Proje Başlığı: Makine Öğrenmesi ile Chicago Suç Tahmini ve Analizi

- **Kerem Aydın - 090200315**

---

1. Giriş

Son yıllarda suç oranlarının arttığı ve kamu güvenliği sorunlarının ciddi şekilde tartışıldığı büyük şehirlerden biri de Chicago'dur. Şehir, Amerika Birleşik Devletleri'nin en büyük ve en kalabalık şehirlerinden biri olarak, suç oranları konusunda da dikkat çekmektedir. Suçları tahmin etmek ve önceden önlem almak, şehir yönetimleri ve güvenlik güçleri için kritik öneme sahiptir. Ancak suç tahminleri yapmak, sadece geçmiş verilerin analizi ile sınırlı kalmaz; aynı zamanda çevresel faktörlerin, ekonomik durumların ve sosyo-kültürel etmenlerin de dikkate alınması gerekir. Bu noktada, makine öğrenmesi ve yapay zeka, suç verilerini analiz etmek, suç oranlarını tahmin etmek ve suçların önlenmesine yönelik stratejiler geliştirmek için etkili araçlar sunmaktadır.

Chicago'daki suç verileri, çok sayıda faktörün suçları etkileyebileceğini gösteriyor. Bu faktörler arasında zaman, yer, hava durumu, tatil günleri ve toplumsal etkinlikler gibi değişkenler bulunur. Bu projede, Chicago'daki suç verisini analiz ederek, suç oranlarını etkileyen temel faktörleri inceleyecek, bu faktörlerle ilgili modeller geliştirecek ve gelecekteki suç oranlarını tahmin etmeye çalışacağız. Proje, suç tahmin modelleri geliştirmek ve bu modellerin doğruluğunu test etmek için makine öğrenmesi algoritmalarından yararlanacaktır.

---

2. Proje Hedefleri

Projenin temel hedefi, Chicago'daki suçları anlamak ve gelecekteki suçları tahmin etmek için makine öğrenmesi tekniklerini kullanmaktır. Projenin ayrıntılı hedefleri aşağıda sıralanmıştır:

- **Veri Toplama ve Ön İşleme:** Suç verisi, Chicago'daki suç türlerinin, zamanlarının ve coğrafi yerlerinin detaylı verilerini içermektedir. Bu veriler, proje için ilk aşamada toplanacak ve daha sonra analiz için işlenecektir. Eksik veriler belirlenecek ve uygun şekilde doldurulacak ya da çıkarılacaktır. Ayrıca, hatalı veriler temizlenecek ve modelleme için hazır hale getirilecektir.
- **Veri Analizi ve Keşifsel Veri Analizi (EDA):** EDA, verilerdeki ilişkileri ve anlamlı desenleri ortaya çıkarmak için yapılacaktır. Suç türlerinin, suçların yoğunlaştığı bölgelerin ve suçların zaman içindeki değişiminin analizi gerçekleştirilecektir. EDA, suçların hangi faktörlere bağlı olarak arttığını ve hangi bölgelerde daha fazla suç işlediğini keşfetmeye yardımcı olacaktır.
- **Özellik Mühendisliği:** Bu adımda, veri setindeki mevcut özelliklerin yanı sıra, modelin doğruluğunu artırmak için yeni özellikler yaratılacaktır. Örneğin, suçların yoğun olduğu zaman dilimleriyle ilişkili yeni özellikler eklenebilir. Ayrıca, hava durumu, tatil günleri ve sosyal etkinlikler gibi çevresel faktörler de suçlarla ilişkilendirilecek.
- **Model Geliştirme ve Sınıflandırma:** Bu projede, suç türlerinin tahmin edilmesi ve gelecekteki suç oranlarının sınıflandırılması için çeşitli makine öğrenmesi algoritmaları kullanılacaktır. Bu modeller, suç oranlarındaki değişimleri ve suçların gelecekteki olasılıklarını tahmin etmek için geliştirilecektir.

- **Zaman Serisi Analizi ve Tahmin:** Geçmiş verilere dayanarak, suçların gelecekteki olasılıkları zaman serisi analizleri ve diğer tahmin yöntemleriyle öngörülecektir. Bu tahminler, suçların hangi zaman dilimlerinde artabileceği ve hangi bölgelerde yoğunlaşacağı konusunda bilgi verecektir.

---

3. Veri Seti ve Kaynaklar

Proje, Chicago'daki suçları analiz etmek için kamuya açık olan **Chicago Crime Data** veri setini kullanacaktır. Bu veri seti, 2001 yılından itibaren Chicago'daki suçlara ait ayrıntılı verileri içermektedir. Veri setinde, suç türleri, suçların işlendiği tarih ve saat bilgileri, suçların coğrafi yerleri gibi bilgiler yer almaktadır. Ayrıca, suçların coğrafi yerleri GPS koordinatlarıyla belirtilmiştir.

Veri seti, aşağıdaki bilgilere sahiptir:

- Suç türleri (hırsızlık, cinayet, tecavüz, şiddet vb.)
- Suçların tarih ve saat bilgileri
- Suçların meydana geldiği mahalleler ve coğrafi bölgeler
- Suçların çevresel faktörlerle olan ilişkileri
- Date: Suçun işlendiği tarih ve saat
- Primary Type: Suçun genel türü (örn. THEFT, BATTERY)
- Description: Daha spesifik açıklama
- Location Description: Olayın gerçekleştiği yerin tanımı
- Arrest: Tutuklama yapılıp yapılmadığı
- Domestic: Aile içi şiddet olup olmadığı
- Beat, District, Ward, Community Area: İdari ve coğrafi sınırlara göre konum bilgileri
- Latitude / Longitude: Coğrafi koordinatlar
- Year: Olayın gerçekleştiği yıl
- Updated On: Verinin güncellenme zamanı.

---

4. Araştırma Soruları

Bu projede, Chicago'daki suçların analizi ve tahmini üzerine aşağıdaki araştırma soruları şekillendirilecektir:

1. **Suçların Zamanla Değişimi:** Suç oranları zaman dilimlerinde nasıl değişiyor? Örneğin, haftanın hangi günlerinde ve günün hangi saatlerinde suç oranları en yüksek? Hangi mevsimlerde suç oranları daha fazla?
2. **Coğrafi Dağılım ve Bölgesel Faktörler:** Suçlar Chicago'nun hangi bölgelerinde yoğunlaşıyor? Bu yoğunlaşmalar, demografik, ekonomik ve sosyal faktörlerle nasıl bir ilişki kuruyor?
3. **Suç ve Çevresel Faktörler:** Hava durumu, tatil günleri ve toplumsal etkinlikler gibi çevresel faktörler suç oranları üzerinde nasıl bir etki yaratıyor?
4. **Gelecekteki Suçlar:** Geçmiş verilere dayalı olarak, gelecekteki suç oranları nasıl tahmin edilebilir?
5. **Suçları Azaltma Stratejileri:** Suçların yoğunlaştığı bölgelerde suç oranlarını azaltmaya yönelik hangi önlemler etkili olabilir? Model, bu tür önlemlerin tahminleri üzerinde nasıl bir etki yaratabilir.

5. Yöntemoloji

Veri temizleme ve ön işleme, herhangi bir veri analizi sürecinin temel adımlarından biridir. Bu adımların amacı, veriyi modele uygun hale getirmek ve modelin doğruluğunu artırmaktır. Chicago'daki suç verisi üzerinde yapılacak işlemler şu şekilde detaylandırılacaktır:

1. Eksik Değerlerin İşlenmesi:
Veri setinde eksik değerlerin bulunması oldukça yaygındır. Eksik değerler, modelin eğitim sürecinde hatalı sonuçlara yol açabileceğinden, bunların doğru şekilde işlenmesi gerekmektedir. Eksik değerlerin işlenmesi için kullanılan bazı yöntemler:

- **Eksik Verilerin Silinmesi:** Eğer eksik değerler veri setinin büyük bir kısmını etkilemiyorsa, eksik verilerle olan satırlar tamamen silinebilir.
- **Eksik Verilerin Doldurulması:** Eksik veriler, ortalama, medyan, en yakın komşu gibi tekniklerle doldurulabilir. Özellikle kategorik verilerde, en sık görülen kategori ile doldurma veya o kategoriye ait mevcut verilerden türetilmiş değerler kullanılabilir.

2. Hatalı Verilerin Tespiti ve Düzeltilmesi:
Veri setinde yer alan hatalı değerler (örneğin, geçersiz GPS koordinatları veya mantıksız suç türü etiketleri) doğru şekilde temizlenmelidir. Hatalı veri tespiti için outlier (aşırı uç) analizleri yapılabilir. Çıkartılacak değerler, veri setindeki genel dağılımı büyük ölçüde etkilemeyecek şekilde seçilmelidir.

3. Kategorik Verilerin Sayısallaştırılması:
Makine öğrenmesi algoritmaları sayısal verilerle daha etkin çalıştığından, kategorik verilerin sayısallaştırılması gereklidir. Bu amaçla:

- **One-Hot Encoding:** Özellikle suç türleri gibi kategorik verilerde, her kategoriye bir sütun atanarak 0 veya 1 şeklinde değerler verilecektir.
- **Label Encoding:** Eğer kategoriler sıralıysa (örneğin suç ciddiyeti seviyesi: düşük, orta, yüksek gibi), her kategoriye bir etiket atanarak sayısal değerlere dönüştürülebilir.

4. Zaman Bilgilerinin İşlenmesi:
Suç verileri zaman damgasıyla ilişkilidir. Suçların zaman dilimlerine göre dağılımını incelemek için:

- **Zaman Özelliklerinin Çıkartılması:** Suç verilerindeki tarih ve saat bilgisi, yıl, ay, hafta, gün, saat gibi alt özelliklere ayrılacaktır. Bu, suçların zamanla nasıl değiştiğini anlamak için önemlidir.

5. Coğrafi Verilerin İşlenmesi:
Veri seti, suçların coğrafi yerlerini GPS koordinatlarıyla içeriyor olabilir. Bu koordinatlar, suçların yoğunlaştığı bölgeleri tespit etmek için kullanılır. GPS verisi, genellikle etkileşimli haritalarla görselleştirilir ve bu da suçların bölgesel dağılımını anlamamıza yardımcı olur.

Veri setinde yer alan tüm özellikler, modelin performansına doğrudan etki etmeyebilir. Bazı özellikler gereksiz olabilir ya da çok düşük bilgi taşıyor olabilir. Özellik mühendisliği, verinin daha anlamlı hale getirilmesi ve modelin başarısını artırmak için yeni özellikler yaratma sürecidir. Bu süreçte şu işlemler yapılacaktır:

1. Kategorik Verilerin Dönüştürülmesi:

- **Zaman Faktörleri:** Verinin içerisinde yer alan zaman bilgisi (gün, hafta, saat gibi) suçların zaman dilimleriyle ilişkisini anlamak açısından oldukça değerli olabilir. Suçların yoğun olduğu günler, hafta içi/hafta sonu, tatil günleri ve hava durumu gibi faktörler de göz önünde bulundurulacaktır.
- **Yerel Etmenler:** Suçların gerçekleştiği mahalle veya bölge bilgisi de önemli bir özelliktir. Çeşitli mahallelerin sosyo-ekonomik durumu, suç oranları üzerinde etkili olabilir. Bu nedenle, her mahalleye ait sosyo-ekonomik veriler ile etkileşimli özellikler oluşturulabilir.

2. Yeni Özellikler:
Veri setinde yer alan mevcut özellikler arasında ilişkiler kurarak, yeni özellikler üretilir. Örneğin:

- **Suç Yoğunluğu:** Suçların yoğun olduğu bölgeler ile çevresel faktörlerin (örneğin hava durumu, özel etkinlikler) etkileşimi.
- **Mevsimsel Suçlar:** Suç oranlarının mevsimsel değişikliklere nasıl tepki verdiği üzerine yeni özellikler oluşturulabilir.

3. Özellik Seçimi:
Veri setindeki bazı özellikler, modelin doğruluğunu artırmak yerine, aşırı uyum (overfitting) riskine yol açabilir. Bu yüzden, daha anlamlı özellikler seçilerek modelin performansı artırılacaktır. Özellik seçimi için:

- **Korelasyon Analizi:** Özellikler arasındaki güçlü korelasyonlar tespit edilerek, aynı bilgiyi taşıyan gereksiz sütunlar çıkarılacaktır.
- **Özellik Önem Sıralaması:** Random Forest gibi modellerin özellik önem sıralaması kullanılarak, önemli özellikler seçilecektir.

Projede suç tahmini yapmak için farklı makine öğrenmesi modelleri kullanılacaktır. Her model, verinin belirli özelliklerine ve problem tanımına göre farklı performans gösterebilir. Bu nedenle, modellerin karşılaştırılması önemlidir. Kullanılacak modeller ve performans ölçütleri şu şekildedir:

1. Kullanılacak Modeller:

- **Lojistik Regresyon (Logistic Regression):** Suç türlerinin sınıflandırılması için kullanılacak, basit ancak etkili bir modeldir.
- **Random Forest**:Karar ağaçlarının topluluğu üzerinden çalışarak yüksek doğruluk sağlar. Özellik önem sıralaması sunması modelin yorumlanabilirliğini artırır. Gürültülü verilere karşı dayanıklıdır.
- **Destek Vektör Makineleri (SVM):** Lineer ve doğrusal olmayan veri kümeleri için güçlü bir sınıflandırma modelidir. Suçların sınıflandırılmasında kullanılacaktır.
- **XGBoost / LightGBM**: Boosting yaklaşımı ile hataları düzeltir. Hızlı ve hafızada verimli çalıştığı için büyük veri setleri için uygundur. Özellikle dengesiz sınıflarda iyi performans gösterir.
- **Gradient Boosting:** Hata düzeltmeleri yaparak daha güçlü tahminler üreten bir başka ensemble modelidir. LightGBM veya XGBoost gibi varyasyonları ile performansı daha da artırılabilir.

- **Sinir Ağları (Neural Networks):** Derin öğrenme teknikleri kullanarak, karmaşık ilişkileri öğrenebilen sinir ağları, daha büyük veri setleri ve daha karmaşık veriler için idealdir.

2. Performans Değerlendirme:

Modelin performansı, doğruluk (accuracy), precision, recall ve F1 skoru gibi metriklerle değerlendirilecektir. Ayrıca, aşağıdaki metrikler de göz önünde bulundurulacaktır:

- **Hassasiyet (Precision):** Suç tahminlerinin ne kadar doğru olduğunu gösterir. Yüksek hassasiyet, modelin yanlış pozitif (false positive) tahminlerinin az olduğu anlamına gelir.
- **Accuracy**: Genel doğruluk oranı.
- **Precision/Recall:** Modelin suçları ne kadar doğru tespit edebildiğini gösterir. Yüksek duyarlılık, modelin yanlış negatif (false negative) tahminlerinin az olduğu anlamına gelir.
- **F1 Skoru:** Hassasiyet ve duyarlılığın harmonik ortalamasıdır ve modelin genel başarısını gösterir.
- **ROC-AUC:** Modelin sınıflandırma başarısını, özellikle dengesiz veri setlerinde, değerlendiren bir metriktir.
- **RMSE / MAE**: Zaman serisi modellerinde hata ölçümleri için kullanılacaktır.

Modelin başarısını artırmak için hiperparametre optimizasyonu (grid search veya random search) yapılacak ve modelin overfitting veya underfitting yapmaması sağlanacaktır.

---

6. Beklenen Sonuçlar

Bu proje ile Chicago'daki suç oranlarını tahmin edebilecek bir model geliştirilmesi amaçlanmaktadır. Modellerin karşılaştırılması, hangi modelin daha doğru tahminler yaptığına dair bir analiz sağlayacak ve suçları tahmin etme doğruluğu artırılacaktır. Ayrıca, suçların zamanla nasıl değişeceği ve hangi bölgelerde yoğunlaşacağına dair önemli bilgiler elde edilecektir. Bu bilgiler, güvenlik güçlerinin ve şehir yönetimlerinin suçları önceden tahmin etmelerini sağlayacak ve toplumun güvenliğini artırmaya yönelik stratejik planlar geliştirilmesine yardımcı olacaktır.

---

7. Gelecek Çalışmalar ve Öneriler

Projenin tamamlanmasının ardından, suçu etkileyen diğer çevresel faktörler ve veri kaynakları eklenerek modelin doğruluğu artırılabilir. Ayrıca, suç tahmini modelleri başka şehirlerde de uygulanabilir. Gelecekteki çalışmalarda, daha derinlemesine analizler ve daha büyük veri setleriyle çalışarak modelin kapsamı genişletilebilir.

8. Kaynakça

1. https://data.cityofchicago.org