

VERİ MADENCİLİĞİ (FET445)

FİNAL PROJESİ

GRUP ADI: TRINITY

Ekip Üyeleri:

Elif Yalınkaya 22040101031

Melisa Arslantaş 22040101032

Ezgi Yıldırım 22040101048

GitHub Linki: <https://github.com/ezgy22/Veri-Madenciligi-TMDB-Proje>.

Youtube Linki: <https://youtu.be/arzc4vGaj0s>

Problem Tanımı

1

Dizi Yatırımlarındaki Belirsizlik: Her yıl binlerce yeni dizi çekiliyor ancak hangilerinin popüler olacağını öngörmek büyük bir finansal risk taşıyor.

2

Karmaşık Veri Yapısı: Tür, platform (Netflix vb.), oyuncu kadrosu ve bölüm sayısı gibi çok sayıda faktörün popülerliği nasıl etkilediği doğrusal modellerle (Linear Regression) tam olarak açıklanamıyor.

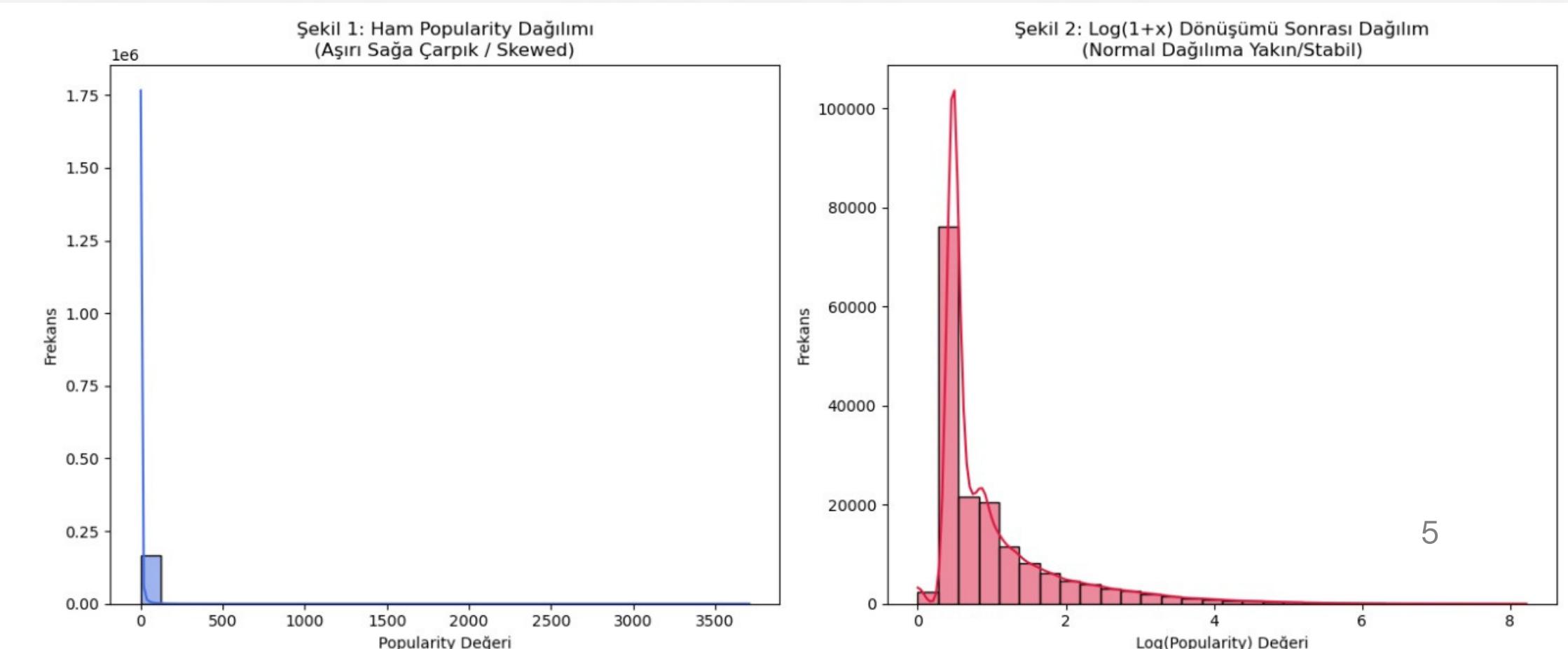
3

Sektörel İhtiyaç: Yayın platformları ve yapımcılar için "başarıyı getiren formülü" matematiksel verilere dayalı olarak ortaya koyma ihtiyacı bulunmaktadır.

Veri Seti ve Hedef

- Veri Kaynağı: TMDB (The Movie Database) TV verileri (≈ 168.000 Satır).
- Özellik Sayısı: 126 öznitelik (One-Hot Encoding sonrası).
- Temel Hedef: Dizilerin "Popülerlik Skorunu" tahmin eden regresyon ve dizileri "Popüler/Değil" olarak ayıran sınıflandırma modelleri kurmak.
- Sınıf Dengesi: Medyan değere göre bölünmüş, dengeli bir veri dağılımı ($\approx 50\%-50$).

Ham popülerlik verisindeki aşırı sağa çarpıklık (skewness), modelin öğrenme kapasitesini ve tahmin tutarlığını artırmak amacıyla $\log(1+x)$ dönüşümü ile stabilize edilmiştir. Bu işlem, özellikle regresyon modellerindeki hata payını minimize etmek için kritik bir adımdır.



Çözüm Yaklaşımı ve Metodoloji

Aşamalı Modelleme:
Sadece tek bir model değil;
Lineer, Mesafe Tabanlı, Ağaç Tabanlı ve Olasılıksal olmak üzere 6 farklı model ailesi test edildi.

Özellik Mühendisliği:
PCA, RFE ve Variance Threshold gibi yöntemlerle verideki gürültü temizlendi.

Derin Öğrenme:
Klasik modellerin ötesine geçmek için PyTorch ile ANN, ResNet-MLP ve Wide & Deep mimarileri geliştirildi.

Gelişmiş Sınıflandırma Stratejileri

Algoritmik Altyapı:

- Çalışmada Gradient Boosting (GBC), Random Forest ve XGBoost gibi güçlü ağaç tabanlı algoritmalar temel alınmıştır.

Stacking (İstifleme) Mimarisi:

- Tekil modellerin zayıf noktalarını telafi etmek amacıyla, meta-model olarak Lojistik Regresyon'un kullanıldığı bir Stacking yapısı kurgulanmıştır.

Kolektif Başarı:

- Stacking modeli, hataları minimize ederek %82.44 doğruluk oranına ulaşmış ve çalışmanın "Sınıflandırma Şampiyonu" olmuştur.

İleri Seviye Regresyon Modelleri

Hiperparametre Ayarlama:

Randomized SearchCV yöntemiyle öğrenme hızı (0.05), derinlik (10) ve ağaç sayısı (300) gibi parametreler en iyi seviyeye getirilmiştir.

Hata Metrikleri:

Tuned XGBoost modeli, $R^2 = 0.7370$ değeri ile klasik makine öğrenmesi modelleri arasında en yüksek varyans açıklama oranını yakalamıştır

Kararlılık:

- Modelin RMSE (0.4570) ve MAE (0.2739) değerlerindeki düşüş, tahmin tutarlılığının vize aşamasına göre anlamlı derecede arttığını kanıtlamaktadır.

Derin Öğrenme Mimarileri (PyTorch)

- Mimariler: Verideki doğrusal olmayan karmaşık ilişkileri modellemek için PyTorch tabanlı Optimized DNN ve Wide & Deep yapıları geliştirilmiştir.
- Optimized DNN: 128-64-32 nöronlu katmanlar, Batch Normalization ve Dropout (%20) kullanımıyla $R^2 = 0.8917$ gibi çok yüksek bir başarı oranına ulaşmıştır .
- Wide & Deep: Kategorik seyrek verilerdeki bariz kuralları (Wide) ve sayısal verilerdeki gizli desenleri (Deep) hibrit olarak işleyen mimari uygulanmıştır .
- Analiz: Derin öğrenme modellerinin AUC skorunun 0.90 olması, sınıfları ayırt etme yeteneğinin "mükemmel" seviyede olduğunu göstermektedir.

Performans Karşılaştırma Tablosu

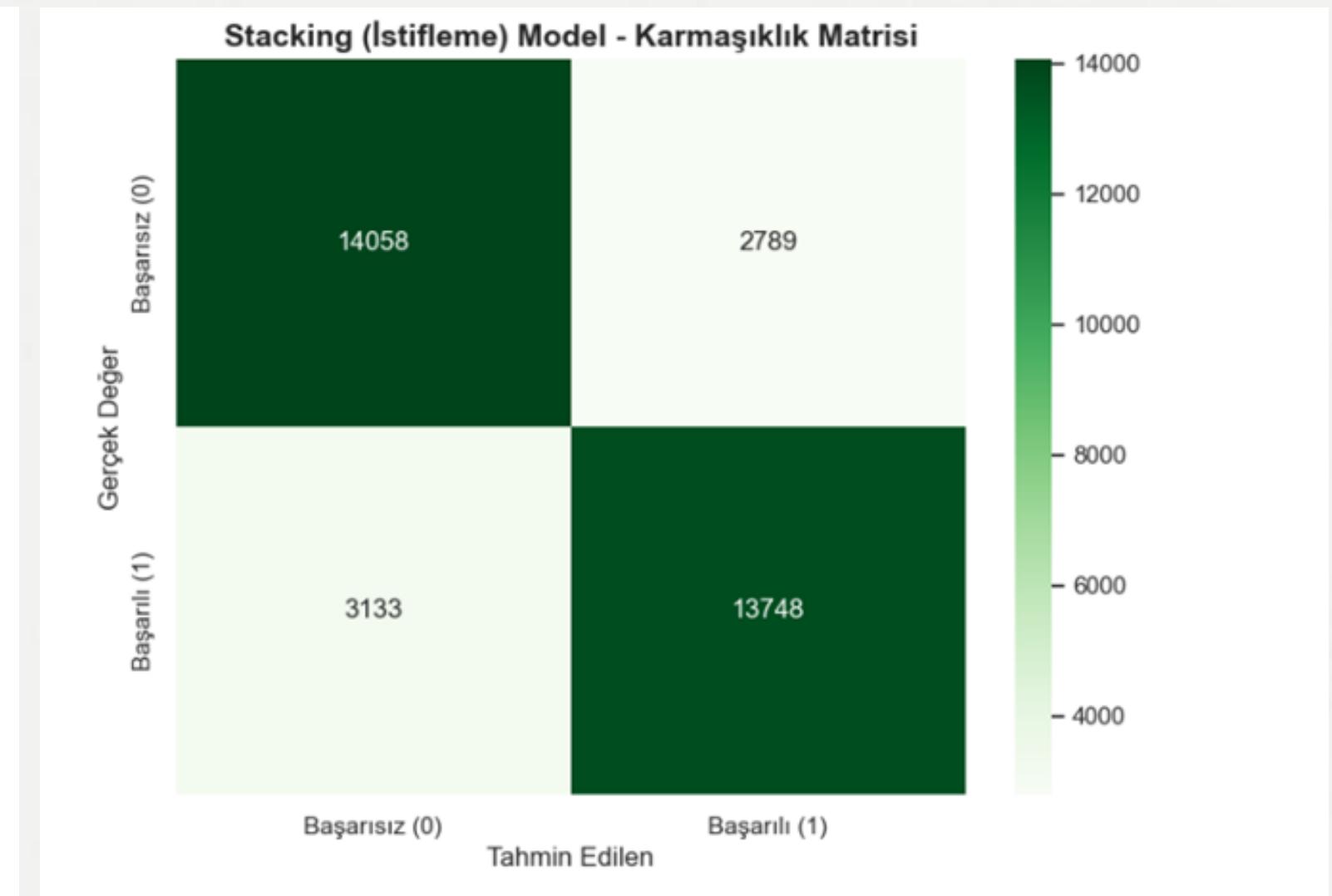
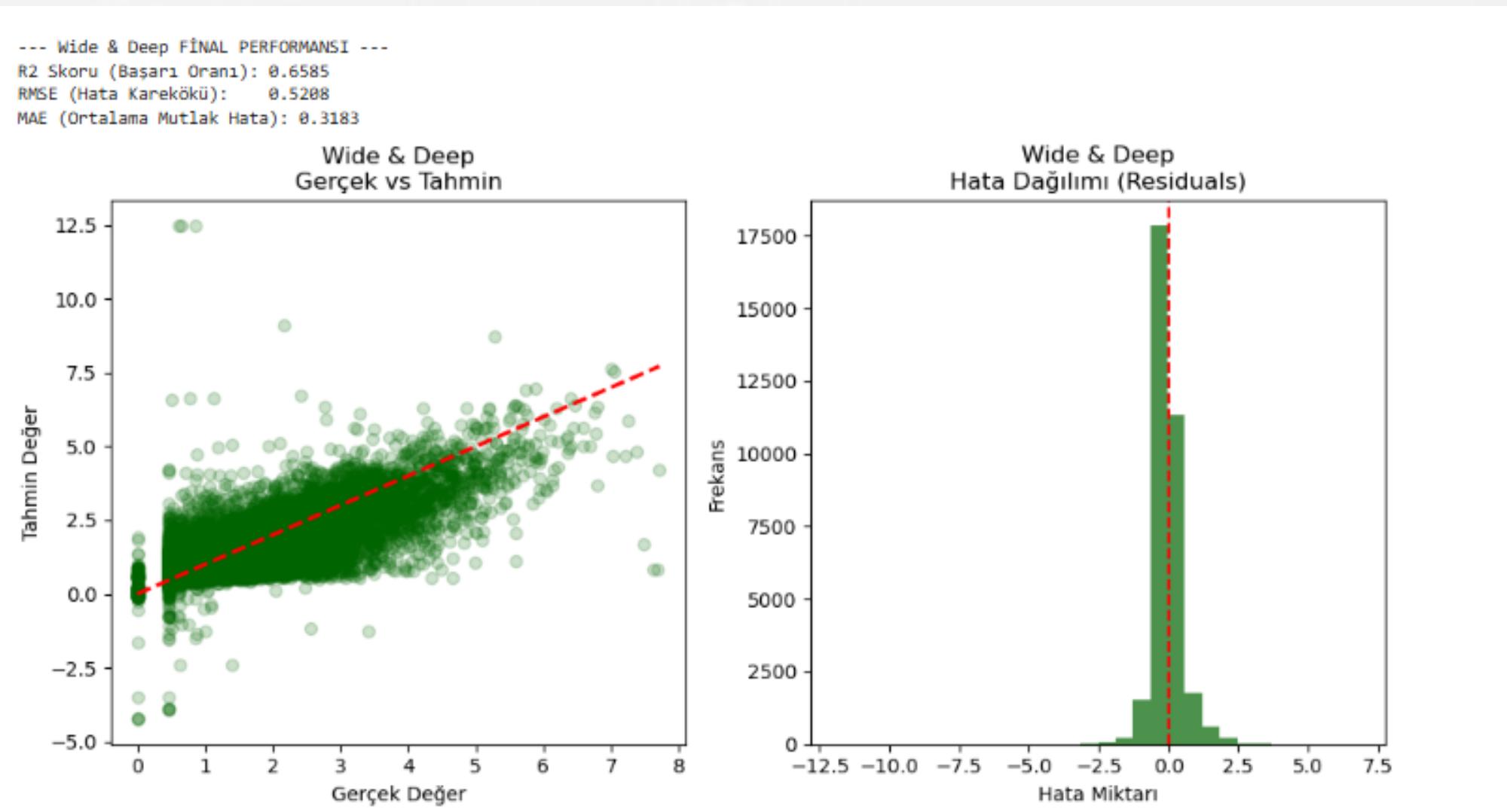
Final modelleri, vize aşamasındaki baz modellere göre %274.49 oranında devasa bir iyileşme sergilemiştir.

Model Kategorisi	En İyi Model	R2 / Accuracy
Vize (Baseline)	Linear Regression	0.1966
Final (Gelişmiş)	Tuned XGBoost	0.7370
Final (Deep Learning)	Optimized DNN	0.8917
Final (Ensemble)	Stacking (Champ)	%82.44

Hata Analizi ve Model Sınırları

- Eksik Veri Etkisi: Veri setinde oyuncu kadrosu (Cast) ve yönetmen bilgisi bulunmaması, modelin ünlü isimlerin etkisini öngörememesine neden olmaktadır .
- Aykırı Değerler: Bölüm sayısı az olmasına rağmen viral hale gelerek yüksek puan alan istisnai yapımlar, modelin öğrendiği "genel kuralları" bozabilmektedir .
- Veri Gürültüsü: Düşük oy sayısına sahip ama yüksek puan ortalamalı belgesel türleri, sınıflandırma hassasiyetini zorlaştıran temel unsurdur.

Öğrenci Adı	Gelişmiş Modeller	HiperParametreler	
Melisa Arslantaş	ML: Gradient Boosting Classifier, Stacking DL: Optimized ANN, PyTorch ResNet-MLP	RandomizedSearchCV , GridSearch	Grup üyelerinin çoklu model denemeleri sonucunda, geleneksel topluluk öğrenmesi (Ensemble) ve modern derin öğrenme (Deep Learning) tekniklerini birleştiren hibrit bir yaklaşım sergilenmiştir. Sınıflandırma görevinde hata telafisi sağlayan Stacking mimarisi %82.44 doğrulukla "Şampiyon" olurken; regresyon görevinde doğrusal olmayan karmaşık ilişkileri çözümleyen Optimized DNN mimarisi $R^2=0.8917$ başarısı ile nihai performansı maksimize etmiştir
Elif Yalınkaya	ML: Tuned XGBoost Regressor, Tuned LightGBM DL: Optimized DNN, Wide & Deep Learning	RandomizedSearchCV	
Ezgi Yıldırım	ML: Random Forest, Extra Trees DL: PyTorch Wide & Deep, PyTorch Simple MLP	Performance Comparison & Tuning	



Derin Öğrenme ve Topluluk (Ensemble) Modelleri Analizi: Geliştirilen Wide & Deep mimarisi, regresyon görevinde doğrusal olmayan karmaşık ilişkileri başarıyla modellerken; Stacking (İstifleme) yapısı, karmaşıklık matrisinde görüldüğü üzere hataları minimize ederek %82.44 sınıflandırma doğruluğuna ulaşmıştır. Bu sonuçlar, modern derin öğrenme teknikleri ile topluluk öğrenmesinin birleştirilmesinin başarısını kanıtlamaktadır.

Sonuç ve Genel Değerlendirme

Bu çalışma, TV dizisi popülerliği gibi dinamik bir parametrenin veri madenciliği teknikleriyle yüksek doğrulukla tahmin edilebileceğini kanıtlamıştır:

- **Şampiyon Modellerin Başarısı:** Sınıflandırma görevinde hata telfisi sağlayan Stacking mimarisi %82.44 doğrulukla , regresyon görevinde ise doğrusal olmayan karmaşık ilişkileri çözümleyen Optimized DNN mimarisi $R^2=0.8917$ başarısı ile nihai performansı maksimize etmiştir.
- Vize-Final Gelişimi: Final aşamasında uygulanan ileri seviye teknikler, vize aşamasındaki baz modellere göre %274.49 oranında devasa bir iyileşme sağlamıştır.
- **Sektörel Çıkarım:** Yapısal (tablo) verilerde toplu öğrenme (Ensemble) modellerinin kararlılığı görülmüş; ancak derinlemesine optimize edilmiş sinir ağlarının regresyon hassasiyetinde daha üstün olduğu saptanmıştır.
- **Öğrenilenler ve Limitler:** Tahminlerdeki %18-19'luk hata payının, veri setinde yer almayan "oyuncu kadrosu ve yönetmen" gibi gizli özniteliklerden kaynaklandığı analiz edilmiştir.
- **Final Notu:** Geliştirilen modeller, yayın platformları için yatırım risklerini minimize edebilecek ve popülerliği belirleyen kritik faktörleri (bölüm sayısı, puan vb.) matematiksel olarak ortaya koyabilecek kapasitededir.