İSTANBUL TOPKAPI ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ

Ders/Dönem: SWE307 - Web Tasarımı ve Programlama / 2025-2026 Güz Dönemi

Proje Başlığı: Makine Öğrenmesi Algoritmaları ile Kalp Hastalığı Risk Tahmini ve Karşılaştırmalı Performans Analizi

Öğrenci Ad Soyad: Umut Torun

Öğrenci Numara: 23040101063

Öğrenci E-Posta: umuttorun@stu.topkapi.edu.tr

Öğrenci İmza:

1. PROBLEM TANIMI & MOTİVASYON

1.1. İş/Bilimsel Soru

Kalp ve damar hastalıkları, dünya genelinde ölümlerin bir numaralı sebebi olarak kabul edilmektedir. Geleneksel tanı yöntemleri, genellikle girişimsel, maliyetli ve uzman hekim yorumuna yüksek oranda bağımlıdır. Hastaların klinik verilerindeki karmaşık ve doğrusal olmayan ilişkilerin insan gözüyle her zaman doğru analiz edilememesi, erken teşhis fırsatlarının kaçırılmasına veya gereksiz tetkiklerin yapılmasına yol açabilmektedir. Bu projenin temel motivasyonu, makine öğrenmesi algoritmalarının örüntü tanıma yeteneklerini kullanarak, klinik veriler ışığında kalp hastalığı riskini yüksek doğrulukla tahmin etmektir. Geliştirilen modelin, hekimler için bir Karar Destek Sistemi olarak kullanılması; teşhis sürecini hızlandırmayı, maliyetleri düşürmeyi ve erken müdahale ile hasta sağkalım oranlarını artırmayı hedeflemektedir.

Temel Bilimsel Soru: Rutin hastane kontrollerinde elde edilen demografik ve klinik veriler kullanılarak, bir hastanın kalp hastalığına sahip olup olmadığı makine öğrenmesi yöntemleriyle ne kadar yüksek bir doğrulukla öngörülebilir?

1.2. Görev Türü

Bu proje, makine öğrenmesi literatüründe Gözetimli Öğrenme kapsamında bir İkili Sınıflandırma problemidir. Model, etiketlenmiş geçmiş hasta verileriyle eğitilerek, yeni gelen bir hasta verisini önceden tanımlanmış iki sınıftan birine atamayı öğrenmektedir.

1.3. Hedef Değişkenler

- Hedef Değişken: target
- Değer Aralığı: Binary (0: Sağlıklı, 1: Hasta)
- Negatif Sınıf: Sağlıklı
- Pozitif Sınıf: Hasta
- Veri Dengesi: %51.3 Sağlıklı, %48.7 Hasta

1.4. Başarı Kriterleri

- Doğruluk (Accuracy): > %90
- Duyarlılık (Recall) >: %95: Hayati risk taşıyan hasta bireylerin gözden kaçırılmaması
- Modelin kararlarının tıbbi gerçeklerle örtüşmesi ve doktorlara anlamlı bir içgörü sunması gerekmektedir.

2. VERİ AÇIKLAMASI VE YÖNETİMİ

2.1. Veri Kümesi Açıklaması

Veri Seti Adı: Heart Disease Dataset

Kaynak: Kaggle (kaynak: UCI Machine Learning Repository verilerinin birleşmiş halidir).

Bağlantı: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Lisans: Public Domain (Kamu Malı)

Açıklama: Bu çalışma kapsamında kullanılan veri seti, kalp hastalığı risk faktörlerini analiz etmek amacıyla oluşturulmuş kapsamlı bir veritabanıdır. Veri seti, farklı yaş ve demografik özelliklere sahip bireylerden toplanan toplam 1025 adet hasta kaydından oluşmaktadır. Her bir kayıt 13 bağımsız klinik öznitelik içermektedir. Hedef değişken, hastanın kalp hastalığı taşıyıp taşımadığını ifade eden ikili bir yapıdadır.

2.2. Veri Şeması

Değişken Adı	Açıklama
age	Hastanın yıl cinsinden yaşı.
sex	Cinsiyet (1: Erkek, 0: Kadın).
cp	Göğüs ağrısı tipi (0-3 arası 4 farklı kategori).
trestbps	Dinlenme kan basıncı (mm/Hg).
chol	Serum kolesterol ölçümü (mg/dl).
fbs	Açlık kan şekeri > 120 mg/dl durumu (1: Evet, 0: Hayır).
restecg	Dinlenme EKG sonuçları (0: Normal, 1-2: Anormal).
thalach	Ulaşılan maksimum kalp atış hızı.
exang	Egzersize bağlı anjina (göğüs ağrısı) varlığı (1: Var, 0: Yok).
oldpeak	Egzersiz sırasında oluşan ST depresyonu.
slope	ST segmentinin eğimi (0, 1, 2).
ca	Floroskopide görülen damar sayısı (0-3).
thal	Talasemi kan hastalığı türü (1, 2, 3).
target	**Teşhis Sonucu (1: Kalp Hastası, 0: Sağlıklı).**

2.3. Veri Boyutu

Toplam Satır Sayısı: 1025 Hasta Kaydı

Sütun Sayısı: 14 (13 Bağımsız Değişken + 1 Hedef Değişken)

2.4. Etik ve Gizlilik

Kullanılan veri seti, kamuya açık lisansa sahip olup, hasta mahremiyetini korumak amacıyla tamamen anonimleştirilmiştir. Veri setinde hastaların adı, kimlik numarası veya adres bilgisi gibi kişisel tanımlayıcılar bulunmamaktadır. Bu çalışma, Kişisel Verilerin Korunması prensiplerine uygun olarak, sadece istatistiksel ve klinik öznitelikler üzerinden yürütülmüştür.

3. YÖNTEMLER VE MİMARİ

Veri Ön İşleme: Eksik verilerin kontrolü ve sürekli değişkenlerdeki gürültüyü azaltmak için IQR yöntemiyle aykırı değer baskılama.

Öznitelik Mühendisliği: Kategorik değişkenlerin modelin anlayabileceği sayısal formata getirilmesi ve tüm özniteliklerin standart normal dağılıma çekilmesi.

Model Optimizasyonu: Algoritmaların en iyi performans gösteren parametrelerini bulmak için GridSearchCV tekniğinin uygulanması.

Topluluk Öğrenmesi: Tekil modellerin zayıf yanlarını kapatmak için Voting Classifier mekanizmasının kurulması.

Tahmin Katmanı: Kullanıcıdan alınan gerçek zamanlı verileri işleyip teşhis koyan arayüz.

4. DENEY TASARIMI

4.1. Ana Amaç

Deneyin temel amacı; kalp hastalığı teşhisinde insan hayatı göz önünde bulundurularak, "Yanlış Negatif" oranını sıfıra yaklaştıran en kararlı algoritmayı tespit etmektir. Deney kapsamında doğrusal (Logistic Regression), ağaç tabanlı (Random Forest, XGBoost) ve vektör tabanlı (SVM) yaklaşımlar birbirleriyle kıyaslanmıştır.

4.2. Değerlendirme Kriterleri

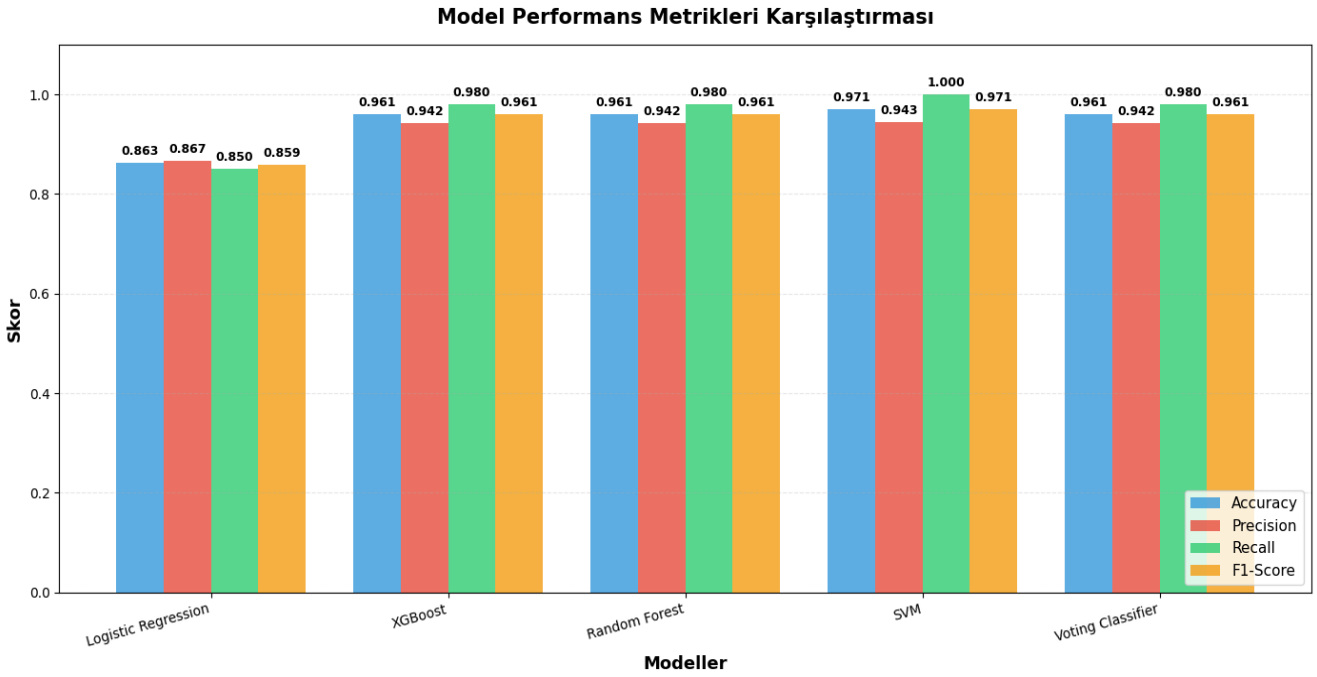
Modellerin başarısı sadece doğruluk oranıyla değil, aşağıdaki dört temel kriterle ölçülmüştür:

- **Doğruluk (Accuracy):** Sistemin genel doğru tahmin oranı.
- **Duyarlılık (Recall):** Gerçek hastaların ne kadarının doğru tespit edildiği.
- **Kesinlik (Precision):** "Hasta" denilen kişilerin gerçekte ne kadarının hasta olduğu.
- **F1-Skoru:** Precision ve Recall metriklerinin harmonik ortalaması.
- **Çapraz Doğrulama (k-Fold CV) Skoru:** Modelin verinin farklı bölümlerindeki tutarlılığı ve genelleme yeteneği.

5. MODEL DEĞERLENDİRME VE BULGULAR

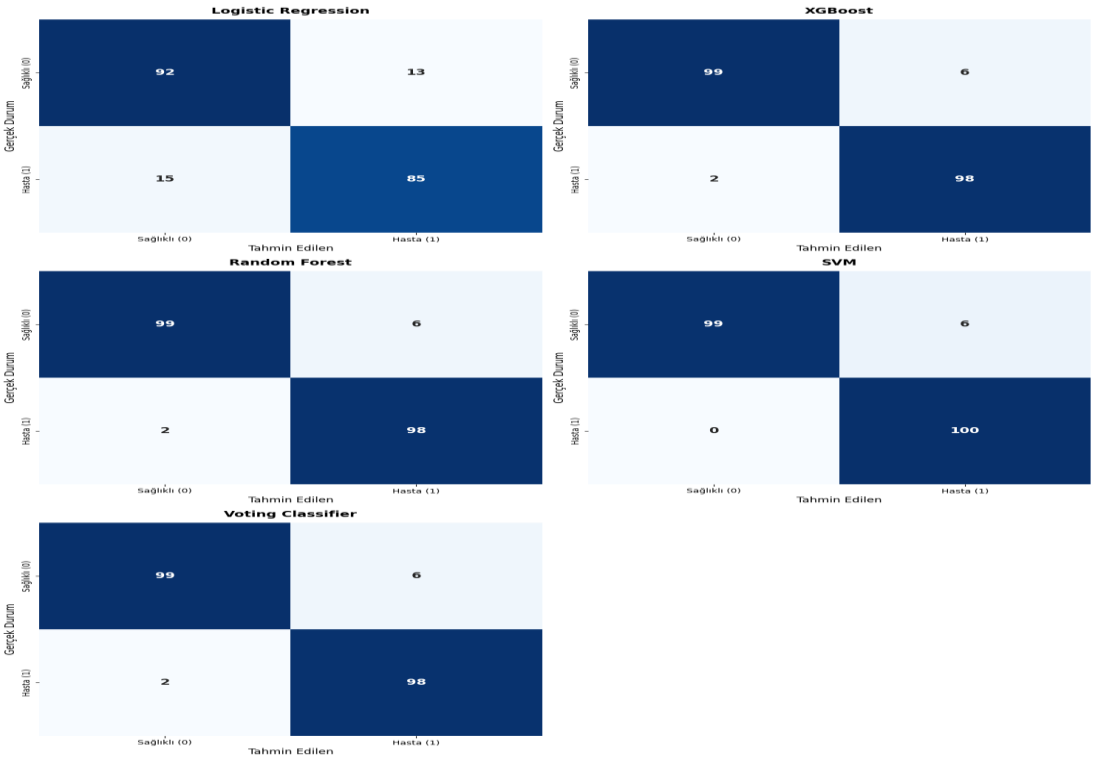
5.1. Model Performans Karşılaştırılması

Eğitilen beş farklı modelin performans metrikleri aşağıda sunulmuştur. Sağlık teşhislerinde en kritik başarı kriteri olan Recall (Duyarlılık), hasta bireylerin kaçırılma riskini gösterdiği için öncelikli olarak değerlendirilmiştir.

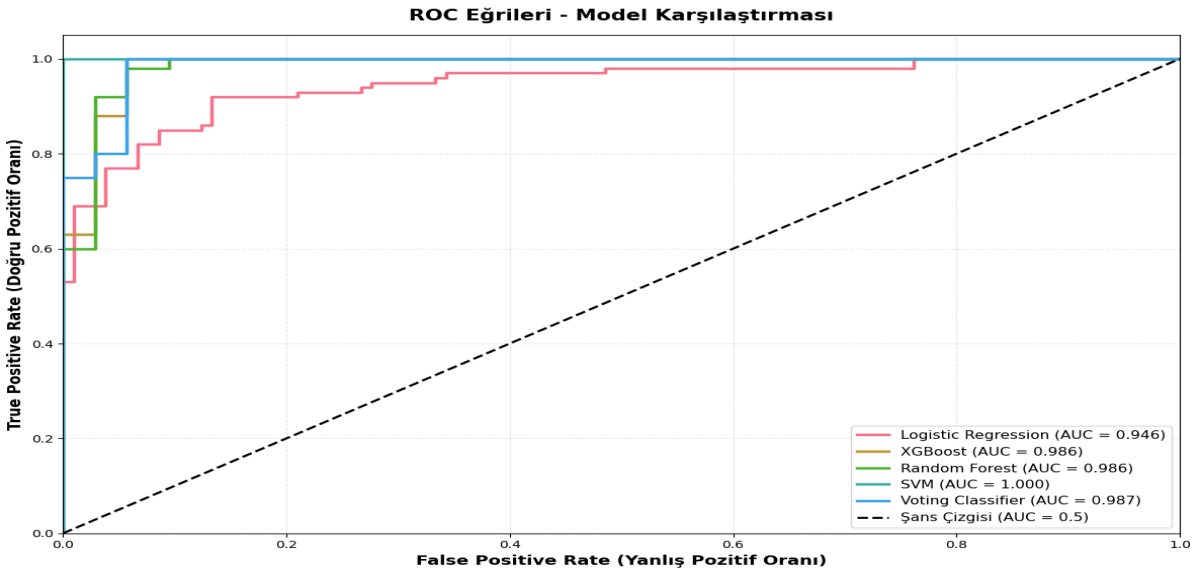


5.2. Karmaşıklık Matrisi

Modeller incelendiğinde Yanlış Negatif değerlerinin düşük olduğu görülmektedir. Bu durum, modelin duyarlılık oranının +%90 olmasını sağlamıştır. Sağlık alanında yapılan teşhislerde en kritik başarı kriteri, gerçek hastaların sistem tarafından atlanmamasıdır. Bu sonuç, geliştirilen modellerin yüksek güvenilirlikte bir tarama aracı olduğunu kanıtlamaktadır.

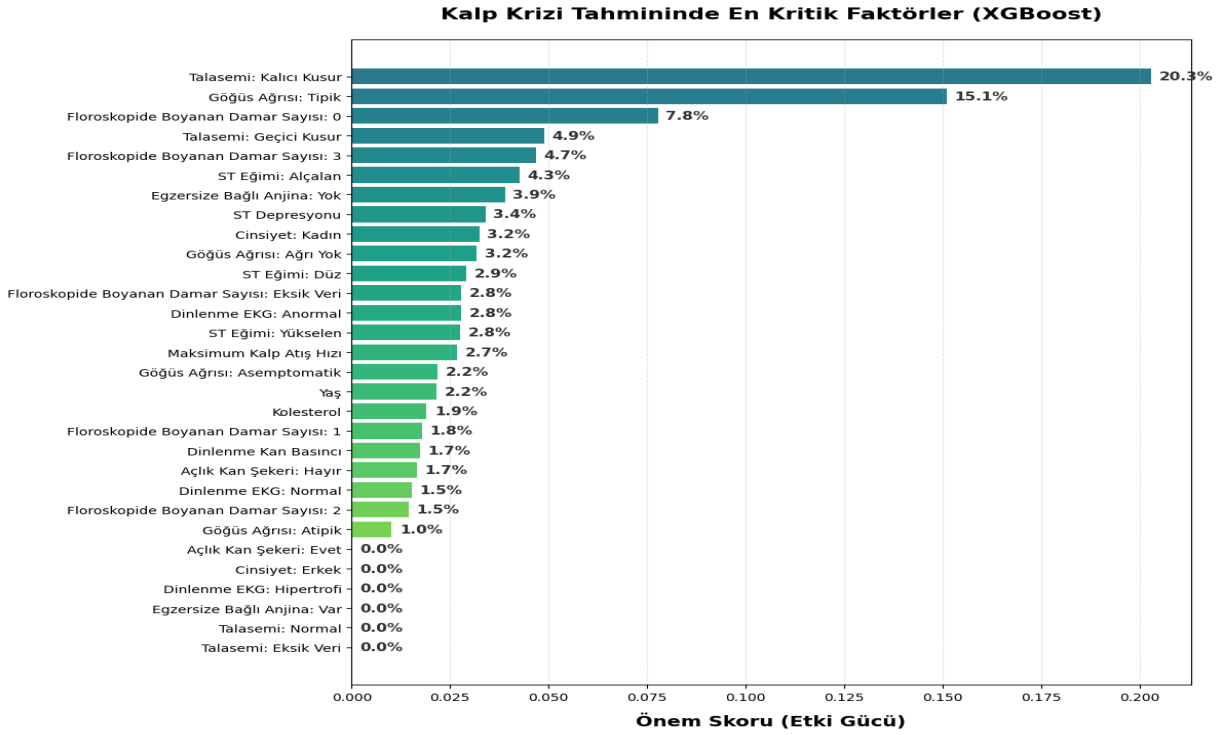


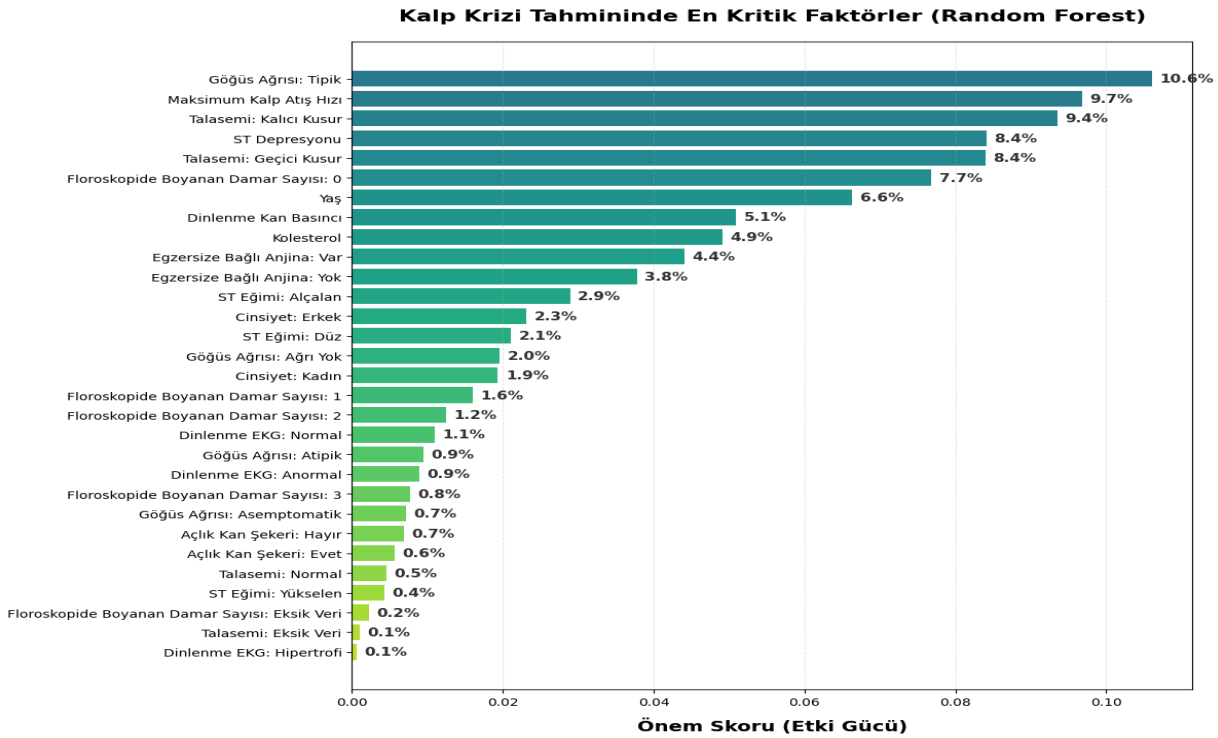
5.3. ROC Eğrisi Analizi



5.4. Özellik Önem Grafiği

Modelin tahminleme sürecinde hangi klinik parametrelere öncelik verdiği, ağaç tabanlı algoritmalar üzerinden analiz edilmiştir. Bu analiz, geliştirilen yapay zekanın tıbbi literatürle ne kadar uyumlu çalıştığını kanıtlamaktadır.





5.5. Kalp Krizi Risk Tahmini

Eğitilen modellerin klinik ortamda kullanılabilirliğini test etmek amacıyla, veri setinden bağımsız, rastgele seçilen bir hasta profili oluşturulmuş ve sisteme girdi olarak verilmiştir. Bu senaryo, modelin karar destek mekanizması olarak nasıl çalıştığını simüle etmektedir.

6. GELİŞTİRME ORTAMI

Projenin geliştirme, modelleme ve analiz aşamalarında; Python veri bilimi ekosisteminin endüstri standardı kabul edilen, yüksek performanslı ve açık kaynaklı kütüphaneleri tercih edilmiştir. Kullanılan temel araçlar ve projeredeki işlevleri aşağıdadır:

- Python 3.x: Projenin geliştirildiği temel programlama dili ve çalışma ortamı.
- Pandas & NumPy: Veri setinin yüklenmesi, eksik verilerin işlenmesi, matris hesaplamaları ve vektörel işlemler gibi veri manipülasyonu süreçlerinde kullanılmıştır.
- Scikit-Learn (Sklearn): Makine öğrenmesinin ana omurgasını oluşturur. Veri ölçeklendirme, eğitim-test ayrımı, model eğitimi ve hiperparametre optimizasyonu bu kütüphane ile gerçekleştirilmiştir.
- XGBoost: Gradyan artırma tabanlı, yüksek hız ve performans sunan sınıflandırma algoritması olarak projeye entegre edilmiştir.
- Matplotlib & Seaborn: Keşifsel veri analizi sırasında veri dağılımlarının incelenmesi ve sonuç bölümünde yer alan Karmaşıklık Matrisi, ROC Eğrisi ve Özellik Önemi grafiklerinin oluşturulması için kullanılmıştır.

7. KAYNAKLAR

[1] "Heart Disease Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset.

[2] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease Data Set," UCI Machine Learning Repository, 1988. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease.

[3] M. Rutecki, "GridSearchCV & K-Fold CV: The Right Way," Kaggle. [Online]. Available: https://www.kaggle.com/code/marcinrutecki/gridsearchcv-kfold-cv-the-right-way. [Erişim: 24 Aralık 2025].

[4] N. Bhat, "Heart Attack Prediction using Different ML Models," Kaggle. [Online]. Available: https://www.kaggle.com/code/nareshbhat/heart-attack-prediction-using-different-ml-models. [Erişim: 25 Aralık 2025].

[5] N. M. Iuddin, "UCI Heart Disease Dataset Analysis," GitHub. [Online]. Available: https://github.com/nmiuddin/UCI-Heart-Disease-Dataset/blob/master/UCI-heart-disease.ipynb. [Erişim: 01 Ocak 2026].