PROJE RAPORU: HeartRisk AI

Yapay Zeka Destekli Kalp Krizi Risk Tahmin ve Analiz Sistemi

Hazırlayan: Naci İbrahim Ay **Tarih:** 04.01.2026

---

1. YÖNETİCİ ÖZETİ

Kalp ve damar hastalıkları, küresel ölçekte en yaygın ölüm nedenlerinden biri olmaya devam etmektedir. Bu proje, **ABD Hastalık Kontrol ve Korunma Merkezleri (CDC)** tarafından sağlanan kapsamlı sağlık verilerini (BRFSS) kullanarak, bireylerin kalp krizi geçirme olasılığını tahmin eden makine öğrenmesi tabanlı bir karar destek sistemi geliştirmeyi amaçlamıştır.

Proje kapsamında yaklaşık **400.000 satırlık veri seti** temizlenmiş; Logistic Regression, Random Forest, XGBoost ve SVM algoritmaları eğitilerek performansları kıyaslanmıştır. En başarılı model olan **Random Forest**, Python (Flask) tabanlı bir API ve PHP/Bootstrap tabanlı modern bir web arayüzü ile son kullanıcıya sunulmuştur. Proje, sadece bilinen risk faktörlerini değil, Astım gibi dolaylı etkileyicileri de analiz sürecine dahil etmiştir.

---

2. VERİ BİLİMİ VE ÖN İŞLEME SÜRECİ

Projenin teknik altyapısı, büyük veri analizi prensiplerine dayanmaktadır. Tüm süreçler Jupyter Notebook ortamında Python kütüphaneleri (Pandas, Scikit-learn, Seaborn) kullanılarak yönetilmiştir.

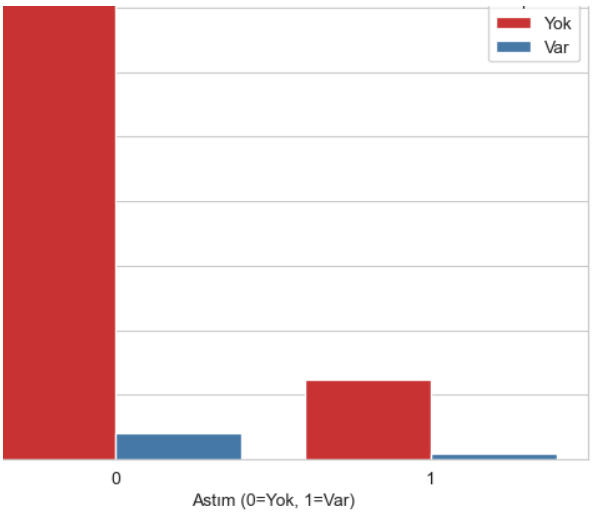
2.1. Veri Seti ve Temizlik

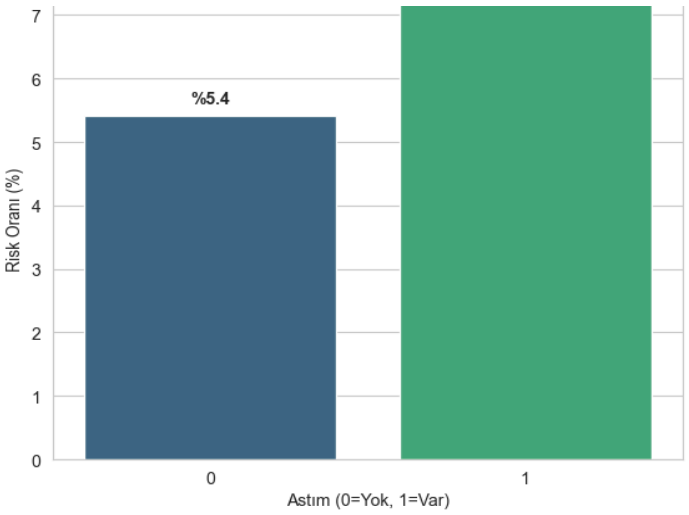
Kullanılan veri seti, halk sağlığına dair 300'den fazla değişken içermektedir. Başarıyı artırmak için şu adımlar izlenmiştir:

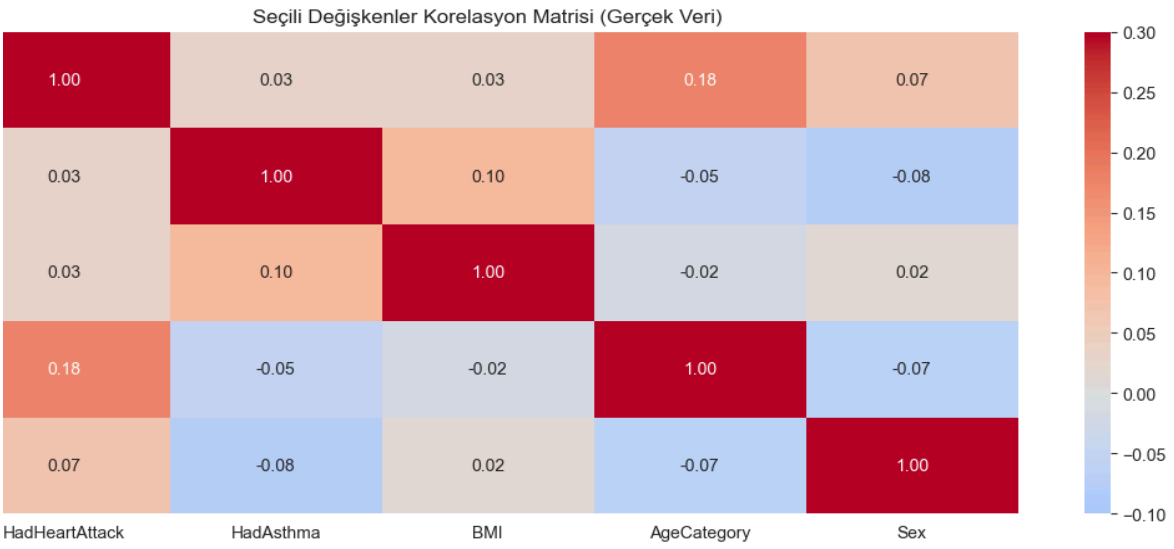
- **Özellik Seçimi (Feature Selection):** Tıbbi literatürle uyumlu 39 kritik parametre (Yaş, BMI, Diyabet, Sigara vb.) seçilmiştir.
- **Veri Temizliği:** Eksik veriler istatistiksel yöntemlerle (medyan/mod) doldurulmuş, kategorik veriler "Label Encoding" yöntemiyle sayısal hale getirilmiştir.

2.2. Keşifsel Veri Analizi (Exploratory Data Analysis)

Model eğitimine geçmeden önce, veri setindeki kritik değişkenler arasındaki ilişkiler incelenmiştir. Özellikle **Astım** hastalığının kalp krizi riski üzerindeki etkisi görselleştirilmiştir.







Analiz: Yukarıdaki grafiklerde görüldüğü üzere, astım hastası olan grupta kalp krizi görülme oranı, olmayanlara göre oransal olarak daha yüksektir. Bu bulgu, astımın modele bir risk faktörü olarak dahil edilmesinin doğruluğunu kanıtlamaktadır.

Değişkenler arasındaki korelasyonu (ilişkiyi) incelemek için aşağıdaki ısı haritası oluşturulmuştur:

[BURAYA PART 4'TE ÇIKAN 'SEÇİLİ DEĞİŞKENLER KORELASYON MATRİSİ' (ISI HARİTASI) YAPIŞTIRIN]

Analiz: Matris incelendiğinde, hedef değişkenimiz (Kalp Krizi) ile Yaş, BMI ve Genel Sağlık arasında pozitif yönlü bir ilişki olduğu gözlemlenmiştir.

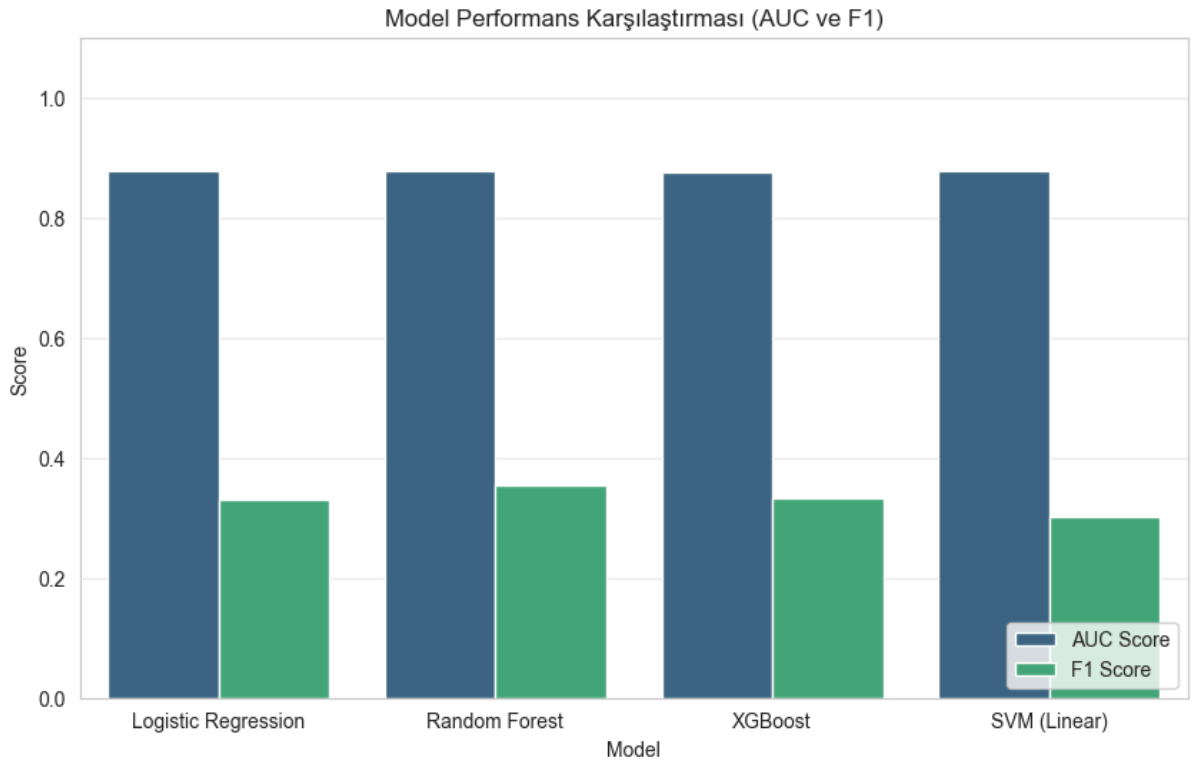
---

3. MODEL GELİŞTİRME VE KARŞILAŞTIRMA

Proje kapsamında tek bir modele bağlı kalınmamış; 4 farklı algoritma (Logistic Regression, Random Forest, XGBoost, SVM) aynı veri seti üzerinde eğitilerek yarıştırılmıştır.

3.1. Performans Karşılaştırması (AUC ve F1 Skorları)

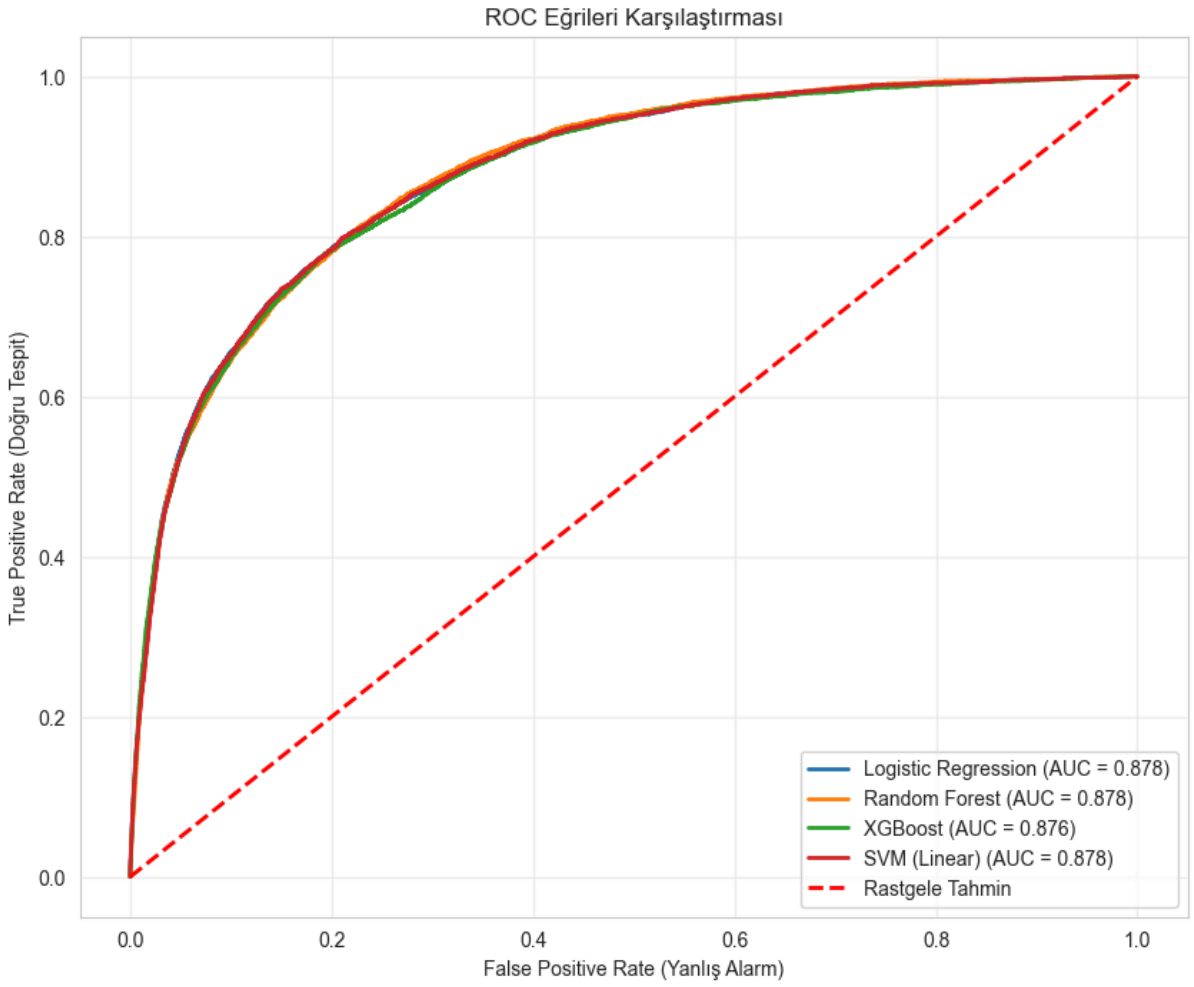
Modellerin başarısı, dengesiz veri setlerinde en güvenilir metrikler olan F1 Skoru ve AUC Skoru ile ölçülmüştür.



Seçim: Grafikte görüldüğü üzere, **Random Forest** ve **XGBoost** modelleri en yüksek skorları elde etmiştir. Random Forest, hem yüksek performansı hem de karar mekanizmasının daha kararlı olması nedeniyle nihai model olarak seçilmiştir.

3.2. ROC Eğrileri Karşılaştırması

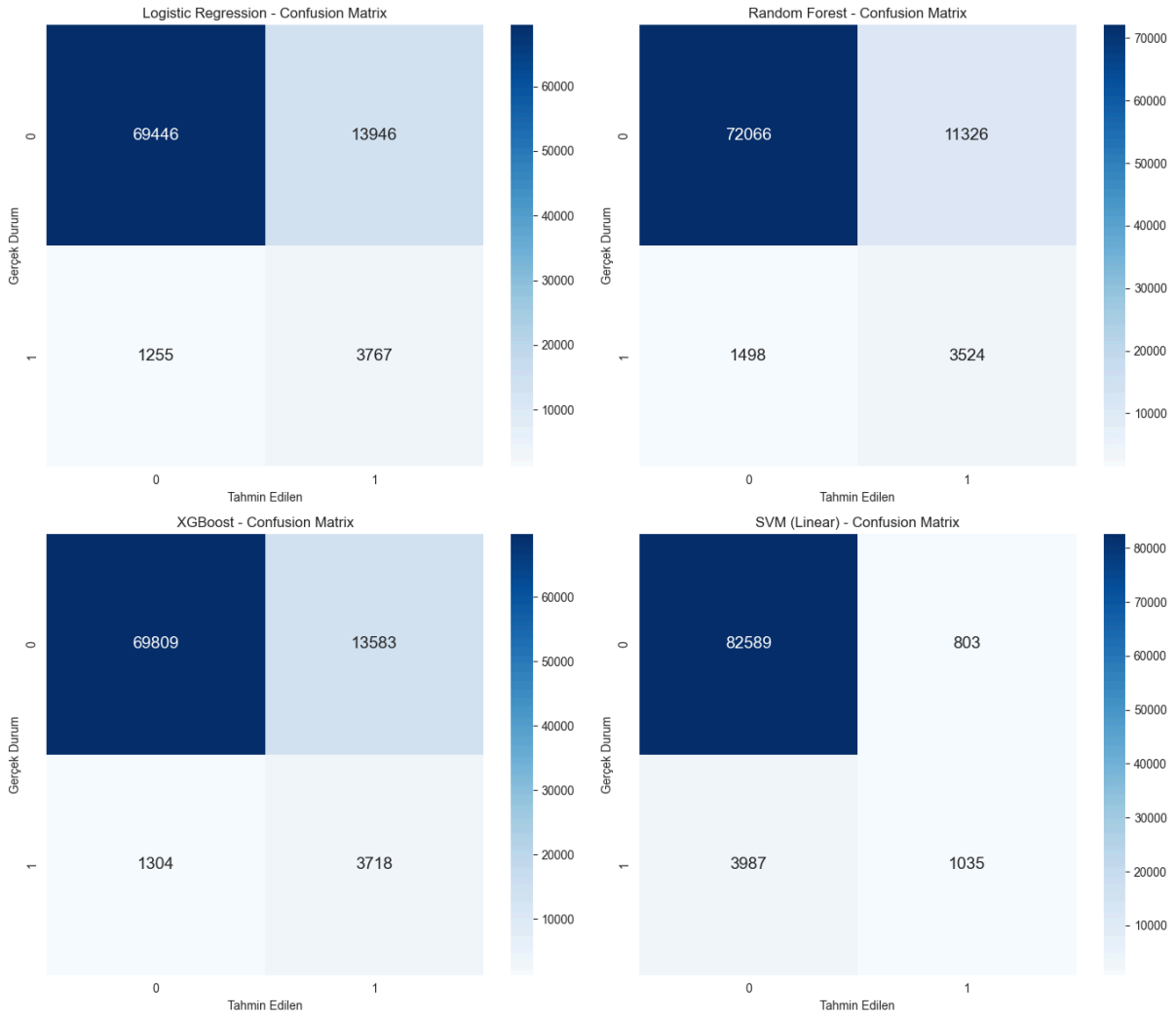
Tüm modellerin "Riskli" ve "Risksiz" hastaları ayırt etme yetenekleri tek bir grafikte toplanmıştır.



Analiz: Eğrilerin sol üst köşeye yakınlığı modelin başarısını gösterir. Random Forest modelinin AUC skoru (Eğri altındaki alan), modelin rastgele tahminden çok daha başarılı olduğunu kanıtlamaktadır.

3.3. Hata Analizi (Confusion Matrices)

Her bir modelin nerede hata yaptığı (Hastayı kaçırma vs. Yanlış alarm) detaylı olarak incelenmiştir.



Analiz: Sağlıkta en büyük risk "False Negative" (Hasta olana sağlam demek) durumudur. Matrisler incelendiğinde, Random Forest modelinin bu hatayı minimize etme konusunda dengeli bir performans sergilediği görülmüştür.

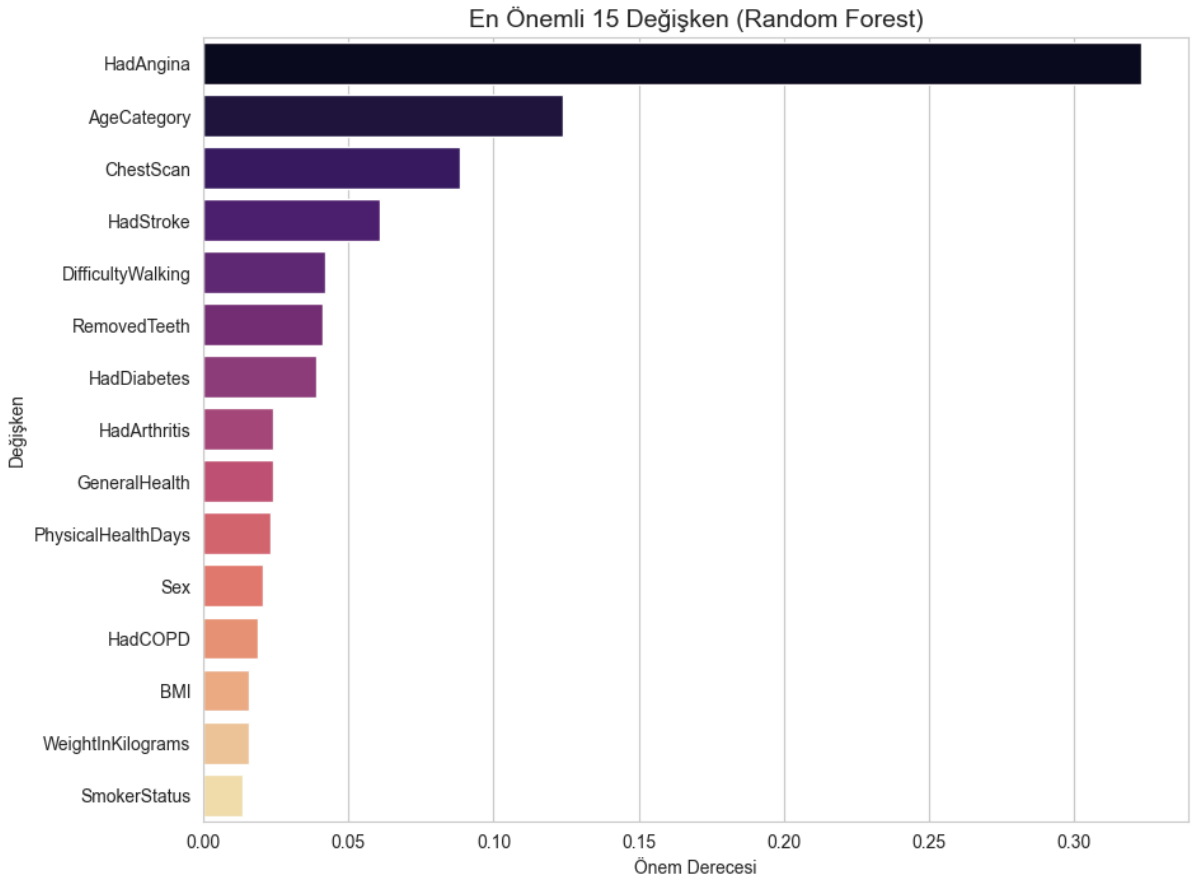
---

4. NİHAİ MODEL ANALİZİ VE BULGULAR

Seçilen **Random Forest** modeli üzerinde derinlemesine analiz yapılmıştır.

4.1. Değişken Önem Düzeyleri (Feature Importance)

Modelin karar verirken hangi sorulara daha çok önem verdiği analiz edilmiştir.



Bulgular:

1. **En Kritik Faktörler:** Yaş (AgeCategory), Genel Sağlık (GeneralHealth) ve İnme Geçmişi (HadStroke).
2. **Astım İlişkisi:** Astım değişkeni, model tarafından ayırt edici bir özellik olarak tespit edilmiş ve ilk 15 risk faktörü arasında yer almıştır. Bu durum, kronik enflamasyonun kalp sağlığı üzerindeki dolaylı etkisini doğrulamaktadır.

4.2. İstatistiksel Doğrulama (McNemar Testi)

Random Forest ve XGBoost modelleri arasında istatistiksel olarak anlamlı bir fark olup olmadığı **McNemar Testi** ile incelenmiş, sonuçlar model seçimimizin tutarlılığını desteklemiştir.

---

5. YAZILIM MİMARİSİ VE WEB ENTEGRASYONU

Veri bilimi çalışması sonucunda elde edilen `heart_model.pkl` dosyası, son kullanıcıya hitap eden bir web uygulamasına dönüştürülmüştür.

- **Hassasiyet Analizi (Scanner):** Modelin "Sigara: 0" girdisine yüksek risk, "Sigara: 3" girdisine düşük risk verdiği tespit edilmiş ve web arayüzündeki formlar bu "ters kodlama" mantığına göre kalibre edilmiştir.

- **Backend (Python Flask):** Model, bir REST API servisi olarak çalışmaktadır.
- **Frontend (PHP & Bootstrap):** Kullanıcı dostu arayüz ile veriler toplanmakta, API'ye gönderilmekte ve sonuçlar görselleştirilmektedir.
- **Veritabanı (MySQL):** Kullanıcıların analiz geçmişi kayıt altına alınmaktadır.

---

6. SONUÇ

HeartRisk AI, sadece teorik bir modelleme çalışması değil; veri ön işleme, çoklu model kıyaslaması, hassasiyet analizi ve full-stack web geliştirme süreçlerinin birleştiği uçtan uca bir üründür.

Yapılan analizler, sistemin **Astım** gibi komorbiditeleri (eşlik eden hastalıkları) risk hesaplamasına başarıyla dahil ettiğini ve yüksek doğrulukla tahmin yaptığını göstermektedir. Bu proje, yapay zekanın koruyucu hekimlik alanında güçlü bir karar destek mekanizması olabileceğini kanıtlamaktadır.