

## 1. Metodoloji

Bu veri bilimi projesinde, kredi başvuruları ile ilgili veriler üzerinde analiz ve modelleme yapılmıştır. Verilerin işlenmesi ve modelleme süreci aşağıdaki adımlarla gerçekleştirilmiştir:

- **Veri Ön İşleme:**
  - Veriler öncelikle **eksik değerler** ve **gereksiz kolonlar** bakımından temizlenmiştir.
  - **Kategorik veriler**, sayısal verilere dönüştürülmek için **One-Hot Encoding** ve **Mapping** kullanılarak işlenmiştir.
  - Sayısal özellikler için **özellik ölçekleme** işlemi gerçekleştirilmiştir (StandardScaler).
- **Keşifsel Veri Analizi (EDA):**
  - Verinin genel yapısı incelenmiş ve **aykırı değerler** ile **dağılımlar** görselleştirilmiştir.
  - **Korelasyon analizi** ile sayısal özellikler arasındaki ilişkiler tespit edilmiştir.
- **Modelleme:**
  - İki farklı model (K-Nearest Neighbors (KNN) ve Lojistik Regresyon) kullanılarak **sınıflandırma** yapılmıştır.
  - Her iki modelin performansı çeşitli metriklerle (doğruluk, kesinlik, duyarlılık, F1-skoru) değerlendirilmiştir.
- **Model Seçimi ve Hiperparametre Ayarı:**
  - KNN modelinin en iyi performans gösteren "k" parametresi, **GridSearchCV** kullanılarak optimize edilmiştir.
  - Lojistik Regresyon için varsayılan parametreler kullanılarak model eğitilmiştir.

## 2. Sonuçlar ve Kilit Bulgular

- **Keşifsel Veri Analizi (EDA) Sonuçları:**
  - **Yaş, Kredi Miktarı ve Kredi Süresi** gibi sayısal özelliklerin dağılımı incelenmiş, bazı aykırı değerlerin varlığı gözlemlenmiştir. Özellikle kredi miktarı ve süresi gibi değişkenlerde uç değerlerin olduğu saptanmıştır.
  - **Korelasyon Matrisi**, bazı değişkenler arasında yüksek korelasyonlar göstermiştir. Örneğin, kredi süresi ile kredi miktarı arasında pozitif bir ilişki olduğu tespit edilmiştir.
- **Model Performansları:**
  - **KNN Modeli:** GridSearchCV ile yapılan hiperparametre optimizasyonu sonucunda, KNN modelinin en iyi performans gösterdiği "k" değeri bulunmuş ve test seti üzerinde yüksek doğruluk oranına ulaşılmıştır.
  - **Lojistik Regresyon Modeli:** Lojistik Regresyon modelinin de güçlü bir performans sergilediği gözlemlenmiştir, ancak doğruluk oranı KNN modeline kıyasla biraz daha düşüktür.

- **Kilit Bulgular:**
  - **Kredi Miktarı** ve **Kredi Süresi** gibi değişkenler kredi riskini belirlemede önemli bir rol oynamaktadır.
  - Lojistik Regresyon, genellikle daha hızlı ve daha kolay yorumlanabilirken, KNN modeli daha esnektir ve daha iyi doğruluk sağlamaktadır.

### 3. EDA Bölümündeki Soruları Yanıtlamak:

- **Verinin Dağılımı Nasıldır?**
  - Sayısal değişkenlerin dağılımı, genellikle **normal dağılmamaktadır** ve birkaç uç değer içermektedir. Özellikle **kredi miktarı** ve **kredi süresi** gibi değişkenler, daha geniş bir yelpazeye sahip ve aykırı değerlere sahiptir.
- **Veri Arasında Korelasyon Var mı?**
  - Sayısal özellikler arasında yüksek korelasyonlar mevcuttur. Örneğin, **kredi süresi** ile **kredi miktarı** arasında güçlü bir pozitif korelasyon gözlemlenmiştir.
- **Aykırı Değerler Var mı?**
  - Evet, özellikle **kredi miktarı** ve **kredi süresi** gibi değişkenlerde aykırı değerler (outliers) bulunmaktadır. Bu aykırı değerler, model performansını olumsuz etkileyebilir.

### 4. Modelin Avantajları ve Sınırlılıkları

- **K-Nearest Neighbors (KNN) Modeli:**
  - **Avantajlar:**
    - KNN, **basit** ve **esnek** bir modeldir. Parametre ayarlarına bağlı olarak performansını optimize etmek mümkündür.
    - **Non-parametrik** bir modeldir, yani verinin dağılımı hakkında herhangi bir varsayımda bulunmaz.
  - **Sınırlılıklar:**
    - KNN, büyük veri setlerinde **yavaş** çalışabilir çünkü her tahmin için tüm veri setini tarar.
    - KNN, **uzun mesafelerde** (özellikle büyük veri setlerinde) düşük performans gösterebilir.
- **Lojistik Regresyon Modeli:**
  - **Avantajlar:**
    - Lojistik Regresyon, **hızlı** ve **kolay yorumlanabilir** bir modeldir. Her bir değişkenin etkisi doğrudan görülebilir.
    - **Çok sayıda özelliğe sahip veri setlerinde** iyi çalışabilir.
  - **Sınırlılıklar:**
    - Lojistik Regresyon, **lineer ilişkiler** varsayar, bu yüzden verilerde doğrusal olmayan ilişkiler varsa performansı düşebilir.

- **Outliers** ve çok büyük veri setlerinde performans sorunları yaşayabilir.

## 5. Sonuç

Bu çalışmada, kredi başvuruları ile ilgili verilerin analiz edilmesi ve iki farklı modelin (KNN ve Lojistik Regresyon) sınıflandırma başarısının karşılaştırılması yapılmıştır. Hem KNN hem de Lojistik Regresyon, güçlü performanslar sergilemiş ancak her modelin kendine has avantajları ve sınırlılıkları vardır. KNN, daha esnek ve doğruluk açısından daha iyi sonuçlar verirken, Lojistik Regresyon hızlı ve yorumlanabilir bir modeldir.

Bu bulgular ışığında, veri setinde yer alan bazı aykırı değerlerin modele etkisi göz önünde bulundurularak, daha ileri düzeyde model optimizasyonu ve aykırı değer işleme adımları yapılabilir.