Banknot Sahteciliği Tespiti Projesi Raporu

# 1. Kullanılan Veri Seti Hakkında Bilgi

Bu çalışmada, UCI Machine Learning Repository üzerinden alınan Banknote Authentication veri seti kullanılmıştır. Veri seti, bir banknotun sahte mi yoksa gerçek mi olduğunu anlamak için, görüntü işleme teknikleriyle çıkarılan istatistiksel özellikleri içerir. Bu bir ikili sınıflandırma problemidir.

* Veri setinde toplam 5 sütun bulunmaktadır:
* variance: Dalgacık dönüşümünden elde edilen varyans
* skewness: Simetri bozukluğu (skewness)
* curtosis: Basıklık (kurtosis)
* entropy: Görüntünün bilgi entropisi
* class: Etiket sütunu – 0 = Gerçek banknot, 1 = Sahte banknot

Veri seti 1372 satırdan oluşmaktadır ve her satır bir banknota aittir. Tüm özellikler sayısaldır, kategorik veri bulunmamaktadır.

# 2. Uygulanan Veri Ön İşleme Adımları

Veri ön işleme sürecinde aşağıdaki işlemler gerçekleştirilmiştir:

**- Eksik Veri Kontrolü:**

df.isnull().sum() ile yapılan kontroller sonucunda veri setinde hiçbir eksik değer bulunmamıştır. Bu nedenle herhangi bir doldurma veya silme işlemi uygulanmamıştır.

**- Kategorik Veri Dönüştürme:**

Veri setindeki tüm değişkenler sayısal olduğu için One-Hot Encoding veya Label Encoding gibi işlemler uygulanmamıştır.

**- Standardizasyon (StandardScaler):**

Özellikle KNN ve Lojistik Regresyon gibi özelliğe duyarlı modeller için veriler, ortalaması 0 ve standart sapması 1 olacak şekilde ölçeklenmiştir. Ölçeklenmiş veri setleri X\_train\_scaled ve X\_test\_scaled olarak adlandırılmıştır.

**- Eğitim-Test Ayrımı:**

Veri seti %80 eğitim ve %20 test olacak şekilde ayrılmıştır. train\_test\_split() fonksiyonu ve random\_state=42 parametresi kullanılmıştır.

# 3. Seçilen Model ve Sonuçları

Bu proje kapsamında 3 farklı sınıflandırma modeli uygulanmıştır:

**1. Lojistik Regresyon**

* Basit ve yorumlanabilir bir modeldir.
* Doğrusal yapıya uygun olmayan verilerde performansı düşebilir.
* Test seti doğruluğu: %98

**2. K-Nearest Neighbors (KNN)**

* k=5 değeri ile uygulanmıştır.
* Eğitim verisine en yakın komşular baz alınarak tahmin yapılır.
* En başarılı model olmuştur.
* Test seti doğruluğu: %100

**3. Karar Ağacı (Decision Tree)**

* Hızlı ve sezgisel bir modeldir.
* Overfitting (aşırı öğrenme) riski taşır.
* Basit veri yapılarında iyi çalışabilir.
* Test seti doğruluğu: %98

# 4. Modelin Değerlendirilmesi ve Yorumlar

Kullanılan Değerlendirme Metrikleri:

* Accuracy (Doğruluk): Genel başarı oranı
* Precision: Pozitif sınıfa ait tahminlerin doğruluğu
* Recall: Gerçek pozitiflerin ne kadarını doğru tahmin ettiğimiz
* F1-Score: Precision ve Recall'un dengeli ortalaması
* ROC-AUC Skoru: Sınıflar arasında ayrım yapabilme gücü

**Gözlemler:**

* KNN modeli, hem doğruluk hem de ROC eğrisi açısından en başarılı modeldir.
* Lojistik regresyon, daha yorumlanabilir olmasına rağmen doğrusal olmayan sınırlarda yetersiz kalmıştır.
* Karar ağacı modeli overfitting'e eğilimli olması nedeniyle bazı sınıflarda daha düşük performans göstermiştir.