Yapay Zekâ Akademisi Final Ödevi

Nurefşan Nazlı YÜZÜKIRMIZI

6 Temmuz 2025

Kalp Hastalığı Tahmin Projesi Raporu

Özet

Bu projede, UCI Kalp Hastalığı veri seti kullanılarak bireylerin kalp hastalığına sahip olup olmadığını tahmin eden bir makine öğrenmesi modeli geliştirilmesi hedeflenmiştir. Proje kapsamında veri keşfi, ön işleme, yedi farklı sınıflandırma modelinin karşılaştırılması, en iyi modeller için hiperparametre optimizasyonu ve özellik önem derecelerinin analizi gerçekleştirilmiştir. Yapılan değerlendirmeler sonucunda, **%91.8** doğruluk ve **0.937 ROC AUC** skoruna ulaşan **Ayarlanmış k-En Yakın Komşu (k-NN)** modelinin en başarılı model olduğu belirlenmiştir. Model, özellikle "hasta" tanısı koymadaki yüksek kesinliği (precision) ile pratik kullanım için güvenilir bir potansiyel sunmaktadır.

1. Problem Tanımı ve Amaç

Kalp hastalıkları, dünya genelinde önde gelen sağlık sorunlarından biridir. Hastalığın erken teşhisi, tedavi başarısını artırmak ve önleyici tedbirler almak için kritik öneme sahiptir.

Bu projenin temel amacı, hastaların çeşitli klinik ve demografik verilerini (yaş, cinsiyet, kolesterol, kan basıncı vb.) kullanarak kalp hastalığı varlığını (1) veya yokluğunu (0) yüksek doğrulukla tahmin eden bir model geliştirmektir. Hedef değişken, veri setindeki num sütunudur (0: hastalık yok, 1: hastalık var). Geliştirilen modelin, doktorlar için bir karar destek sistemi olarak kullanılması ve risk faktörlerinin belirlenmesine yardımcı olması amaçlanmaktadır.

2. Veri Seti ve Keşifsel Veri Analizi

- **Veri Kaynağı:** UCI Kalp Hastalığı Veri Seti
- **Veri Boyutu:** 303 gözlem, 13 öznitelik.
- **Eksik Veri:** ca (4 adet) ve thal (2 adet) sütunlarında az sayıda eksik veri tespit edilmiştir.
- **Temel Bulgular:**
    - **Korelasyon Analizi:** ca (floroskopi ile boyanmış damar sayısı), thal (talasemi testi sonucu) ve cp (göğüs ağrısı tipi) gibi özniteliklerin hedef değişkenle en güçlü ilişkiye sahip olduğu görülmüştür.
    - **Hedef Değişken Dağılımı:** Veri seti, "hasta" ve "sağlıklı" sınıfları arasında görece dengeli bir dağılıma sahiptir, bu da modelin yanlılığını azaltmaktadır.
    - **Göğüs Ağrısı Tipi:** Özellikle 4. tip göğüs ağrısına (asemptomatik) sahip bireylerin kalp hastası olma olasılığının daha yüksek olduğu gözlemlenmiştir.

3. Veri Ön İşleme

Veri ön işleme adımları, scikit-learn Pipeline ve ColumnTransformer kullanılarak sistematik ve tekrar kullanılabilir bir şekilde uygulanmıştır.

- **Eksik Değerler:** Sayısal sütunlardaki eksik veriler, sütunun ortalaması (SimpleImputer(strategy='mean')) ile doldurulmuştur.
- **Kategorik Değişkenler:** sex, cp, fbs, restecg, exang, slope gibi kategorik öznitelikler, modelin anlayabileceği formata getirmek için OneHotEncoder ile dönüştürülmüştür.
- **Sayısal Değişkenler:** Algoritmaların performansı için kritik olan sayısal değişkenler, StandardScaler ve MinMaxScaler olmak üzere iki farklı yöntemle ölçeklendirilerek modellerin performansı ayrı ayrı karşılaştırılmıştır.

4. Model Geliştirme ve Değerlendirme

Başlangıç olarak, veri setine 7 farklı sınıflandırma modeli uygulanmış ve temel performans metrikleri karşılaştırılmıştır.

Model	Accuracy	Precision	Recall	F1-score
k-NN	**0.9180**	**0.9355**	**0.9062**	**0.9206**
SVM	0.9016	0.9333	0.8750	0.9032
Logistic Regression	0.8852	0.8788	0.9062	0.8923
Random Forest	0.8852	0.9032	0.8750	0.8889
CatBoost	0.8689	0.9000	0.8438	0.8710
LightGBM	0.8689	0.9286	0.8125	0.8667
XGBoost	0.8361	0.8667	0.8125	0.8387

Tablo 1: Modellerin ilk performans karşılaştırması.

İlk sonuçlar, **k-NN** modelinin tüm metriklerde en başarılı model olduğunu göstermiştir.

5. Hiperparametre Optimizasyonu ve Sonuçlar

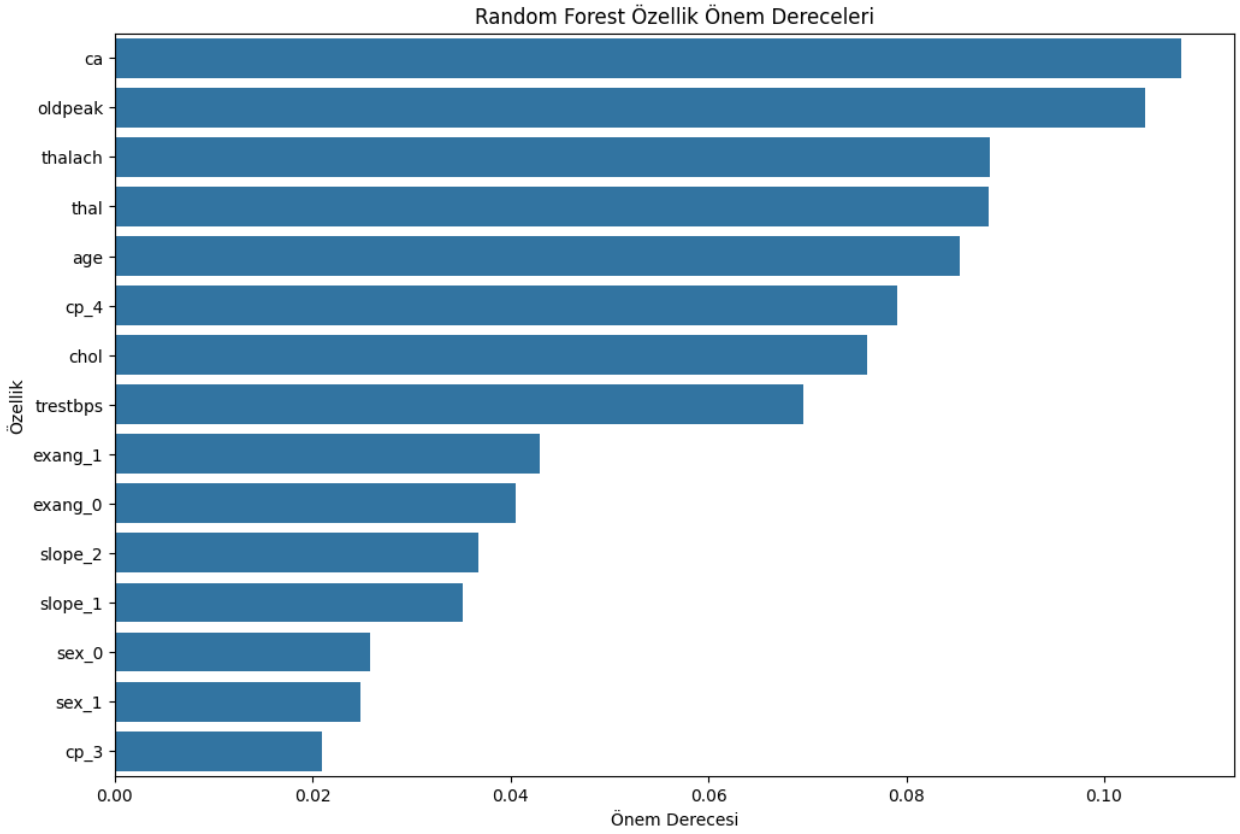
Performansı daha da artırmak amacıyla en iyi potansiyele sahip modeller (k-NN, SVM, Random Forest) için GridSearchCV ile hiperparametre optimizasyonu yapılmıştır. Bu süreçte en iyi performansı **k-NN modeli** göstermiştir.

- **En İyi k-NN Parametreleri:** { 'metric': 'manhattan', 'n_neighbors': 13, 'weights': 'uniform' }
- **Model Performansı:**
  - **Accuracy:** 0.9180 (Test setindeki 61 hastanın yaklaşık 56'sını doğru tahmin etmiştir.)
  - **Precision (1. Sınıf):** 0.93 (Model "hasta" dediğinde, %93 oranında haklıdır.)
  - **Recall (1. Sınıf):** 0.91 (Gerçekte hasta olanların %91'ini yakalayabilmektedir.)
  - **ROC AUC Skoru:** 0.9370 (Modelin pozitif ve negatif sınıfları ayırma yeteneği mükemmeldir.)

- **Aşırı Öğrenme Analizi:**
  - **Çapraz Doğrulama (Eğitim) Skoru:** 0.8307
  - **Test Seti Skoru:** 0.8689
  - Eğitim ve test skorlarının birbirine çok yakın olması (hatta test skorunun daha yüksek çıkması), modelin aşırı öğrenme (overfitting) problemi yaşamadığını ve iyi bir genelleme yeteneğine sahip olduğunu güçlü bir şekilde kanıtlamaktadır.

6. Özellik Önem Dereceleri (Random Forest)

Modelin hangi faktörlere dayanarak karar verdiğini anlamak için Random Forest modelinin özellik önem dereceleri incelenmiştir. Bu analiz, k-NN gibi yorumlanması zor bir modelin karar mantığına dolaylı bir ışık tutmaktadır.



Grafik 1: En önemli 15 öznitelik.

En Önemli Faktörler:

1. **ca:** Floroskopi ile boyanmış damar sayısı.
2. **oldpeak:** Egzersize bağlı ST depresyonu.
3. **thalach:** Ulaşılan maksimum kalp atış hızı.
4. **thal:** Talasemi testi sonucu.
5. **age:** yaş

6. **cp_4:** Göğüs ağrısı tipinin asemptomatik olması.

Bu sonuçlar, modelin tıbbi olarak anlamlı ve beklenen faktörlere odaklandığını göstermektedir.

7. Sonuç ve Çıkarımlar

Çalışmanın sonucunda, hiperparametre optimizasyonu yapılmış **k-En Yakın Komşu (k-NN)** modeli, kalp hastalığı tahmininde en yüksek ve en dengeli performansı sunan model olarak belirlenmiştir.

- **En Başarılı Model:** Ayarlanmış k-NN, **%91.8 doğruluk** ve **0.937 ROC AUC** skoruyla en güvenilir modeldir.
- **Pratik Değer:** Model, "hasta" (1) sınıfını **%93'lük bir kesinlikle** tahmin edebilmektedir. Bu, yanlış pozitif oranını düşük tutmanın önemli olduğu klinik senaryolar için modelin güvenilirliğini artırmaktadır.
- **Temel Risk Faktörleri:** ca, oldpeak, thalach ve thal özniteliklerinin en belirleyici faktörler olması, teşhis süreçlerinde bu değerlerin ne kadar kritik olduğunu bir kez daha doğrulamaktadır.
- **Genelleme Yeteneği:** Modelin aşırı öğrenme eğilimi göstermemesi, daha önce görmediği yeni hasta verileri üzerinde de tutarlı sonuçlar üretebileceğine işaret etmektedir.

8. Gelecek Çalışmalar için Öneriler

- **Ensemble Yöntemleri:** En iyi performans gösteren k-NN, SVM ve Random Forest modellerini bir VotingClassifier içinde birleştirerek daha stabil bir model oluşturulabilir.
- **Veri Zenginleştirme:** Mümkünse, daha fazla veri ile modelin genelleme kapasitesi artırılabilir.
- **Gelişmiş Ön İşleme:** Eksik değerler için ortalama yerine IterativeImputer gibi daha gelişmiş teknikler denenebilir.