Sisterslab Bitirme Projesi - Heart Disease Classification

Elif Şevval Güler

6 Temmuz 2025

1 Problem Tanımı

1.1 Projenin Amacı

Bu projenin temel amacı, bireylerin sahip olduğu klinik ve demografik özelliklere göre kalp hastalığı riski taşıyıp taşımadığını tahmin edebilen bir makine öğrenmesi modeli geliştirmektir.

1.2 Hedef Değişken (target)

- 0: Kalp hastalığı yok
- 1: Kalp hastalığı var

Modelin temel görevi, bu ikili sınıflandırmayı mümkün olduğunca doğru yapmaktır.

1.3 Problem Türü

İkili Sınıflandırma (Binary Classification)
Kullanılacak algoritmalar: Logistic Regression, Random Forest, XGBoost, vb.

1.4 Veri Seti Kaynağı

- Kaynak: Kaggle – UCI Heart Disease Dataset
- Klinikler: Cleveland, Hungary, Switzerland, VA Long Beach

1.5 Medikal Amaç

Kalp hastalıkları, dünya çapında ölüm nedenlerinin başında gelmektedir. Klinik verilerle geliştirilen bu model, erken uyarı sistemleri için yorumlanabilir ve güvenilir bir tahmin mekanizması sunmayı hedeflemektedir.

2 Veri Seti Özellikleri

2.1 Genel Bilgiler

- 920 satır × 16 sütun

2.2 Değişken Türleri

2.2.1 Sayısal Değişkenler

- age: Yaş (yıl)
- trestbps: Dinlenme sırasındaki kan basıncı (mm Hg)
- chol: Serum kolestrol (mg/dl)
- thalach: Maksimum kalp atış hızı
- oldpeak: ST segmenti depresyonu

2.2.2 İkili Kategorik Değişkenler

- sex: 0= Kadın, 1= Erkek
- fbs: Açlık kan şekeri > 120 mg/dl (1= Evet, 0= Hayır)
- exang: Egzersize bağlı anjina (1= Evet, 0= Hayır)

2.2.3 Çok Kategorili Değişkenler

- cp: Göğüs ağrısı tipi (0–3)
- restecg: Dinlenme EKG sonuçları
- slope: ST segment eğimi
- ca: Büyük damar sayısı (0–3)
- thal: Talasemi tipi

3 Eksik Veri Değerlendirmesi

3.1 Eksik Veri Türleri

- **MCAR**: Tamamen rastgele
- **MAR**: Başka gözlemlenebilir değişkenlerle ilişkili
- **MNAR**: Değişkenin kendisiyle ilişkili

3.2 Eksikliği Yüksek Değişkenler (Çıkarıldı)

- ca: %66 eksik
- thal: %53 eksik

3.3 Doldurma Stratejileri

Tür	Değişkenler	Yöntem
Sayısal	trestbps, chol, thalach, oldpeak	KNNImputer (k=5)
Kategorik	slope, fbs, restecg, exang	Mod (en sık)

4 Aykırı Değer Analizi

Değişken	Aykırı Sayısı	Yöntem
age	0	-
trestbps	28	Medyan ile değiştirildi
chol	185	Winsorization
thalach	2	Korundu
oldpeak	16	Medyan ile değiştirildi

5 Korelasyon Analizi (Sayısal Değişkenler)

Değişken	Korelasyon	Açıklama
thalach	-0.38	Daha düşük kalp atış hızı → daha yüksek risk
oldpeak	+0.37	ST depresyonu arttıkça risk artıyor
age	+0.28	Yaş arttıkça risk artıyor
chol	-0.22	Zayıf ters ilişki

6 Encoding İşlemleri

6.1 Label Encoding

Değişken	Önce	Sonra
sex	Male, Female	1, 0
fbs	True, False	1, 0
exang	True, False	1, 0

6.2 One-Hot Encoding

- dataset: 4 sütun (Hungary, Switzerland, vb.)
- cp, restecg, slope: her biri için 2–3 ek sütun oluşturuldu

7 Sayısal Değişkenlerin Ölçeklendirilmesi

- Yöntem: `StandardScaler`
- Tüm sayısal değişkenler normalize edildi (ortalama ≈ 0, std ≈ 1)

Tablo 1: Tuning Öncesi Model Performans Karşılaştırması

Model	AUC	FN	FP	Mean Acc. (CV)	Test Acc.	Train Acc.	Overfit Farkı	Notlar
Logistic Regression	0.91	13	19	0.77	0.83	0.83	0.00	Dengeli ve overfitting yok
Random Forest	0.91	13	16	0.73	0.84	1.00	0.16	Aşırı öğrenme eğilimi
KNN	0.88	9	21	0.72	0.84	0.85	0.01	En düşük FN, stabil
SVM	0.90	10	20	0.75	0.84	0.86	0.02	Dengeli, genel başarı yüksek
XGBoost	0.90	17	16	0.68	0.82	1.00	0.18	Aşırı öğrenme net
Gradient Boosting	0.91	12	16	0.65	0.85	0.92	0.07	ROC güçlü, overfitting orta
CatBoost	0.91	12	18	0.71	0.84	0.96	0.12	Güçlü ayırıcı ama overfit eğilimli

Tablo 2: Tuning Sonrası Model Performans Karşılaştırması

Model	AUC	FN	FP	Mean Acc. (CV)	Test Acc.	Train Acc.	Overfit Farkı	Notlar
Logistic Regression	0.91	13	19	0.819	0.83	0.83	0.00	Dengeli, overfitting yok
Random Forest	0.91	13	16	0.818	0.84	1.00	0.16	Aşırı öğrenme riski var
KNN	0.88	9	21	0.804	0.84	0.85	0.01	En düşük FN, istikrarlı
SVM	0.90	10	20	0.818	0.84	0.86	0.02	Dengeli, düşük fark
XGBoost	0.90	17	16	0.814	0.82	1.00	0.18	Aşırı öğrenme gözlenmiş
Gradient Boosting	0.91	12	16	0.803	0.85	0.92	0.07	Orta düzey overfitting
CatBoost	0.91	12	18	0.807	0.84	0.96	0.12	Overfitting riski var

8 Klinik Değerlendirme

Bu projede, kalp hastalığı riskinin tahmini amacıyla farklı makine öğrenmesi sınıflandırma algoritmaları karşılaştırılmıştır. Veri ön işleme adımlarından başlayarak hiperparametre tuning süreci dahil tüm modelleme adımları detaylıca uygulanmıştır.

Elde edilen sonuçlara göre:

- En iyi genel performans Gradient Boosting modeline aittir (Test doğruluğu %85).
- En yüksek güvenilirlik ve istikrar: Logistic Regression ve SVM.
- Hasta kaçırma oranı açısından en iyi model KNN'dir (en düşük FN).
- XGBoost ve Random Forest aşırı öğrenme göstermektedir.

False Negative (FN) sayısı medikal uygulamalarda kritik bir metriktir. FN'nin düşük olduğu modeller, özellikle sağlık alanında daha değerli olabilir.

Not: Gerçek dünyada dengesiz veri setleriyle karşılaşılabileceğinden, ROC-AUC ve Recall gibi metrikler öncelikli değerlendirme kriterleri olmalıdır.