SPOTIFY SONG POPULARITY PREDICTION

by Alican Tunç

MODEL TYPE
Machine Learning

INPUT FEATURES
Audio Features, Temporal Context

TARGET
Popularity (0–100 Scale)

EVALUATION
Accuracy and F1 Score

IMPACT
Understanding Factors Behind Song Success

İÇERİK

Problem Tanımı ve Senaryo
Spotify 2020 şarkı verisiyle müzikal özellikler üzerinden popülerlik sınıflandırması (Low–Medium–High).

Veri Analizi ve Sınıf Dağılımları
Şarkıların popülerlik dağılımı, türlere ve yıllara göre değişimi; sınıf dengesizliği tespiti.

Ses Özellikleri ve Tür Analizleri
Audio feature korelasyonları, tür bazında ses profilleri (PCA & genre similarity).

Özellik Mühendisliği (Feature Engineering)
Zaman, sanatçı, playlist ve müzik profili temelli yeni değişkenlerin oluşturulması.

Modelleme Süreci
V1–V12A arası deneysel evrim: genre-based başlangıçtan artist intelligence modeline geçiş.

Model Mimarisi ve Performans Sonuçları
XGBoost tabanlı V12A mimarisi, sınıflandırma raporu, doğruluk ve F1 skorları.

Model Yorumlama (SHAP Analizi)
Modelin karar süreçlerinde en etkili özelliklerin açıklanması.

Tavsiyeler ve Sonuçlar
Playlist ve artist görünürlüğü vurgusu, trend temelli yeni özellik önerileri.

GUI Prototip (Spotify Popularity Predictor)
Modelin etkileşimli arayüzle (Streamlit) uygulanabilir prototip sunumu

Problem Tanımı ve Senaryo

Problem Tanımı ve Senaryo

- **Spotify 2020 Songs Dataset** kullanılarak, 32.000'den fazla şarkıya ait veriler analiz edilmiştir.
- Veri setinde her şarkının **müzikal özellikleri** *(energy, valence, danceability, acousticness, tempo)* ile **sanatçı bilgileri**, **tür (genre)** ve **yayın yılı** gibi değişkenler yer almaktadır.
- **Amaç:** Her şarkı için, müzikal ve yapısal özelliklere dayanarak **popülerlik düzeyini** *(Low / Medium / High)* tahmin etmektir.
- Böylece yeni çıkan bir parçanın **hit olma potansiyeli** veriyle ön görülebilmektedir.
- **Veri Kaynağı:** *TidyTuesday Spotify Songs Dataset (rfordatascience, 2020)*

Veri Analizi

Bu projede, Spotify üzerindeki şarkıların ses ve içerik özelliklerini analiz ederek popülerliklerini belirleyen faktörleri anlamaya çalıştık.

Spotify'ın sunduğu 13 temel müzikal özellik arasından özellikle **danceability, energy, loudness, valence ve tempo** değişkenleri, şarkının dinleyici üzerindeki enerjik ve duygusal etkisini temsil eden en belirgin göstergeler olarak öne çıktı.

Kategori	Popülerlik Aralığı	Açıklama
Low	0-50	Daha az dinlenmiş veya az bilinen şarkılar
Medium	50-75	Ortalama dinlenme düzeyine sahip, belirli bir kitleye ulaşmış şarkılar
High	75-100	Geniş kitlelerce dinlenmiş, viral veya listelerde yer alan şarkılar

Bu sınıflandırma, hem **makine öğrenmesi modellerinin hedef değişkenini dengelemek** hem de **görsel analizlerde eğilimleri daha net ortaya koymak** amacıyla yapıldı

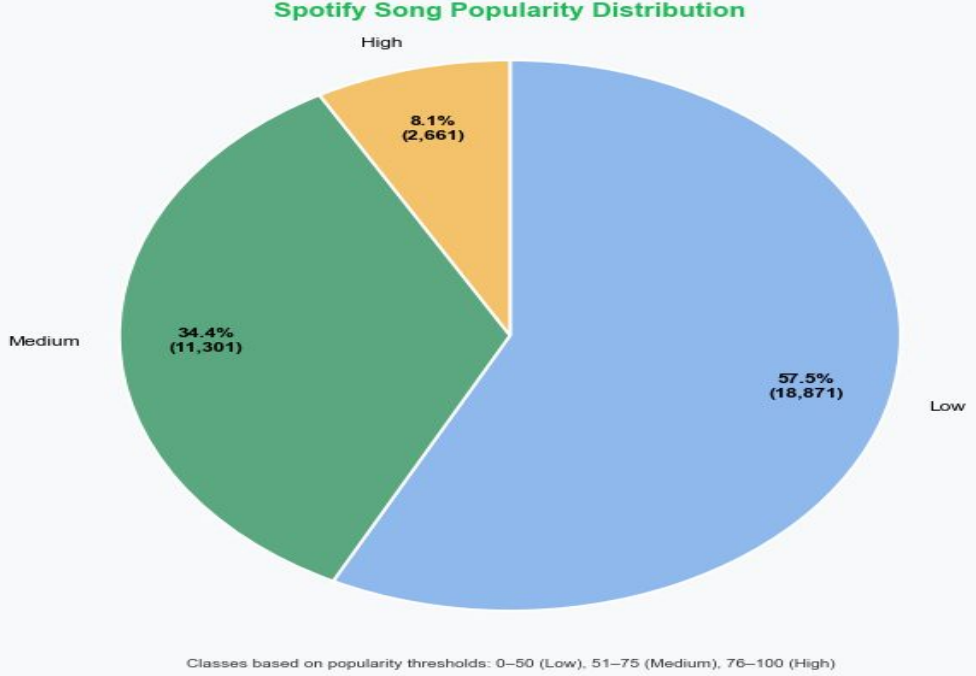
Sınıf Dağılımları

Spotify'daki şarkı popülerlikleri incelendiğinde, **veri dağılımının oldukça dengesiz** olduğu açıkça görülüyor.

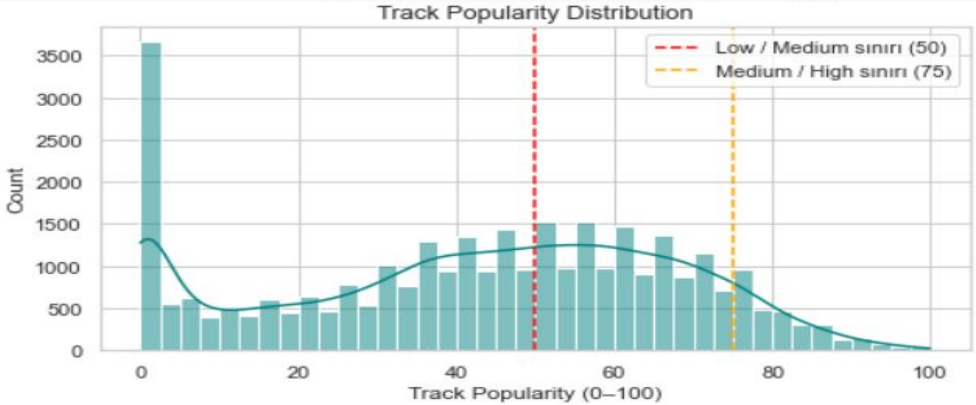
Şarkıların **%57,5'i düşük**, **%34,4'ü orta** ve yalnızca **%8,1'i yüksek popülerlik** düzeyinde yer alıyor.

Bu tablo, müzik dünyasında popülerliğin her şarkıya eşit dağılmadığını, az sayıdaki "hit" parçanın çoğunluğa oranla çok daha fazla öne çıktığını gösteriyor. Özellikle 0 puanlı yani unutulmuş şarkılar çok fazla iken 100'e yakın hit şarkılar neredeyse yok denecek kadar az.

Bu nedenle analizlerde ve modelleme aşamalarında, **sınıf dengesizliğine dikkat edilmesi** gerektiği önemli bir bulgu olarak karşımıza çıkıyor.



Classes based on popularity thresholds: 0–50 (Low), 51–75 (Medium), 76–100 (High)



Zaman ve türe bağlı popülerlik

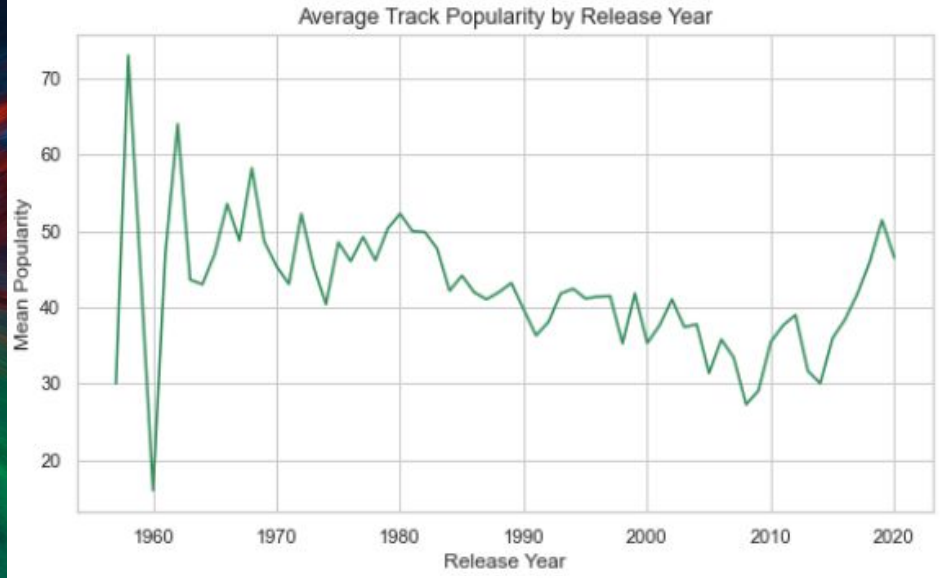
Spotify'daki şarkı popülerlikleri yıllara ve türlere göre incelendiğinde, müzik dünyasında popülerliğin dönemsel olarak değiştiği açıkça görülüyor.

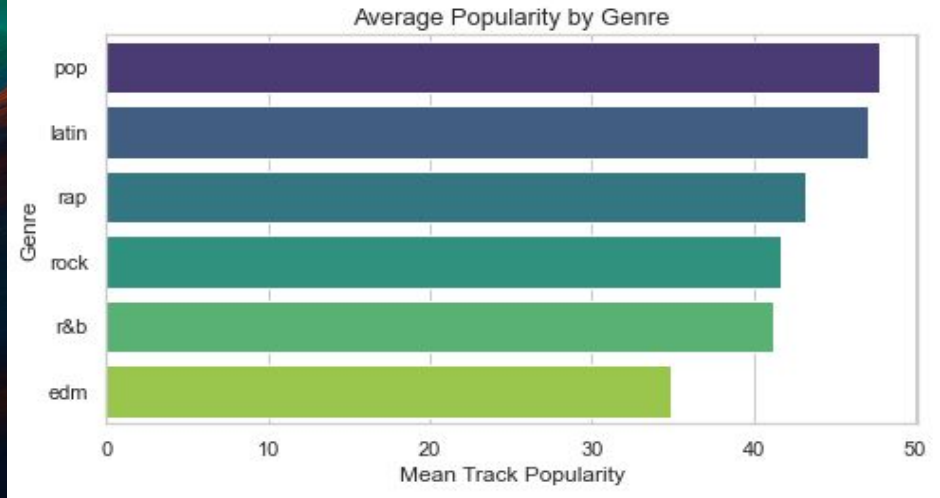
Yıllar ilerledikçe ortalama popülerlik zaman zaman dalgalanma gösterse de, özellikle **2010 sonrası dönemde dijital platformların yükselişiyle yeniden artış eğilimi** dikkat çekiyor.

Türler özelinde bakıldığında ise **Pop** ve **Latin müzik**, geniş dinleyici kitlesi sayesinde en yüksek ortalama popülerliğe sahip türler olarak öne çıkıyor.

Buna karşın **EDM** ve bazı niş türlerin ortalama popülerliği daha düşük kalmış durumda.

Bu tablo, müzikte popülerliğin hem **zamana hem de türe bağlı olarak dengesiz dağıldığını** ve belirli dönemlerde bazı türlerin öne çıktığını gösteriyor. Yani modelimizi eğitirken hem zamana bağlı hem de türe bağlı değişkenlere dikkat etmeliyiz.





How Spotify Knows You

Quantifiable attributes

Speechiness

Energy

Tempo

Duration

Loudness

Ses özellikleri bazında analizler

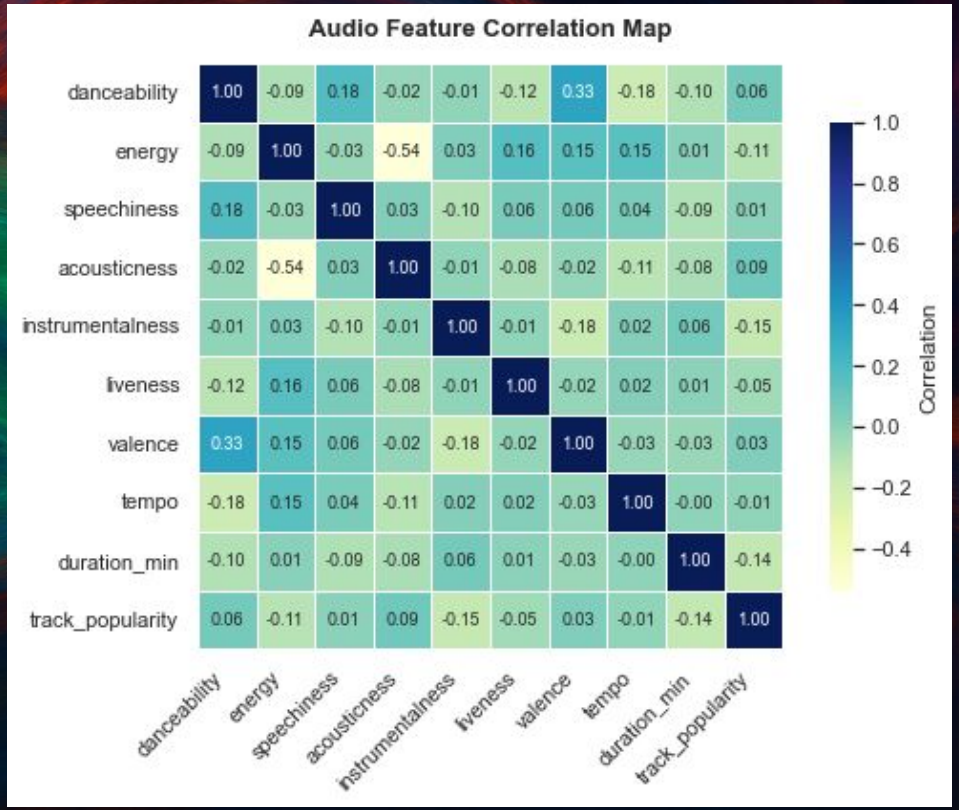
Kolerasyon matrisi bize açıkça gösteriyor ki spotify popülerliği, tek bir müzikal özelliğe bağlı değil;

“enerji”, “dans edilebilirlik” ve “şarkı uzunluğu” gibi birçok faktörün birleşiminden etkileniyor.

Bu tablo, popülerliğin **karmaşık ve çok boyutlu bir olgu** olduğunu, müzikal çeşitliliğin önemli bir rol oynadığını açıkça gösteriyor. Tablo incelendiğinde hiçbir özelliğin popülerlikle yüksek doğrusal ilişkiye sahip olmadığı görülüyor (|r| < 0.2).

Ama bazı müzikal özellikler birbiri alakalı ve hem pozitif hem de negatif olarak etkililer. Bu bize her şarkının kendi içerisindeki türü ve karakteristiğini yansıttığını gösteriyor.

→ Yani popülerlik, tek bir müzikal parametreyle açıklanamaz; **çok boyutlu bir kavram.**

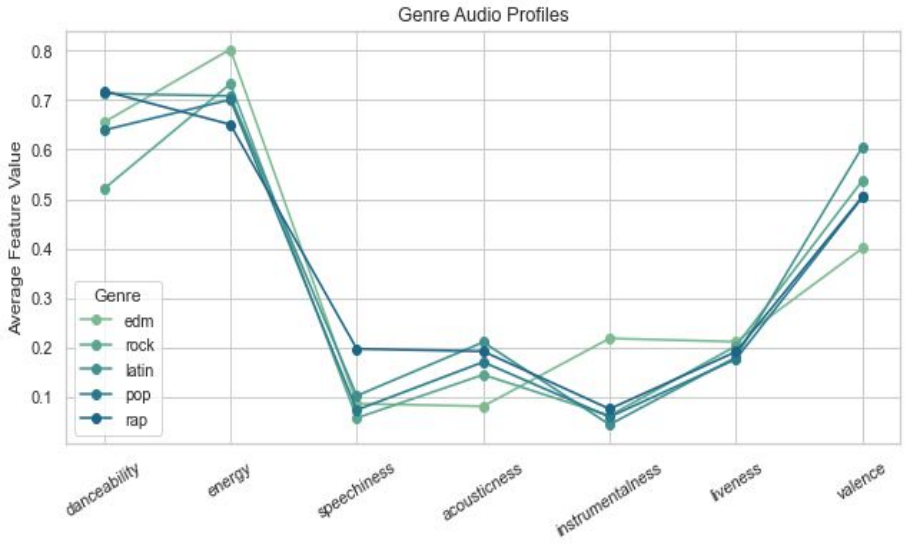


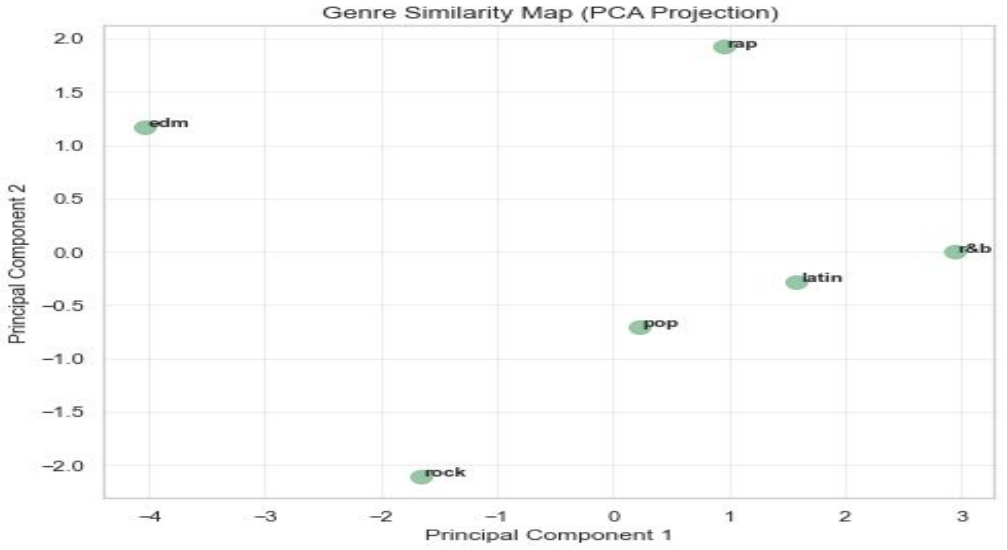
Tür bazında ses özellikleri ile analizler

Bu iki grafik birlikte değerlendirildiğinde,

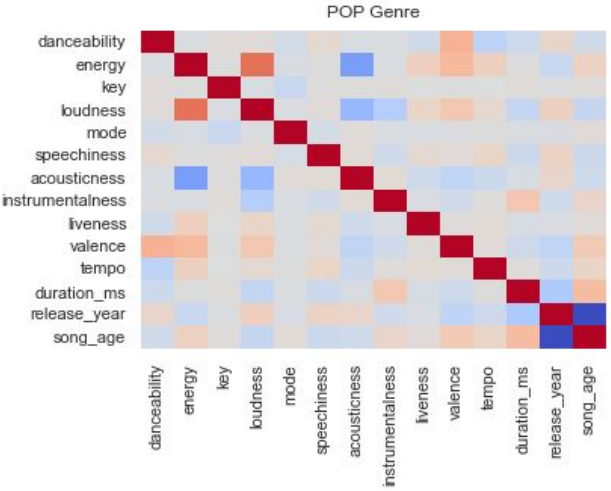
müzik türlerinin hem **dahili ses imzalarının (audio profile)** hem de **benzerlik kümelerinin (PCA space)**, Spotify ekosisteminde oldukça belirgin biçimde ayrıştığı görülüyor.

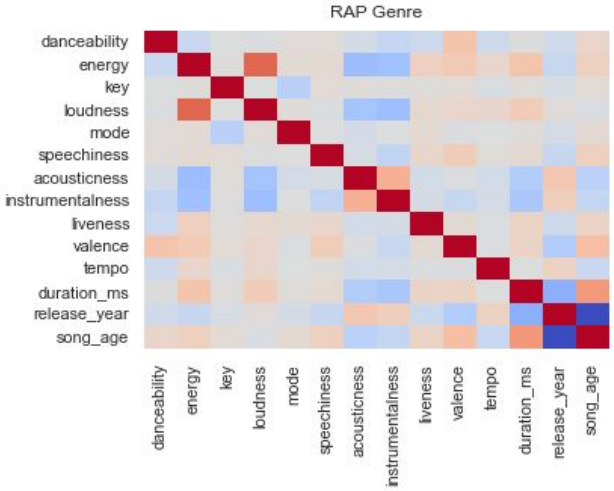
Bu da modelin tür bazlı özellik mühendisliği (genre-aware feature engineering) açısından güçlü bir temel sunduğunu gösteriyor.

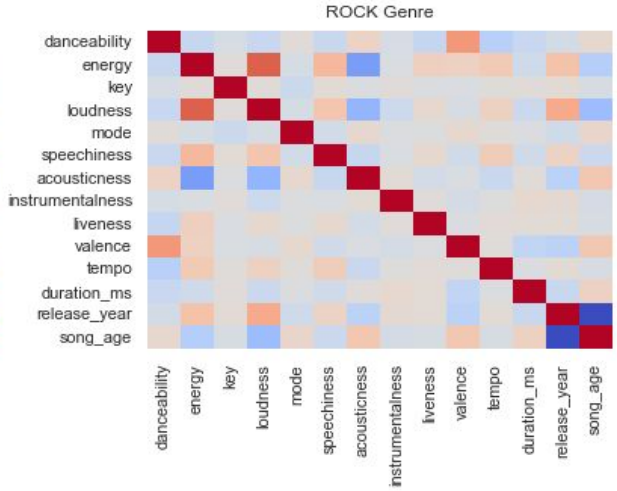


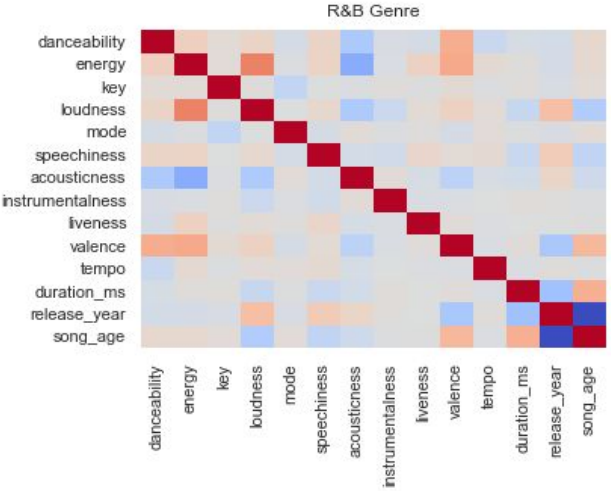


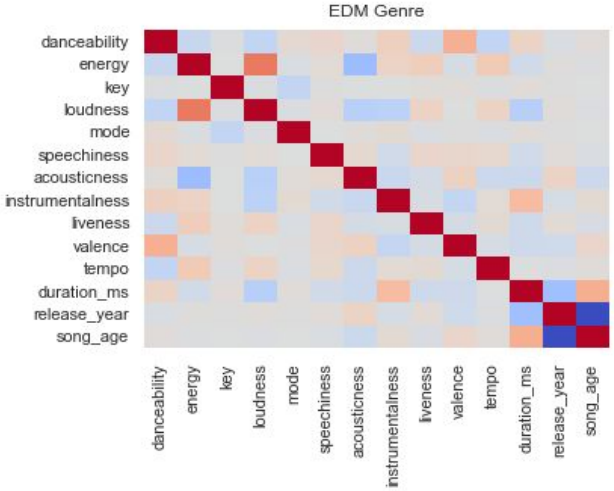
Genre-wise Audio Feature Correlations & Top Influencers

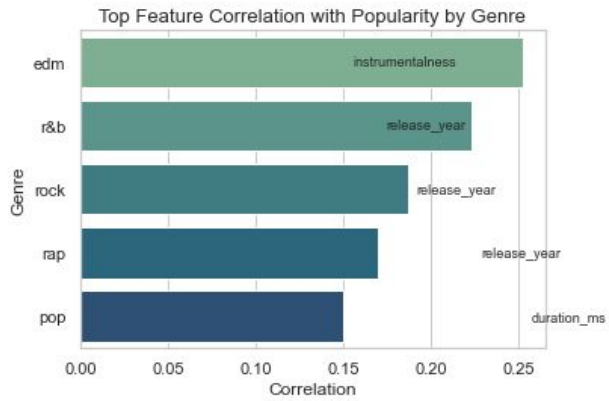












Analiz Çıktıları

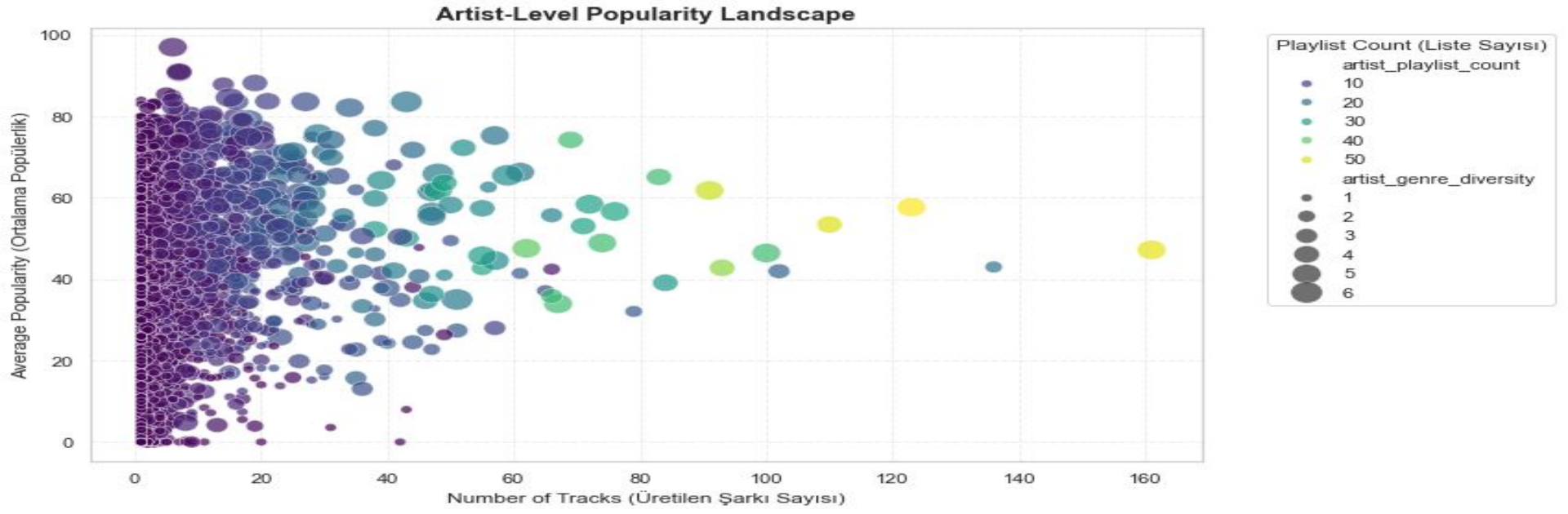
Yapılan analizler, müzik türlerinin ses özellikleri açısından belirgin biçimde ayrıştığını, ancak bir şarkının popülerliğini yalnızca bu özelliklerle açıklamanın yeterli olmadığını ortaya koymaktadır. Energy, valence ve danceability gibi faktörler popülerlikle ilişkili olsa da, türler arasındaki farklılıklar bu etkinin sınırlı kaldığını göstermektedir. Bu sonuç, popülerliğin sadece ses veya zaman temelli unsurlardan değil; sanatçı özellikleri, playlist yapısı ve dinleyici eğilimleri gibi daha geniş bağlamsal faktörlerden de etkilendiğini göstermekte, dolayısıyla modelin bu tür bilgileri de içermesi gerektiğine işaret etmektedir.

Özellik mühendisliği ve sanatçı bazında analizler

Grafik, sanatçıların üretkenlik, tür çeşitliliği ve playlist görünürlüğü açısından nasıl konumlandığını gösteriyor.

Genel olarak, **çok sayıda şarkısı ve yüksek playlist görünürlüğü** olan sanatçılar daha yüksek ortalama popülerliğe sahip.

Ancak az sayıda şarkısı olup **çok sayıda listede yer alan sanatçılar** da dikkat çekiyor — bu da *kalite ve vitrin etkisinin* popülerlik üzerinde belirleyici olduğunu gösteriyor. Yani açıkça sanatçı bazında değişkenlerin popülerliği etkilediğini görebiliyoruz.



Özellik Mühendisliği (Feature Engineering)

🕑 Temporal Features — Zaman Tabanlı Bilgiler

Amaç: Müzik trendlerinin zamanla değişimini yakalamak

- `release_year`: Şarkının çıkış yılı
- `song_age`: 2020 tabanlı yaş (2020 - release_year)
- `is_pre_spotify_era`: 2010 öncesi dönemi işaretler
- `is_2010s`: 2010–2019 arası
- `is_recent`: 2019–2020 arası yakın dönem

Zaman faktörü: Trend değişimlerinin popülerliğe etkisini modeller.

🎧2️⃣ Playlist Features — Görünürlük & Vitrin Etkisi

Amaç: Şarkının Spotify'da ne kadar "öne çıkarıldığını" ölçmek

- `playlist_size`: Playlist'teki toplam şarkı sayısı
- `playlist_count`: Aynı şarkının kaç playlist'te geçtiği
- `is_editorial`: Spotify resmi listelerinde yer alma durumu

Playlist görünürlüğü, popülerlik için doğrudan vitrin etkisi yaratır.

Özellik Mühendisliği (Feature Engineering)

🎤 Artist Intelligence Features — Sanatçı Bilgileri

Amaç: Sanatçının üretkenliğini ve çok yönlülüğünü modellemek

- `artist_track_count`: Toplam şarkı sayısı
- `artist_genre_diversity`: Yer aldığı farklı tür sayısı
- `artist_career_length`: İlk ve son şarkı yılı farkı
- `artist_playlist_count`: Farklı playlist sayısı
- `artist_exposure_score_log`: Normalize edilmiş bileşik görünürlük skoru

Popülerliği "sanatçı ayak izi" üzerinden dolaylı olarak tahmin eder. Ayrıca sanatçıların ses özelliklerinin ortalama değerleri de veri setine eklendi.

🎼 Musical Profile (PCA) — Müzikal Öz Nitelikler

Amaç: Çok boyutlu müzik verisini pca ile sadeleştirmek:

- `music_pca_1`: Enerji yoğunluğu
- `music_pca_2`: Duygusal ton
- `music_pca_3`: Ritim karmaşıklığı

Interaction Features — Etkileşimli Özellikler

🔗 **Amaç:** Müzikal değişkenlerin birlikte etkisini göstermek

- `energy_dance = energy × danceability`
- `valence_energy = valence × energy`
- `speech_live = speechiness × liveness`
- `acoustic_energy = acousticness × energy`

Gerçek müzik karakteristiği, bu etkileşimlerle daha net yakalanır.

Modelleme

V1–V5 | Genre-Based Başlangıç Dönemi

- Yalnızca audio ve genre bilgileri (danceability, energy, valence, tempo, playlist_genre) kullanıldı.
- Genre similarity için cosine distance ve genre_map koordinatları üretildi.
- Amaç: Müzikal yakınlık popülerliği açıklar mı?
- Sonuç: Accuracy ≈ 0.4-0.6 — türler arası fark anlamlı ama sınırlı etki.

V6–V8 | Çoklu Model & Neural Embedding Dönemi

- LightGBM, CatBoost, Random Forest ve ilk neural network denemeleri yapıldı.
- Audio + Genre özellikleriyle modellerin kıyaslaması yapıldı.
- Sonuç: Accuracy 0.6 bandında, model varyansı yüksek.
- Farkındalık: Yalnızca ses temelli bilgi yeterli değil

V9–V10 | Artist & Playlist Bilgisi Eklenmesi

- Sanatçı (track_artist) ve playlist bazlı (playlist_name, playlist_genre) bilgiler dahil edildi.
- Temporal özellikler eklendi: release_year, is_2010s, is_pre_spotify_era.
- Playlist vitrin etkisi incelendi: playlist_count, is_editorial.
- V10A (Random Split): 0.78 accuracy- V10B (Artist Split): 0.69 accuracy fakat data leak sebepli yüksek oranlar belirlendi.
- Fark: Model aynı sanatçıyı görünce ezber yapabiliyor → generalization sorunu.

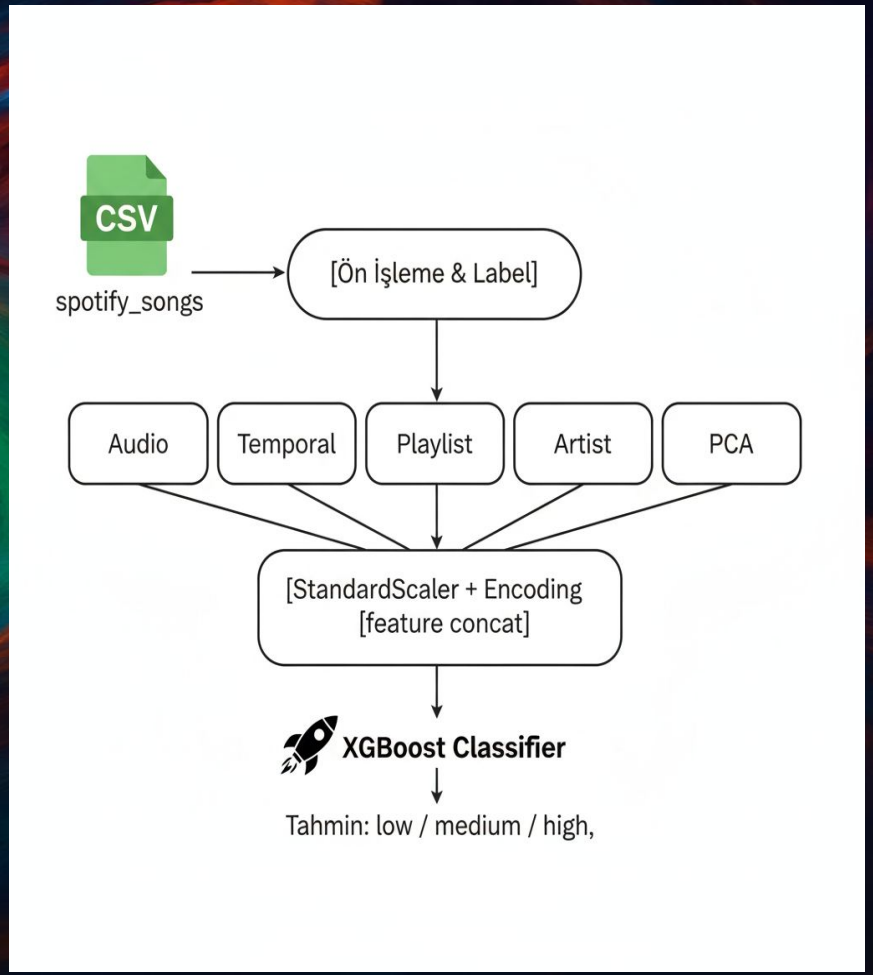
V11 | Artist Intelligence & Leak-Free Trendler

- Doğrudan popülerlik kullanılmadan artist bazlı özellikler üretildi:
    - artist_exposure_score, artist_career_length, artist_recent_activity
- Model artık “sanatçının geçmişi” ve “üretkenliği”ni de anlamlandırabiliyor.
- V11A (Random Split): 0.80 acc / 0.81 F1- V11B (Artist Split): 0.71 acc / 0.65 F1

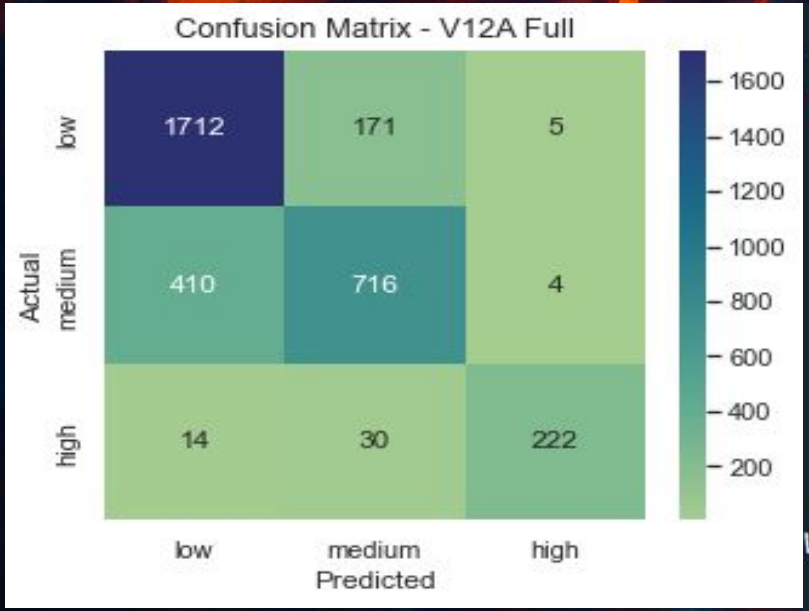
Model Mimarisi

V12 | Trend + PCA + Exposure Log (Final Model)

- Zaman düzeltmesi: Song age = 2020 – release_year
- PCA (3 bileşen): müzikal profili sıkıştırılmış temsille gösterdi.
- Artist görünürlüğü: exposure_score_log
- Interaction terms: energy_dance, valence_energy, acoustic_energy, speech_live
- Final Results: Accuracy 0.809 | Macro F1 0.816
- Model, az “high” örneğe rağmen sınıf dengesini korudu.



Model Çıktıları



📊 Classification Report:

	precision	recall	f1-score	support
low	0.80	0.91	0.85	1888
medium	0.78	0.63	0.70	1130
high	0.96	0.83	0.89	266
accuracy			0.81	3284
macro avg	0.85	0.79	0.81	3284
weighted avg	0.81	0.81	0.80	3284

- Modelin genel doğruluk oranı **%81** olup, bu tür çok sınıflı popülerlik tahminlerinde literatürdeki benzer çalışmalarla karşılaştırıldığında oldukça **başarılı bir sonuç** olarak değerlendirilebilir.
- Model, özellikle **yüksek** ve **düşük popülerlikteki şarkıları** güçlü biçimde ayırt ederken, **orta düzey popülerlikteki** şarkılarda doğal belirsizlikler gözlenmiştir. Bu durum, müzik popülerliği tahminlerinde yaygın bir zorluktur çünkü orta popülerlik grubundaki parçalar genellikle hem ses hem de dinlenme özellikleri bakımından heterojendir.
- Genel olarak, modelin bu düzeydeki performansı; oluşturulan **özellik mühendisliği stratejisinin**, müzik verisindeki karmaşık örüntüleri anlamada **etkili ve literatüre uygun** bir yaklaşım sunduğunu göstermektedir.

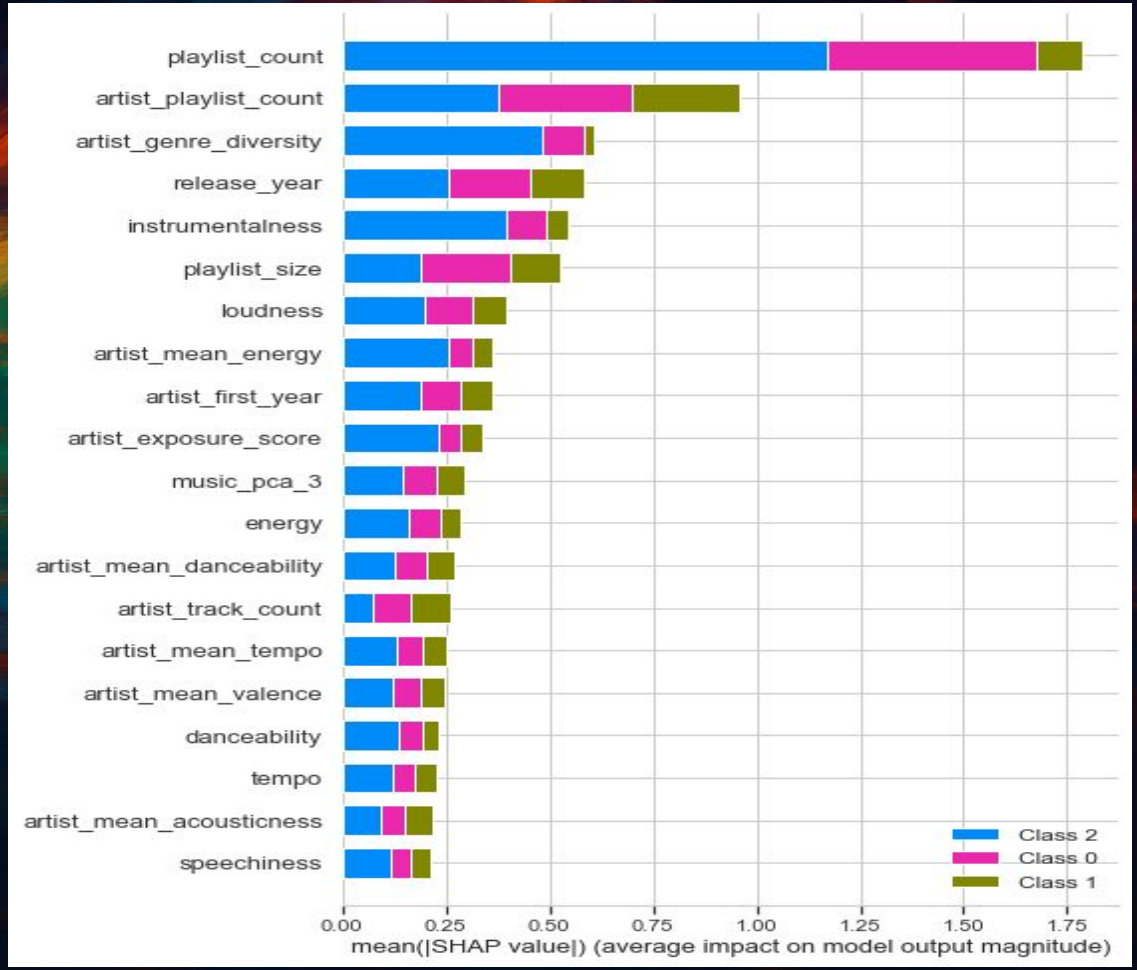
SHAP

Bu SHAP analizi, modelin hangi özelliklerden en çok etkilendiğini açıkça gösteriyor.

Grafiğe göre, **modelin kararlarında en belirleyici faktörler** şunlardır:

- 🎧 **Playlist Count**: Bir şarkının kaç farklı çalma listesinde yer aldığı, popülerlik üzerinde en büyük etkiye sahip. Bu, şarkının Spotify ekosisteminde ne kadar görünür olduğunun güçlü bir göstergesi.
- 🧜‍♀️ **Artist Playlist Count** ve **Artist Genre Diversity**: Sanatçının ne kadar farklı türde listelerde bulunduğu ve müzikal çeşitliliği de popülerlik üzerinde önemli bir rol oynuyor.
- 📅 **Release Year**: Yeni dönem şarkıların (özellikle 2019–2020 arası) popülerliğe katkısı yüksek, bu da zaman faktörünün önemini destekliyor.
- 🎵 **Instrumentalness**, **Loudness** ve **Energy** gibi ses temelli değişkenler ise popülerliğe katkı sağlasa da, liste ve sanatçı bazlı özellikler kadar belirleyici değil.

Özetle, model **popülerliği sadece müzikal özelliklere değil**, aynı zamanda **şarkının ekosistem içindeki konumuna** (örneğin playlist ve sanatçı görünürlüğü) dayalı olarak açıklıyor. Bu durum, müzik endüstrisinde popülerliğin yalnızca “nasıl duyulduğuyla” değil, aynı zamanda “nerede ve ne kadar sık duyduğuyla” da belirlendiğini ortaya koyuyor.



Tavsiyeler ve Sonuç

🎯 Model Özeti

Analiz sonuçlarına göre modelin en güçlü belirleyicileri **playlist görünürlüğü** ve **sanatçı çeşitliliği** oldu.

- **Playlist Count**: Şarkının popülerliğinde en etkili faktör, platform içi görünürlük.
- **Artist Playlist Count** ve **Genre Diversity**: Sanatçının çok yönlülüğü ve erişim alanı da önemli rol oynuyor.
- **Audio feature'lar** (energy, tempo, valence) popülerliği tek başına açıklamakta sınırlı kaldı.

💡 Tavsiyeler

Modelin öngörü gücünü artırmak için bağlamsal ve trend tabanlı özellikler eklenebilir:

- **Artist Popularity Index** (takipçi veya dinleyici sayısı)
- **Song Trend Momentum** (zaman içi popülerlik değişimi)
- **Google Trends / Sosyal Medya Verisi** (şarkı veya sanatçı aranma hacimleri)
- **Release Recency Score** (yenilik etkisi)

Bu tür ek sinyaller, modelin sadece müzikal veriye değil, **dinleyici davranışına ve kültürel trendlere** de duyarlı hale gelmesini sağlayacaktır.

GUI PROTOTİP

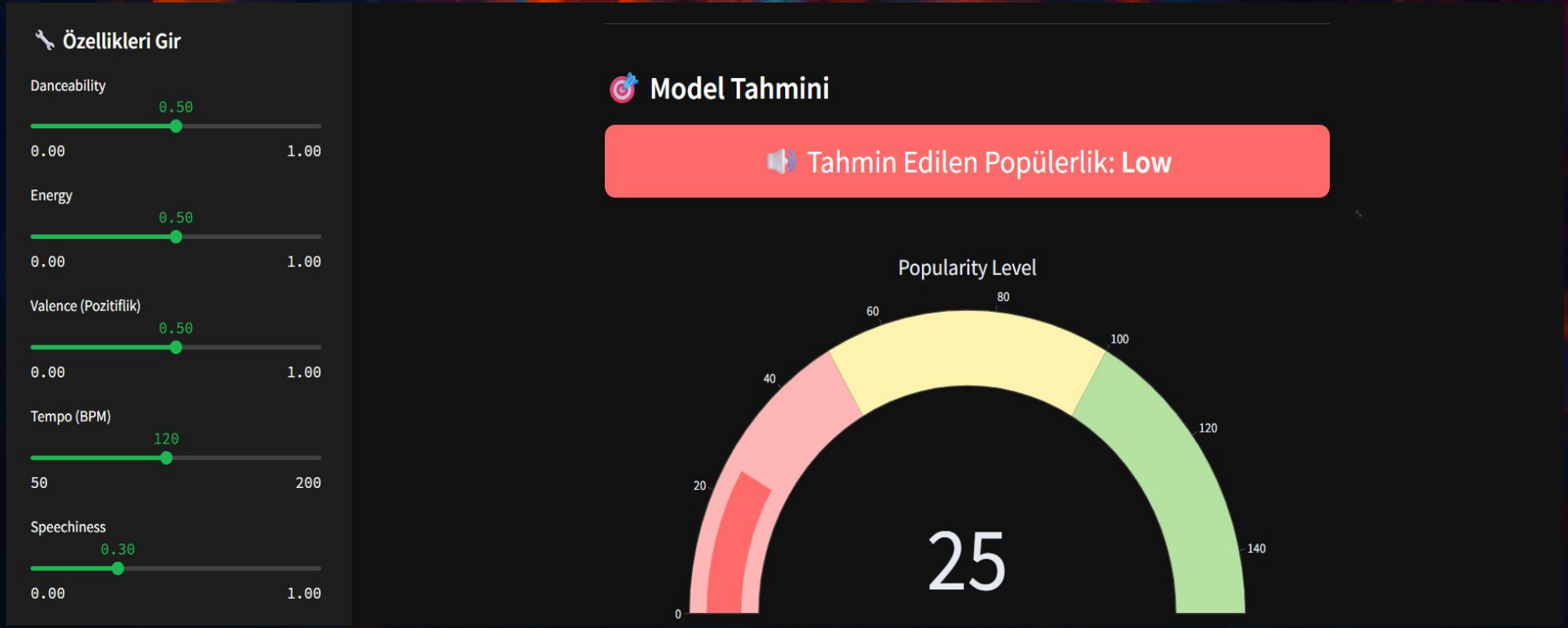
Bu aşamada modelimizin ilk etkileşimli prototipi geliştirilmiştir.
Kullanıcı, şarkıya ait temel müzikal özellikleri ve sanatçı bilgilerini girerek **anlık olarak popülerlik sınıfı (Low / Medium / High)** tahmini alabilmektedir. Şu anda güncel olarak çalışmamaktadır.

Arayüz, **Streamlit** altyapısı kullanılarak oluşturulmuş olup görsel olarak sade ve işlevsel bir yapıdadır.
Prototip şu anda temel demo işlevlerini sunmakta olup, ilerleyen aşamalarda:

- 🎧 Daha gelişmiş görselleştirmeler (ör. feature etkileri, SHAP özetleri)
- 🌐 Gerçek zamanlı Spotify / trend verisi entegrasyonu
- 📱 Kullanıcı dostu tasarım geliştirmeleri

ile genişletilmesi planlanmaktadır.

GUI PROTOTİP



“Bir şarkıyı hit yapan sadece notalar değil, onu taşıyan bağlamdır.”

Sorularınız için alicanbecoder@gmail.com