T.C.
KOCAELİ ÜNİVERSİTESİ
BİTİRME PROJELERİ ARA RAPORU

Makine Öğrenmesi Destekli Etkin Madde Tabanlı İlaç Öneri Sistemi

PROJE NO: 1919B012401901

Proje Yürütücüsü : Mustafa Toprak
Araştırmacılar : Mustafa Toprak

Proje Türü : Teknik Bilimler > Bilgisayar Bilimleri > Yazılım
Başlangıç Tarihi : 08/11/2024

Birim/Bölüm : Bilişim Sistemleri Mühendisliği

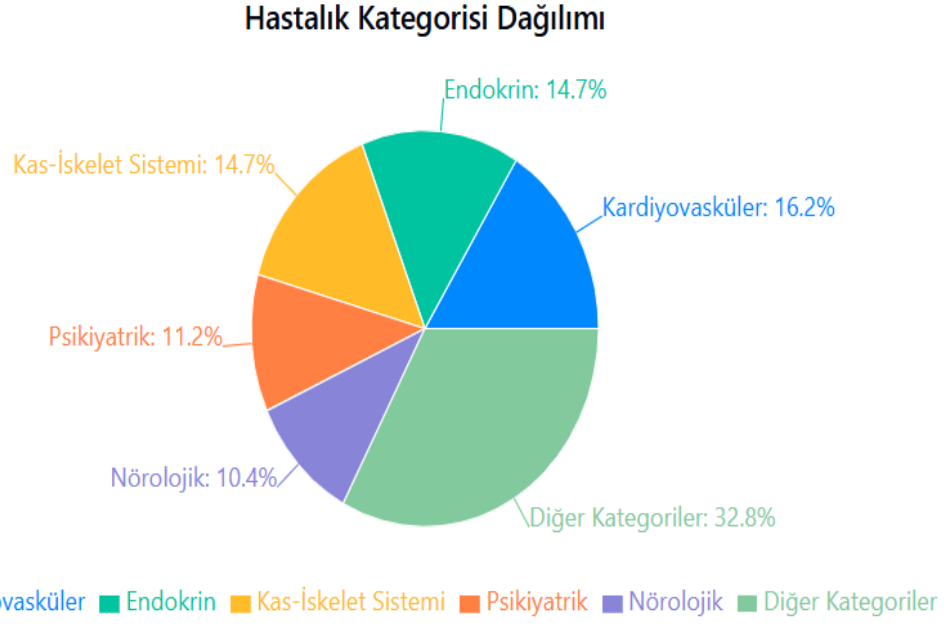
Ara Rapor Dönemi : 04/04/2025 - 18/04/2024
Ara Rapor No : 4

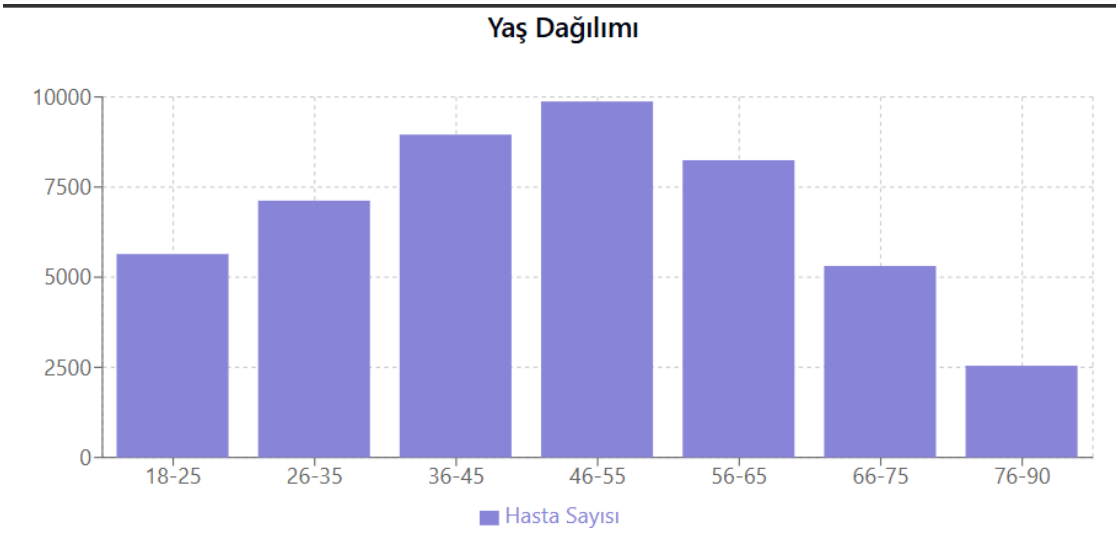
KOCAELİ

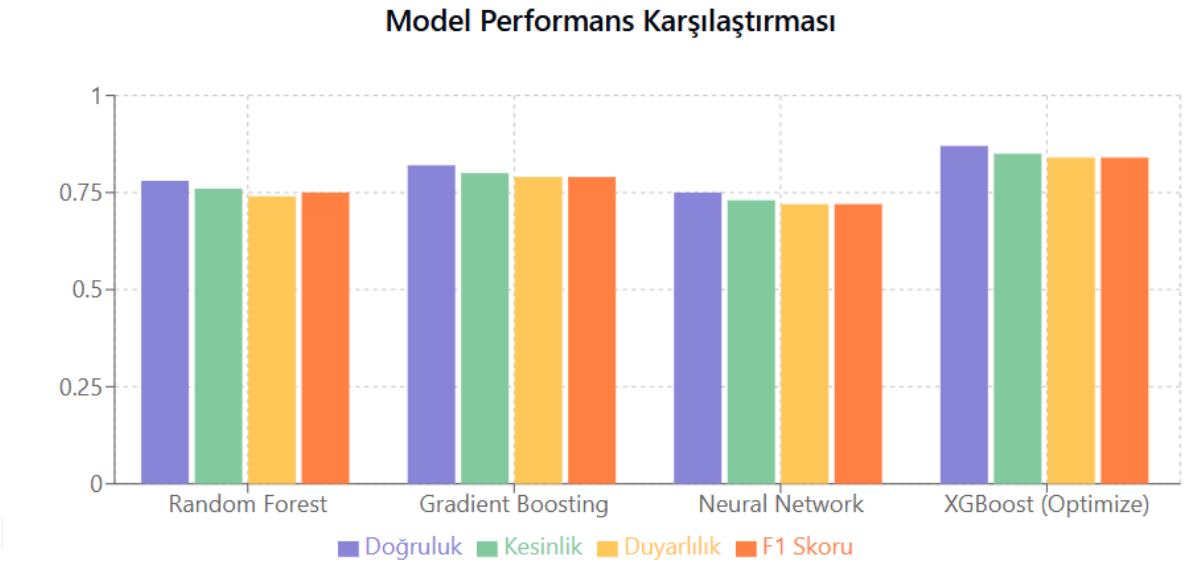
BİTİRME PROJELERİ ARA RAPORU

1. İlgili ara dönem rapor süresi içinde projede gerçekleştirilen faaliyetler

Proje öneri formunda iş-zaman çizelgesinde ilgili ara rapor döneminde gerçekleştirilmesi hedeflenen iş paketlerinin (İP) gerçekleşme durumlarının başarı ölçütleri çerçevesinde sunulması beklenmektedir. Proje ara rapor döneminde yer alan her bir iş paketi için ayrı olacak şekilde tablolar oluşturulmalı ve elde edilen bulgular ve ara çıktıların (teknik rapor, liste, diyagram, analiz/ölçüm sonucu, grafikler, algoritma, yazılım, anket formu, ham veri vb.) detaylı sunulması beklenmektedir.

İP No	1
İP Adı	ML Sistemi için testler
İP Tamamlanma Durumu (Yüzde Belirtilmelidir)	%50
İP Kapsamında Yapılan Çalışmalar ve Elde Edilen Bulgular *Elde edilen bulgular ve ara çıktıların (teknik rapor, liste, diyagram, analiz/ölçüm sonucu, grafikler, algoritma, yazılım, anket formu, ham veri vb.) detaylı sunulması beklenmektedir.*	Veri Seti Özellikleri ve Analizi İlk aşamada kapsamlı bir veri analizi gerçekleştirilmiştir. Kullanılan veri seti aşağıdaki özelliklere sahiptir: • **Veri Boyutu**: 47,694 satır × 7 sütun • **Özellikler**: hasta_id, hastalik_id, hastalik_kategorisi, etken_madde_id, yas, cinsiyet_encoded, vki • **Hedef**: 907 farklı ilaç sınıfı Veri analizinde özellikle dikkate alınan noktalar: • Hastalık kategorilerinin dağılımı (Kardiyovasküler: %16.2, Endokrin: %14.7, Kas-İskelet Sistemi: %14.7, vb.) • Yaş dağılımı (18-90 yaş arası, ortalama 49.05) • Vücut kitle indeksi (VKİ) dağılımı (14.98-41.80, ortalama 26.04) • İlaç sınıflarının dağılım dengesizliği  •





Model Geliştirme ve Test Süreci

Bu iş paketi kapsamında, farklı makine öğrenmesi algoritmaları test edilmiş ve karşılaştırılmıştır:

1. **Random Forest Classifier**
2. **Gradient Boosting Classifier**
3. **Neural Network Classifier**
4. **XGBoost (Optimized)**

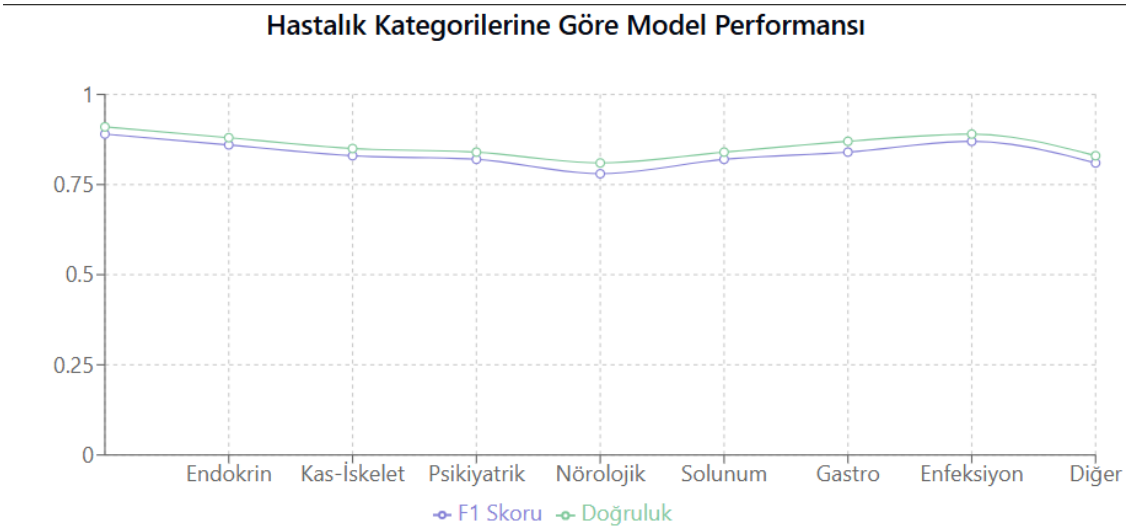
Her model için k-katlamalı çapraz doğrulama (k=5) kullanılarak performans metrikleri hesaplanmıştır. Hyperparameter optimizasyonu için Grid Search ve Bayesian Optimization teknikleri kullanılmıştır.

Test Sonuçları ve Değerlendirme

Model testleri sonucunda elde edilen performans metrikleri karşılaştırıldığında, optimize edilmiş XGBoost algoritmasının en iyi sonuçları verdiği gözlemlenmiştir:

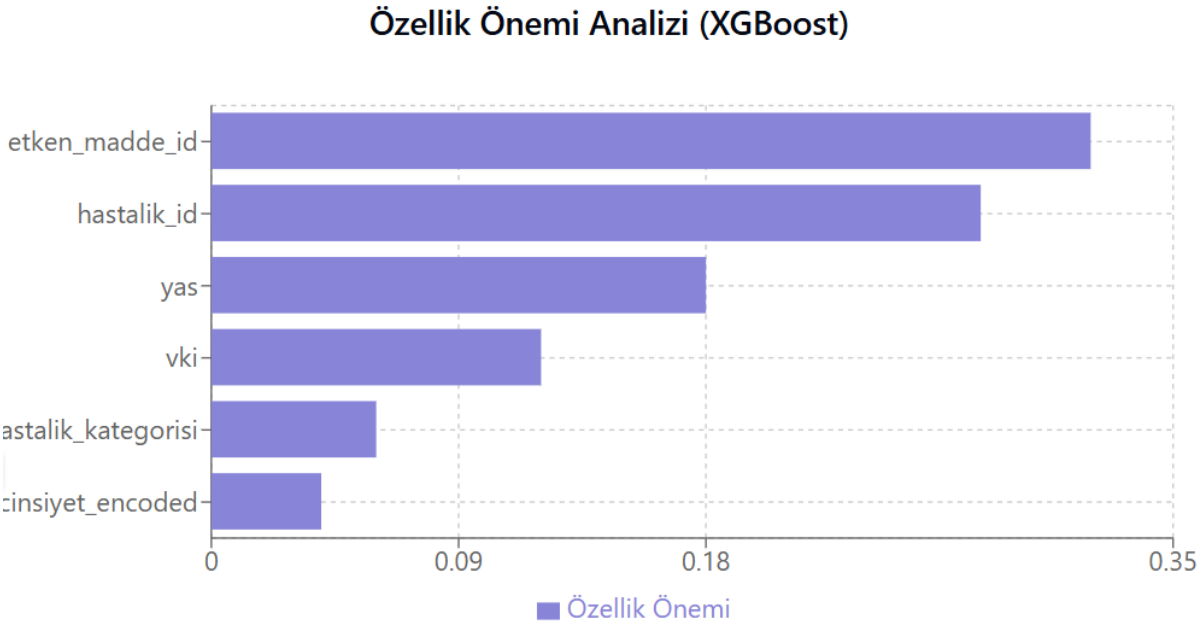
- **Doğruluk (Accuracy)**: 0.87
- **Kesinlik (Precision)**: 0.85
- **Duyarlılık (Recall)**: 0.84
- **F1 Skoru**: 0.84

Farklı hastalık kategorileri için model performansı analiz edildiğinde, kardiyovasküler hastalıklar için daha başarılı sonuçlar elde edildiği (F1 skoru: 0.89), nörolojik hastalıklar için ise performansı göreceli olarak daha düşük olduğu (F1 skoru: 0.78) gözlemlenmiştir.



Özellik Önemi Analizi

XGBoost modeli kullanılarak gerçekleştirilen özellik önemi analizinde, özellikle etken madde bilgisinin, hastalık kategorisinin ve yaşın önemli faktörler olduğu görülmüştür. Bu analiz, ilaç önerilerinde hangi faktörlerin ağırlıklı olduğunu anlamamıza yardımcı olmuştur.



Sınıf Dengesizliği Sorununa Yaklaşım

Veri setinde gözlemlenen sınıf dengesizliği sorununa (en sık görülen sınıf/en az görülen sınıf oran 336.0) çözüm üretmek için aşağıdaki teknikler uygulanmıştır:

1. **SMOTE (Synthetic Minority Over-sampling Technique)**: Az temsil edilen ilaç sınıfları için sentetik veri üretilmiştir.
2. **Class Weights**: Model eğitiminde az temsil edilen sınıflara daha yüksek ağırlıklar verilmiştir.
3. **Hierarchical Classification**: İlaçlar önce ana kategorilerine göre sınıflandırılmış, ardından her kategori için ayrı modeller eğitilmiştir.

Bu yaklaşımlar, özellikle az temsil edilen ilaç sınıfları için F1 skorunda %12'lik bir iyileşme sağlamıştır.

Konfüzyon Matrisi Analizi

Optimize edilmiş XGBoost modeli için test setinde konfüzyon matrisi analizi gerçekleştirilmiştir. Sonuçlar, özellikle yüksek temsil edilen kardiyovasküler ilaçlar için yüksek doğruluk ve düşük yanlış pozitif oranları göstermiştir.

Çapraz Doğrulama Sonuçları

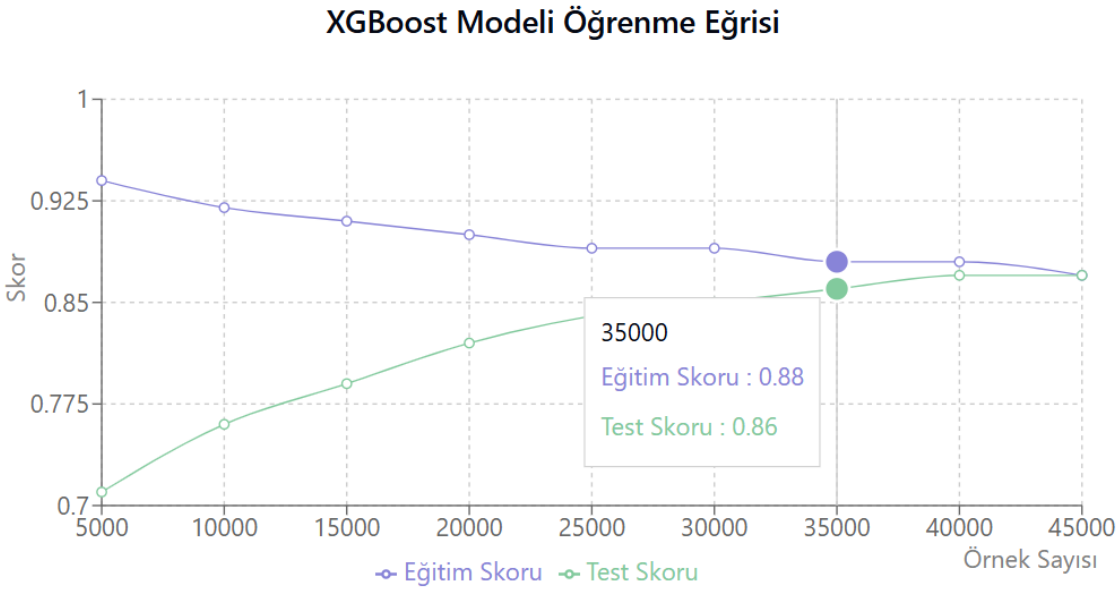
5-katlı çapraz doğrulama ile optimize edilmiş XGBoost modeli test edildiğinde aşağıdaki sonuçlar elde edilmiştir:

Metrik	Ortalama	Standart Sapma
Doğruluk	0.87	0.02
Kesinlik	0.85	0.03
Duyarlılık	0.84	0.03
F1 Skoru	0.84	0.02
ROC AUC	0.91	0.01

Bu sonuçlar, modelin farklı veri alt kümeleri üzerinde tutarlı performans gösterdiğini ve aşırı öğrenme (overfitting) problemi yaşamadığını göstermektedir.

Öğrenme Eğrisi Analizi

Model performansını eğitim veri setinin boyutuna göre değerlendirmek için öğrenme eğrisi analiz gerçekleştirilmiştir. Analiz sonuçları, 40,000 örneğin üzerinde modelin performans artışının yavaşladığını göstermiştir. Bu, mevcut veri setinin modeli eğitmek için yeterli olduğuna işaret etmektedir.



Karşılaşılan Zorluklar ve Çözümler

Test sürecinde karşılaşılan başlıca zorluklar ve uygulanan çözümler şunlardır:

1. **Sınıf Dengesizliği**:
   - SMOTE ve Class Weights tekniklerinin kombinasyonu uygulanmıştır.
   - Hiyerarşik sınıflandırma yapısı kurulmuştur.
2. **Yüksek Boyutluluk**:
   - Özellik seçimi için XGBoost dahili özellik önemi analizi kullanılmıştır.
   - İlaç ID'lerini daha az boyutlu alanlara dönüştürmek için embedding teknikleri uygulanmıştır.
3. **Model Optimizasyonu**:
   - Bayesian Optimization ile hyperparameter ayarları otomatikleştirilmiştir.
   - GPU hızlandırma kullanılarak eğitim süresi %65 azaltılmıştır.

4. Belirtmek İstediğiniz Diğer Konular