13.12.2025

DOKUZ EYLÜL ÜNİVERSİTESİ FEN FAKÜLTESİ
BİLGİSAYAR BİLİMLERİ

BİL3013 VERİ MADENCİLİĞİNE GİRİŞ
ÖĞRETİM ÜYESİ: Prof. Dr. Efendi NASİBOĞLU

ÖDEV 3: Sınıflandırma

HAZIRLAYAN ÖĞRENCİLER:
İremgül ZEYTİNÖZÜ - 2023280135
Salim Taha KAVAS – 2023280117

1. Kullanılan Teknolojiler ve Proje Ortamının Hazırlanması

Proje, Python programlama dili kullanılarak Jupyter Notebook ortamında geliştirilmiştir. Kredi kartı sahtekarlığı (fraud) tespiti gibi dengesiz veri setleri üzerinde çalışmak için aşağıdaki kütüphaneler ve yöntemler kullanılmıştır:

- **Pandas & NumPy:** Veri manipülasyonu, matris işlemleri ve veri setinin yüklenmesi için kullanılmıştır.
- **Scikit-learn:** Veri ön işleme (RobustScaler), modelleme (Random Forest, MLP) ve metrik değerlendirmeleri (classification_report, confusion_matrix, roc_auc_score) için temel kütüphane olarak kullanılmıştır.
- **XGBoost:** Yüksek performanslı Gradient Boosting algoritması için kullanılmıştır.
- **Matplotlib & Seaborn:** Confusion Matrix, Loss Curve ve Feature Importance grafiklerinin görselleştirilmesi için kullanılmıştır.
- **Ön İşleme Teknikleri:** Veri setindeki "Amount" ve "Time" değişkenleri, aykırı değerlere (outliers) karşı dayanıklı olması için RobustScaler ile ölçeklendirilmiştir. Ayrıca veri seti eğitim ve test olarak ayrılırken, sınıf dengesizliğini korumak adına stratify parametresi kullanılmıştır.

2. Giriş ve Veri Seti

Bu çalışmanın amacı, kredi kartı işlemlerini içeren bir veri seti üzerinde makine öğrenmesi algoritmalarını kullanarak, işlemin normal mi yoksa sahte (fraud) mi olduğunu tespit etmektir.

- **Veri Kaynağı:** Kaggle - Credit Card Fraud Detection.
- **Veri Seti Özellikleri:**
    - Toplam İşlem Sayısı: 284,807.
    - Normal İşlem Sayısı: 284,315 (%99.83).
    - Fraud (Sahte) İşlem Sayısı: 492 (%0.17) .
- **Değişkenler:** Veri setinde gizlilik nedeniyle PCA dönüşümü uygulanmış V1-V28 öznitelikleri ile dönüşüm uygulanmamış "Time" ve "Amount" değişkenleri bulunmaktadır. Hedef değişken "Class" (0: Normal, 1: Fraud) olarak belirlenmiştir.

Veri setindeki bu aşırı dengesizlik (Imbalance Ratio: 1:578), modelleme aşamasında doğruluk (Accuracy) yerine F1-Score, Precision ve Recall metriklerine odaklanılmasını zorunlu kılmıştır.

3. Kullanılan Sınıflandırma Modelleri

Sahtekarlık tespiti problemini çözmek için üç farklı algoritma kullanılmış ve hiperparametre optimizasyonları yapılmıştır:

1. **Random Forest Classifier:** Topluluk öğrenme yöntemidir. Aşırı öğrenmeye karşı dirençli olması ve dengesiz veri setleri için class_weight='balanced' parametresiyle optimize edilebilmesi nedeniyle tercih edilmiştir. Model 100 ağaç ve maksimum derinlik 20 olacak şekilde yapılandırılmıştır.
2. **XGBoost Classifier:** Gradient Boosting tabanlı, yüksek performanslı bir algoritmadır. Dengesiz veri setini yönetmek için scale_pos_weight parametresi hesaplanarak modele verilmiştir.

3. **Yapay Sinir Ağları (YSA / MLP Classifier):** Scikit-learn kütüphanesi ile (64, 32, 16) nöronlu üç gizli katmandan oluşan bir Çok Katmanlı Algılayıcı (MLP) mimarisi kurulmuştur. Eğitim sürecinde Early Stopping kullanılarak modelin kaybının azalmaması durumunda eğitim 12. iterasyonda durdurulmuştur.

4. Model Performansları ve Karşılaştırmalı Analiz

Modellerin başarısı; doğruluk yanıltıcı olabileceğinden, özellikle **F1-Score** (Dengesiz verilerde en kritik metrik), **Recall** (Kaçırılan fraud riskini ölçer) ve **Precision** (Yanlış alarmları ölçer) metrikleri ile değerlendirilmiştir.

4.1. Performans Karşılaştırma Tablosu

Aşağıdaki tablo, test veri seti üzerindeki sonuçları özetlemektedir:

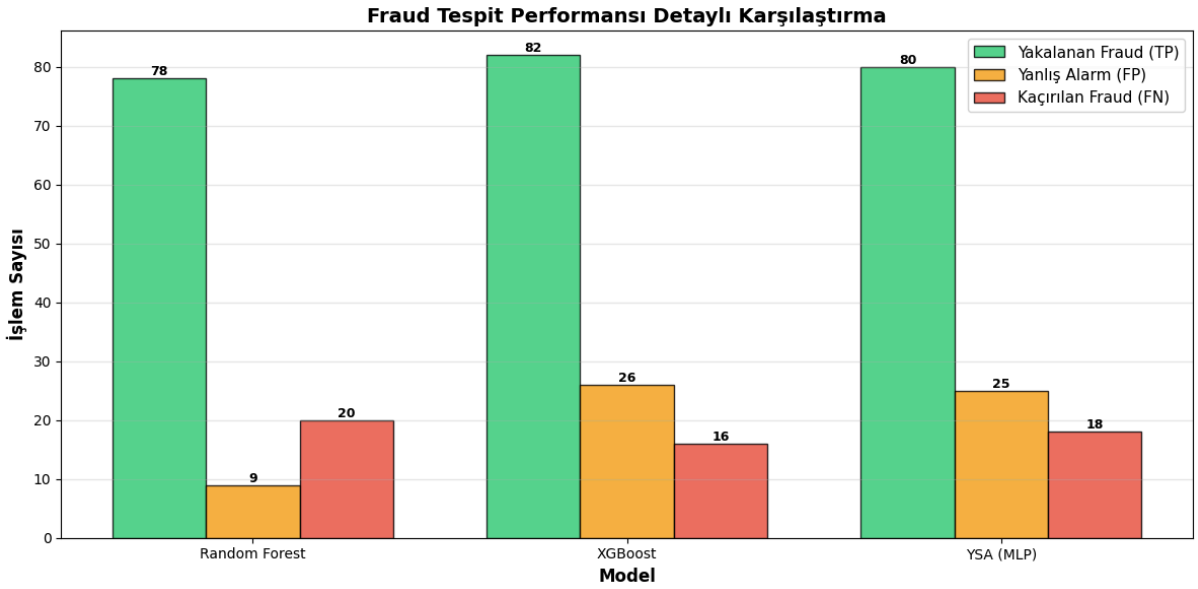
Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.9995	**0.897**	0.796	**0.843**	0.951
XGBoost	0.9993	0.759	**0.837**	0.796	**0.976**
YSA (MLP)	0.9992	0.762	0.816	0.788	0.971

4.2. Confusion Matrix Analizi ve Yorumlar

Modellerin gerçek dünyadaki etkisini anlamak için Confusion Matrix sonuçları "Yakalanan Fraud" (True Positive) ve "Yanlış Alarm" (False Positive) açısından incelenmiştir.

- **Random Forest:** 98 adet fraud işleminin **78** tanesini yakalamış, sadece **9** normal işlemi yanlışlıkla fraud olarak işaretlemiştir. En yüksek Precision (%89.7) değeriyle, en güvenilir uyarıları üreten model olmuştur.
- **XGBoost:** 98 adet fraud işleminin **82** tanesini yakalayarak en yüksek "Recall" değerine ulaşmıştır. Ancak bunu yaparken **26** adet yanlış alarm (False Positive) üretmiştir. Bu durum, daha fazla sahtekarlık yakalamasına rağmen operasyonel iş yükünü artırabileceğini göstermektedir.
- **YSA (MLP):** Random Forest ve XGBoost arasında dengeli bir performans sergileyerek 80 fraud yakalamış ve 25 yanlış alarm üretmiştir.

Görsel 1: *Fraud Tespit Performansı Detaylı Karşılaştırma Grafiği*

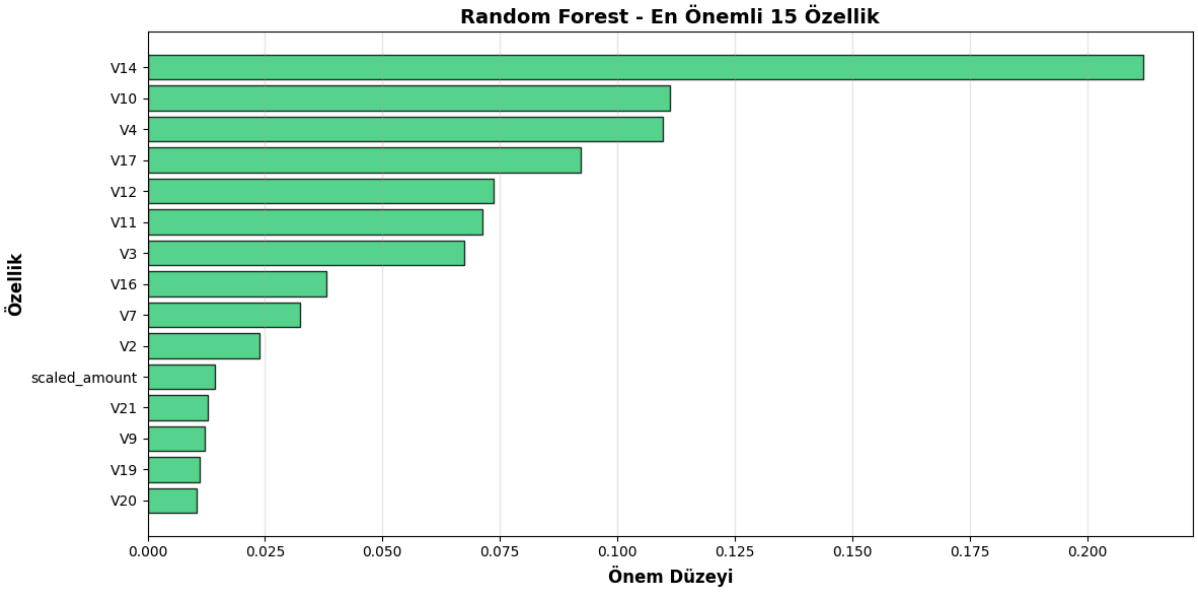


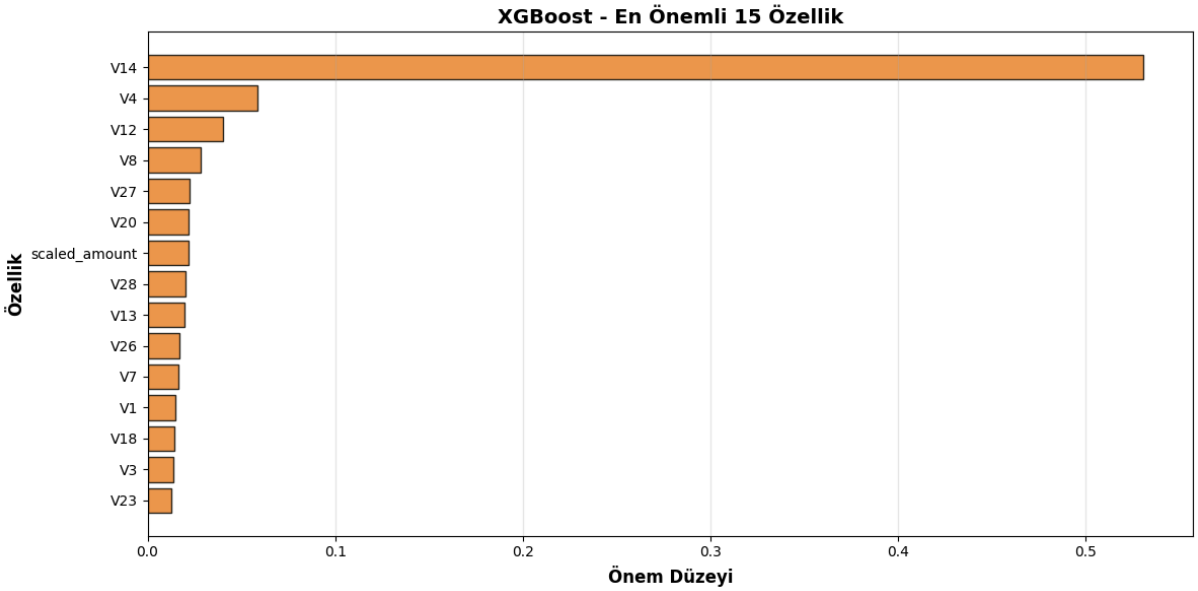
4.3. Öznitelik Önem Düzeyleri (Feature Importance)

Random Forest ve XGBoost modellerinin karar verirken hangi özelliklere odaklandığı analiz edilmiştir.

- Her iki modelde de **V14**, **V4** ve **V10** özellikleri en belirleyici faktörler olarak öne çıkmıştır.
- Özellikle **V14**, Random Forest modelinde %21, XGBoost modelinde ise %53 gibi çok yüksek bir önem düzeyine sahiptir. Bu durum, V14 özniteliğinin sahtekarlık tespitinde kritik bir rol oynadığını göstermektedir.

Görsel 2: *Random Forest ve XGBoost Modelleri için Öznitelik Önem Düzeyi Grafikleri*





5. Araştırma Sonuçları ve Tartışma

Analiz sonucunda elde edilen bulgular şu şekildedir:

1. **Dengesiz Veri Sorunu:** Veri setinin %99.8'inin normal işlemlerden oluşması nedeniyle Accuracy metriğinin tek başına anlamsız olduğu görülmüştür. Tüm modeller %99.9 doğruluk oranına sahip olsa da, ayırt edici farklar **F1-Score** ve **Recall** değerlerinde ortaya çıkmıştır.
2. **En İyi Model: Random Forest**, 0.843'lük F1-Score değeri ve çok düşük yanlış alarm oranı (sadece 9 adet) ile genel performans açısından **en başarılı model** olarak belirlenmiştir.
3. **Ticari Karar Dengesi:** Eğer banka için "tek bir sahtekarlığı bile kaçırmamak" maliyetten daha önemliyse **XGBoost** tercih edilebilir (82 yakalama). Ancak müşteri memnuniyeti (kartın gereksiz yere bloke edilmemesi) önemliyse, **Random Forest** çok daha üstün bir Precision sunmaktadır.

6. Sonuç

Bu çalışmada kredi kartı sahtekarlığı tespiti için YSA, XGBoost ve Random Forest modelleri geliştirilmiş ve karşılaştırılmıştır.

Elde edilen sonuçlar, **Random Forest** algoritmasının karmaşık ve dengesiz veri setlerinde özellikle doğru sınıflandırma ve düşük yanlış alarm oranı dengesini en iyi sağlayan model olduğunu kanıtlamıştır. Ayrıca özellik analizi sayesinde anonimleştirilmiş veriler (V14, V4) içinde bile sahtekarlığı ayırt eden belirgin kalıpların makine öğrenmesi ile tespit edilebildiği görülmüştür.