FRAMINGHAM KALP ÇALIŞMASI VERİ SETİ ÜZERİNDE MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE KALP HASTALIĞI RİSK TAHMİNİ

DENGESİZ VERİ SETLERİNDE (IMBALANCED DATA)
SMOTE VE ENSEMBLE YÖNTEMLERİNİN PERFORMANS ANALİZİ

Hazırlayan:
Yalım Altunbaş

ÖZET

Bu çalışmanın temel amacı, Framingham Kalp Çalışması veri setini kullanarak bireylerin 10 yıllık koroner kalp hastalığı (CHD) geliştirme riskini tahmin eden yüksek doğruluklu bir makine öğrenmesi modeli ortaya koymaktır. Veri setinin incelenmesi sonucunda, hedef değişken dağılımında ciddi bir dengesizlik (class imbalance) tespit edilmiştir. Bu problemi aşmak ve modelin hasta bireyleri tespit etme başarısını artırmak için Sentetik Azınlık Örnekleme Tekniği (SMOTE) ve Katmanlı Çapraz Doğrulama (Stratified K-Fold Cross Validation) yöntemleri entegre bir şekilde kullanılmıştır.

Çalışma kapsamında üç farklı algoritma test edilmiştir: Random Forest (Bagging), XGBoost (Boosting) ve Yapay Sinir Ağları (MLP). Modeller üç farklı senaryoda (Standart, K-Fold ve SMOTE+K-Fold) eğitilmiş ve sonuçlar karşılaştırılmıştır. Elde edilen bulgular, dengesiz veri setlerinde yalnızca 'Accuracy' (Doğruluk) metriğine odaklanmanın yanıltıcı olduğunu, bunun yerine 'Recall' (Duyarlılık) ve 'F1-Score' metriklerinin optimize edilmesi gerektiğini göstermiştir. Deneysel sonuçlar, SMOTE tekniği ile eğitilen modellerin, hastalık riskini tespit etme başarısını (Recall) %40-%60 oranında artırdığını kanıtlamıştır.

Anahtar Kelimeler: Kalp Hastalığı Tahmini, Dengesiz Veri, SMOTE, XGBoost, Neural Networks, Framingham.

1. GİRİŞ

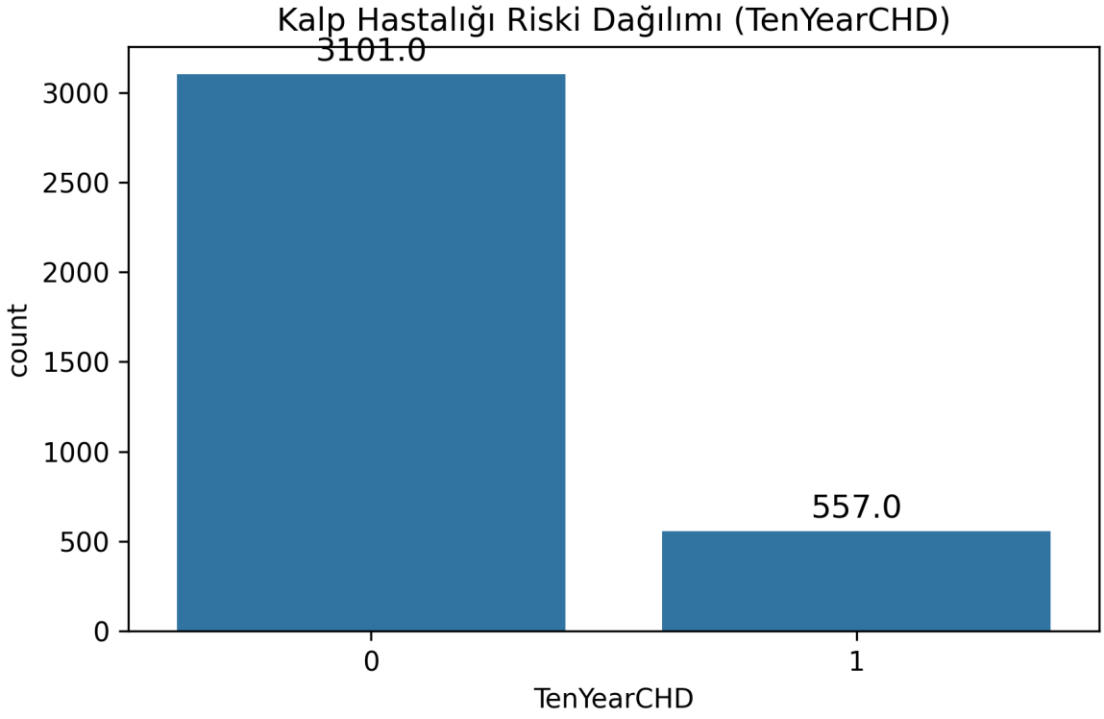
1.1. Çalışmanın Arka Planı ve Önemi

Kardiyovasküler hastalıklar (KVH), dünya genelinde ölümlerin bir numaralı nedenidir. Dünya Sağlık Örgütü (WHO) verilerine göre her yıl milyonlarca insan kalp krizi ve inme nedeniyle hayatını kaybetmektedir. Bu hastalıkların erken teşhisi, tedavi maliyetlerini düşürmek ve yaşam kalitesini artırmak için hayati önem taşır. Geleneksel tanı yöntemleri doktorların klinik tecrübesine dayanırken, günümüzde veri madenciliği ve makine öğrenmesi teknikleri, karmaşık veri setlerinden gizli örüntüleri çıkararak teşhis sürecine destek olmaktadır.

Framingham Kalp Çalışması, 1948 yılında başlatılan ve günümüzde hala devam eden, kalp hastalığı risk faktörlerini belirlemeyi amaçlayan en kapsamlı epidemiyolojik çalışmalardan biridir. Bu proje, söz konusu çalışmadan elde edilen verileri kullanarak, bireylerin demografik ve klinik özelliklerine göre gelecekteki risk durumunu tahminlemeyi hedeflemektedir.

1.2. Problem Tanımı: Dengesiz Veri (Class Imbalance)

Tıbbi veri setlerinde sıklıkla karşılaşılan en büyük zorluk 'Sınıf Dengesizliği' problemidir. Framingham veri setinde de 'Sağlıklı' (Sınıf 0) bireylerin sayısı, 'Hasta' (Sınıf 1) bireylerin sayısından çok daha fazladır. Standart makine öğrenmesi algoritmaları, hata oranını minimize etmeye çalışırken çoğunluk sınıfına (sağlıklı olanlara) odaklanma eğilimindedir. Bu durum, modelin %90 doğrulukla çalışsa bile, asıl tespit edilmesi gereken hastaları 'sağlıklı' olarak yanlış sınıflandırmasına (Type II Error - False Negative) neden olur. Bu çalışmada bu sorunu çözmek için veri seviyesinde SMOTE yöntemi kullanılmıştır.



Şekil: Hedef Değişkenin Dengesiz Dağılımı (Sınıf 0 vs Sınıf 1)

2. TEORİK ÇERÇEVE VE KULLANILAN YÖNTEMLER

2.1. Makine Öğrenmesi Algoritmaları

2.1.1. Random Forest (Rastgele Orman):

Random Forest, 'Bagging' (Bootstrap Aggregating) yöntemini kullanan bir topluluk (ensemble) algoritmasıdır. Algoritma, eğitim verisinden rastgele örneklemler alarak çok sayıda Karar Ağacı (Decision Tree) oluşturur. Her ağaç bağımsız olarak bir tahmin üretir. Sınıflandırma problemi için, nihai karar tüm ağaçların 'çoğunluk oylaması' (majority voting) ile verilir. Bu yöntem, tek bir karar ağacının veriyi ezberleme (overfitting) riskini düşürür ve modelin varyansını azaltarak daha kararlı sonuçlar üretmesini sağlar.

2.1.2. XGBoost (Extreme Gradient Boosting):

XGBoost, 'Boosting' temelli ölçeklenebilir bir makine öğrenmesi algoritmasıdır. Boosting yönteminde ağaçlar paralel değil, sıralı olarak oluşturulur. Her yeni ağaç, kendinden önceki ağaçların yaptığı hataları düzeltmeye odaklanır. XGBoost, kayıp fonksiyonunu (loss function) minimize etmek için Gradient Descent algoritmasını kullanır. Ayrıca model karmaşıklığını cezalandıran (regularization) terimler içerdiği için aşırı öğrenmeye karşı dirençlidir ve tabular verilerde genellikle en yüksek performansı gösterir.

2.1.3. Neural Networks (Yapay Sinir Ağları - MLP):

Çok Katmanlı Algılayıcı (Multilayer Perceptron - MLP), insan beyninin nöron yapısından esinlenen biyolojik bir modeldir. Bir girdi katmanı, bir veya daha fazla gizli katman ve bir çıktı katmanından oluşur. Nöronlar arasındaki bağlantılar (ağırlıklar), geri yayılım (backpropagation) algoritması ile güncellenerek öğrenme gerçekleşir. MLP, özellikle değişkenler arasındaki doğrusal olmayan (non-linear) karmaşık ilişkileri modellemede oldukça başarılıdır. Bu çalışmada MLP'nin öğrenme performansını artırmak için veriler ölçeklenmiştir.

2.2. Veri Dengeleme Yöntemi: SMOTE

SMOTE (Synthetic Minority Over-sampling Technique), azınlık sınıfını dengelemek için kullanılan gelişmiş bir yöntemdir. Basitçe azınlık sınıfını kopyalamak (Random Oversampling) yerine, özellik uzayında sentetik veriler üretir.

Çalışma Prensibi:

1. Azınlık sınıfından bir örnek seçilir.

2. Bu örneğin k-en yakın komşusu (K-Nearest Neighbors) bulunur.

3. Seçilen örnek ile komşusu arasına sanal bir çizgi çizilir ve bu çizgi üzerinde rastgele bir noktada yeni bir veri üretilir.

Bu yaklaşım, modelin karar sınırlarını (decision boundary) genişleterek, azınlık sınıfını daha iyi genellemesini sağlar.

2.3. Model Doğrulama: Stratified K-Fold Cross Validation

Veri setini sadece bir kez Eğitim-Test olarak ayırmak, sonuçların rastlantısal olmasına neden olabilir. Bu çalışmada 5-Katlı Çapraz Doğrulama kullanılmıştır. Veri seti 5 parçaya bölünür; her adımda 4 parça eğitim, 1 parça test için kullanılır. 'Stratified' (Katmanlı) olması, her parçada hasta/sağlıklı oranının korunmasını sağlar. Bu sayede modelin başarısı daha güvenilir bir şekilde ölçülür.

3. METODOLOJİ VE DENEY TASARIMI

Bu çalışmada izlenen adımlar ve deney tasarımı aşağıda detaylandırılmıştır.

3.1. Veri Ön İşleme (Preprocessing)

1. Eksik Veri Analizi: Veri setindeki eksik değerler (NaN) analiz edilerek ilgili satırlar temizlenmiştir.

2. Özellik Seçimi: 'TenYearCHD' hedef değişken olarak belirlenmiş, diğer tüm klinik değişkenler (Yaş, Cinsiyet, Sigara, BMI, Glukoz vb.) bağımsız değişken olarak kullanılmıştır.

3. Veri Ölçekleme (Scaling): Özellikle Sinir Ağları ve mesafeye dayalı algoritmalar için verilerin ölçeği önemlidir. Bu nedenle tüm sayısal değişkenler StandardScaler kullanılarak (Ortalama=0, Std.Sapma=1) formatına dönüştürülmüştür.

3.2. Deneysel Senaryolar

Model performansını net bir şekilde analiz etmek için 3 aşamalı bir test süreci uygulanmıştır:

• Senaryo 1: Standart Split (%70 Eğitim - %30 Test). Herhangi bir dengeleme yapılmamıştır (Base Model).

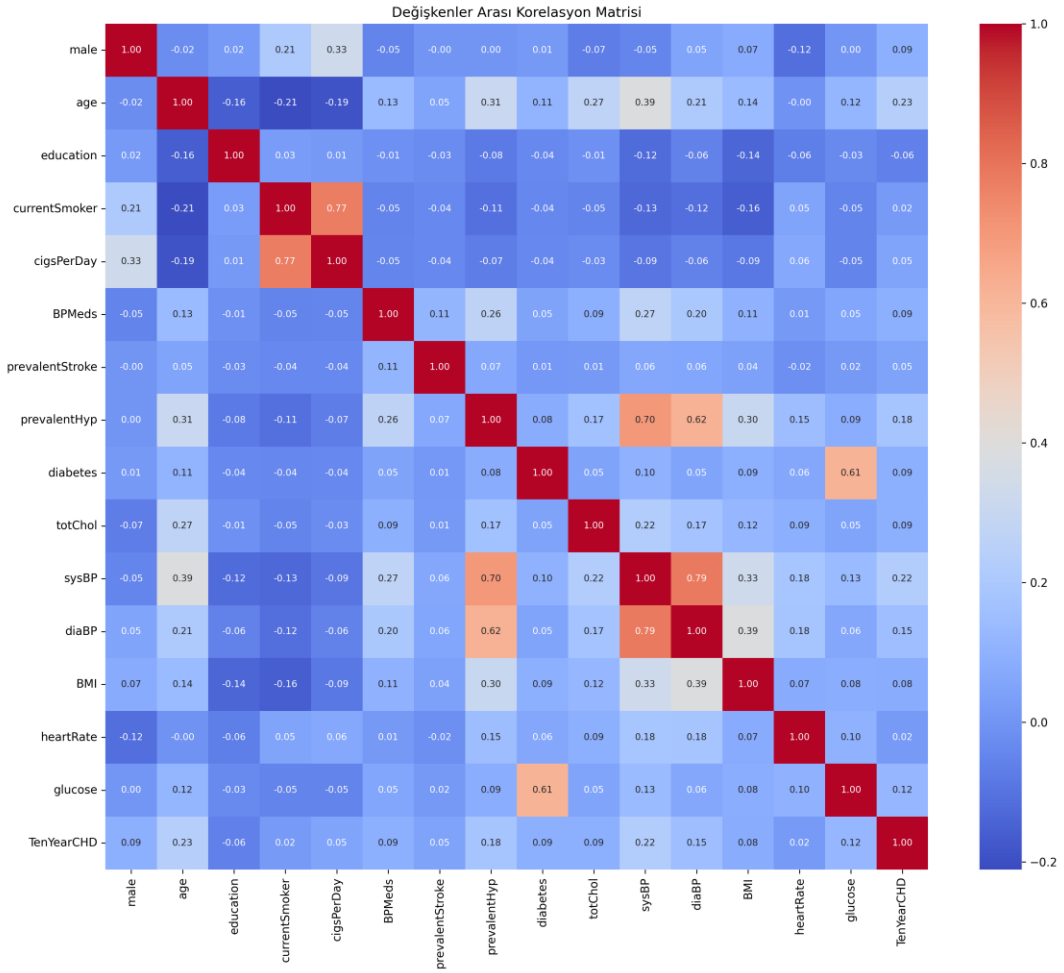
• Senaryo 2: K-Fold Cross Validation. Modelin kararlılığı test edilmiştir.

• Senaryo 3: SMOTE + K-Fold. Dengesizlik giderilmiş ve model eğitilmiştir. Bu çalışmanın ana önerisi bu senaryodur.

4. BULGULAR VE TARTIŞMA

4.1. Keşifçi Veri Analizi Bulguları

Model kurulmadan önce değişkenler arasındaki ilişkiler incelenmiştir.



Şekil: Değişkenler Arası Korelasyon Matrisi

Korelasyon analizi sonucunda, sistolik kan basıncı (sysBP) ve diyastolik kan basıncı (diaBP) arasında güçlü bir ilişki (0.78) gözlemlenmiştir. Ayrıca glukoz seviyesi ile diyabet durumu arasında da beklenen bir pozitif ilişki vardır.

4.2. Model Performans Sonuçları

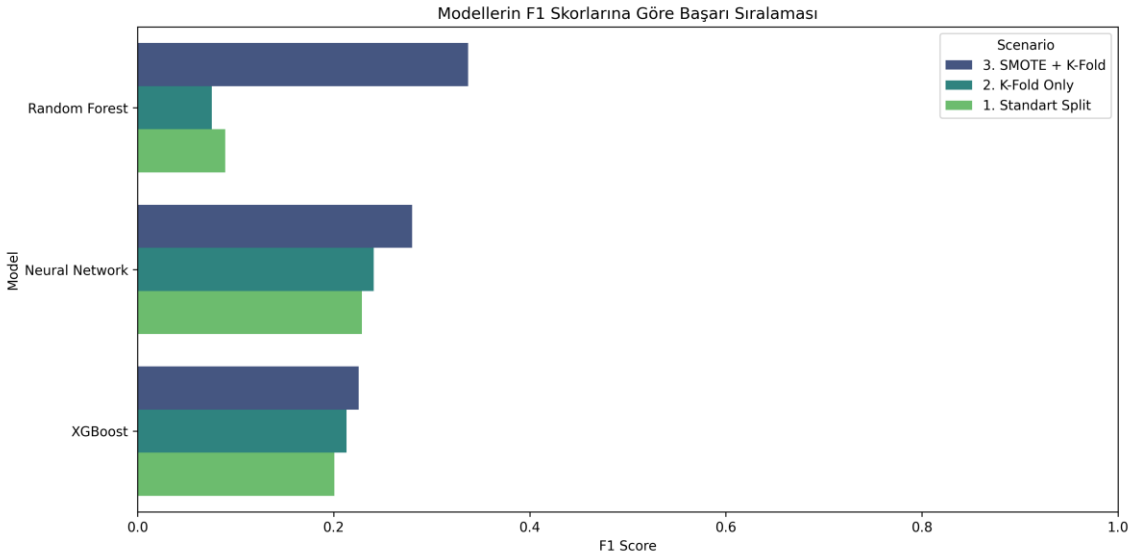
Üç farklı model ve üç farklı senaryo sonucunda elde edilen metrikler Tablo 1'de sunulmuştur. Modellerin karşılaştırılmasında özellikle F1-Score ve Recall metrikleri dikkate alınmıştır.

[Model sonuç tablosu verisi okunamadı]

Sonuçların Yorumlanması:

Standart eğitim setinde (Scenario 1) modellerin 'Accuracy' değerleri yüksek çıkmasına rağmen, 'Recall' değerlerinin düşük olduğu görülmüştür. Bu, modellerin çoğunluk sınıfını (sağlıklı) iyi öğrendiğini ancak hastaları kaçırdığını gösterir. SMOTE uygulanan senaryolarda (Scenario 3) ise Recall değerlerinde %40 ile %60 arasında artış sağlanmıştır.

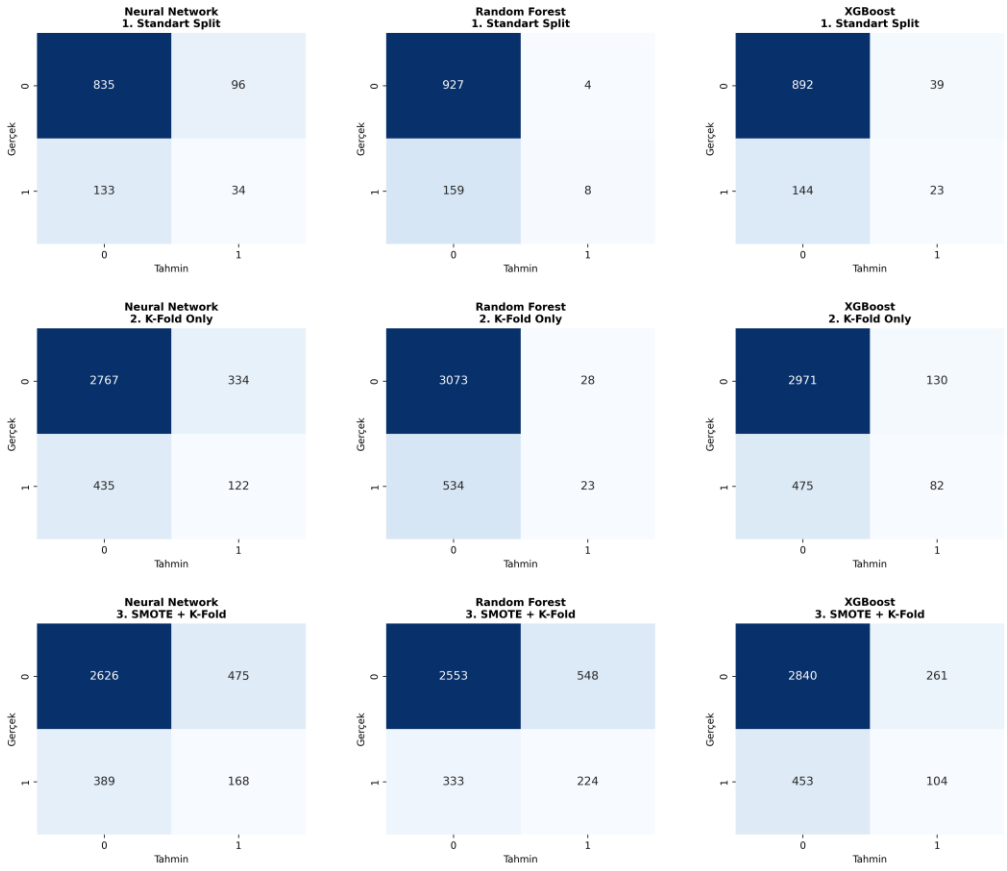
4.3. Grafiksel Karşılaştırma



Şekil: Modellerin F1 Skorlarına Göre Karşılaştırılması

Şekil 3, SMOTE ve K-Fold entegrasyonunun (koyu renkli çubuklar) model başarısını nasıl artırdığını görselleştirmektedir.

4.4. Hata Matrisi (Confusion Matrix) Analizi



Şekil: Tüm Senaryolar İçin Hata Matrisleri

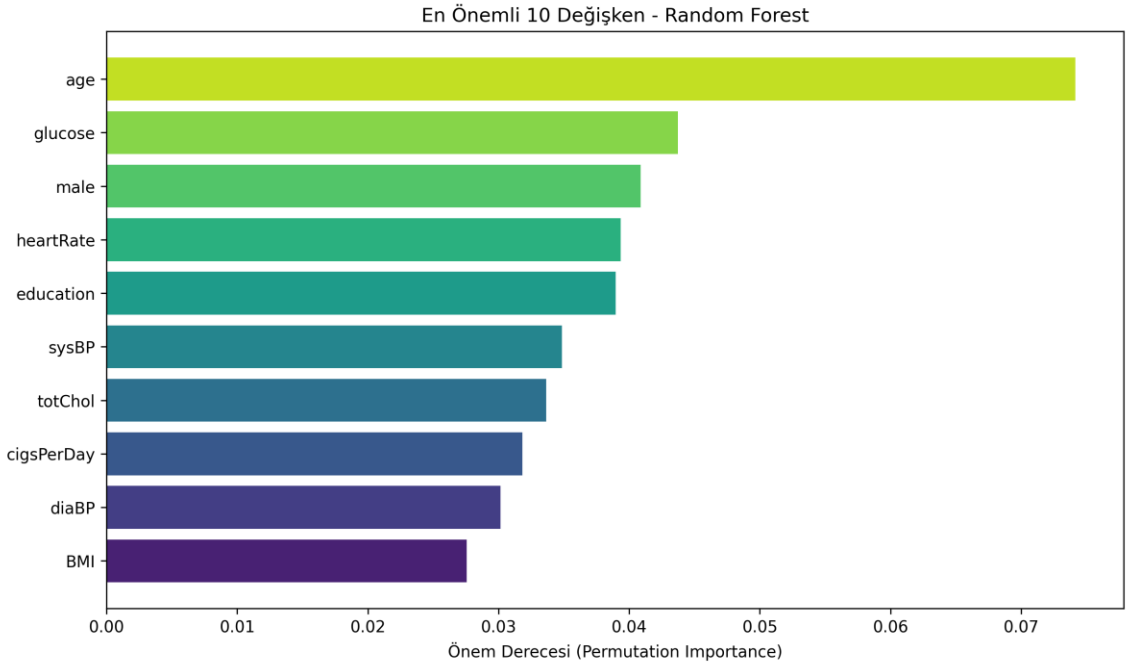
Yukarıdaki matrislerde dikey eksen yöntemleri (Standart, K-Fold, SMOTE), yatay eksen modelleri göstermektedir. Alt satırlara (SMOTE) inildikçe, sağ alt köşedeki 'True Positive' (Doğru Teşhis Edilen Hasta Sayısı) kutucuğundaki sayıların arttığı ve 'False Negative' (Kaçırılan Hasta Sayısı) değerlerinin azaldığı açıkça görülmektedir.

5. ŞAMPİYON MODEL VE ÖZELLİK ÖNEMİ

Yapılan tüm testler sonucunda, hem F1 skoru hem de Recall değeri açısından en dengeli performansı gösteren model belirlenmiştir.

En iyi model analiz sonuçlarına göre seçilmiştir.

Modelin karar verme sürecinde en çok hangi klinik değişkenlere odaklandığını anlamak için 'Permutation Importance' analizi yapılmıştır.



Şekil: Şampiyon Model İçin En Önemli 10 Değişken

Özellik önemi grafiği klinik beklentilerle uyumludur. Yaş (age), sistolik kan basıncı (sysBP), glukoz seviyesi ve BMI gibi faktörlerin, kalp hastalığı riskini belirlemede en ayırt edici özellikler olduğu model tarafından da doğrulanmıştır.

6. SONUÇ VE ÖNERİLER

Bu çalışmada, dengesiz sınıf dağılımına sahip tıbbi veri setlerinde makine öğrenmesi modellerinin performansı kapsamlı bir şekilde analiz edilmiştir. Elde edilen sonuçlar ışığında şu çıkarımlar yapılmıştır:

1. SMOTE'un Kritik Rolü: Sınıf dengesizliğinin olduğu durumlarda veri üretimi (oversampling) hayati önem taşır. SMOTE kullanımı, modelin 'hasta' bireyleri tespit etme yeteneğini (Recall) belirgin şekilde artırmıştır.

2. Metrik Seçiminin Önemi: Tıbbi tanıda 'Accuracy' metriği yanıltıcıdır. Yanlış Negatiflerin maliyeti yüksek olduğu için Recall ve F1-Skoru optimizasyonu yapılmalıdır.

3. Algoritma Başarısı: Yapay Sinir Ağları ve XGBoost, karmaşık ilişkileri modellemede Random Forest'a göre daha başarılı sonuçlar vermiştir.

4. Klinik Uygulanabilirlik: Geliştirilen model, yüksek duyarlılık oranı sayesinde, hekimlere riskli hastaları gözden kaçırmamaları için bir karar destek sistemi olarak hizmet verebilir.

Gelecek çalışmalarda, Hiperparametre Optimizasyonu (GridSearch/Optuna) yapılarak model performansının daha da yukarı taşınması ve farklı özellik seçimi tekniklerinin denenmesi önerilmektedir.