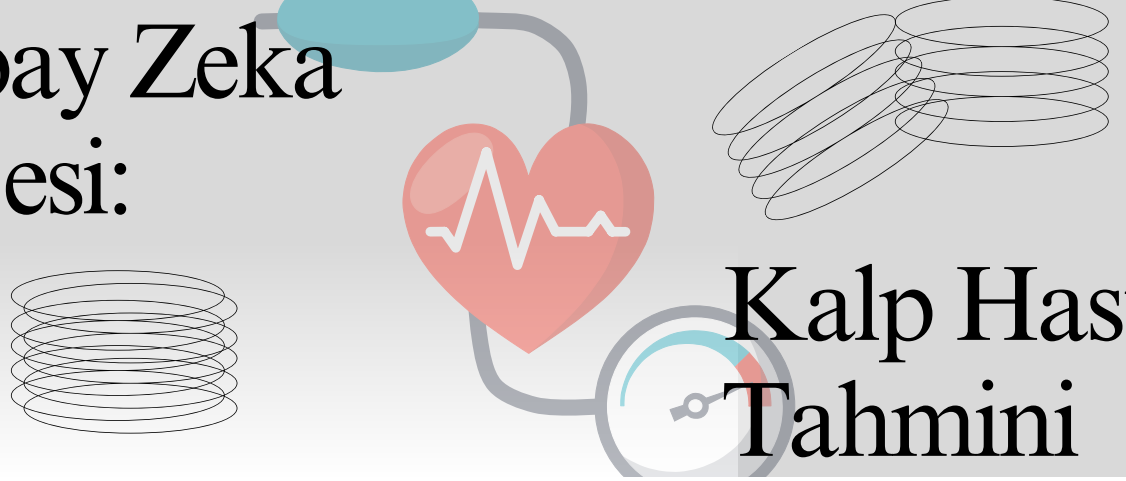
Yapay Zeka Projesi:



Kalp Hastalığı Tahmini

Proje Sahibi: Elif Karadeniz
Proje Türü: Yapay Zeka Akademisi – Bitirme Projesi

Proje Özeti

Problemin Arka Planı ve Alan İçindeki Yeri

Kalp-damar hastalıkları, dünyada ve Türkiye'de en yaygın ölüm nedenleri arasında yer almakta ve halk sağlığı için büyük bir tehdit oluşturmaktadır. Dünya Sağlık Örgütü'nün 2019 verilerine göre, her yıl yaklaşık **17,9 milyon** kişi kalp hastalıkları nedeniyle hayatını kaybetmektedir. Türkiye İstatistik Kurumu'nun 2023 verileri, ölümlerin **%33,4'**ünün kalp ve damar hastalıklarından kaynaklandığını göstermektedir.

Hipertansiyon gibi sessiz ilerleyen risk faktörleri, hastaların çoğunlukla hastalık belirtileri ortaya çıktıktan sonra teşhis edilmesine yol açmaktadır. Bu nedenle, kalp hastalıklarının erken teşhisi ve risk sınıflandırması halk sağlığı açısından kritik öneme sahiptir. Geleneksel yöntemler zaman alıcı ve sınırlı uygulanabilirken, yapay zeka destekli sistemler daha hızlı, doğru ve ölçeklenebilir çözümler sunmaktadır.

Problem Tanımı

Bu projede, UCI Machine Learning Repository'den alınan Heart Disease veri seti kullanılarak bireylerin kalp hastalığı riskinin tahmin edilmesi hedeflenmiştir. Amaç yaş, cinsiyet, kan basıncı, kolesterol seviyesi, EKG bulguları, maksimum kalp atış hızı, egzersiz sonrası ST depresyonu, koroner damar tıkanıklığı gibi önemli sağlık verilerini kullanarak hastalık riskini sınıflandırmaktır.

Veri Seti ve Ön İşleme

- **Veri seti:** UCI Heart Disease 920 Satır, 16 Sütun(Özellik)
- **Ön işleme adımları:**
  - Kategorik değişkenler, etiketleme (label encoding) ve one-hot encoding yöntemleriyle sayısal verilere dönüştürülmüştür.
  - Eksik veriler için çeşitli imputasyon yöntemleri denenmiş ve en yüksek doğruluk sağlayan yöntem tercih edilmiştir.
  - Sayısal özellikler, mesafe tabanlı algoritmaların gereksinimlerine uygun olarak pipeline kullanılarak StandardScaler ile standartlaştırılmıştır.

- Özellikle kolesterol ve kan basıncı gibi tıbbi verilerdeki aykırı değerlerin etkisini azaltmak amacıyla Winsorizing yöntemi uygulanmıştır. Böylece uç değerlerin etkisi sınırlandırılırken, klinik anlam kaybı önlenmiştir.

Veri setinde özellikle ca (%66,4), thal (%52,8) ve slope (%33,6) sütunlarında yüksek oranda eksik veri bulunmakta olup, bu sütunlar için özel imputasyon yöntemleri uygulanmıştır. Aykırı değerler detaylıca analiz edilerek biyolojik ve klinik açıdan anlamlı olmayan veriler düzeltilmiştir. Hedef değişken, çok sınıflı olmasına rağmen ikili sınıflandırma için hastalığın varlığı (num > 0) temelinde dönüştürülmüş ve böylece sınıf dengesizliği azaltılmıştır.

Klinik perspektiften bakıldığında, veri setindeki bazı özelliklerde (örneğin kolesterol [chol] ve kan basıncı [trestbps]) biyolojik olarak mümkün olmayan veya klinik açıdan anlamlı olmayan uç değerler yer almaktadır. Örneğin, kolesterolde 0 ya da aşırı yüksek değerler hatalı veri olarak kabul edilmektedir ve model performansını olumsuz etkileyebilir. Ancak ST depresyonu (oldpeak) değişkenindeki negatif değerler klinik olarak anlamlıdır ve bu nedenle korunmuştur. Bu doğrultuda, aykırı değerlerin model üzerindeki etkisini azaltmak amacıyla Winsorizing yöntemi uygulanmıştır; böylece aşırı uç değerlerin zararlı etkisi sınırlandırılırken, klinik açıdan anlamlı veriler korunmuş ve veri kalitesi ile tıbbi gerçeklikler dengelenmiştir.

Uygulanan Yöntemler

Projenin modelleme aşamasında farklı makine öğrenimi algoritmaları karşılaştırılmıştır:

- Random Forest
- Logistic Regression
- CatBoost
- XGBoost
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)

Model başarısı için doğruluk (accuracy), F1-score, hassasiyet (precision), duyarlılık (recall) ve ROC-AUC metrikleri kullanılmıştır.

Model Performans Sonuçları

Projede kullanılan modeller ve performansları aşağıdaki tabloda özetlenmiştir:

Model Karşılaştırması

Model	Doğruluk	ROC-AUC	Precision (Pozitif)	Recall (Pozitif)	F1-Score (Pozitif)
CatBoost	**0.8478**	**0.925**	**0.89**	0.84	**0.87**
Random Forest	**0.8478**	0.920	**0.89**	**0.85**	**0.87**
SVM	0.8424	0.918	**0.89**	0.83	0.86
XGBoost	0.8315	0.915	0.88	0.83	0.85
KNN	0.8207	0.912	0.87	0.83	0.85
Logistic Regression	0.8043	0.902	0.88	0.78	0.83

Hangi Model Daha İyi Performans Verdi?

Random Forest modeli, hem genel doğruluk hem de recall açısından en iyi performansı göstermiştir. Bu durum, hasta olan bireylerin doğru şekilde tespit edilmesinde önemli bir avantaj sağlamaktadır.

Hangi Metrik Üzerinde Öne Çıktı?

- **ROC-AUC**, sınıflar arasındaki ayrımı ölçen önemli bir metriktir ve CatBoost, Random Forest ile SVM modelleri bu alanda öne çıkmaktadır.
- **F1-Score (Target = 1)** dengesiz hataları dengelemek için kritik bir metriktir; bu metrikte CatBoost ve Random Forest modelleri en iyi performansı sergilemiştir.
- **Recall** ise, özellikle kalp krizi gibi kritik durumlarda yanlış negatifleri azaltmak için büyük önem taşımaktadır ve bu metrikte en yüksek başarı Random Forest modeline aittir.
- **Accuracy (doğruluk)** açısından ise CatBoost ve Random Forest modelleri birinciliği paylaşmaktadır.

Öne Çıkan Özellikler ve Model Farklılıkları

Projede birçok model kullanılmış ve farklı modellerde öne çıkan özellikler değişiklik göstermiştir. Çoğu modelde kolesterol seviyesi (chol_winsorized), göğüs ağrısı tipi (cp_asymptomatic) ve yaş (age) gibi tıbbi faktörler tahmin başarısına güçlü katkı sağlamıştır. CatBoost modelinde ST segment değişimi (oldpeak_winsorized), Random Forest modelinde ise kolesterol seviyesi (chol_winsorized) öne çıkmıştır.

SVM ve KNN modellerinde ise doğrudan özellik önemi analiz edilmemiştir; çünkü bu algoritmalar doğası gereği, her bir özelliğin modele katkısını ayrı ayrı skorlayacak bir iç mekanizma sunmaz. SVM modeli karar sınırını destek vektörleri üzerinden oluştururken, KNN modeli tamamen komşuluk temelli bir yapı kullandığı için bu tür bir analiz mümkün değildir.

Overfitting ve Underfitting Durumu

- Güçlü ensemble yöntemler olan **Random Forest, XGBoost ve CatBoost** modellerinde ilk parametre konfigürasyonlarında **overfitting (aşırı öğrenme)** durumu gözlemlenmiştir.
- Bu aşırı öğrenme durumu, model karmaşık veriye aşırı uyum sağladığında ortaya çıkar ve modelin genelleme kabiliyetini düşürür.
- Ancak, yapılan **hiperparametre ayarları ile overfitting başarılı şekilde baskılanmıştır**.
- **Underfitting (yetersiz öğrenme)** ise hiçbir modelde gözlenmemiştir; tüm modeller eğitim verisini makul derecede öğrenmiştir.
- ROC-AUC değerleri, doğruluk ve diğer metriklerle uyumlu olup, sınıfların dengeli olması sebebiyle ek dengeleme yapılmamıştır.

Confusion Matrix

Modellerin hata matrisi sonuçlarına göre, **Random Forest** dengeli sınıflandırma yaparak en düşük yanlış negatif (16) değerine sahiptir ve kritik vakaların atlanmasını önler. **Logistic Regression** ise yüksek yanlış negatif (24) nedeniyle erken tanıda dezavantajlıdır. **SVM** yüksek doğruluk ve dengeli hata oranlarıyla güvenilir bir performans sunar. **XGBoost** biraz daha fazla yanlış pozitif üretse de düşük yanlış negatif ve yüksek doğruluk ile güçlü bir alternatiftir. **CatBoost**, SVM ile birlikte en düşük yanlış pozitif (11) ve düşük yanlış negatif (17) oranlarına sahip olup dengeli ve güvenilir sonuçlar verir. **KNN** genel olarak başarılı olmakla beraber, diğer modellere kıyasla biraz daha yüksek yanlış pozitif ve negatif değerleri göstermektedir.

Kritik Çıkarım ve Model Tercihi

Kalp krizi gibi hayati risk taşıyan durumlarda, yanlış negatiflerin (FN) en aza indirilmesi büyük önem taşımaktadır. Bu nedenle tercih sıralaması şu şekildedir:

- **En iyi tercih:** Random Forest modeli; en düşük yanlış negatif oranı ve en yüksek recall değerine sahiptir.
- **Alternatif tercih:** CatBoost modeli; Random Forest'tan sonra en düşük yanlış negatif sayısına (FN=17) sahiptir ve ayrıca daha az yanlış pozitif üretmektedir.

Özetle:

- Hasta kaçırmamak öncelikliyse **Random Forest** modeli tercih edilmelidir.
- Yanlış alarmı azaltmak daha önemliyse **CatBoost** modeli daha uygun bir seçenektir.

Sonuç ve Yorumlar

Modelin Pratik Kullanımı Hakkında Değerlendirme:
Random Forest modeli, yüksek recall değeri sayesinde kalp hastalığı riskini doğru tespit ederek klinik uygulamalarda faydalı olabilir. Yanlış negatif oranının düşük olması, kritik vakaların gözden kaçmasını önler.

Daha İyi Sonuçlar İçin Neler Yapılabilir?
Veri çeşitliliğinin artırılması, yeni özelliklerin eklenmesi ve hiperparametre optimizasyonunun genişletilmesi model performansını yükseltebilir. Ayrıca farklı model kombinasyonları ve farklı ensemble yöntemleri de değerlendirilebilir.

Kaynakça

https://dergipark.org.tr/tr/download/article-file/3576727#:~:text=Bir%C3%A7ok%20insan%20kalp%20krizi%20ve,2'sine%20denk%20gelmektedir.

https://hsgm.saglik.gov.tr/tr/haberler/29-eylul-dunya-kalp-gunu-2.html