**Proje Raporu: Sağlık Verisiyle Kardiyovasküler Hastalık Risk Tahmini**

**Hazırlayan:** Cansu Kahve

**Tarih:** 21.10.2025

**1. Proje Özeti**

Bu proje, Kaggle'dan temin edilen kardiyovasküler hastalık veri setini kullanarak bireylerin kalp hastalığı riskini tahmin etmeyi amaçlamıştır. Proje; veri temizleme, keşifsel veri analizi (EDA), özellik mühendisliği, modelleme ve model değerlendirme aşamalarını kapsamaktadır. Python programlama dili ve Pandas, Scikit-learn, XGBoost gibi kütüphaneler kullanılarak **Logistic Regression**, **Random Forest** ve **XGBoost** algoritmalarıyla modeller eğitilmiştir. Analiz ve model sonuçları, Microsoft Power BI üzerinde interaktif dashboard'lar ile görselleştirilmiştir.

Yapılan değerlendirmeler sonucunda, **%73.6 doğruluk (Accuracy)** ve **0.801 AUC** skoru ile **XGBoost Classifier** modelinin en iyi performansı sergilediği tespit edilmiştir.

**2. Giriş ve Amaç**

Kardiyovasküler hastalıklar, dünya genelinde önde gelen ölüm nedenleri arasında yer almaktadır. Hastalığın erken teşhisi, tedavi başarısını artırmakta ve sağlık maliyetlerini önemli ölçüde düşürmektedir. Bu projenin temel amacı, bireylerin demografik ve klinik sağlık verilerini kullanarak kalp hastalığı riskini makine öğrenmesi modelleriyle tahmin etmek ve böylece erken teşhis için bir karar destek sistemi oluşturmaktır.

**3. Veri Seti ve Veri Hazırlama**

* **Kaynak:** Kaggle - Cardiovascular Disease Dataset
* **Kayıt Sayısı:** Projeye 70.000 kayıt ile başlanmış, veri temizleme adımları sonucunda 68.647 kayıt ile çalışılmıştır.
* **Hedef Değişken:** cardio (0: Hastalık Yok, 1: Hastalık Var).
* **Veri Hazırlama Süreci:**
  + age sütunu gün formatından yıl formatına dönüştürülmüştür.
  + Kan basıncı (ap\_hi, ap\_lo), boy (height) ve kilo (weight) sütunlarında bulunan mantıksız ve aykırı değerler (örneğin negatif tansiyon, gerçek dışı boy/kilo) veri setinden çıkarılmıştır. Bu işlem sonucunda veri setinde %1.93'lük bir veri kaybı yaşanmıştır.
  + Veri setinde eksik değere rastlanmamıştır.

**4. Keşifsel Veri Analizi (EDA) ve Bulgular**

Temizlenmiş veri seti üzerinde yapılan analizlerde aşağıdaki temel bulgulara ulaşılmıştır:

* **Yaş ve Kan Basıncı:** Kalp hastalığı olan bireylerin medyan yaşının ve ortalama kan basıncı değerlerinin (hem sistolik hem diyastolik), hastalığı olmayanlara göre belirgin şekilde daha yüksek olduğu gözlemlenmiştir.
* **Kolesterol ve Glikoz:** Kolesterol ve glikoz seviyeleri arttıkça kalp hastalığı görülme oranı da artmaktadır. Özellikle "Çok Yüksek" (kategori 3) seviyelerde riskin ciddi oranda arttığı tespit edilmiştir.
* **Korelasyon:** Yaş, kilo, kan basıncı ve kolesterol seviyeleri, kalp hastalığı ile en güçlü pozitif korelasyona sahip değişkenler olarak öne çıkmıştır.
* Bu bulguların tamamı, projenin bir parçası olarak hazırlanan **Power BI - Veri Keşfi Dashboard**'unda interaktif grafiklerle detaylı olarak sunulmuştur.

**5. Modelleme ve Değerlendirme**

Modelleme aşamasında veri seti, özellik mühendisliği (One-Hot Encoding), eğitim-test ayrımı (%80-%20) ve sayısal verilerin ölçeklendirilmesi (StandardScaler) adımlarından geçirilmiştir. Eğitilen üç modelin test verisi üzerindeki performans metrikleri aşağıdaki gibidir:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy (Doğruluk)** | **Precision (Kesinlik)** | **Recall (Duyarlılık)** | **F1-Score** | **AUC Skoru** |
| Logistic Regression | 0.727 | 0.73 | 0.72 | 0.72 | 0.791 |
| Random Forest | 0.714 | 0.72 | 0.70 | 0.71 | 0.772 |
| **XGBoost** | **0.736** | **0.75** | **0.72** | **0.73** | **0.801** |

* **En İyi Modelin Seçimi:** **XGBoost Classifier**, hem en yüksek doğruluk oranını hem de en yüksek AUC skorunu sunarak en başarılı model olarak belirlenmiştir. 0.801'lik AUC skoru, modelin hasta ve sağlıklı bireyleri ayırt etme yeteneğinin "iyi" seviyede olduğunu göstermektedir.
* **ROC Eğrisi Analizi:** Modellerin ROC eğrileri tek bir grafikte karşılaştırıldığında, XGBoost eğrisinin diğerlerine göre sol üst köşeye daha yakın olduğu ve bu sayede sınıflandırma yeteneğinin daha üstün olduğu görsel olarak da teyit edilmiştir.

**6. Sonuç ve Gelecek Çalışmalar**

Bu proje, makine öğrenmesi kullanarak kardiyovasküler hastalık riskini tahmin etme potansiyelini başarıyla ortaya koymuştur.

* **Sonuç:** XGBoost modeli, %73.6 doğruluk oranıyla, özellikle yaş, kan basıncı ve kolesterol gibi bilinen risk faktörlerini kullanarak anlamlı tahminler yapabildiğini göstermiştir. Proje başında hedeflenen %80'lik doğruluk oranına ulaşılmamış olsa da, elde edilen sonuçlar modelin pratik bir karar destek aracı olarak kullanılabileceğine işaret etmektedir.
* **Gelecek Çalışmalar İçin Öneriler:**
  1. **Hiperparametre Optimizasyonu:** XGBoost ve Random Forest modellerinin hiperparametreleri, GridSearchCV gibi yöntemlerle optimize edilerek doğruluk oranı artırılabilir.
  2. **Yeni Özellikler Türetme:** Vücut Kitle İndeksi (BMI) gibi mevcut verilerden yeni ve anlamlı özellikler türetmek, modelin performansını iyileştirebilir.
  3. **Farklı Algoritmalar:** LightGBM, CatBoost gibi daha modern gradient boosting algoritmaları veya basit sinir ağları denenebilir.
  4. **Power BI Geliştirmeleri:** Projenin son çıktısı olan "Model Sonuçları Dashboard"una, seçilen en iyi modelin özellik önemlerini (feature importance) gösteren bir görsel eklenerek, hangi faktörlerin riski ne kadar etkilediği sunulabilir.