkları içerebildiği için ad üzerinden birleştirme hatalı eşleşmelere yol açacaktır. GEOID her ilçe için tekil ve standart bir tanımlayıcıdır.

Nüfus yoğunluğu, Census nüfus verisinin Gazetteer kara alanı verisine bölünmesiyle hesaplanmıştır. Bu değişken, yalnızca nüfusun taşımadığı bir bilgiyi içerir: aynı nüfusa sahip iki ilçeden biri yoğun kentsel bir alan, diğeri geniş bir kırsal bölge olabilir.

3.1.2. Bağımsız Şehir Çakışmalarının Çözülmesi

Birleştirme sırasında altı kayıta çakışma tespit edilmiştir. Bu kayıtlar, ABD idari yapısında "bağımsız şehir" (independent city) statüsündeki yerleşimlere aittir; bu yerleşimler bir ilçeye bağlı olmadıkları hâlde aynı GEOID aralığında yer alabilmektedir.

Çakışmalar groupby işlemiyle çözülmüştür: aynı GEOID'e ait kayıtlar tek satırda toplanmış, nüfus değerleri toplanmış ve alan değerleri birleştirilmiştir. Kayıtların silinmesi yerine toplanması tercih edilmiştir; silme durumunda ilgili bölgelerdeki kazalar referans tablosuyla eşleşemeyecekti.

İşlem sonucunda 3.138 satırlık county_referans_kendi.csv dosyası üretilmiştir.

3.2. Veri Hazırlama

3.2.1. Parçalı Okuma ve Eyalet Süzme

7,7 milyon satırlık dosya belleğe sığmadığından, pandas kütüphanesinin chunksize parametresi kullanılarak veri 500.000 satırlık parçalar hâlinde okunmuştur. Her parça okunduğu anda Florida, New York ve Minnesota kayıtları süzölmüş, geri kalan satırlar bellekten çıkarılmıştır.

Bu yaklaşım, tüm veriyi belleğe alıp sonra süzmeye kıyasla belirgin bir bellek tasarrufu sağlamıştır. Süzme sonrasında 1.278.270 satır elde edilmiştir.

3.2.2. Hedef Sızıntısı Taşıyan Sütunların Çıkarılması

Hedef sızıntısı (data leakage), bir girdi değişkeninin hedef değişken hakkında doğrudan bilgi taşıması durumudur. Bu tür değişkenler modele verildiğinde eğitim ve test başarımı yapay olarak yükselir; ancak model gerçek kullanım koşullarında geçersiz hâle gelir.

Ham veri setinde iki sütun bu niteliktedir:

- Distance(mi): Kazanın trafiği kaç mil boyunca etkilediğini gösterir. Severity zaten trafiğe etki derecesini ölçtüğünden, bu iki değişken aynı olguyu ifade etmektedir.
- End_Time: Kazanın etkisinin sona erdiği zamandır. Kaza anında bilinemez; ancak kaza sonlandıktan sonra ölçülebilir.

Her iki sütun da veri hazırlama aşamasında çıkarılmıştır. Bu sütunlarla birlikte, çalışmanın kapsamı dışında kalan veya bilgi taşımayan toplam 11 sütun veri setinden çıkarılmıştır.

3.2.3. Eksik Değer Yönetimi

Eksik değerler, değişkenin niteliğine göre üç farklı yaklaşımla doldurulmuştur.

Değişken Türü	Doldurma Yöntemi	Gerekçe
Yağış miktarı	0 ile doldurma	Yağış ölçümünün bulunmaması, çoğunlukla yağış olmadığı anlamına gelmektedir
Diğer sayısal değişkenler	Medyan	Ortalamanın aksine medyan uç değerlerden etkilenmez
Metinsel değişkenler	"Bilinmiyor" etiketi	Bilginin yokluğu ayrı bir kategori olarak korunmuştur

Tablo 3.2.3.1. Eksik değer doldurma stratejisi

İşlem sonrasında veri setinde eksik değer kalmamıştır.

3.2.4. Referans Tablosuyla Birleştirme

Temizlenmiş kaza verisi, ilçe referans tablosuyla GEOID üzerinden birleştirilmiştir. Eşleşme oranı %99,99 olarak gerçekleşmiştir. Eşleşmeyen kayıtlar, referans tablosunda karşılığı bulunmayan ilçelere ait olup toplam veri içindeki payları binde birin altındadır.

Birleştirme sonucunda veri setine üç yeni değişken eklenmiştir: nufus_2022, nufus_yogunlugu ve kentsel_kirsal.

3.2.5. Türetilen Değişkenler

Ham veri setindeki bilgilerden modele doğrudan verilmeye uygun olmayanlar, kullanılabilir değişkenlere dönüştürülmüştür.

Değişken	Kaynak	Açıklama
saat	Start_Time	Kazanın gerçekleştiği saat (0–23)
gun	Start_Time	Haftanın günü (0–6)
ay	Start_Time	Ay bilgisi (1–12)
yil	Start_Time	Yıl bilgisi
hafta_sonu	gun	Cumartesi/Pazar ise doğru
yogun_saat	saat	İşe gidiş-geliş saatleri
mevsim	ay	Dört mevsim kategorisi
yol_tipi	Street	Yol adından çıkarılan yol sınıfı

Tablo 3.2.5.1. Türetilen değişkenler

Zaman damgasının parçalanması gereklidir; çünkü ham zaman damgası model tarafından tek bir büyük sayı olarak yorumlanır ve "hangi saatte" ya da "hangi mevsimde" gibi anlamlı örüntüler bu biçimde öğrenilemez.

Yol tipi türetmesi. Yol tipi bilgisi veri setinde ayrı bir sütun olarak bulunmamaktadır; yol adı içerisinde örtük biçimde yer almaktadır ("I-95" otoyol, "US-1" federal yol, "Main St" şehir içi yol gibi). Bu bilgi düzenli ifadeler (regex) kullanılarak çıkarılmıştır.

İlk denemede yol adlarının %27,2'si hiçbir kalıba uymamış ve "Bilinmiyor" kategorisinde kalmıştır. Eşleşmeyen adlar incelenerek kalıp listesi genişletilmiş, bu oran %3,0'a düşürülmüştür. Elde edilen değişken altı kategori içermektedir: Otoyol, Federal yol, Eyalet yolu, İlçe yolu, Şehir içi ve Diğer.

3.3. Keşifsel Veri Analizi

3.3.1. Aykırı Değer Tespiti ve Fiziksel Sınır Kırpması

Betimleyici istatistikler incelendiğinde fiziksel olarak imkânsız değerler tespit edilmiştir: 984 mph rüzgâr hızı ve 174 °F sıcaklık gibi kayıtlar, ölçüm cihazı hatalarına işaret etmektedir.

Aykırı değerlerin temizlenmesinde ilk olarak IQR (çeyrekler açıklığı) yöntemi denenmiştir. Ancak bu yöntem yanıltıcı sonuçlar üretmiştir: yağış miktarı gibi değerlerin büyük çoğunluğunun sıfır olduğu değişkenlerde Q1 ve Q3 değerleri eşit çıkmakta, bu durumda sıfırdan farklı her ölçüm "aykırı" olarak işaretlenmektedir. Yöntem bazı sütunlarda verinin %15'ine varan bir kısmını aykırı olarak nitelendirmiştir.

IQR yerine, her değişken için fiziksel olarak mümkün alt ve üst sınırlar tanımlanarak kırpmaya (clipping) uygulanmıştır. Bu yaklaşımla yalnızca 128 hatalı değer düzeltilmiş, hiçbir satır veri setinden çıkarılmamıştır.

İstatistiksel bir kuralın her veri setine uygulanamayacağı, değişkenin fiziksel anlamının dikkate alınması gerektiği bu adımda ortaya çıkmıştır.

3.3.2. Hedef Değişken Dağılımı

Severity dağılımı ileri derecede dengesizdir (Tablo 2.3.1). Kayıtların %81,69'u Severity 2 sınıfında toplanmaktadır. Eyalet bazında ciddi kaza oranları (Severity 3 ve 4) arasında da farklılık gözlenmiştir; New York en yüksek, Florida en düşük orana sahiptir.

3.3.3. Saat Kırılımı

Saat kırılımı iki farklı örüntü ortaya koymuştur. Kaza sayısı bakımından tepe noktaları 07.00–08.00 ve 16.00–17.00 aralıklarındadır; bu, işe gidiş-geliş trafiğiyle örtüşmektedir.

Ancak ciddi kaza oranı farklı bir dağılım göstermektedir: en yüksek oran akşam saatlerinde, en düşük oran ise öğleden sonra gözlenmiştir. Kazaların en yoğun olduğu saatler ile en ciddi olduğu saatler örtüşmemektedir. Yoğun trafikte kaza sayısı artmakta ancak düşük hız nedeniyle şiddet azalmakta; gece saatlerinde ise kaza sayısı düşerken şiddet yükselmektedir.

Bu bulgu, saat değişkeni ile kaza şiddeti arasındaki ilişkinin doğrusal olmadığını göstermektedir: oran gün içinde önce düşmekte, sonra yükselmektedir.

3.3.4. Yol Tipi Kırılımı

Yol tipi, ciddi kaza oranı bakımından en belirgin ayrımı üreten değişkendir.

Yol Tipi	Ciddi Değil (%)	Ciddi (%)
Federal yol	91,10	8,90
Şehir içi	90,01	9,99
İlçe yolu	83,14	16,86
Eyalet yolu	78,54	21,46
Diğer	73,78	26,22
Otoyol	71,80	28,20

Tablo 3.3.4.1. Yol tipine göre ciddi kaza oranları

Otoyollardaki ciddi kaza oranı, şehir içi yollara kıyasla yaklaşık üç kat yüksektir. Bu fark hız farkıyla açıklanabilir: otoyollarda ortalama hız daha yüksektir ve kaza anındaki kinetik enerji trafiğe etkiyi

artırmaktadır. Ayrıca otoyollarda alternatif güzergâh bulunmaması, tek bir kazanın trafik akışını daha uzun süre etkilemesine yol açmaktadır.

3.3.5. Hava Durumu Kırılımı

Hava durumu değişkeni ham hâliyle 101 farklı kategori içermektedir. Bu kategorilerin önemli bir kısmı birbirinin yakın varyantıdır ("Light Rain", "Light Rain Shower", "Light Drizzle" gibi).

Analiz sırasında bir tutarsızlık tespit edilmiştir: "Clear" ve "Fair" kategorileri aynı hava koşulunu ifade etmelerine rağmen ayrı ayrı kaydedilmiştir. Yıl bazında inceleme, bu ayrımın coğrafi veya zamansal bir farklılıktan değil, veri sağlayıcının terminolojisindeki bir değişiklikten kaynaklandığını göstermiştir: "Clear" ifadesi yalnızca 2016–2019 aralığında, "Fair" ifadesi ise 2019 sonrasında kullanılmıştır.

Bu bulgu, veri kalitesi kontrolünün yalnızca eksik ve aykırı değerlerle sınırlı olmadığını; kategorik değişkenlerin zaman içindeki tutarlılığının da denetlenmesi gerektiğini göstermektedir.

Modelleme aşamasında bu 101 kategori, anahtar kelime eşleştirmesiyle sekiz anlamlı gruba indirgenmiştir: Açık, Bulutlu, Yağmur, Kar/Buz, Fırtına, Sis, Rüzgâr/Toz ve Diğer. Bu gruplama hem seyrek kategorilerin ezberlenmesini önlemiş hem de Clear/Fair ayrımını ortadan kaldırmıştır.

3.3.6. Kentsel/Kırsal Kırılımı

Kentsel/kırsal sınıflandırmasında en yüksek ciddi kaza oranı banliyö bölgelerinde, en düşük oran ise kırsal bölgelerde gözlenmiştir. Bu ilişki tek yönlü değildir: oran şehir merkezinden banliyöye doğru yükselmekte, banliyöden kırsala doğru ise düşmektedir.

Ters-U biçimindeki bu örüntü, doğrusal bir modelin yakalayamayacağı bir yapıdır ve model seçimi açısından belirleyici olmuştur.

3.3.7. Korelasyon Analizi

Sayısal değişkenler arasındaki Pearson korelasyonları incelenmiştir. Elde edilen bulgular üç başlıkta toplanabilir.

Çoklu doğrusal bağlantı bulunmamaktadır. Değişkenler arasındaki en yüksek mutlak korelasyon 0,38 (sıcaklık – basınç) olarak ölçülmüştür. Yaygın olarak kullanılan 0,70 eşiği aşılmadığından, bu gerekçeyle herhangi bir sütun veri setinden çıkarılmamıştır.

Meteorolojik ilişkiler beklenen yöndedir. Nem – görüş mesafesi (–0,38), nem – saat (–0,28) ve sıcaklık – görüş mesafesi (0,29) ilişkileri fiziksel beklentilerle uyumludur. Bu tutarlılık, aykırı değer kırpması sonrasında verinin fiziksel anlamını koruduğunu doğrulamaktadır.

Hedef değişkenle doğrusal ilişkiler zayıftır. Severity ile hiçbir sayısal değişken arasında $|r| \geq 0,09$ düzeyinde bir korelasyon bulunmamıştır. En yüksek değer basınç değişkenine aittir (0,088).

Bu üçüncü bulgu ilk bakışta veride öngörü gücü bulunmadığı izlenimi vermektedir. Ancak kategori kırılımlarındaki farklar bunun aksini göstermektedir: kentsel/kırsal değişkeninin Severity ile korelasyonu –0,004 olmasına rağmen kategoriler arasında yaklaşık iki katlık bir oran farkı bulunmaktadır. Benzer biçimde saat değişkeninin korelasyonu 0,015 iken, saatler arasındaki ciddi kaza oranı farkı belirgindir.

Bu çelişkinin nedeni, Pearson korelasyonunun yalnızca doğrusal (monoton) ilişkileri ölçebilmesidir. Saat değişkenindeki dalga biçimli ve kentsel/kırsal değişkenindeki ters-U biçimli ilişkilerde, yükselen ve alçalan bölümler hesaplamada birbirini götürmekte ve katsayı sıfıra yaklaşmaktadır.

Ayrıca basınç değişkeninin Severity ile göreceli olarak yüksek korelasyonu doğrudan meteorolojik bir etki olarak yorumlanmamalıdır. Basınç ile kentsel/kırsal (–0,32) ve basınç ile sıcaklık (0,38) ilişkileri, bu değişkenin rakım ve konum bilgisini dolaylı olarak taşıdığını göstermektedir.

Bu bölümün model seçimine etkisi: doğrusal ilişkilerin zayıf, kategorik farkların ise belirgin olması, doğrusal olmayan yapıları öğrenebilen ağaç tabanlı bir modelin gerekli olduğu hipotezini doğrumuştur. Bu hipotez modelleme aşamasında sınanmıştır.

3.4. Modelleme

3.4.1. Özellik Seçimi

Modelleme öncesinde, öğrenmeye katkı sağlamayacak veya modeli yanıltacak sütunlar veri setinden çıkarılmıştır. Çıkarma gerekçeleri iki başlıkta toplanmaktadır.

Yüksek kardinalite. Bir kategorik değişkendeki farklı değer sayısı çok yüksek olduğunda, model genellenebilir bir örüntü öğrenmek yerine tekil kayıtları ezberler. Eğitim başarımları yükselirken test başarımları düşer (aşırı öğrenme).

Bilgi tekrarı. Taşıdığı bilgi başka bir değişken tarafından zaten temsil edilen sütunlar modele ek katkı sağlamaz.

Sütun	Farklı Değer	Çıkarma Gerekçesi
ID	1.278.270	Her kayıt tekil; tamamen ezberleme
Zipcode	135.500	Posta kodu başına ortalama on kayıt; örüntü oluşmaz
Street	68.825	Aynı gerekçe; yol tipi bilgisi zaten türetilmiştir
City	2.334	Bilgisi nüfus ve nüfus yoğunluğu ile temsil edilmektedir
County	207	Bilgisi nüfus ve kentsel/kırsal değişkenleriyle temsil edilmektedir
GEOID, county_key	—	Birleştirme anahtarları; sayısal değerleri anlamsızdır
Timezone	3	State değişkeniyle aynı bilgiyi taşımaktadır
Wind_Direction	25	Kaza şiddetiyle nedensel bağı bulunmamaktadır
Start_Time	—	Bileşenleri ayrı değişkenler olarak türetilmiştir
Start_Lat, Start_Lng	—	Coğrafi ezberleme riski; konum başka değişkenlerle temsil edilmektedir

Tablo 3.4.1.1. Modelden çıkarılan sütunlar

Şehir ve ilçe adlarının çıkarılması bilgi kaybı anlamına gelmemektedir. Referans tablosunun oluşturulmasındaki amaç tam olarak budur: modelin bölge adını ezberlemesi yerine, o bölgenin ölçülebilir özelliklerini (nüfus, yoğunluk, kentsel/kırsal sınıf) öğrenmesi hedeflenmiştir. Ad ezberlenir, özellik genellenir.

Sayısal değişkenlerdeki yüksek farklı değer sayıları (örneğin sıcaklık için 713) sorun oluşturmamaktadır; ağaç tabanlı modeller sayısal değişkenlerde eşik bölmesi yapmakta, kategori ezberlememektedir.

Bu işlemler sonucunda modele 35 değişken verilmiştir.

3.4.2. Kategorik Değişkenlerin Kodlanması

Makine öğrenmesi algoritmaları metin girdisi işleyemediğinden, kategorik değişkenlerin sayısal biçime dönüştürülmesi gerekmektedir. İki yaygın yöntem değerlendirilmiştir.

One-hot kodlama her kategori için ayrı bir sütun oluşturur. 1,28 milyon satırlık bir veri setinde bu yöntem bellek açısından uygulanabilir değildir.

Ordinal kodlamada her kategoriye bir tam sayı atanır ve sütun sayısı değişmez. Bu yöntemin bilinen sakıncası, modelin atanan sayılar arasında bir sıralama olduğunu varsayabilmesidir. Ancak ağaç tabanlı modeller eşitlik sorgusu ("kategori değeri 3 mü?") yaptıklarından bu varsayımı taşımaz. Bu nedenle ordinal kodlama tercih edilmiştir.

Eğitim setinde görülmeyen bir kategoriyle test aşamasında karşılaşılmaması durumunda hata oluşmaması için `handle_unknown` parametresi yapılandırılmıştır.

3.4.3. Eğitim/Test Ayrımı

Veri seti %80 eğitim ve %20 test olarak ayrılmıştır. Test seti model eğitimi sırasında hiçbir biçimde kullanılmamış; yalnızca nihai başarımlar ölçümünde devreye alınmıştır. Bu ayırım, modelin daha önce görmediği veriler üzerindeki performansını ölçmeyi sağlar.

Küme	Satır Sayısı	Değişken Sayısı
Eğitim	1.022.616	35
Test	255.654	35

Tablo 3.4.3.1. Eğitim ve test kümesi boyutları

Bölme işleminde stratify parametresi kullanılmıştır. Severity 1 sınıfı verinin yalnızca %0,66'sını oluşturduğundan, rastgele bölme bu sınıfın kümeler arasında dengesiz dağılmasına yol açabilirdi. Stratify, her iki kümede de orijinal sınıf oranlarını korumaktadır.

Ayrıca `random_state` parametresi sabitlenmiştir. Bölme işlemi rastgeledir; sabitlenmediğinde her çalıştırmada farklı sonuçlar elde edilir ve raporlanan değerler yeniden üretilemez hâle gelir.

3.4.4. Model 1 — Baseline (DummyClassifier)

İlk model bir öğrenme modeli değil, karşılaştırma referansıdır. `DummyClassifier` girdi değişkenlerine hiç bakmaz ve her kayda çoğunluk sınıfını atar. Amacı, diğer modellerin başarımının anlamlı olup olmadığını değerlendirebilmek için bir alt sınır oluşturmaktır.

Ölçüt	Değer
Accuracy	0,8169
Macro-F1	0,2248

Tablo 3.4.4.1. Baseline model başarımı

Bu iki değer arasındaki fark, çalışmanın başında yapılan ölçüt seçimini doğrudan doğrulamaktadır. Hiçbir öğrenme gerçekleştirilmeyen bir model %81,69 doğruluk elde etmektedir; aynı modelin macro-F1 değeri ise 0,2248'dir. Doğruluk ölçütü bu veri setinde model başarısını değil sınıf dağılımını ölçmektedir.

Sonraki modeller için aşılması gereken eşik 0,2248 olarak belirlenmiştir.

3.4.5. Model 2 — Lojistik Regresyon

İkinci model, keşifsel analiz aşamasında kurulan "ilişkiler doğrusal değildir" hipotezini sınamak amacıyla kurulmuş bir kontrol grubudur. Lojistik regresyon, sınıflar arasında doğrusal bir ayırım sınırı çizer.

Doğrusal modeller değişken ölçeğine duyarlı olduğundan, eğitim öncesinde standardizasyon uygulanmıştır. Nüfus değişkeni milyonlar mertebesinde, görüş mesafesi ise 0–10 aralığındadır; ölçekleme yapılmadığında model büyük sayılı değişkeni yapay olarak daha önemli sayacaktır.

Standardizasyon istatistikleri (ortalama ve standart sapma) yalnızca eğitim setinden hesaplanmış, test setine bu istatistiklerle dönüşüm uygulanmıştır. Aksi hâlde test verisine ait bilgi eğitim sürecine sızacaktı.

Ölçüt	Eğitim	Test
Macro-F1	0,3093	0,3099
Accuracy	—	0,4861

Tablo 3.4.5.1. Lojistik regresyon başarımı

Model baseline değerini aşmıştır ($0,3099 > 0,2248$); veride bir miktar doğrusal sinyal bulunmaktadır. Eğitim ve test skorlarının neredeyse eşit olması, modelin ezberleme yapmadığını göstermektedir. Ancak bu durum aynı zamanda modelin yetersiz öğrenme (underfitting) durumunda olduğuna işaret etmektedir: doğrusal yapı, veri setindeki örüntüleri temsil edebilecek karmaşıklıkta değildir.

3.4.6. Model 3 — Random Forest

Üçüncü model ağaç tabanlı bir topluluk yöntemidir. Random Forest, birbirinden bağımsız çok sayıda karar ağacı kurar ve tahmini bu ağaçların oylamasıyla belirler (bagging). Tek bir karar ağacının ezberleme eğilimi, bu şekilde dengelenir.

Ağaç tabanlı modeller veriyi ardışık eşik değerleriyle böler. Bu yapı, doğrusal olmayan ilişkilerin öğrenilmesini mümkün kılar; keşifsel analizde tespit edilen dalga biçimli (saat) ve ters-U biçimli (kentsel/kırsal) örüntüler bu yöntemle temsil edilebilmektedir.

Parametre	Değer	Gerekçe
n_estimators	100	Oylamaya katılacak ağaç sayısı
max_depth	20	Ağaç derinliğini sınırlayarak ezberlemeyi kısıtlar
min_samples_leaf	50	Her yaprakta en az 50 örnek; tekil kayıtlar için kural üretilmesini engeller
class_weight	balanced	Seyrek sınıfların çoğunluk sınıfı tarafından bastırılmasını önler
random_state	42	Sonuçların yeniden üretilebilirliği

Tablo 3.4.6.1. Random Forest model parametreleri

Ölçüt	Eğitim	Test
Macro-F1	0,4522	0,4331
Accuracy	—	0,7019

Tablo 3.4.6.2. Random Forest başarımı

Eğitim ve test macro-F1 değerleri arasındaki fark 0,019'dur. Bu düşük fark, modelin aşırı öğrenme göstermediğini doğrulamaktadır; derinlik ve yaprak büyüklüğü kısıtları amacına ulaşmıştır.

Sınıf bazında başarımların değerleri aşağıdaki tabloda sunulmuştur.

Severity	Precision	Recall	F1-Skoru	Örnek Sayısı
1	0,113	0,927	0,201	1.690
2	0,956	0,702	0,810	208.839
3	0,459	0,712	0,558	40.342
4	0,097	0,522	0,164	4.783
Makro ortalama	0,406	0,716	0,433	255.654
Ağırlıklı ortalama	0,856	0,702	0,754	255.654

Tablo 3.4.6.3. Random Forest sınıf bazında başarımlar değerleri

Tablo, precision ve recall arasındaki dengeyi göstermektedir. Precision, "bu sınıfa atanan kayıtların ne kadarı gerçekten o sınıfa aittir" sorusunu; recall ise "o sınıfa ait kayıtların ne kadarı yakalanabilmiştir" sorusunu yanıtlar.

Severity 3 sınıfında model dengeli bir başarımlar göstermiştir ($F1 = 0,558$). Severity 2 sınıfında precision yüksek (0,956), recall ise görece düşüktür (0,702); bu, class_weight parametresinin beklenen etkisidir. Model çoğunluk sınıfını atama konusunda daha seçici davranmakta, böylece seyrek sınıflara alan açmaktadır.

Severity 1 ve 4 sınıflarında recall yüksek (0,927 ve 0,522), precision ise düşüktür (0,113 ve 0,097). Model bu sınıflara ait kayıtların büyük bölümünü yakalamakta, ancak çok sayıda yanlış pozitif üretmektedir. Bu, seyrek sınıflara yüksek ağırlık verilmesinin doğal bir sonucudur ve sınıf dengesizliğinin bu ölçekte olduğu veri setlerinde kaçınılmaz bir ödünleşimdir.

3.4.7. Model Karşılaştırması

Model	Tür	Test Macro-F1	Accuracy
Baseline (DummyClassifier)	Referans	0,2248	0,8169
Lojistik Regresyon	Doğrusal	0,3099	0,4861
Random Forest	Ağaç tabanlı	0,4331	0,7019

Tablo 3.4.7.1. Model karşılaştırma tablosu

Tablo üç ayrı sonucu bir arada göstermektedir.

Ölçüt seçiminin doğrulanması. En yüksek doğruluk değeri (0,8169) hiçbir öğrenme yapmayan baseline modele aittir. Random Forest'ın doğruluğu daha düşük (0,7019) olmasına rağmen macro-F1 değeri yaklaşık iki katıdır. Baseline, seyrek sınıfları tamamen göz ardı ederek doğruluk kazanmakta; Random Forest ise bu sınıfları tahmin etmeye çalışarak doğruluktan ödün vermektedir. Bu, doğruluk ölçütünün dengesiz veri setlerinde neden kullanılmaması gerektiğinin somut göstergesidir.

Doğrusal olmayan yapının doğrulanması. Ağaç tabanlı model, doğrusal modele kıyasla %39,8 daha yüksek macro-F1 değeri üretmiştir (0,4331 / 0,3099). Bu fark, keşifsel analiz aşamasında ortaya konan hipotezin — korelasyonların sıfıra yakın olmasına rağmen kategorik farkların belirgin olması, dolayısıyla ilişkilerin doğrusal olmaması — sayısal doğrulamasıdır.

Aşırı öğrenme kontrolü. Random Forest için eğitim ve test macro-F1 değerleri arasındaki fark 0,019'dur. Model, eğitim verisini ezberlemek yerine genellenebilir örüntüler öğrenmiştir.

Bu bulgular doğrultusunda nihai model olarak Random Forest seçilmiştir.

4. SONUÇ

Bu çalışmada, ABD trafik kazaları veri seti kullanılarak kaza anındaki çevresel koşullardan kaza şiddetinin tahmin edilmesi amaçlanmıştır. Florida, New York ve Minnesota eyaletlerine ait 1.278.270 kayıt üzerinde, veri hazırlamadan model değerlendirmesine uzanan eksiksiz bir makine öğrenmesi süreci yürütülmüştür.

Ham veri seti, resmî nüfus ve idari sınıflandırma kaynaklarıyla ilçe bazında zenginleştirilmiş; zaman damgasından ve yol adından yeni değişkenler türetilmiştir. Hedef sızıntısı taşıyan sütunlar model kurulmadan önce çıkarılmış, aykırı değerler istatistiksel bir kural yerine fiziksel sınırlar dikkate alınarak düzeltilmiştir.

Keşifsel veri analizi iki temel bulgu üretmiştir. Birincisi, yol tipinin ciddi kaza oranı üzerinde belirgin bir etkiye sahip olmasıdır; otoyollardaki oran şehir içi yollara kıyasla yaklaşık üç kat yüksektir. İkincisi ve yöntemsel açıdan daha belirleyici olanı, değişkenlerle hedef arasındaki doğrusal korelasyonların tamamının sıfıra yakın çıkmasına rağmen kategori kırılımlarında iki katına varan farkların gözlenmesidir. Bu çelişki, ilişkilerin var olduğunu ancak doğrusal olmadığını göstermiştir.

Modelleme aşamasında bu hipotez sınanmıştır. Öğrenme yapmayan referans model 0,2248 macro-F1 değeri üretmiş; doğrusal model bu değeri 0,3099'a, ağaç tabanlı model ise 0,4331'e yükseltmiştir. Doğrusal ve doğrusal olmayan modeller arasındaki yaklaşık %40'lık fark, keşifsel analizde ortaya konan öngörüğü doğrulamıştır.

Çalışma ayrıca başarımlar ölçütü seçiminin önemini somut biçimde ortaya koymuştur. Hiçbir öğrenme gerçekleştirilmeyen referans model %81,69 doğruluk elde etmiş, buna karşın macro-F1 değeri 0,2248'de kalmıştır. Nihai model, doğruluk bakımından referans modelin gerisinde kalmasına rağmen macro-F1 bakımından yaklaşık iki katı başarımlar göstermiştir. Dengesiz veri setlerinde doğruluk ölçütünün model başarısını değil sınıf dağılımını yansıttığı bu karşılaştırmayla gösterilmiştir.

Sınırlılıklar ve Gelecek Çalışmalar

Modelin seyrek sınıflardaki başarımları sınırlıdır. Severity 1 ve 4 sınıflarında recall değerleri yüksek olmasına rağmen precision düşüktür; model bu sınıfları yakalamakta ancak çok sayıda yanlış pozitif üretmektedir. Sınıf dengesizliğinin bu ölçekte olduğu veri setlerinde bu bir ödünleşimdir.

Çalışma üç eyaletle sınırlandırılmıştır. Modelin farklı iklim ve yol yapısına sahip bölgelerdeki genellenebilirliği ayrıca sınanmalıdır.

Veri setinde tespit edilen terminoloji değişikliği (Clear/Fair ayrımı), veri toplama yönteminin zaman içinde değiştiğine işaret etmektedir. Bu tür ölçüm kaynaklı tutarsızlıkların model başarımına etkisinin ayrı bir çalışmada incelenmesi yararlı olacaktır.

Gelecek çalışmalarda hiperparametre optimizasyonu (GridSearchCV veya RandomizedSearchCV), gradyan artırılmalı yöntemler (LightGBM, XGBoost) ve sınıf dengesizliğine yönelik örnekleme teknikleri (SMOTE gibi) denenerek başarımın artırılması hedeflenebilir.

5. KAYNAKÇA

- [1] Moosavi, S. (2023). US Accidents (2016 – 2023). Kaggle. <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>
- [2] Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A Countrywide Traffic Accident Dataset. arXiv:1906.05409.
- [3] United States Census Bureau. County Population Totals and Components of Change: 2020–2024. <https://www.census.gov/programs-surveys/popest.html>
- [4] United States Census Bureau. Gazetteer Files — Counties. <https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html>
- [5] National Center for Health Statistics (NCHS). NCHS Urban–Rural Classification Scheme for Counties. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/data_access/urban_rural.htm
- [6] Pedregosa, F. ve diğerleri (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
- [7] McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51–56.
- [8] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
- [9] Sazan, D. (2026). ABD Kaza Tahmini — Proje Deposu. GitHub. <https://github.com/Dilan-Sazan/abd-kaza-tahmini>