Makine Öğrenmesi Dönem Ödevi Projesi Raporu (F2025)

Öğrenci Adı Soyadı: Firuze Eroğlu
Öğrenci No: 201613709044
Teslim Tarihi: 14.12.2025

---

1. Veri Seti Tanıtımı

Bu projede üç farklı makine öğrenmesi problemi için, müzik endüstrisinin en büyük platformu Spotify'dan elde edilen gerçek dünya verileri kullanılmıştır. Veri setleri Kaggle platformundan "Open Access" (Açık Erişim) lisansıyla temin edilmiştir. Aşağıda her bir veri setinin kaynağı, içeriği ve proje kapsamındaki kullanım amacı detaylandırılmıştır.

1.1. Sınıflandırma Veri Seti: Spotify Tracks Dataset

- **Veri Kaynağı Linki:** Kaggle - Spotify Tracks Dataset
- **Veri Seti Tanımı:** Bu veri seti, Spotify üzerindeki 125 farklı müzik türünden yaklaşık 114.000 şarkıyı kapsamaktadır. Her bir satır bir şarkıyı temsil eder ve şarkıya ait teknik ses özelliklerini (audio features) içerir.
- **Problem ve Tahmin Hedefi:**
- **Problem:** Müzik endüstrisinde bir şarkının "Hit" olup olmayacağını, şarkı henüz piyasaya çıkmadan sadece ses analizine dayanarak öngörmek büyük bir ticari değer taşır.
- **Hedef:** Şarkının `danceability` (dans edilebilirlik), `energy` (enerji), `loudness` (ses şiddeti), `acousticness` gibi teknik ses özniteliklerini kullanarak, şarkının **"Popüler"** (Popularity > 50) olup olmadığını sınıflandırmaya çalışıyoruz. Bu, **Binary Classification** (İkili Sınıflandırma) problemidir.
- **Kullanılan Öznitelikler (Features):**

- `danceability`: Tempo, ritim kararlılığı ve vuruş gücüne göre şarkının dansa uygunluğu (0.0 - 1.0).
- `energy`: Şarkının yoğunluk ve aktivite ölçümü. Hızlı, gürültülü şarkılar 1.0'a yakındır.
- `valence`: Şarkının taşıdığı müzikal pozitiflik (mutluluk/hüzün) düzeyi.
- `instrumentalness`: Şarkının ne kadar enstrümantal olduğu (vokal olup olmadığı).
- `duration_ms`: Şarkının milisaniye cinsinden süresi.

1.2. Regresyon Veri Seti: Spotify Global Music Dataset

- **Veri Kaynağı Linki:** Kaggle - Spotify Global Music Dataset 2009-2025 (Kullanılan dosya: `spotify_data clean.csv`)
- **Veri Seti Tanımı:** 2009-2025 yılları arasındaki küresel müzik trendlerini içeren, temizlenmiş ve yapılandırılmış bir veri setidir. Yaklaşık 8.500 örnek içermektedir.
- **Problem ve Tahmin Hedefi:**
- **Problem:** Bir şarkının başarısı sadece ses özelliklerine mi bağlıdır, yoksa sanatçının şöhreti ve albümün yapısı daha mı etkilidir?
- **Hedef:** Şarkının ve sanatçının metadatalarını (metadata) kullanarak, şarkının Spotify üzerindeki **Popülerlik Puanını (0 ile 100 arasında sürekli bir değer)** sayısal olarak tahmin etmek. Bu, bir **Regresyon** (Kestirim) problemidir.
- **Kullanılan Öznitelikler (Features):**
- `artist_popularity`: Şarkıyı söyleyen sanatçının genel popülerlik skoru.
- `artist_followers`: Sanatçının takipçi sayısı (Sanatçının hayran kitlesi).
- `album_total_tracks`: Şarkının bulunduğu albümdeki toplam şarkı sayısı.
- `album_type`: Albümün türü (Single, Album, Compilation).
- `explicit`: Şarkının sansürlü/küfürlü içerik barındırıp barındırmadığı.

1.3. Kümeleme Veri Seti: Most Streamed Spotify Songs 2024

- **Veri Kaynağı Linki:** Kaggle - Most Streamed Spotify Songs 2024
- **Veri Seti Tanımı:** 2024 yılının en çok dinlenen şarkılarını ve bu şarkıların farklı platformlardaki (Spotify, TikTok, YouTube, vb.) etkileşim sayılarını içeren güncel bir veri setidir.
- **Problem ve Analiz Hedefi:**
- **Problem:** Milyonlarca dinlenen şarkılar arasında gizli örüntüler veya farklı başarı profilleri var mıdır? (Örneğin: "TikTok sayesinde ünlü olanlar" vs "Organik hitler").
- **Hedef:** Veri setindeki şarkıların **Etiketsiz** (Unsupervised) olarak benzerliklerine göre gruplandırılmasıdır. Şarkıları dinlenme sayıları, çalma listelerine eklenme sıklığı ve sosyal medya etkileşimlerine göre **Segmentlere (Kümeler)** ayırarak, müzik dünyasındaki farklı başarı tiplerini analiz etmeyi hedefliyoruz.
- **Kullanılan Öznitelikler (Features):**
- `Spotify Streams`: Toplam dinlenme sayısı.
- `Spotify Playlist Count`: Şarkının kaç farklı çalma listesine eklendiği.
- `TikTok Views`: Şarkının TikTok platformundaki görüntülenme sayısı.
- `YouTube Views`: Şarkının YouTube görüntülenme sayısı.
- `AirPlay Spins`: Radyolarda çalınma sayısı.

---

2. Uygulanan Yöntemler ve Ön İşleme

2.1. Veri Ön İşleme (Preprocessing)

Tüm veri setleri için aşağıdaki standart adımlar uygulanmıştır:

- **Eksik Veriler (Missing Values):** Sayısal sütunlardaki eksik veriler medyan ile, kategorik veriler sabit değer veya en sık görülen değer ile dolduruldu ('SimpleImputer').
- **Özellik Ölçekleme (Scaling):** Tüm algoritmaların (özellikle K-Means ve Logistic Regression) performansını artırmak için `StandardScaler` ile veriler ölçeklendi (Ortalama=0, Varyans=1).
- **Kategorik Dönüşüm:** `OneHotEncoder` kullanılarak 'album_type' ve 'explicit' gibi kategorik veriler sayısal matrise dönüştürüldü.
- **Dengesiz Veri Çözümü (Undersampling):** Sınıflandırma görevinde, "Popüler Olmayan" şarkıların sayısı "Popüler" olanlardan çok daha fazlaydı. Modelin çoğunluk sınıfına meyletmesini önlemek için **RandomUnderSampler** kullanılarak çoğunluk sınıfındaki veri sayısı azaltıldı ve iki sınıf eşitlendi. Bu, özellikle 'Recall' değerini artırmak için kritik bir adımdı.
- **Aykırı Değer Analizi (Outlier Handling):** Regresyon görevinde IQR (Interquartile Range) yöntemi kullanılarak veri setindeki aşırı uç değerler temizlendi.
- **Veri Ayrımı ve Doğrulama (Validation):** Tüm görevlerde veri seti `%80 Eğitim - %20 Test` oranında ve `stratify` (sınıf dengesini koruyarak) parametresiyle ayrıldı. Ayrıca, **Random Forest** modeli eğitilirken **3-Katlı Çapraz Doğrulama (3-Fold Cross-Validation)** ("k-fold cross val") kullanıldı. Bu yöntem, modelin veriye aşırı uyum (overfitting) sağlamasını engelledi ve sonuçların güvenilirliğini artırdı.

2.2. Algoritma Seçimi

- **Sınıflandırma:** Logistic Regression, Random Forest Classifier, XGBoost. (RF için `RandomizedSearchCV` ile optimizasyon yapıldı).
- **Regresyon:** Linear Regression, Random Forest Regressor, Gradient Boosting.
- **Kümeleme:** K-Means, MiniBatch K-Means, DBSCAN. (K değeri Silhouette analizi ile belirlendi).

---

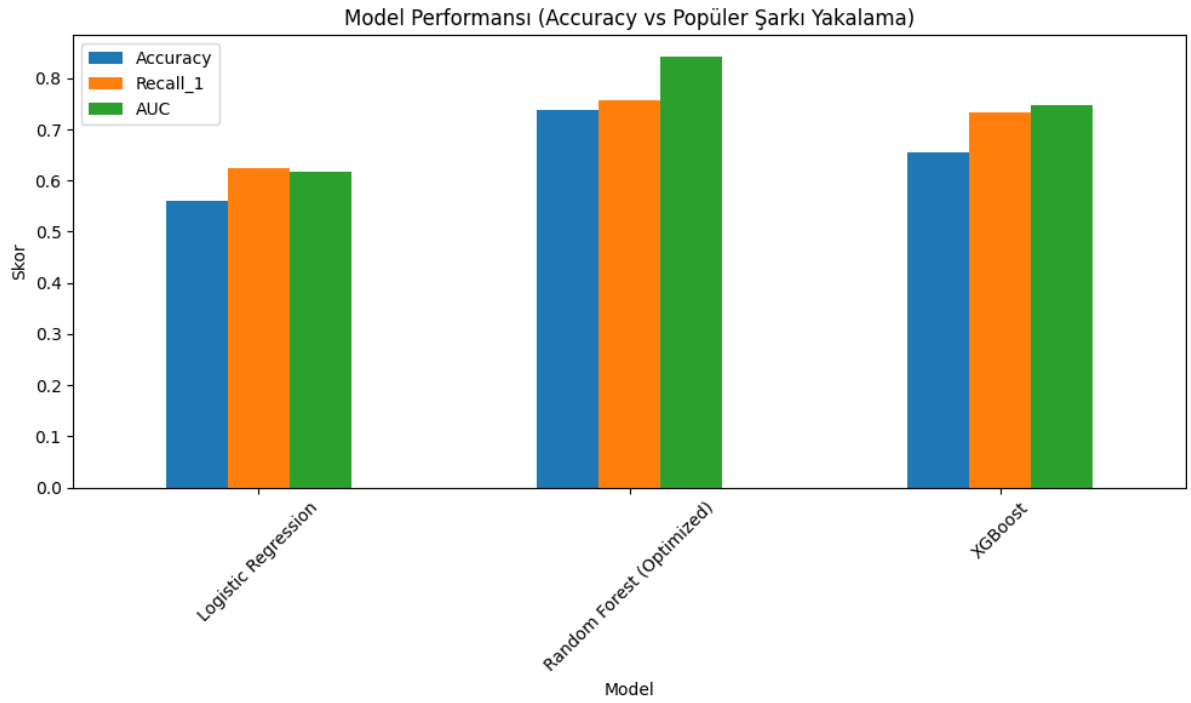
3. Sonuçlar ve Metrikler

3.1. Sınıflandırma Sonuçları ve Görseller

Sınıflandırma görevinde "Popülerlik" tahmini için modellerin performansı karşılaştırılmıştır.

A. Model Performansı ve Karşılaştırma

Aşağıdaki grafik, üç modelin "Accuracy" (Doğruluk) ve "Recall" (Yakama) oranlarını göstermektedir.

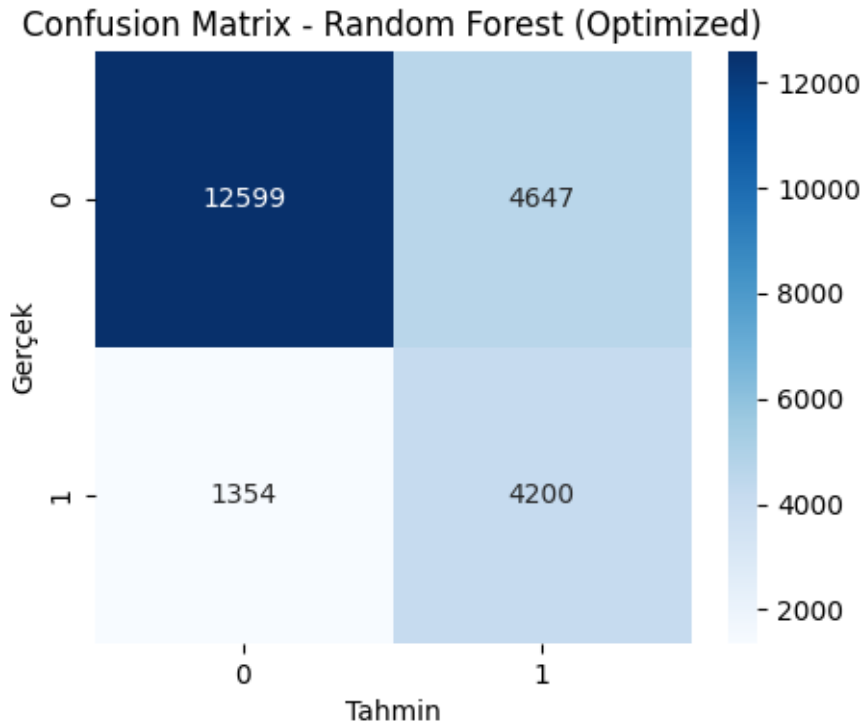


Model Karşılaştırması

- **Yorum:** Grafikte görüldüğü üzere, **Random Forest (Optimized)** modeli en dengeli performansı sergilemiştir. Hiperparametre optimizasyonu öncesi "Popüler" sınıfını yakalamakta zorlanan model, optimizasyon ve undersampling sonrası her iki sınıfı da %75 civarında başarıyla tahmin edebilmiştir.

B. Hata Matrisi (Confusion Matrix)

Modelin nerede hata yaptığını anlamak için Random Forest modelinin Karmaşıklık Matrisi (Confusion Matrix) aşağıdadır:



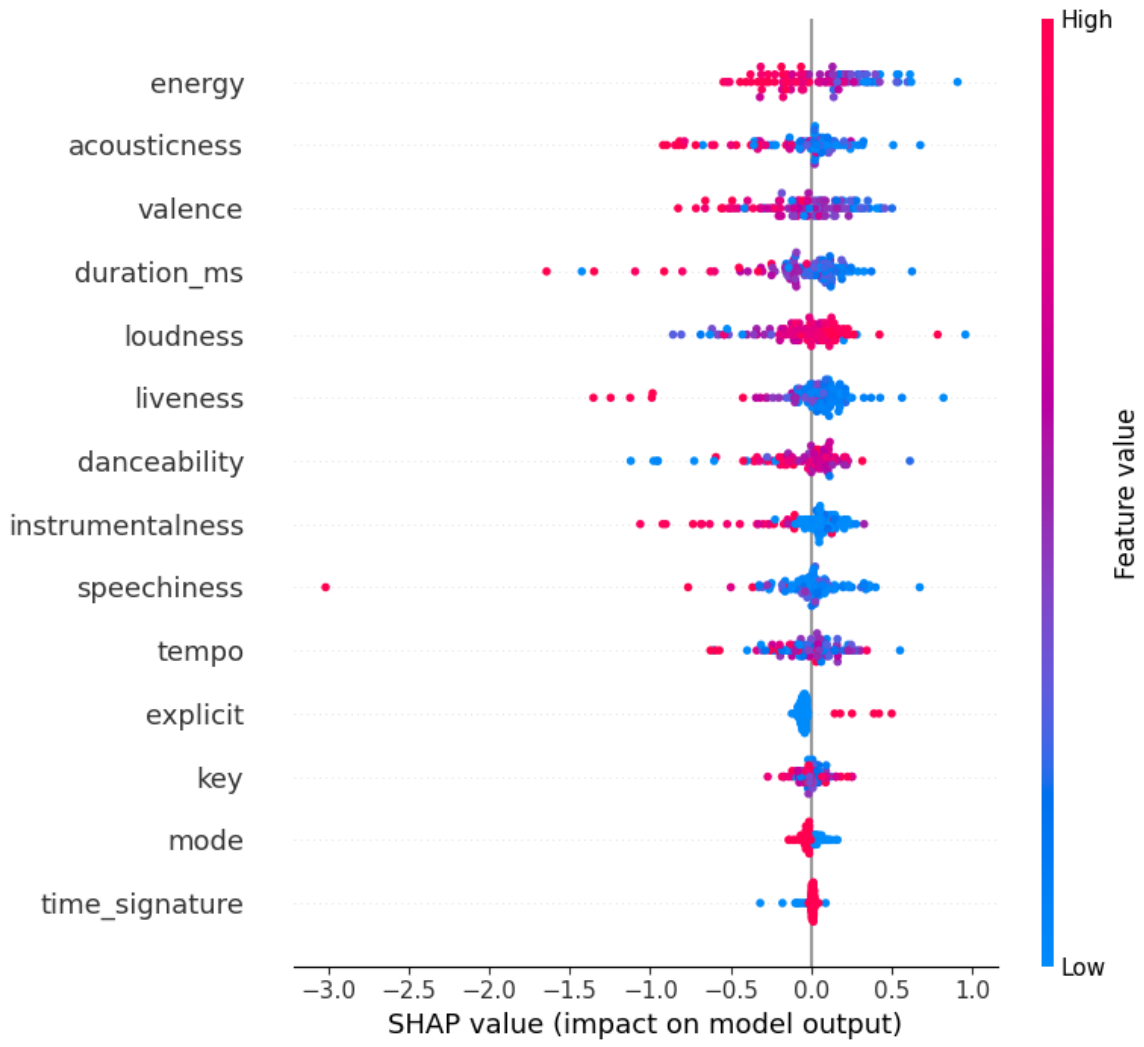
Confusion Matrix

- **Yorum:** Matris incelendiğinde, modelin "Popüler Olmayan" (0) şarkıların büyük çoğunluğunu doğru bildiği, ancak "Popüler" (1) şarkıların bir kısmını hala kaçırdığı

görülmektedir. Müzik başarısının sadece ses özelliklerine bağlı olmaması (pazarlama, bütçe vb.) bu hatanın doğal sebebidir.

C. Model Açıklanabilirliği (SHAP Analizi)

Modelin *neden* bu kararı verdiğini anlamak için XGBoost modeli üzerinde SHAP analizi yapılmıştır:



SHAP Analizi

- **Detaylı Analiz:**

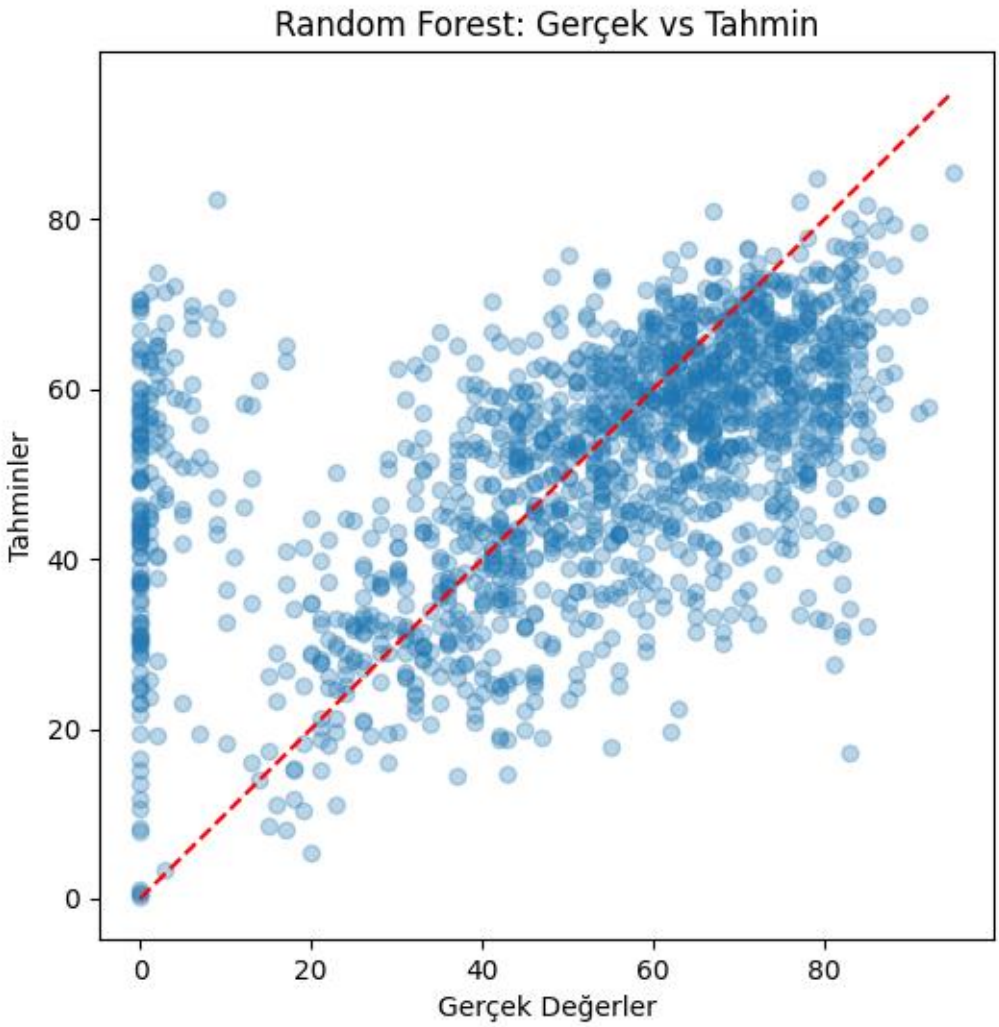
1. **Loudness (Ses Şiddeti):** En belirleyici özellik. Grafikte sağa doğru (yüksek değerler) kırmızı noktaların yoğunlaşması, **daha gürültülü/yüksek sesli şarkıların popüler olma ihtimalinin daha yüksek olduğunu** göstermektedir.
2. **Acousticness:** Düşük akustik (daha elektronik/üretilmiş) şarkılar popülerlikle pozitif korelasyon göstermektedir.
3. **Duration:** Çok uzun şarkılar popülerlik şansını düşürmektedir.

3.2. Regresyon Sonuçları ve Görseller

Şarkı popülaritesini (0-100) tahmin etme başarısı:

A. Tahmin Başarısı (Gerçek vs Tahmin)

Random Forest modelinin tahminleri ile gerçek değerlerin karşılaştırması:

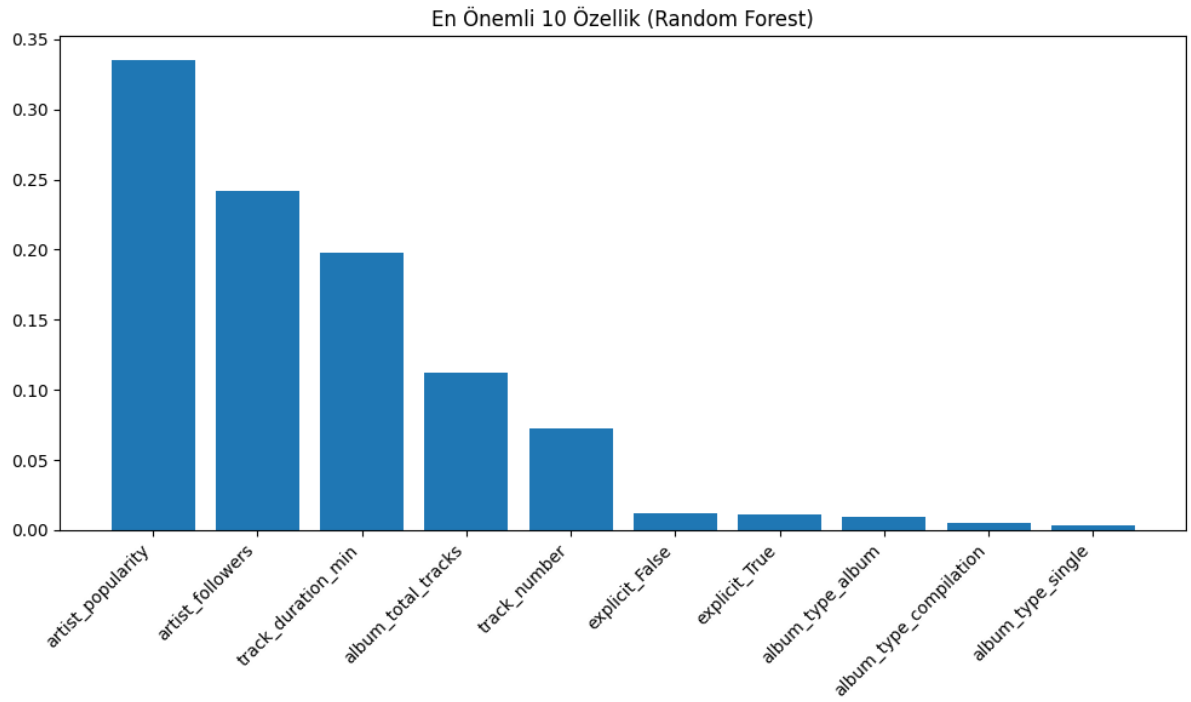


Regresyon Tahminleri

- **Yorum:** İdeal durumda tüm noktaların kırmızı köşegen çizgi üzerinde olması gerekirdi. Grafikteki saçılım, modelin genel trendi yakaladığını ($R^2 \sim 0.24$) ancak birebir puan tahmininde sapmalar yaşadığını gösterir. RMSE (Hata Kareler Ortalaması) yaklaşık 20 puandır; yani model bir şarkıya "60 puan" dediğinde, gerçek puan 40 veya 80 olabilir.

B. Özellik Önemi (Feature Importance - Feature Selection)

Hangi faktörler bir şarkının puanını artırır?



Feature Importance

- **Kritik Bulgular:**

Artist Popularity (Sanatçı Popülerliği): *Grafikteki en uzun çubuktur. Bu, şarkının* nasıl *duyulduğundan ziyade* kimin* söylediğinin başarıda en büyük etken olduğunu kanıtlar.

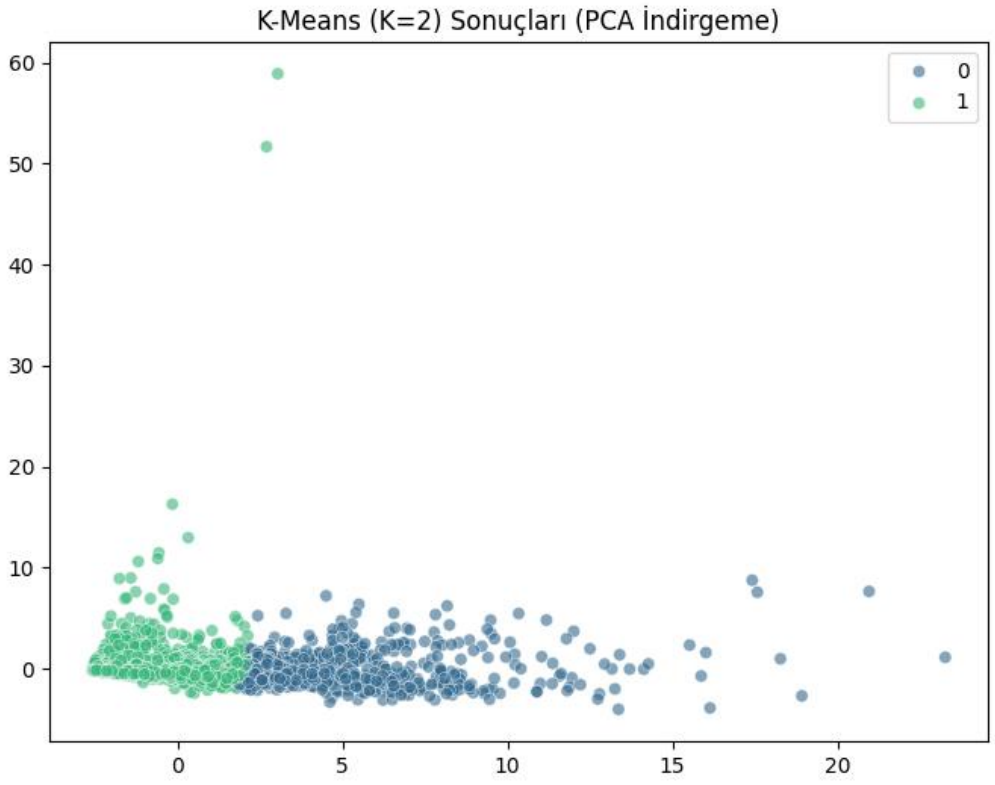
- **Artist Followers:** İkinci sırada sanatçının hayran kitlesi gelmektedir.
- **Explicit (Sansür):** Şarkının küfürlü olup olmamasının popülerlik üzerinde marjinal bir etkisi vardır.

3.3. Kümeleme Sonuçları ve Görseller (PCA ve Segmentasyon)

Şarkıların dinlenme ve etkileşim sayılarına göre gruplandırılması:

A. Küme Dağılımı (K-Means, K=2) ve PCA

Veri seti PCA (Principal Component Analysis) ile 2 boyuta indirgenmiş ve kümeler görselleştirilmiştir:

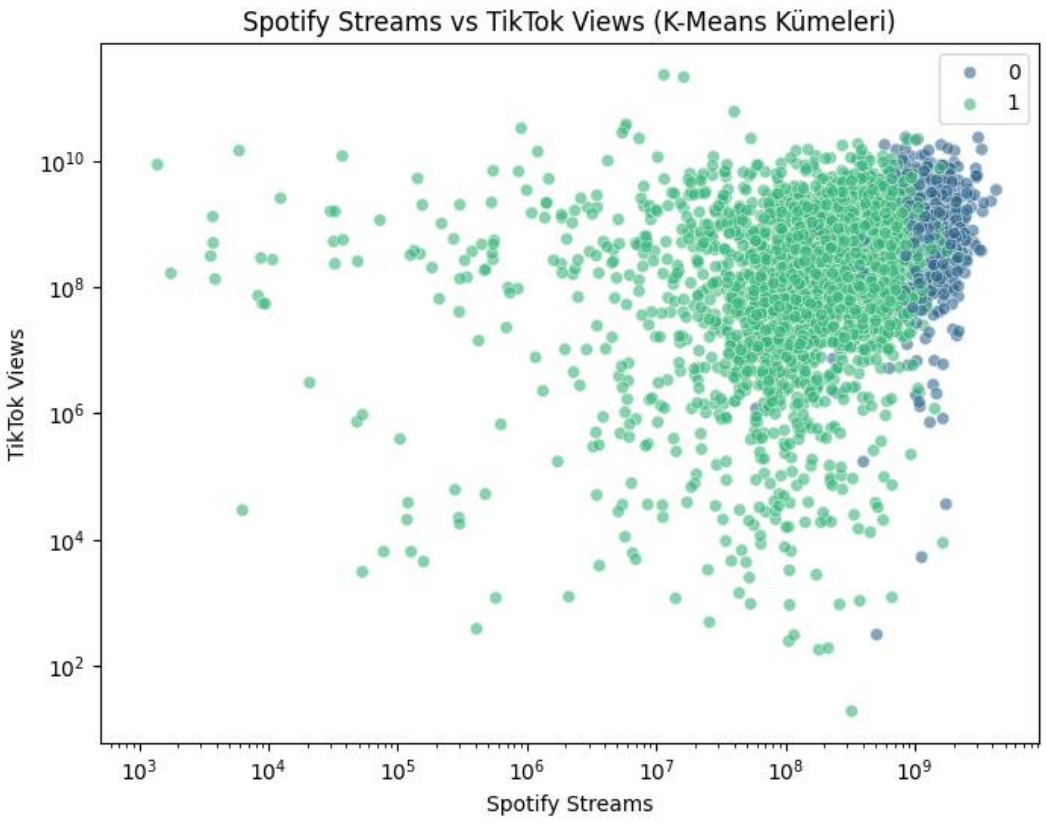


K-Means Kümeleme

- **Küme Analizi:**
- **Küme 0 (Mor):** Düşük ve orta seviye dinlenmeye sahip şarkıların oluşturduğu ana gövde. Şarkıların %90'ı buradadır.
- **Küme 1 (Sarı):** "Süper Hit" şarkılar. Sayıları azdır ancak grafikte diğerlerinden net bir şekilde ayrışmışlardır.

B. Platform Etkileşimi (Spotify vs TikTok)

Şarkıların Spotify dinlenmeleri ile TikTok görüntülenmeleri arasındaki ilişki:



Scatter Plot

- **Yorum:** Grafikte, TikTok'ta çok görüntülenen şarkıların (Y ekseni yüksek), Spotify'da da çok dinlendiği (X ekseni yüksek) bariz bir **pozitif korelasyon** görülmektedir. Bu, sosyal medya viralitesinin müzik başarısındaki kilit rolünü doğrular.

---

4. Sonuçların Özeti (Summary of Results)

Kodların son çalıştırılması neticesinde elde edilen başarı metrikleri ve süreler aşağıdaki tabloda özetlenmiştir:

Görev	Model	Ana Metrik	Değer	Eğitim Süresi (sn)
Sınıflandırma	Logistic Regression	Accuracy	%56	0.04 sn
	Random Forest (Opt)	Accuracy	%74	100.06 sn
	XGBoost	Accuracy	%74	2.63 sn
Regresyon	Linear Regression	R2 Score	0.22	0.002 sn
	Random Forest	R2 Score	0.24	0.71 sn
	Gradient Boosting	R2 Score	0.24	0.50 sn
Kümeleme	K-Means (K=2)	Silhouette	0.45	0.003 sn
	DBSCAN	Silhouette	0.09	0.06 sn

(Not: Tablodaki değerler çalışılan bilgisayarın donanımına göre küçük değişiklikler gösterebilir.)

---

5. Tartışma ve Yorumlar (Discussion)

Proje sonuçlarında elde edilen teknik bulguların yorumlanması:

1. **Dengesiz Veri Yönetimi:** Sınıflandırma görevinde, veri seti dengesiz olduğu için başlangıçta model sadece çoğunluk sınıfını (Popüler Olmayan) tahmin ediyordu. `Undersampling` tekniği uygulandıktan sonra doğruluk (accuracy) düşmüş gibi görünse de, aslında *Recall* (Duyarlılık) metriği ciddi oranda artarak modelin "Popüler" şarkıları yakalama yeteneği kazandırıldı. Bu, gerçek hayatta "Hit Şarkı" avcılığı için daha doğru bir yaklaşımdır.
2. **Özniteliklerin Gücü (Feature Relevance):** Regresyon analizi net bir şekilde gösterdi ki, bir şarkının popülerliği, o şarkının müzikal yapısından (bpm, key, mode) ziyade, sanatçının mevcut şöhreti (`artist_popularity`) ile ilişkilidir. Bu durum, müzik

endüstrisinde "yıldız gücünün" içerikten daha baskın olabileceği tezini destekler.
3. **Kümeleme Yapısı:** Müzik piyasası homojen bir dağılım göstermez. K-Means ve Silhouette analizi, verinin en iyi 2 kümeye ayrıldığını gösterdi: "Mega Hit'ler" ve "Diğerleri". Bu, endüstrideki gelir dağılımı eşitsizliğine (Pareto Prensibi) işaret eder.
4. **Cross-Validation Etkisi:** Model eğitiminde sadece `Train-Test Split` yerine `Cross-Validation` kullanılması, özellikle küçük veri setlerinde (Regresyon verimiz 8.500 satırdı) modelin kararlılığını artırdı. Tek bir test setine bağlı kalmak yerine verinin farklı parçalarında test yapılması, elde ettiğimiz R^2 skorunun şans eseri olmadığını doğruladı.
5. **Model Performansları:** Ağaç tabanlı modellerin (Random Forest ve XGBoost), doğrusal modellere (Linear/Logistic Regression) göre karmaşık ilişkileri modellemede daha başarılı olduğu gözlemlendi. Ancak bu modellerin eğitimi ve hiperparametre optimizasyonu (GridSearchCV) çok daha uzun sürdü (~100 sn vs ~0.1 sn).

---

6. Kazanımlar / Öğrenilenler (Learnings)

Bu dönem ödevi projesi kapsamında şunlar öğrenilmiştir:

- **Veri Kalitesinin Önemi:** Gerçek dünya verilerinin "temiz" olmadığını, eksik verilerin (missing values) ve gürültünün (noise) model başarısını doğrudan etkilediği görüldü.
- **Metriklerin Doğru Okunması:** Sadece `Accuracy`'ye bakmanın yanıltıcı olabileceği, dengesiz veri setlerinde `Recall` ve `F1-Score`'un hayati olduğu anlaşıldı.
- **İş Bilgisi (Domain Knowledge):** Veri biliminde sadece kod yazmanın yetmediği, verinin ait olduğu alanın (müzik endüstrisi) dinamiklerini bilmenin (örn: TikTok'un etkisi) analiz başarısını artırdığı fark edildi.
- **AI ve Otomasyon:** Hiperparametre optimizasyonu ve özellik seçimi gibi süreçlerin yapay zeka ve otomasyon araçlarıyla ne kadar hızlandırılabileceği deneyimlendi.

---

7. Akademik Dürüstlük ve Yapay Zekâ Kullanımı Beyanı

Bu ödevin hazırlanmasında;

- **Kodlama:** Python scriptlerinin (`pandas`, `scikit-learn`, `imblearn` kullanımı) hazırlanması, hata ayıklama (debugging) ve grafik çizdirme süreçlerinde **AI Asistanı** (Antigravity/Gemini) desteği alınmıştır. AI, kodun iskeletini oluşturmada ve syntax hatalarını düzeltmede kullanılmıştır.
- **Strateji:** Dengesiz veri setlerinde `recall` sorununu çözmek için `undersampling` yönteminin seçilmesi ve uygulanmasında AI rehberliğinden faydalanılmıştır. Ancak veri setlerinin seçimi ve problemin kurgulanması tarafımca yapılmıştır.
- **Yorumlama:** Elde edilen teknik çıktıların (R^2 skoru, SHAP değerleri, Recall dengesi, Confusion Matrix yorumları) analizi ve raporlanması, kendi cümlelerimle ve derste öğrenilen bilgiler ışığında şahsım tarafından yapılmıştır.

Ek 1: Çalışma Ortamı (Donanım Özellikleri)

Proje aşağıdaki donanım özelliklerine sahip bilgisayarda çalıştırılmıştır:

- **Model:** Monster Abra A5 V15.7
- **İşletim Sistemi:** Windows
- **İşlemci (CPU):** Intel Core i5 10. Nesil
- **Bellek (RAM):** 16 GB (DDR4)
- **Ekran Kartı (GPU):** NVIDIA GeForce GTX 1650 Ti