simge, sembol, grafik, yazı tipi, logo içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

SAMSUN ÜNİVERSİTESİ MÜHENDİSLİK FAKÜLTESİ

YAZILIM MÜHENDİSLİĞİ BÖLÜMÜ

Makine Öğrenmesi Yaklaşımı ile Kredi Kartı İşlemlerinde Sahtekarlık Tespiti

Hazırlayan: EMRE YILDIZ

Öğrenci numarası: 211118060

Danışman: Özgür Tonkal

İçindekiler

[Bölüm 1: Proje Başlığı ve Tanımı 3](#_Toc199164262)

[Bölüm 2: Giriş 4](#_Toc199164263)

[Bölüm 3: Veri Toplama ve Temizleme 5](#_Toc199164266)

[3.1. Kullanılan veri seti 5](#_Toc199164267)

[3.2. Veri temizleme süreci 6](#_Toc199164268)

[3.3. Sınıf dengesizliği sorunu ve SMOTE uygulaması 6](#_Toc199164271)

[Bölüm 4: Veri Analizi Yöntemleri 7](#_Toc199164272)

[4.1. Kullanılan analiz yöntemleri 7](#_Toc199164273)

[4.2. Yöntem seçiminin gerekçesi 8](#_Toc199164275)

[Bölüm 5: Analiz Sonuçları 9](#_Toc199164279)

[5.1. Görselleştirme temelli analiz bulguları 9](#_Toc199164275)

[5.1.1 İşlem türü analizi 9](#_Toc199164275)

[5.1.2 Lokasyon analizi 10](#_Toc199164275)

[5.1.3 Zaman analizi 10](#_Toc199164275)

[5.1.4 İşlem tutarı analizi 11](#_Toc199164275)

[5.2. Model karşılaştırma bulguları 11](#_Toc199164275)

[Bölüm 6:Öneriler ve Sonuçlar 14](#_Toc199164280)

[6.1. Öneriler 14](#_Toc199164275)

[6.2.Sonuçlar 14](#_Toc199164275)

[Bölüm 7: Kaynakça 15](#_Toc199164281)

# Bölüm 1: Proje Başlığı ve Tanımı

* 1. **Proje Başlığı:**

Makine Öğrenmesi Yaklaşımı ile Kredi Kartı İşlemlerinde Sahtekarlık Tespiti

* 1. **Proje Tanımı:**

Bu proje kapsamında, kredi kartı işlem verileri üzerinde sahtekârlık (fraud) tespiti yapmak amacıyla bir makine öğrenmesi çözümü geliştirilmiştir. Günümüzde çevrim içi ve fiziksel kart işlemlerinde artan dolandırıcılık faaliyetleri, bankalar ve finansal kuruluşlar için önemli bir güvenlik sorununa dönüşmüştür. Bu nedenle, işlemleri otomatik olarak analiz edip şüpheli olanları tespit edebilecek akıllı sistemlere ihtiyaç duyulmaktadır.

Projede kullanılan veri seti, her kredi kartı işlemi hakkında tarih, saat, tutar, işlem türü, konum ve mağaza bilgileri içermektedir. Bu bilgiler üzerinden işlemin sahtekârlık olup olmadığı belirlenmeye çalışılmıştır. Amaç, farklı makine öğrenmesi algoritmalarını kullanarak en yüksek doğrulukla çalışan ve gerçek dünyada uygulanabilir bir model elde etmektir.

* 1. **Proje Amaçları ve Hedefleri:**
* Kredi kartı işlemleri üzerinde sahtekârlık örüntülerini analiz etmek,
* Verideki dengesiz sınıf dağılımını dengeleyerek daha adil bir model eğitmek,
* Farklı algoritmalar (Random Forest, XGBoost, LightGBM) kullanarak başarılarını karşılaştırmak,
* Sahtekârlık açısından en riskli işlem türlerini, saat dilimlerini ve lokasyonları belirlemek,
* Gerçek dünyaya entegre edilebilecek, güvenilir bir sahtekârlık tespit sistemi elde etmek.

# Bölüm 2: Giriş

Teknolojinin gelişmesiyle birlikte dijital ödeme sistemlerinin kullanımı hızla artmıştır. Özellikle kredi kartı ile yapılan işlemler hem kullanıcılar hem de kurumlar için büyük kolaylık sağlarken, aynı zamanda güvenlik risklerini de beraberinde getirmektedir. Bu risklerin başında kredi kartı sahtekârlıkları gelmektedir. Günümüzde birçok kullanıcı, kart bilgileri çalınarak bilgisi dışında işlem yapılması gibi durumlarla karşılaşmaktadır. Bu tür sahtekârlıklar hem bireylerin hem de finans kuruluşlarının ciddi maddi kayıplar yaşamasına neden olmaktadır.

Bu proje, kredi kartı işlemleri üzerinde sahtekârlık (fraud) tespitini otomatikleştirmek amacıyla geliştirilmiştir. Kredi kartı işlemlerinden oluşan veri seti kullanılarak, her işlemin sahte olup olmadığını sınıflandırmaya yönelik bir makine öğrenmesi modeli oluşturulmuştur.

Projede ele alınan temel problem, sahtekârlık işlemlerinin veride çok az sayıda bulunması (sınıf dengesizliği) ve bu azınlık sınıfın doğru şekilde tespit edilebilmesidir. Çünkü bu işlemler nadir olduğu için modelin genelde bu işlemleri göz ardı etme riski bulunmaktadır. Bu sorunu çözmek ve güvenilir bir model elde etmek için dengesizlik giderici yöntemler ve birden fazla makine öğrenmesi algoritması birlikte değerlendirilmiştir.

Bu rapor kapsamında veri analizi, modelleme süreci, elde edilen sonuçlar ve yapılan yorumlar detaylı olarak sunulmuştur. Gerçek hayat senaryolarına uygun bir sistem geliştirmek hedeflenmiştir.

# Bölüm 3: Veri Toplama ve Temizleme

## 3.1. Kullanılan Veri Seti

Bu projede credit\_card\_fraud\_dataset veriseti üzerinde çalışılmıştır, kredi kartı işlemlerinden oluşan ve sahtekârlık (fraud) tespiti amacıyla düzenlenmiş yapay ama gerçekçi bir veri kümesidir. Veri seti, her işlem için şu bilgileri içermektedir:

* TransactionID: İşlem numarası (benzersiz ID),
* TransactionDate: İşlemin tarihi ve saati,
* Amount: İşlem tutarı,
* MerchantID: İşlemin yapıldığı mağaza veya firma kodu,
* TransactionType: İşlemin türü (satın alma, iade, transfer vs.),
* Location: İşlemin yapıldığı yer,
* IsFraud: İşlemin sahtekârlık içerip içermediğini gösteren etiket (0 = normal, 1 = sahtekârlık).

## 3.2. Veri Temizleme Süreci

Veri seti ilk yüklendiğinde temel temizlik adımları uygulanmıştır. Bu adımlar aşağıdaki gibidir:

* TransactionDate sütunu tarih-zaman (datetime) formatına dönüştürülmüştür.
* Kategorik veriler (TransactionType, Location, MerchantID) makine öğrenmesi algoritmalarının anlayabileceği şekilde **LabelEncoder** yöntemiyle sayısal değerlere dönüştürülmüştür.
* Boş (NaN) veya eksik değer içeren satır sayısı oldukça az olduğu için bu satırlar veri setinden çıkarılmıştır.
* Tutar (Amount) sütununda aykırı (uç) değerlerin etkisini azaltmak adına gerekli analizler yapılmış, ancak modelin doğallığını korumak amacıyla işlem dışı bırakılmamıştır.

## 3.3. Sınıf Dengesizliği Sorunu ve SMOTE Uygulaması

Veri setinde **sahtekârlık içeren işlemler (IsFraud = 1)**, tüm işlemlerin sadece küçük bir kısmını oluşturmaktadır. Bu durum, makine öğrenmesi modellerinin sahtekârlık sınıfını öğrenmesini zorlaştırır. Bu problemi çözmek için **SMOTE (Synthetic Minority Over-sampling Technique)** yöntemi kullanılmıştır.

SMOTE, azınlık sınıfına ait yeni örnekler oluşturarak eğitim veri setini dengelemiş ve modelin her iki sınıfı da adil şekilde öğrenmesini sağlamıştır. Böylece model sadece normal işlemleri değil, sahtekârlıkları da başarıyla tanıyabilecek hale getirilmiştir.

# Bölüm 4: Veri Analizi Yöntemleri

# Bu projede kredi kartı işlemleri üzerinden sahtekârlık tespiti yapılabilmesi için çeşitli veri analizi teknikleri ve makine öğrenmesi algoritmaları uygulanmıştır. Bu yöntemler, hem veri setinin yapısını anlamak hem de doğru sınıflandırma modelleri geliştirmek amacıyla seçilmiştir.

## 4.1 Kullanılan Analiz Yöntemleri

* **Keşifsel Veri Analizi (Exploratory Data Analysis-EDA)**

Projeye başlamadan önce veri seti üzerinde görselleştirme temelli analizler yapılmıştır. Bu analizler sayesinde:

* Sahtekârlık oranı (%),
* İşlem türleri ve saatlere göre dağılım,
* Tutar ile sahtekârlık ilişkisi,
* Lokasyon ve merchant bazlı risk analizi gibi bilgiler elde edilmiştir.

Bu görselleştirmeler sayesinde hangi özelliklerin sahtekârlık üzerinde daha etkili olduğu anlaşılmıştır.

* **Sınıf Dengesizliği İçin SMOTE**

Sahtekârlık işlemleri veride oldukça az sayıda bulunduğu için sınıflar arasında dengesizlik söz konusuydu. Bu problemi çözmek için SMOTE algoritması kullanılarak, azınlık sınıfa (IsFraud = 1) ait yeni örnekler üretilmiş ve eğitim verisi dengelenmiştir. Bu sayede modellerin her iki sınıfı da öğrenebilmesi sağlanmıştır.

* **Makine Öğrenmesi Algoritmaları**

Veri analizi ve sınıflandırma aşamasında üç farklı denetimli makine öğrenmesi algoritması kullanılmıştır:

* **Random Forest:** Karar ağaçları topluluğu üzerinden çalışan, hem doğruluğu yüksek hem de aşırı öğrenmeye karşı dirençli bir modeldir.
* **XGBoost:** Hataları ardışık olarak azaltarak öğrenen güçlü bir boosting algoritmasıdır.
* **LightGBM:** Özellikle büyük veri setlerinde hızlı ve etkili çalışan, hafif bir gradient boosting algoritmasıdır.

Tüm bu modeller eğitim ve test setlerine ayrılmış dengeli veri seti üzerinde eğitilmiş ve doğruluk, precision, recall, f1-score ve ROC AUC gibi metriklerle değerlendirilmiştir.

## 4.2 Yöntem Seçiminin Gerekçesi

Bu yöntemler seçilirken dikkate alınan temel kriterler şunlardır:

* Sınıf dengesizliğiyle baş edebilme yetenekleri (SMOTE ile uyum),
* Karmaşık veri yapılarını öğrenme yetenekleri,
* Yorumlanabilirlik ve model başarısı,
* Eğitim süresi ve kaynak kullanımı.

Sonuç olarak her modelin avantaj ve dezavantajları göz önüne alınarak hem analiz hem de karşılaştırma amacıyla birlikte kullanılmışlardır.

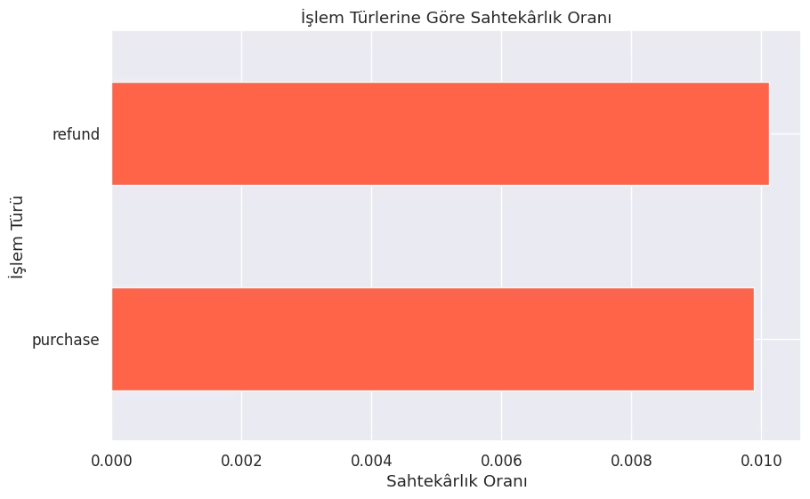
# Bölüm 5: Analiz Sonuçları

Bu bölümde hem görselleştirme analizlerinden hem de makine öğrenmesi modellerinin çıktılarından elde edilen bulgular sunulmuştur.

## 5.1 Görselleştirme Temelli Analiz Bulguları

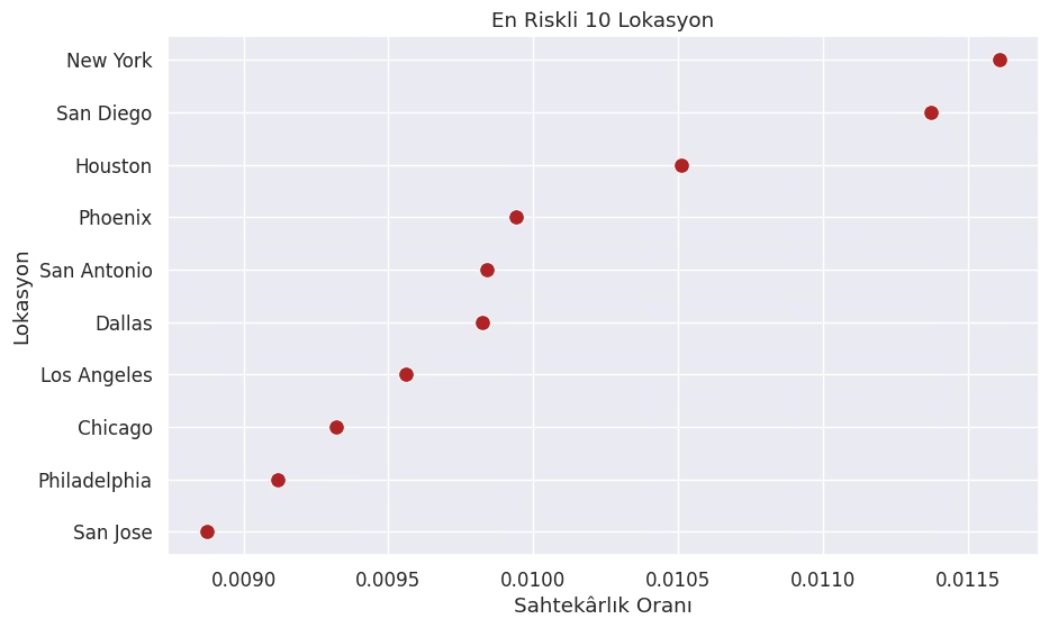
Projede yapılan keşifsel veri analizleri sonucunda şu önemli çıkarımlar elde edilmiştir:

### 5.1.1 İşlem Türü Analizi



Refund (geri ödeme) işlemlerinin, Purchase (satın alma) işlemlerine kıyasla daha yüksek sahtekârlık oranına sahip olduğu görülmüştür. Bu durum, iade süreçlerinin sahtekârlar tarafından daha fazla istismar edildiğini düşündürmektedir.

### 5.1.2 Lokasyon Analizi



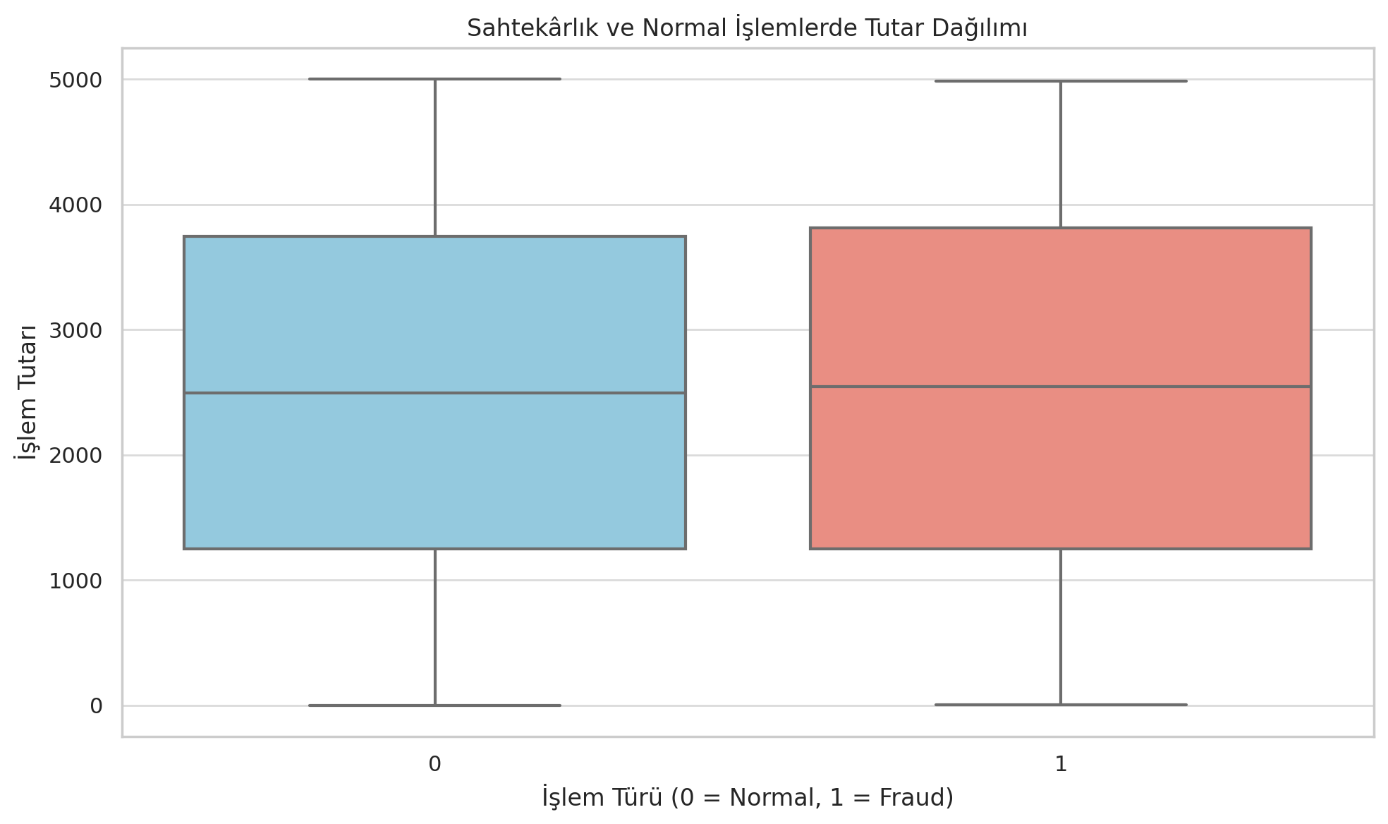
Belirli şehirler (örneğin New York, San Diego, Houston vb.) diğer lokasyonlara göre daha yüksek sahtekârlık oranına sahiptir. Bu şehirlerdeki işlem hacminin yüksek olması veya sistem açıklarının daha fazla olması bu duruma etki ediyor olabilir.

### 5.1.3 Zaman Analizi

## 

Sahtekârlık işlemleri günün her saatine yayılmış olsa da, bazı saatlerde belirgin artışlar var (örneğin 1, 8 ve 18. saatler). Bu da belirli zaman dilimlerinin daha riskli olabileceğini gösteriyor.

### 5.1.4 İşlem Tutarı Analizi

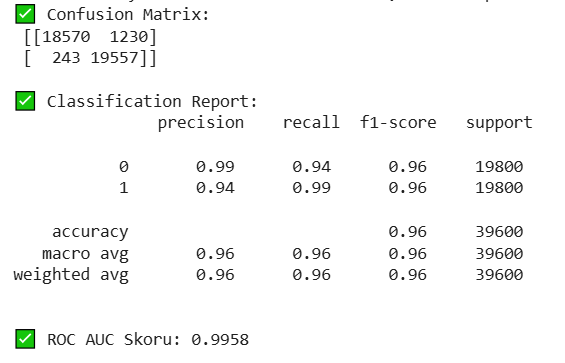


Her iki sınıfta da işlem tutarları geniş bir aralığa yayılmış durumda. Sahtekârlık işlemleri, ortalama olarak normal işlemlerden biraz daha yüksek tutarlarda gerçekleşiyor gibi görünüyor.

## 5.2. Model Karşılaştırma Bulguları

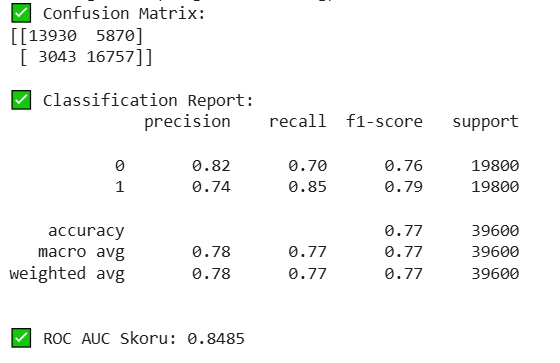
Üç farklı makine öğrenmesi algoritması ile yapılan sınıflandırma sonuçları karşılaştırıldığında aşağıdaki gibi bir sonuç ortaya çıkmıştır:

* **Random Forrest:**



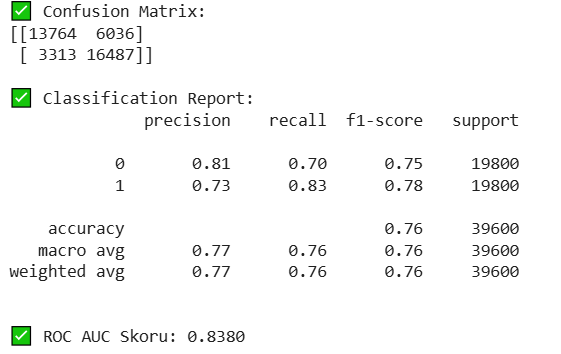
**Yorum:** Model oldukça yüksek performans göstermiştir. İki sınıfı ayırmadaki gücünün neredeyse mükemmel olduğu Confusion Matrixde okunmaktadır. Bu nedenle en güvenilir model olarak öne çıkmaktadır.

* **XGBoost:**

****

**Yorum:** XGBoost modeli de makul bir başarı göstermiştir. Recall oranı %85 ile fraud işlemleri fena olmayan bir oranla tespit edebilmiştir ancak precision değeri %74 olduğu için model, sahtekârlık olmayan bazı işlemleri de yanlışlıkla sahte olarak işaretlemiştir. Performansı Random Forest’a göre belirgin şekilde düşüktür.

* **LightGBM:**



**Yorum:** LightGBM modeli, XGBoost’a yakın ama biraz daha düşük performans göstermiştir. Recall %83, precision %73 düzeyindedir. Yani sahtekârlıkları %83 oranında yakalayabilmiş ama %27 oranında normal işlemi yanlış alarm olarak işaretlemiştir. Bu model daha hızlı çalışmaktadır ancak performans açısından Random Forest’ın gerisindedir.

**Karşılaştırmalı Değerlendirme:**

Random Forest modeli, yüksek doğruluk (%96) ve özellikle sahtekârlıkları yakalamadaki başarısıyla (%99 recall) en iyi performansı gösteren model olmuştur. XGBoost ve LightGBM ise daha düşük doğruluk ve ROC AUC skorlarıyla, sahtekârlıkları tespit etme konusunda Random Forest’a göre daha zayıf kalmıştır. Bu nedenle, hem güvenilirlik hem de genel başarı açısından **Random Forest** modeli, sahtekârlık tespiti için en uygun tercih olarak öne çıkmıştır.

# Bölüm 6: Öneriler ve Sonuçlar

## 6.1 Öneriler

* Gelecekte daha büyük ve gerçek zamanlı veri kümeleriyle çalışılarak modelin genel performansı artırılabilir.
* Ek özellik mühendisliği (feature engineering) yapılarak modele katkı sağlayacak yeni sütunlar (kart tipi, cihaz ID’si, IP adresi vb.) oluşturulabilir.
* Sahtekârlık vakalarının gerçek zamanlı olarak tespit edilmesi için geliştirilen model, bir API aracılığıyla mevcut sistemlere entegre edilebilir.
* Model sonuçları, kullanıcıya veya sisteme “risk skoru” şeklinde sunularak esnek aksiyon mekanizmaları geliştirilebilir (örneğin: şüpheli işlemde doğrulama adımı eklenmesi).

## 6.2 Sonuçlar

Bu projede derin öğrenme yaklaşımları ile kredi kartı sahtekarlığını tahmin etme başarılı bir şekilde gerçekleştirilmiştir. Özellikle LSTM gibi zaman serisi odaklı modellerin, dinamik finansal verilerde daha başarılı olduğu gözlemlenmiştir. Yapay zeka destekli bu yaklaşım, bankalar veya fintech şirketleri için erken uyarı sistemleri geliştirmek adına önemli bir temel oluşturmaktadır.

# Bölüm 7: Kaynakça

1. Kaggle - Credit Card Fraud Dataset
2. Scikit-learn dokümantasyonu
3. TensorFlow / Keras dökümantasyonu
4. Python resmi belgeleri