**VERİ MADENCİLİĞİNE GİRİŞ DÖNEM ÖDEVİ**

**daire, amblem, ticari marka, logo içeren bir resim

Yapay zeka tarafından oluşturulmuş içerik yanlış olabilir.**

**T.C.**

**YILDIZ TEKNİK ÜNİVERSİTESİ**

**FEN EDEBİYAT FAKÜLTESİ**

**İSTATİSTİK BÖLÜMÜ**

**VERİ MADENCİLİĞİNE GİRİŞ 2024-2025**

Los Angeles Suç Verisi Analizi

**Hazırlayanlar:**

**Selin ÖZKAN 24023652**

**Şükrü Kaan ÖZCAN 24023650**

**Dersi Veren Öğretim Üyesi**

**Prof. Dr. Ali Hakan BÜYÜKLÜ**

1. **Giriş**

**• Konunun Tanıtımı**

Bu proje kapsamında Los Angeles'ta 2020–2023 yılları arasında işlenmiş suçlara ilişkin kamuya açık bir veri seti analiz edilmiştir. Projede temel amaç, suçların mekânsal ve zamansal dağılımlarını tespit etmek ve veri madenciliği yöntemleriyle anlamlı desenler ortaya çıkarmaktır.

**• Konunun Önemi**

Kent güvenliği, kaynak planlaması ve toplum refahı açısından suç verilerinin analizi büyük önem taşımaktadır. Bu analizlerle suçun türü, zamanı ve bölgesi belirlenerek karar vericilere katkı sağlanabilir.

**• Ödevin Amacı ve Hedefleri**

* En sık işlenen suçların belirlenmesi
* Zaman serisi analiziyle mevsimsel değişimlerin izlenmesi
* Anomali tespiti ve kümelenme ile örüntü çıkarımı
* Görselleştirme ile bilgilerin daha açık sunumu

**• Kapsam**

Proje 2020–2023 arası Los Angeles suç verisiyle sınırlıdır. Şahıs bazlı veriler (fail bilgileri) dışta bırakılmıştır.

1. **Literatür Özeti**

Suç verisi analizi, veri madenciliği ve makine öğrenmesi tekniklerinin önemli uygulama alanlarından biridir. Literatürde suç tahmini, mekânsal suç yoğunluğu ve zamansal desenlerin keşfi gibi konular sıklıkla ele alınmıştır.

**Ratcliffe (2004)**, suçların belirli bölgelerde ve zamanlarda yoğunlaştığını ortaya koyarak “hotspot” kavramını öne çıkarmıştır. **Wang et al. (2013)** ise K-means algoritması ile şehir içi suç bölgelerini kümelendirerek mekânsal analizlerin önemini göstermiştir.

**Chen et al. (2017)** tarafından yapılan zaman serisi analizinde, mevsimsel dalgalanmaların suç oranlarını etkilediği tespit edilmiştir. Aynı şekilde **McKinney (2022)** ve **Géron (2019)** gibi kaynaklar, Python ortamında veri analizi ve makine öğrenmesi yöntemlerinin suç verisi üzerindeki uygulamalarını detaylı şekilde açıklamaktadır.

Random Forest algoritmasının suç verilerinde yüksek performans göstermesi nedeniyle son yıllarda birçok çalışmada tercih edildiği gözlemlenmiştir (**Breiman, 2001**).

Bu projede literatürde sık kullanılan yöntemler uygulanmış, ancak özellikle zaman, mağdur profili ve mekânsal bilgiler birleştirilerek daha kapsamlı bir analiz gerçekleştirilmiştir.

1. **Teorik Arka Plan**

**• Temel Kavramlar**

* **Sınıflandırma:** Girdi değişkenlere göre suç türünü tahmin etme
* **Kümeleme:** Benzer olayların gruplandırılması
* **Zaman Serisi Analizi:** Suçların zaman içindeki değişimini inceleme
* **Anomali Tespiti:** Alışılmışın dışındaki olayların belirlenmesi

**• Kullanılan Teknik ve Algoritmalar**

* K-Means Kümeleme
* Zaman Serisi Trend ve Anomali Analizi
* Korelasyon analizi
* Regresyon analizi

1. **Veri Seti Açıklaması**

**• Veri Setinin Adı ve Kaynağı**

* **Adı:** Los Angeles Crime Data (2020–2023)
* **Kaynağı:** Kaggle – Open Data LA

**• Gözlem ve Değişken Sayısı**

* **Gözlem Sayısı:** ~800.000+ satır
* **Değişken Sayısı:** ~20 değişken
  + **Öne Çıkan Değişkenler ve Türleri**

|  |  |  |
| --- | --- | --- |
| **Değişken Adı** | **Açıklama** | **Veri Türü** |
| DR\_NO | Olay numarası | Nominal |
| Date OCC | Olayın tarihi | Tarih |
| Time OCC | Olay saati (saat ve dakika) | Sayısal |
| AREA NAME | Suç bölgesi | Kategorik (nominal) |
| Crm Cd | Suç kodu | Sayısal (kategorik kod) |
| Crm Cd Desc | Suç açıklaması | Kategorik (nominal) |
| Vict Age | Mağdurun yaşı | Sayısal (nümerik) |
| Vict Sex | Mağdurun cinsiyeti | Kategorik (nominal) |
| Vict Descent | Mağdurun etnik kökeni | Kategorik(nominal) |
| Weapon Used Cd | Kullanılan silah kodu (varsa) | Kategorik (kod) |
| Weapon Desc | Silah türü açıklaması | Kategorik |
| Premis Desc | Olayın gerçekleştiği yer (ev, sokak vb.) | Kategorik |
| Status | Olay durumu (tamamlandı, çözülmedi vs.) | Kategorik |
| LAT, LON | Enlem, boylam (konum) | Sürekli (float) |
| Date Rptd | Olayın rapor edildiği tarih | Tarih |
| MO Codes | Suça dair jargon kodları (modus operandi) | Kategorik |

* **Üretilen Değişken**

|  |  |  |
| --- | --- | --- |
| **Yeni değişken** | **Açıklama** | **Türü** |
| Date Time OCC | Tarih ve saat birleşimi |  |
| Year, Month, Hour, DayOfWeek | Zaman analizi için ayrıştırılan sütunlar | Sayısal/Kategorik |

1. **Veri Temizliği ve Ön İşleme**

* Eksik değerler kontrol edilip Weapon Used Cd, Weapon Desc gibi sütunlardaki boşluklar "Unknown" ile dolduruldu.
* Date OCC ve Time OCC birleştirilerek Date Time OCC sütunu oluşturuldu.
* Vict Age değişkeni için aykırı değerler filtrelendi (örneğin 0 veya 120 üzeri yaşlar çıkarıldı).
* Crm Cd Desc sütunu üzerinden en sık rastlanan suçlar analiz için filtrelendi.

1. **Yöntem ve Uygulama**

* **Kullanılan Teknikler**
  + Zaman Serisi Görselleştirme
  + Anomali Tespiti
  + K-Means Kümeleme
  + Korelasyon Analizi
  + Regresyon Analizi
  + Heatmap, Lineplot gibi grafiksel yöntemler
* **Araçlar** 
  + Pandas: Veri okuma, temizleme ve dönüştürme işlemleri
  + Matplotlib & Seaborn: Grafik ve görselleştirme
  + Scikit-Learn: K-Means algoritması
  + Folium: Harita üzerinde suç yoğunluklarının gösterimi
* **Uygulama**
  + Veriler Pandas ile yüklendi ve datetime sütunları oluşturuldu.
  + `Crm Cd Desc` üzerinden en sık görülen suçlar filtrelendi.
  + Saat, bölge ve suç türüne göre veri kümeleme için hazırlandı.
  + K-Means algoritması uygulanarak en uygun küme sayısı belirlendi (Elbow yöntemi).
  + Kümeleme sonuçları yorumlandı ve grafiklerle desteklendi.
  + Zaman serisi grafikleri ile aylık ve haftalık suç trendleri görselleştirildi.
  + Korelasyon matrisleri ile değişkenler arası ilişkiler incelendi.
  + Anomali analizinde özellikle pandemi döneminde suçlardaki düşüşler tespit edildi.
  + Projede ayrıca zaman serisi verisine dayalı olarak basit doğrusal regresyon analizi uygulanmıştır. Amaç, belirli zaman dilimlerinde suç sayısının trendini incelemek ve geleceğe yönelik tahminlerde bulunabilmektir.

1. **Sonuçlar ve Yorum**

Bu bölümde projeden elde edilen bulgular, görseller eşliğinde değerlendirilmiş; kullanılan modelin başarı düzeyi ve sınırlılıkları analiz edilmiştir

**Zaman ve Mekâna Dayalı Bulgular**

* **Yıllık trende** göre 2020–2022 arasında suç sayısı artmış, 2023'te veri eksikliği nedeniyle keskin bir düşüş gözlemlenmiştir.
* **Aylık dağılımda**, ilkbahar ve yaz aylarında suç oranı daha yüksektir.
* **Cuma ve Cumartesi günleri**, haftalık bazda en yüksek suç oranlarının gözlendiği günlerdir.
* **Saatlik analizde**, suçlar öğle sonrası ve akşam saatlerinde yoğunlaşmaktadır.
* **Central ve 77th Street** bölgeleri suç oranı açısından en riskli alanlardır.

**Suç Türleri ve Mağdur Profili**

* En sık işlenen suçlar arasında *araç hırsızlığı, basit saldırı ve kimlik hırsızlığı* öne çıkmaktadır.
* Mağdurların büyük kısmı *Hispanik ve Beyaz* kökenlidir; erkek mağdurlar sayıca kadınlardan fazladır.

**Regresyon Modeli (Suç Sayısı Tahmini)**

* **R² skoru = 0.0866**, yani model varyansın %8.66’sını açıklayabilmiştir.
* Model, zaman bilgileriyle (gün, ay, yıl) sınırlı olduğu için zayıf performans göstermiştir.
* Sonuç: Suç sayısını sadece zaman değişkenleriyle tahmin etmek yetersizdir.

**Sınıflandırma Modeli (Suç Türü Tahmini)**

* **Model:** Random Forest Classifier
* **Amaç:** Suçun Part 1 (ağır) veya Part 2 (hafif) olmasını tahmin etmek
* **Accuracy:** 0.64
* **F1-score:** Part 1 için 0.70, Part 2 için 0.55
* Model Part 1 suçlarını daha başarılı tahmin etmiş, Part 2 sınıfı için sınırlı kalmıştır.

1. **Sonuç ve Değerlendirme**

**Proje kapsamında veri seti çok boyutlu şekilde analiz edilmiştir. Zamansal, mekânsal, demografik ve tür bazlı dağılımlar görselleştirilerek suç örüntüleri belirlenmiştir. Zaman serisi analizi, regresyon ve sınıflandırma teknikleriyle öngörülebilirlikler test edilmiştir.**

* Görselleştirmeler sayesinde örüntüler açıkça belirlenmiştir.
* Regresyon modeli sınırlı başarı göstermiştir; sınıflandırma modeli daha anlamlı sonuçlar vermiştir.

**Eksiklikler ve Sınırlılıklar**

- 2023 yılı verisi Temmuz ayından sonra eksiktir; bu da analizde dengesizlik yaratmıştır.  
- Kullanılan bazı değişkenler (örneğin dayofweek, hour) suç türünü açıklamada sınırlı kalmıştır.  
- Etnik köken ve cinsiyet gibi değişkenler belirli oranda eksik veri içermektedir.  
- Mekânsal analiz sınırlı sayıda bölgeyle yapılmıştır; daha detaylı bölge verisi sınıflandırmayı güçlendirebilir.

**Gelecek Çalışmalar İçin Öneriler**

**-Coğrafi bilgi sistemleri (GIS)** ve harita tabanlı analizlerle suç kümeleri daha hassas modellenebilir.  
-Model performansını artırmak için **demografik ve sosyoekonomik** değişkenler dahil edilebilir.  
-Feature importance analizi ile hangi değişkenlerin tahminde etkili olduğu daha net ortaya konabilir.

1. **Kaynakça**

* **Kaggle.** (2023). Los Angeles Crime Data (2020–2023).  
  **(**[**https://www.kaggle.com/datasets/venkatsairo4899/los-angeles-crime-data-2020-2023**](https://www.kaggle.com/datasets/venkatsairo4899/los-angeles-crime-data-2020-2023)**)**
* **Kaggle, açık projeler  
  (https://www.kaggle.com/code/vinayak121/eda-los-angles-crime)**
* Scikit-learn documentation. (2024). [*https://scikit-learn.org*](https://scikit-learn.org)
* Github projesi   
  (<https://github.com/rahulodedra30/Los-Angeles-Crime-Data-Analysis-and-Forecasting/blob/main/Jupyter%20Notebooks/Data_Visualization.ipynb>)
* Ratcliffe, J. H. (2004). The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. Police Practice and Research, 5(1), 5–23. <https://doi.org/10.1080/1561426042000181343>
* Wang, F., Luo, W., & Zhang, X. (2013). Spatial-temporal analysis of urban crime using self-organizing maps and K-means. Computers, Environment and Urban Systems, 39, 93–102. <https://doi.org/10.1016/j.compenvurbsys.2013.01.002>
* Chen, P., Yuan, H., & Shu, Y. (2017). Spatiotemporal patterns and detection of crime hotspots in a metropolitan city using space-time scan statistics. ISPRS International Journal of Geo-Information, 6(5), 132. <https://doi.org/10.3390/ijgi6050132>
* McKinney, W. (2022). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media.
* Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.
* Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
* Chatgpt

1. **Ekler**

* Grafik çıktıları
* Kod dosyası
* Haritalar

-EKLER, dosyada bulunmaktadır.