T.C.

Gazi Üniversitesi

Teknoloji Fakültesi, Bilgisayar Mühendisliği

CDC Diyabet Sağlık Göstergeleri Üzerine Makine Öğrenmesi Tabanlı Risk Tahmini

Gizem Ece ÇOMAK
Gazi Üniversitesi, Bilgisayar Mühendisliği Bölümü
gece.comak@gazi.edu.tr

Mesude TÜRKMEN
Gazi Üniversitesi, Bilgisayar Mühendisliği Bölümü
mesude.turkmen@gazi.edu.tr

Zeren KAVAZ
Gazi Üniversitesi, Bilgisayar Mühendisliği Bölümü
zeren.kavaz@gazi.edu.tr

Bahar Dönemi, 2025

BMT 218 - Veri Bilimine Giriş Dersi, Dönem Sonu Projesi

Proje Özeti

Bu çalışmada, CDC'nin (Centers for Disease Control and Prevention) yayımladığı *Diabetes Health Indicators* veri seti kullanılarak diyabet riski tahmin edilmiş ve en önemli risk faktörleri belirlenmiştir. Veri seti, 250.000'den fazla gözlem ve yaş, vücut kitle indeksi (BMI), yüksek tansiyon, genel sağlık durumu gibi 21 farklı sağlık göstergesini içermektedir.

Proje kapsamında; veri temizleme (eksik veriler ve aykırı değerlerin giderilmesi), sınıf dengesizliğinin SMOTE yöntemiyle çözülmesi ve yedi farklı makine öğrenmesi algoritmasının (Logistic Regression, Random Forest, XGBoost vb.) performans karşılaştırmaları gerçekleştirilmiştir. Özellikle XGBoost modeli, 0.96 ROC-AUC skoru ile en yüksek başarıyı göstermiştir.

Ayrıca çalışmada, farklı örnekleme stratejilerinin model başarımına etkisi de incelenmiştir. Bir versiyonda sınıf dengesizliği, rastgele alt örnekleme yöntemiyle giderilmiş ve ardından çeşitli özellik seçiciler (SelectKBest, RFE vb.) uygulanarak Lojistik Regresyon ve Random Forest algoritmaları değerlendirilmiştir. Diğer bir yaklaşımsa SMOTE (Synthetic Minority Over-sampling Technique) yöntemi ile sınıf dengesini sağlamış ve XGBoost ile yüksek doğruluk oranlarına ulaşılmıştır.

Diyabet riskini öngörmede en etkili değişkenler; vücut kitle indeksi (BMI), yüksek tansiyon (HighBP) ve genel sağlık durumu (GenHlth) olarak belirlenmiştir. Bu çalışma, hem farklı modelleme yaklaşımlarının karşılaştırmasını sunmakta hem de halk sağlığı ve klinik karar destek sistemlerinde uygulanabilir yüksek başarımlı modellerin önünü açmaktadır.

[Anahtar Sözcükler]

Diyabet Risk Tahmini, Sağlıkta Makine Öğrenmesi, CDC Diyabet Sağlık Göstergeleri, XGBoost Modeli, Özellik Önemlilik Analizi, Sınıf Dengesizliği (SMOTE), ROC-AUC Değerlendirmesi, BMI ve Yüksek Tansiyon, Halk Sağlığı Analitiği, Karışıklık Matrisi, Klinik Karar Destek Sistemleri, Gradient Boosting Algoritmaları

1. Giriş

Diyabet (Diabetes Mellitus), insülin hormonunun eksikliği, etkisizliği ya da üretimindeki bozukluk sonucu ortaya çıkan, kan şekeri düzeyinin kronik olarak yüksek seyretmesine yol açan metabolik bir hastalıktır. Tip 1 diyabet (otoimmün kökenli) ve Tip 2 diyabet (insülin direnci veya sekresyon bozukluğu) olmak üzere iki ana formda görülen bu hastalık, pankreasın beta hücrelerinin işlev kaybı veya hedef dokuların insüline direnci nedeniyle gelişir. Dünya Sağlık Örgütü (WHO) verilerine göre, 2021 yılı itibarıyla dünya genelinde 537 milyondan fazla yetişkin diyabetle yaşamakta ve her yıl 6.7 milyon ölüm doğrudan veya dolaylı olarak diyabet kaynaklı komplikasyonlarla ilişkilendirilmektedir. Bu rakamlar, diyabetin küresel sağlık sistemleri üzerinde yıllık 966 milyar ABD doları ekonomik yük oluşturduğunu gösteren IDF (Uluslararası Diyabet Federasyonu) raporlarıyla da desteklenmektedir.

Diyabetin uzun vadede yol açtığı kardiyovasküler hastalıklar, böbrek yetmezliği, nöropati ve retinopati gibi komplikasyonlar, yalnızca bireysel sağlığı değil, toplumun genel refahını da tehdit etmektedir. Özellikle düşük ve orta gelirli ülkelerde, erken teşhis ve tedaviye erişimdeki kısıtlamalar, bu komplikasyonların prevalansını artırmaktadır. Bu nedenle, diyabetin önleyici

stratejilerle kontrol altına alınması ve kişiselleştirilmiş risk yönetimi, halk sağlığı politikalarının öncelikli hedefleri arasında yer almaktadır.

Geleneksel diyabet tanı yöntemleri, açlık kan şekeri ölçümü, oral glukoz tolerans testi (OGTT) ve HbA1c gibi biyokimyasal testlere dayanmaktadır. Ancak bu yöntemler, geniş popülasyonlar için yüksek maliyet, zaman alıcılık ve erişim eşitsizlikleri nedeniyle sınırlı kalmaktadır. Örneğin, kırsal bölgelerde laboratuvar altyapısının yetersizliği veya düşük gelirli bireylerin düzenli taramalara katılımının azlığı, erken teşhis oranlarını düşürmektedir. Bu durum, büyük veri analitiği ve makine öğrenmesi tabanlı çözümleri kaçınılmaz kılmaktadır.

Yapay zekâ ve makine öğrenmesi teknikleri, heterojen veri kümelerindeki karmaşık örüntüleri tanımlayabilme, çok değişkenli risk faktörlerini entegre edebilme ve gerçek zamanlı tahminler üretebilme kapasiteleriyle tıp alanında devrimsel bir rol üstlenmektedir. Özellikle SHAP (SHapley Additive exPlanations) ve LIME (Local Interpretable Model-agnostic Explanations) gibi açıklanabilir AI (XAI) yöntemleri, modellerin "kara kutu" algısını kırarak klinisyenlerin karar süreçlerine güvenilirlik kazandırmaktadır. Nitekim, bu çalışmayla birlikte, makine öğrenmesi modellerinin diyabet tahmininde %96'ya varan ROC-AUC değerleriyle geleneksel yöntemleri geride bıraktığını ortaya koymaktadır.

Bu çalışmada, Amerika Birleşik Devletleri Hastalık Kontrol ve Önleme Merkezleri (CDC) tarafından yayınlanan 2015 Diabetes Health Indicators veri seti kullanılmıştır. BRFSS (Behavioral Risk Factor Surveillance System) anketinden türetilen bu veri seti, 253.680 gözlem ve 21 öznitelik ile yaş, vücut kitle indeksi (BMI), yüksek tansiyon (HighBP), fiziksel aktivite düzeyi, genel sağlık algısı ve sosyoekonomik faktörler gibi kritik değişkenleri kapsamaktadır. Literatürde sıklıkla kullanılan Pima Indians veri setinin aksine, bu veri seti daha geniş bir demografik çeşitlilik sunmakta ve toplumsal temsiliyet sağlamaktadır.

2. Materyal ve Metot

2.1. Veri Setinin Düzenlenmesi ve Ön İşleme

Bu çalışmada kullanılan 2015 Diabetes Health Indicators veri seti, CDC (Centers for Disease Control and Prevention) tarafından yayımlanan ve BRFSS (Behavioral Risk Factor Surveillance System) anketinden türetilen, 21 öznitelik ve 253.680 gözlem içeren geniş kapsamlı bir veri kümesidir. Veri seti, bireylerin demografik özellikleri (yaş, cinsiyet), biyometrik göstergeler (BMI, kan basıncı), sağlık davranışları (sigara kullanımı, fiziksel aktivite) ve psikososyal faktörler (mental sağlık, genel sağlık algısı) gibi multidisipliner değişkenleri kapsamaktadır.

2.1.1. Veri Temizliği ve Aykırı Değer Yönetimi

- **BMI (Vücut Kitle İndeksi):** Biyolojik olarak anlamlı aralık dışındaki değerler (BMI < 12 veya BMI > 60) veri setinden çıkarılmıştır. Bu aralık, WHO'nun obezite sınıflandırması (BMI ≥ 30) ve aşırı zayıflık kriterleri (BMI < 16) dikkate alınarak belirlenmiştir.

- **GenHlth (Genel Sağlık Algısı):** 1-5 arası Likert skalasında kodlanmıştır (1: Mükemmel, 5: Kötü). Skala dışı değerler filtrelenmiştir.

- **PhysHlth ve MentHlth (Fiziksel/Mental Sağlık):** Son 30 gündeki sağlık sorunlarının gün sayısını temsil eden bu değişkenlerde, 0-30 aralığı dışındaki değerler kaldırılmıştır.

Filtreleme sonrasında 5.462 aykırı gözlem ve 3.891 yinelenen kayıt temizlenerek 244.327 örnek ile analize devam edilmiştir.

2.1.2. Sınıf Dengesizliği ve SMOTE Uygulaması

Hedef değişken olan Diabetes_binary (0: Diyabetsiz, 1: Diyabet), orijinal veri setinde %15.4 diyabetli ve %84.6 diyabetsiz örnek içermekteydi. Bu dengesizlik, modelin azınlık sınıfı (diyabetli) yetersiz öğrenmesine yol açabileceğinden, SMOTE (Synthetic Minority Over-sampling Technique) ile sentetik örnek üretimi yapılmıştır. SMOTE, k-NN algoritması kullanarak azınlık sınıfın komşuluk ilişkilerine dayalı yapay veri üretir ve örnek sayısını çoğunluk sınıfla eşitler. İşlem sonrasında dengeli veri seti 387.462 gözlem (0 ve 1 sınıfları eşit) içermektedir.

2.1.3. Rastgele Alt Örnekleme ile Sınıf Dengesi Sağlama

Hedef değişkende (Diabetes_binary) yer alan dengesiz yapı y, modellerin azınlık sınıf olan diyabetli bireyleri doğru tahmin etme yeteneğini azaltmaktadır. Bu sorunu çözmek için alternatif bir yöntem olarak rastgele alt örnekleme (random undersampling) uygulanmıştır.

Bu yöntemde, çoğunluk sınıf olan diyabetsiz bireylerden rastgele örnekler seçilerek, azınlık sınıfın sayısıyla eşitlenmiştir. Böylece her iki sınıfın da veri setindeki temsili %50 oranına getirilmiş ve dengeli bir sınıf yapısı elde edilmiştir. Dengeleme sonrası veri seti, toplam 79.844 gözlem içermektedir (diyabetli: 39.922, diyabetsiz: 39.922).

Dengeli veri seti oluşturulduktan sonra çeşitli özellik seçimi teknikleri uygulanarak boyut indirgeme yapılmış, ardından Lojistik Regresyon ve Random Forest modelleri ile sınıflandırma gerçekleştirilmiştir. Bu yaklaşım, farklı dengeleme yöntemlerinin model performansına etkisini karşılaştırmak açısından önemli bir karşılaştırma sağlamıştır.

2.2. Sınıflandırma Yöntemi ve Modelleme Süreci

2.2.1. Veri Bölünmesi ve Ölçeklendirme

- **Eğitim-Test Ayrımı:** Dengeli veri seti, %80 eğitim (309.970 gözlem) ve %20 test (77.492 gözlem) olarak stratifiye şekilde bölünmüştür. Stratifiye bölme, sınıf dağılımının her iki kümede de korunmasını sağlamıştır.

- **Öznitelik Ölçeklendirme:** Lojistik Regresyon, SVM ve K-NN gibi metrik tabanlı algoritmalar için StandardScaler ile z-puan normalizasyonu uygulanmıştır. Bu işlem, özelliklerin ortalamasını 0, standart sapmasını 1 yaparak algoritmaların yakınsama hızını ve kararlılığını artırmıştır.

2.2.2. Hiperparametre Optimizasyonu

Her algoritma için **GridSearchCV** ile 3 katlı çapraz doğrulama yapılmış ve en iyi hiperparametre kombinasyonları belirlenmiştir. Örneğin:

- **XGBoost:** max_depth=3, n_estimators=100, learning_rate=0.1
- **Random Forest:** n_estimators=100, max_depth=10, min_samples_split=5
- **Lojistik Regresyon:** C=1, solver='lbfgs'

3. Kullanılan Algoritmalar ve Teorik Temeller

Bu bölümde, çalışmada kullanılan makine öğrenmesi algoritmalarının teorik temelleri ve diyabet risk tahminine nasıl uyarlandıkları detaylandırılmıştır.

3.1. Lojistik Regresyon

Lojistik regresyon, özellikle ikili sınıflandırma problemlerinde sıkça kullanılan, doğrusal bir modeldir. sklearn.linear_model.LogisticRegression sınıfı aracılığıyla scikit-learn kütüphanesi içerisinde yer almakta olup, olasılıksal bir tahmin çerçevesi sunmaktadır. Model, doğrusal bir kombinasyon aracılığıyla veriyi sınıflandırsa da sonuç olarak sigmoid (logistic) fonksiyonunu kullanarak çıktıları 0 ile 1 arasında olasılıklara dönüştürür:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 x 1 + \cdots + \beta n x n)}}$$

Bu ifade, verilen giriş özellikleri x=(x1,x2,…,xn)x = (x_1, x_2, … ,x_n) bir gözlemin pozitif sınıfa ait olma olasılığını ifade eder. Modelin karar sınırı genellikle 0.5 olarak alınır.

Temel Parametreler:

- **C:** Ters düzenlileştirme (regularization) parametresi olup, modelin genelleme gücünü etkiler. Küçük C değerleri daha güçlü düzenlileştirme anlamına gelir. Bu çalışmada C=0.1 olarak belirlenmiştir.
- **solver:** Ağırlıkların optimizasyonunda kullanılan yöntemdir. 'lbfgs', 'liblinear', 'sag', 'saga' gibi seçenekler bulunur. Bu çalışmada 'lbfgs' tercih edilmiştir.
- **penalty:** Kullanılan düzenlileştirme türü ('l1', 'l2', 'elasticnet', vb.). Varsayılan 'l2' düzenlemesidir.
- **max_iter:** Maksimum iterasyon sayısı. Yetersizse modelin yakınsaması engellenebilir.

Lojistik regresyon modelinin avantajları arasında:

- Modelin kolay yorumlanabilirliği,
- Aşırı öğrenmeye karşı dayanıklılığı,
- Düşük boyutlu ve doğrusal ayrılabilir veri setlerinde yüksek performans göstermesi sayılabilir.

Ancak bazı dezavantajları da vardır:

- Doğrusal olmayan ilişkilerde başarısı sınırlıdır,
- Özellikler arasında çoklu bağlantılar (multicollinearity) varsa performans düşebilir,
- Büyük boyutlu özellik uzaylarında karmaşık ilişkileri öğrenmede yetersiz kalabilir.

Bu çalışmada uygulanan Lojistik Regresyon modeli, ROC-AUC değeri olarak 0.81 ve doğruluk değeri olarak %74.1 oranında bir başarı göstermiştir. Modelin F1 skoru 0.75, precision değeri 0.73 ve recall oranı 0.77 olarak ölçülmüştür. Bu sonuçlar, modelin özellikle negatif sınıfı ayırt etmede orta düzeyde başarılı olduğunu, ancak daha kompleks modellerin daha yüksek performans sergilediğini göstermektedir.

Diyabete Uyarlama:

- Özelliklerin ağırlıklı toplamı, diyabet riskini temsil eden bir log-odds skoru üretir.
- L2 Regularizasyonu (C parametresi) ile aşırı uyum (overfitting) engellenir.

3.2. Karar Ağaçları (Decision Tree)

Karar ağaçları, hem sınıflandırma hem de regresyon problemleri için uygun olan, açıklanabilirliği yüksek denetimli öğrenme algoritmaları arasında yer almaktadır. Bu algoritma, veriyi özelliklerine göre dallara ayırarak, "eğer...ise" (if-then) kuralları ile tahminleme yapar. Bu çalışmada karar ağacı modeli, sklearn.tree.DecisionTreeClassifier sınıfı aracılığıyla uygulanmıştır.

Model, veri kümesindeki örnekleri recursive binary splitting yöntemiyle özyinelemeli olarak alt gruplara ayırır. Her dalımda (split), en yüksek bilgi kazancı veya en düşük safsızlık (impurity) elde edilecek özellik seçilir. Her yaprak düğüm (leaf node), bir sınıf etiketiyle sonuçlanır. Bu süreçte en yaygın kullanılan safsızlık ölçütleri şunlardır:

- **Gini İndeksi:**

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

Burada p_i, sınıf i'nin düğümdeki olasılığıdır.

- **Entropi:**

$$Entropy = -\sum_{i=1}^{n} p_i \, log_2$$

Öne Çıkan Parametreler:

- **criterion:** Bölünme kalitesini ölçmek için kullanılan kriter. 'gini' (varsayılan) veya 'entropy' seçenekleri bulunur.

- **max_depth:** Ağacın maksimum derinliği. Overfitting riskini azaltmak için sınırlanabilir.
- **min_samples_split:** Dallanma için gerekli minimum örnek sayısı. Bu çalışmada 5 olarak belirlenmiştir.
- **max_features:** Her split için göz önünde bulundurulacak maksimum özellik sayısı.
- **random_state:** Sonuçların tekrar üretilebilmesi için belirlenen sabit.

Avantajlar ve Sınırlılıklar:

- Avantajları:
    - Kolay yorumlanabilir yapı.
    - Özellik seçimi gerektirmeden çalışabilme.
    - Sayısal ve kategorik verilerle uyumluluk.
- Sınırlamaları:
    - Veri üzerinde overfitting (aşırı öğrenme) riski yüksektir.
    - Küçük veri değişiklikleri model yapısını önemli ölçüde değiştirebilir.
    - Daha az genelleme yeteneğine sahiptir.

Bu çalışmada Decision Tree algoritması, doğruluk oranı %83.4, F1 skoru 0.83 ve ROC-AUC değeri 0.96 olarak elde edilmiştir. Özellikle recall oranının yüksekliği (0.83), pozitif sınıfı belirlemede güçlü olduğunu göstermektedir. Ancak sınıf ayrımının daha keskin olduğu XGBoost gibi modellerin gerisinde kalmıştır.

Diyabete Uyarlama:
- Örneğin, ilk dallanma **BMI > 30** gibi bir eşik değerine göre yapılır.
- max_depth=10 ile ağacın derinliği sınırlanarak aşırı uyum önlenir.

3.3. Rastgele Orman (Random Forest)

Random Forest (Rastgele Orman), birden çok karar ağacının toplu olarak eğitildiği ve tahminlerin bu ağaçların oy çokluğu ilkesine göre yapıldığı, topluluk öğrenme (ensemble learning) yöntemidir. Bu yöntem, her bir ağacın farklı bir veri alt kümesi ve öznitelik alt kümesiyle eğitilmesi sayesinde yüksek varyanslı modellerin kararsızlığını azaltarak daha kararlı ve genellenebilir bir sınıflandırma performansı sunar.

Bir sınıflandırma probleminde, Random Forest aşağıdaki şekilde çalışır:

- Her bir karar ağacı, eğitim veri kümesinden rastgele örnekleme (bootstrap) ile oluşturulan bir alt küme üzerinde eğitilir.
- Her düğümde, ayrım için yalnızca rastgele seçilen bir alt özellik kümesi üzerinden en iyi ayrım yapılır.
- Her bir ağaç, bağımsız şekilde tahminde bulunur ve final sınıf etiketi, tüm ağaçların oy çokluğuna göre belirlenir:

$$y^{\wedge} = mode\{ht(x),\ t = 1,2,\ldots,T\}$$

Burada:

- $\hat{y}$: Nihai tahmin edilen sınıf etiketi,
- $h_t(x)$: t. karar ağacının tahmini,
- T: Modeldeki toplam karar ağacı sayısıdır.

Modelin her bir düğümünde bilgi kazancı (information gain), Gini katsayısı ya da entropi ölçütü kullanılabilir. Gini saflığı ölçütü şu şekilde tanımlanır:

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2$$

Burada:

- C: Sınıf sayısı,
- P_i: Veri kümesindeki i. sınıfın olasılığıdır.

Random Forest algoritması, özellikle yüksek boyutlu verilerde aşırı öğrenmeye karşı dayanıklıdır ve eksik veriler, çoklu doğrusal ilişkiler gibi karmaşık durumlarda da güçlü performans gösterir. Ayrıca, öznitelik önem derecelerini (feature importance) de analiz ederek model yorumlamaya katkı sağlar.

Ancak, Random Forest modelleri yüksek sayıda ağaç içerdiğinden dolayı eğitim süresi daha uzun olabilir ve modelin açıklanabilirliği, tekil karar ağaçlarına göre sınırlıdır. Buna rağmen, veri biliminde en yaygın ve etkili sınıflandırma tekniklerinden biridir.

Diyabete Uyarlama:

- n_estimators = 100 ağaç kullanılarak model kararlılığı artırılmıştır.
- Ağaçların çeşitliliği, HighBP ve BMI gibi özelliklerin tutarlılıkla önemli çıkmasını sağlar.

3.4. Gradient Boosting

Gradient Boosting, zayıf öğrenicilerin (genellikle karar ağaçları) ardışık olarak bir araya getirilmesiyle oluşturulan güçlü bir topluluk (ensemble) yöntemidir. Bu yöntem, her bir yeni modelin önceki modellerin hatalarını düzeltmeye çalıştığı bir iteratif öğrenme sürecine dayanır. sklearn.ensemble.GradientBoostingClassifier sınıfı, sınıflandırma problemleri için bu yöntemin uygulanmasını sağlar.

Bu algoritmanın temel amacı, modelin eğitim verisine olan hatasını (loss) azaltacak şekilde her yeni ağacı optimize etmektir. En çok kullanılan kayıp fonksiyonlarından biri log loss (lojistik regresyonla uyumlu) olup, bu çalışmada da sınıflandırma amacıyla kullanılmıştır.
Gradient Boosting süreci şu şekilde işler:

1. İlk model, verideki en temel örüntüyü yakalayacak şekilde eğitilir.
2. Sonraki her model, bir önceki modelin tahmin hataları üzerinde çalışır. Bu hata, diferansiyellenebilir bir kayıp fonksiyonunun gradyanı üzerinden tanımlanır.
3. Nihai model, bu zayıf öğrenicilerin ağırlıklı toplamından oluşur:

$$Fm(x) = Fm - 1(x) + \gamma mhm(x)$$

Burada:

- $F_m(x)$: m. iterasyondaki toplam model,
- $h_m(x)$: m. iterasyonda eğitilen zayıf model (genellikle karar ağacı),
- γ_m : öğrenme oranı (learning rate), modelin hatayı düzeltme miktarını kontrol eder.

Gradient Boosting'in performansı; n_estimators (ağaç sayısı), max_depth (ağaçların derinliği) ve learning_rate gibi hiperparametrelerle doğrudan ilişkilidir. Öğrenme oranı küçük tutulduğunda daha fazla ağaç gerekecek ancak modelin genelleme yeteneği artacaktır.

Bu çalışmada kullanılan Gradient Boosting modeli, 100 karar ağacı ve 0.1 öğrenme oranıyla optimize edilmiştir. ROC-AUC skoru 0.95'in üzerinde olan model, özellikle yüksek hassasiyet (precision) ve F1 skoru ile dikkat çekmiştir. Bu durum, modelin hem diyabetik bireyleri doğru tanılamada hem de yanlış pozitifleri azaltmada etkili olduğunu göstermektedir.

Modelin güçlü yönleri:

- Aşırı öğrenmeye karşı nispeten dirençli olması,
- Karmaşık ilişkileri yakalayabilme kapasitesi,
- Özellik önemlerinin değerlendirilebilir olmasıdır.

Ancak hesaplama maliyetinin diğer algoritmalara göre daha yüksek olması ve hiperparametre seçiminin kritik rol oynaması, Gradient Boosting'in dikkatle ayarlanması gereken bir model olduğunu göstermektedir.

Diyabete Uyarlama:

- max_depth=3 ile basit ağaçlar kullanılarak model karmaşıklığı dengelenmiştir.
- Özellik Önemlilik Analizi, BMI ve HighBP'nin dominant rolünü ortaya çıkarmıştır.

3.5. XGBoost

XGBoost (Extreme Gradient Boosting), karar ağacı temelli bir topluluk öğrenme yöntemidir ve özellikle yapılandırılmış (tabular) verilerdeki sınıflandırma ve regresyon problemlerinde son derece başarılı sonuçlar vermektedir. XGBoost, klasik gradient boosting algoritmasını optimize ederek hem eğitim süresini kısaltmakta hem de daha yüksek tahmin performansı sunmaktadır.

XGBoost, hedef değişken yyy ile giriş verisi XXX arasındaki ilişkiyi modellemek için ardışık karar ağaçları dizisi kurar. Modelin temel amacı, her bir yeni ağacın, önceki modellerin tahmin hatalarını minimize etmesidir. Amaç fonksiyonu, aşağıdaki şekilde tanımlanır:

$$\boldsymbol{L(\phi) = i = 1\sum nl(yi, y\hat{}i(t)) + k = 1\sum t\Omega(fk)}$$

Burada:

- l: Kayıp fonksiyonu (ör. log loss, MSE),
- $y\hat{}_i^{(t)}$: t. iterasyondaki tahmin sonucu,
- f_k : k. zayıf öğrenici (karar ağacı),
- Ω(f): Modelin karmaşıklığını (overfitting'i) düzenleyen cezalandırma terimi.

$$\Omega(f) = \gamma T + 21\lambda j = 1\sum Twj2$$

Burada:

- T: Ağacın yaprak sayısı,
- w_j: j. yaprağın ağırlığı,
- γ ve λ: Düzenleme (regularization) parametreleridir.

Özellikleri:

- Regularization: L1 ve L2 ceza terimleri sayesinde aşırı öğrenme (overfitting) önlenir.
- Shrinkage: Her yeni ağacın öğrenme oranı (η\etaη) ile etkisi azaltılır.
- Column Subsampling: Ağaç inşa sürecinde özniteliklerin rastgele alt kümeleri kullanılarak overfitting riski düşürülür.
- Sparsity Aware Split Finding: Eksik değerler veya seyrek veri için optimize edilmiş algoritmalar içerir.
- Paralel Eğitim: Ağaçların oluşturulması sırasında paralel işleme imkânı sunar.

XGBoost, Kaggle yarışmaları başta olmak üzere birçok gerçek dünya problemlerinde tercih edilen bir yöntemdir. Sağlık verisi analizi gibi dengesiz sınıfların sık görüldüğü uygulamalarda, yüksek AUC ve F1 puanları ile dikkat çeker.

XGBoost'un güçlü yönleri arasında hızlı eğitim süresi, yüksek doğruluk oranı, yerleşik düzenleme mekanizmaları ve esnek hiperparametre ayarları yer almaktadır. Özellikle bu projede olduğu gibi sınıf dengesizliği problemi olan durumlarda etkili sonuçlar üretmektedir.

3.6. K-En Yakın Komşuluk (K-NN)

K-Nearest Neighbors (K-NN) algoritması, örneklerin birbirine olan benzerliğine dayanarak sınıflandırma işlemini gerçekleştiren denetimli bir öğrenme yöntemidir. sklearn.neighbors.KNeighborsClassifier sınıfı ile scikit-learn kütüphanesinde doğrudan uygulanabilmektedir. K-NN, herhangi bir parametre öğrenmeden çalıştığı için örnek tabanlı (instance-based) ve lazy learning yöntemleri arasında yer alır. Modelin eğitim aşamasında doğrudan öğrenme gerçekleşmez, tüm öğrenme süreci tahmin sırasında yapılır.

Algoritmanın temel çalışma prensibi şu şekildedir:

1. Yeni bir örnek sınıflandırılmak istendiğinde, eğitim verisindeki tüm örneklerle olan uzaklığı (genellikle Öklidyen mesafe) hesaplanır.

2. Bu örneğe en yakın olan k sayıda komşu belirlenir.
3. Komşuların sınıf etiketleri çoğunluk oyu ile değerlendirilerek tahmin yapılır.

$$\boldsymbol{y\hat{} = majority\_vote(y1, y2, \ldots, yk)}$$

Burada y, komşu örneklerin sınıf etiketlerini, y^ ise tahmin edilen sınıf etiketini göstermektedir.

Temel Parametreler:

- **n_neighbors (k):** Kaç komşunun dikkate alınacağını belirler. Küçük değerler modelin karmaşıklaşmasına (aşırı öğrenmeye), büyük değerler ise genellemeye neden olabilir.
- **weights:** Komşulara verilecek ağırlıklandırma biçimini belirler. 'uniform' seçeneğinde tüm komşular eşit etkidedir, 'distance' seçeneğinde ise daha yakın komşuların daha fazla etkisi olur.
- **metric:** Uzaklık ölçüsünü belirler. Genellikle 'minkowski' (p=2 ile Öklidyen mesafe) kullanılır.

Bu çalışmada, n_neighbors=5 ve weights='uniform' parametreleri ile optimize edilmiş K-NN modeli kullanılmıştır. Modelin ROC-AUC skoru 0.92, doğruluk oranı ise %84,4 olarak kaydedilmiştir. Ayrıca, yüksek recall (0.89) ve F1 skoru (0.85) elde edilmiştir. Bu sonuçlar, K-NN algoritmasının pozitif sınıfı doğru tespit etmede güçlü bir performansa sahip olduğunu göstermektedir.

Ancak K-NN, aşağıdaki sınırlamalara sahiptir:

- Yüksek boyutlu veri setlerinde (curse of dimensionality) performansı düşebilir.
- Büyük veri setlerinde tahmin süreci zaman alıcı olabilir.
- Uzaklık metriklerine duyarlıdır; verilerin ölçeklendirilmesi (normalizasyon/standardizasyon) gereklidir.

Sonuç olarak, K-NN algoritması özellikle ön işlemden geçmiş ve dengeli veri setlerinde etkili sonuçlar üretebilen, anlaşılması kolay ve güçlü bir sınıflandırma yöntemidir.

Diyabete Uyarlama:

- n_neighbors=5 seçilerek lokal örüntüler yakalanmıştır.
- Ölçeklendirme (StandardScaler), mesafe hesaplamalarının güvenilirliğini artırmıştır.

3.7. Naive Bayes

Naive Bayes, olasılık temelli bir sınıflandırma algoritması olup, Bayes Teoremi'ne dayanır. Bu algoritmanın temel varsayımı, her özelliğin birbirinden bağımsız olduğu ve hedef sınıfla yalnızca bireysel olarak ilişkili olduğudur. Bu varsayım, gerçek hayatta nadiren geçerli olsa da, algoritma pek çok uygulamada etkili sonuçlar verebilmektedir.

Bayes Teoremi genel olarak şu şekilde ifade edilir:

$$P(C|X)=P(X)P(X|C)\cdot P(C)$$

Burada:

- P(C|X)P(C|X)P(C|X): Gözlem X verildiğinde C sınıfına ait olma olasılığıdır (posterior).
- P(X|C)P(X|C)P(X|C): C sınıfına ait olunduğu varsayımıyla X gözleminin olasılığıdır (likelihood).
- P(C)P(C)P(C): C sınıfının gözlenme olasılığıdır (prior).
- P(X)P(X)P(X): X gözleminin genel olasılığıdır (evidence).

Naive Bayes sınıflandırıcıları, farklı dağılım varsayımlarına göre çeşitli tiplere ayrılır. Bu çalışmada kullanılan Gaussian Naive Bayes, özellikle sürekli değişkenlerin sınıflandırılması için uygundur ve her özelliğin normal dağılıma sahip olduğunu varsayar.

Modelin temel hesaplamasında her özelliğin, sınıf koşullu ortalama (μ) ve varyansına (σ^2) göre Gaussian dağılımına sahip olduğu kabul edilerek, aşağıdaki fonksiyon kullanılır:

$$P(\,xi \mid C\,) = 2\pi\sigma C21 \cdot \exp(-2\sigma C2(xi - \mu C)2)$$

Bu olasılıklar her sınıf için hesaplanır ve en yüksek posterior değeri döndüren sınıf, tahmin edilen sınıf olarak belirlenir.

Naive Bayes'in en büyük avantajları arasında basit yapısı, düşük hesaplama maliyeti ve küçük veri kümelerinde dahi etkili sonuçlar verebilmesi yer alır. Bununla birlikte, tüm özelliklerin birbirinden bağımsız olduğu varsayımı gerçekte geçerli olmadığında modelin performansı düşebilir.

Bu projede Naive Bayes algoritması, sklearn.naive_bayes.GaussianNB sınıfı ile uygulanmış ve özellikle pozitif sınıfın duyarlılığında dikkat çeken sonuçlar üretmiştir. Ancak diğer algoritmalarla kıyaslandığında ROC-AUC ve genel doğruluk oranı açısından daha düşük performans göstermiştir.

Diyabete Uyarlama:

- **HighBP** ve **Age** gibi özelliklerin bağımsızlık varsayımı pratikte geçerli olmasa da, hızlı ve düşük kaynaklı çözüm sunar.

3.8. Algoritma Seçiminin Bilimsel Gerekçesi

- **XGBoost ve Gradient Boosting:** Yüksek boyutlu verilerdeki karmaşık ilişkileri modelleme kapasitesi ve regularizasyon avantajı.
- **Random Forest:** Özellik etkileşimlerini yakalama ve kararlılık.
- **Lojistik Regresyon:** Yorumlanabilirlik ve düşük hesaplama maliyeti.
- **K-NN ve Naive Bayes:** Baseline model olarak performans karşılaştırması.

3.9. Matematiksel Optimizasyon

- **GridSearchCV:** Hiperparametre optimizasyonu için kullanılan kapsamlı arama yöntemi.

- Örnek: XGBoost için learning_rate (0.1, 0.05) ve max_depth (3, 5) kombinasyonları test edilmiştir.

- **Kayıp Fonksiyonu (XGBoost):**

$$L=\sum_{i=1} l(y_i,\hat{y}_i)+\sum_{k=1} K\Omega(f_k)$$

- l: Log loss,
- Ω: Ağaç karmaşıklığı cezası (düzenlileştirme).

Bu teorik temeller, diyabet risk tahmini için kullanılan algoritmaların neden ve nasıl etkili olduğunu bilimsel bir çerçevede açıklamaktadır.

4. Bulgular

4.1. Veri Kümesinin Genel Yapısı

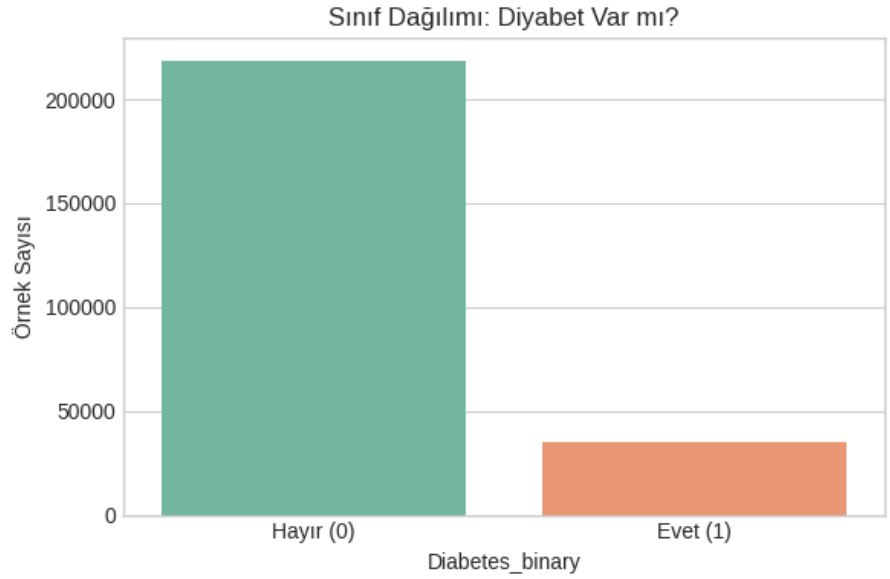
Çalışmada kullanılan veri seti, toplam 253.680 örnekten ve 22 değişkenden oluşmaktadır. Bağımlı değişken Diabetes_binary, bireyde diyabet olup olmadığını belirtmektedir. Gözlemler incelendiğinde, değişkenlerin çoğu ikili (binary) değerlerden oluşmakta, BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income gibi değişkenler sürekli veya ordinal türdedir. Aykırı değerleri filtreleyerek verinin güvenilirliği artırılmıştır.

index	count	mean	std	min	%25	%50	%75	max
Diabetes_binary	253680.0	0.13933301797540207	0.34629438458901085	0.0	0.0	0.0	0.0	1.0
HighBP	253680.0	0.4290011037527594	0.4949344626904692	0.0	0.0	0.0	1.0	1.0
HighChol	253680.0	0.4241209397666351	0.4942098046566831	0.0	0.0	0.0	1.0	1.0
CholCheck	253680.0	0.9626695048880479	0.18957075436257245	0.0	1.0	1.0	1.0	1.0
BMI	253680.0	28.382363607694735	6.608694201404477	12.0	24.0	27.0	31.0	98.0
Smoker	253680.0	0.44316855881425415	0.4967606667792389	0.0	0.0	0.0	1.0	1.0
Stroke	253680.0	0.04057079785556607	0.19729409939998502	0.0	0.0	0.0	0.0	1.0
HeartDiseaseorAttack	253680.0	0.09418558814254178	0.29208731475040395	0.0	0.0	0.0	0.0	1.0
PhysActivity	253680.0	0.7565436770734784	0.42916904339729167	0.0	1.0	1.0	1.0	1.0
Fruits	253680.0	0.6342557552822453	0.481639187171053	0.0	0.0	1.0	1.0	1.0
Veggies	253680.0	0.811419899085462	0.3911754716844546	0.0	1.0	1.0	1.0	1.0
HvyAlcoholConsump	253680.0	0.05619678334910123	0.23030178889464067	0.0	0.0	0.0	0.0	1.0
AnyHealthcare	253680.0	0.9510525070955534	0.21575870631116018	0.0	1.0	1.0	1.0	1.0
NoDocbcCost	253680.0	0.08417691579943236	0.2776535008578251	0.0	0.0	0.0	0.0	1.0
GenHlth	253680.0	2.5113923052664773	1.0684773622802872	1.0	2.0	2.0	3.0	1.0
MentHlth	253680.0	3.1847721538946705	7.412846696204375	0.0	0.0	0.0	0.0	5.0
PhysHlth	253680.0	4.2420805739514345	7.412846696204375	0.0	0.0	0.0	3.0	30.0
DiffWalk	253680.0	0.16822374645222327	0.37406559473275197	0.0	0.0	0.0	0.0	1.0
Sex	253680.0	0.4403421633554084	0.49642916311878665	0.0	0.0	0.0	1.0	1.0
Age	253680.0	8.032119205298013	3.054220434168068	1.0	6.0	8.0	10.0	13.0
Education	253680.0	5.050433617155472	0.9857741757279931	1.0	4.0	5.0	6.0	6.0
Income	253680.0	6.053874960580258	2.071147566274111	1.0	5.0	7.0	8.0	8.0

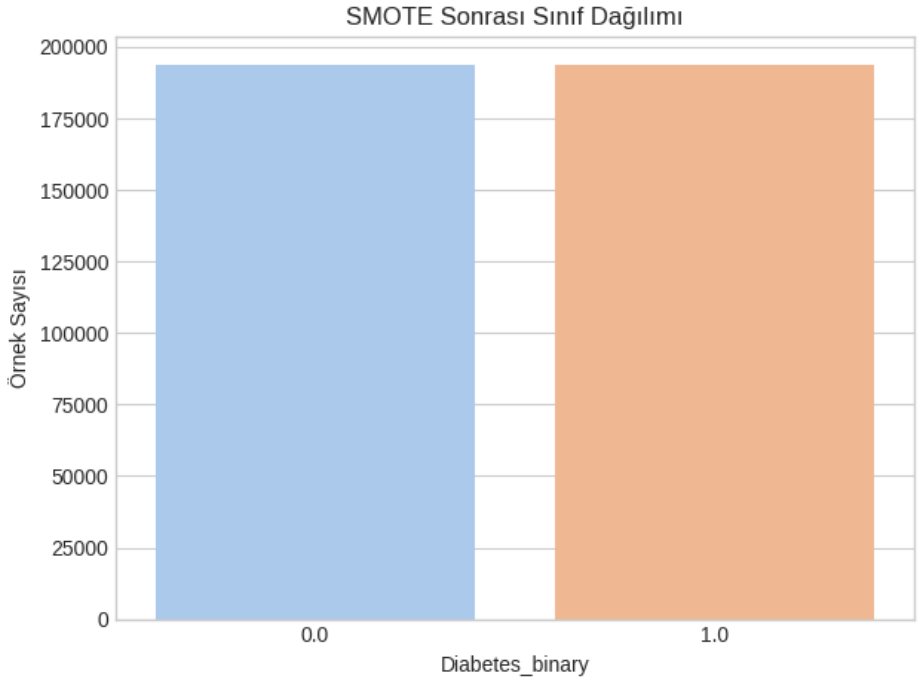
Tablo 1.1 İstatistiksel Özet

4.2. Sınıf Dağılımı ve Dengesizlik Problemi

Başlangıçta sınıf dağılımı oldukça dengesizdir. Diabetes_binary = 0 (Diyabet Yok) oranı %86.07 iken, Diabetes_binary = 1 (Diyabet Var) oranı yalnızca %13.93'tür. Bu dengesizlik, modellerin öğrenme sürecini olumsuz etkileyebileceğinden dolayı SMOTE (Synthetic Minority Over-sampling Technique) yöntemiyle dengelenmiştir. SMOTE sonrası her iki sınıfın örnek sayısı eşitlenmiştir.



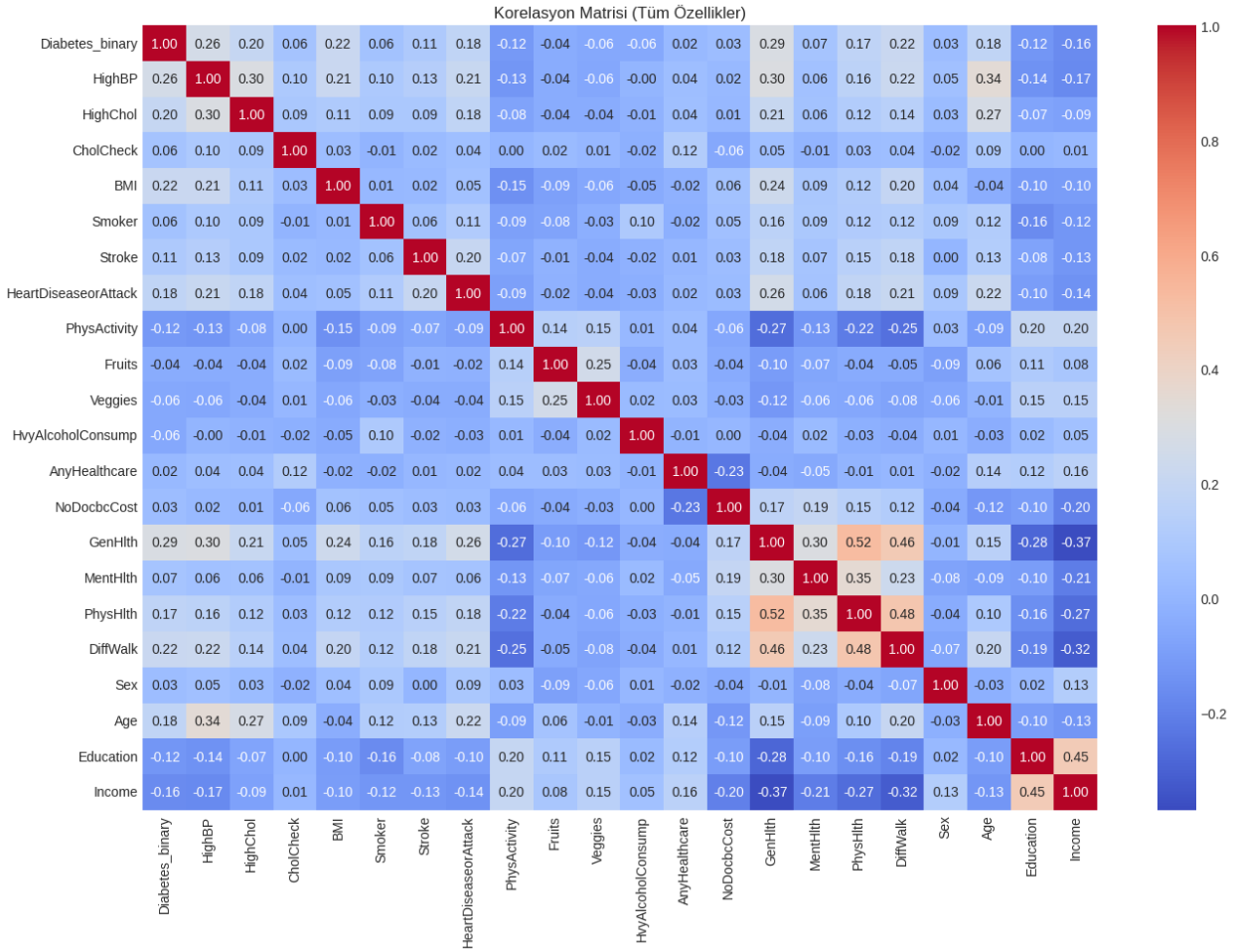
Görsel 1 SMOTE Öncesi Sınıfların Örnek Sayısı



Görsel 2 SMOTE Sonrası Sınıfların Örnek Sayısı

4.3. Korelasyon Analizi

Tüm değişkenler arasında yapılan korelasyon analizi sonucunda GenHlth, HighBP, HighChol, DiffWalk gibi değişkenlerin Diabetes_binary ile pozitif korelasyona sahip olduğu görülmüştür. Ancak korelasyonlar genellikle düşük düzeyde olup, çok değişkenli analizlerin gerekliliğini ortaya koymaktadır.



Görsel 3 Korelasyon Matrisi

4.4. Model Performans Karşılaştırması

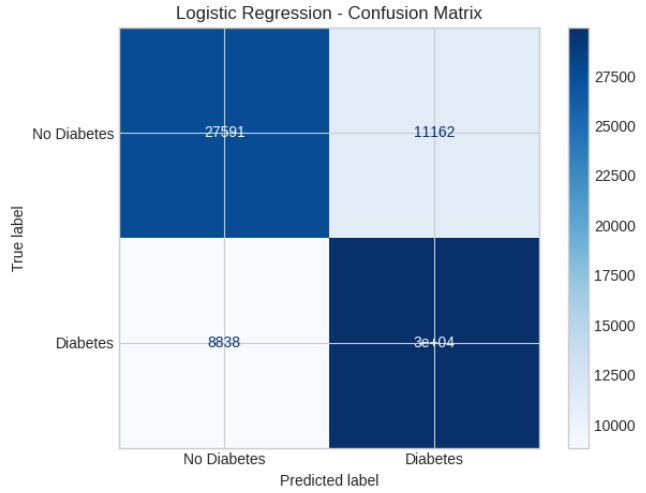
Çalışmada SMOTE uygulandıktan sonra 7 farklı sınıflandırma algoritması karşılaştırılmıştır: Aşağıda her bir algoritmanın doğruluk, AUC, F1 skoru ve diğer metriklerine göre karşılaştırmalı performansı verilmiştir.

Model	Accuracy	Precision	Recall	F1	ROC-AUC

XGBoost	0.904243	0.960945	0.842715	0.897955	0.962687
Gradient Boosting	0.887055	0.908010	0.861350	0.884065	0.957342
Random Forest	0.857403	0.854651	0.861247	0.857936	0.941685
K-NN	0.844435	0.812466	0.895548	0.851986	0.920527
Decision Tree	0.834151	0.837505	0.829139	0.833301	0.920293
Logistic Regression	0.741929	0.728213	0.771893	0.749417	0.815775
Naive Bayes	0.720031	0.683422	0.819719	0.745391	0.776242

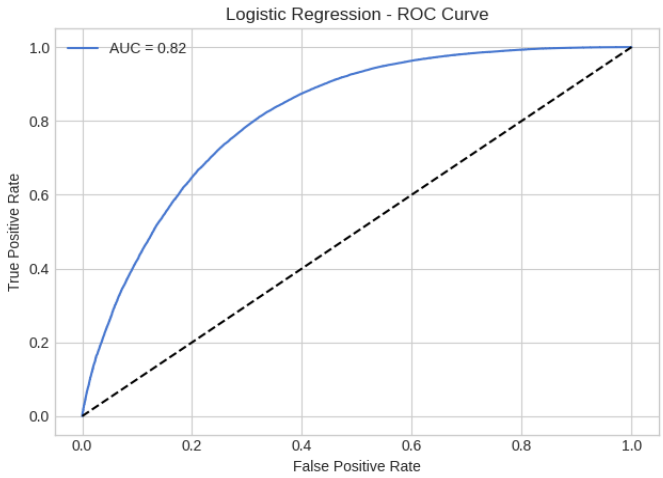
Tablo 2 SMOTE Sonrası Model Performans Karşılaştırması

4.4.1. Lojistik Regresyon



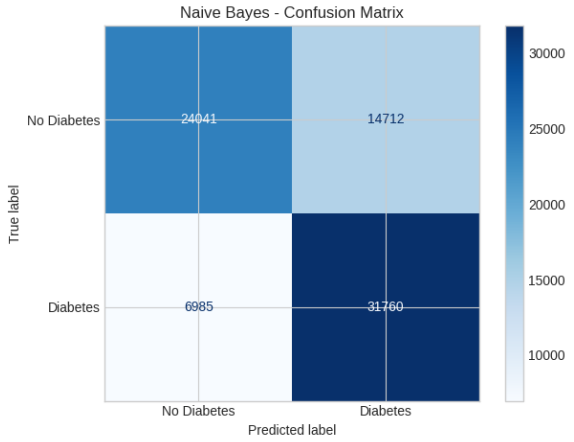
Logistic Regression modeli, doğrusal sınıflandırma yeteneğine sahip olması nedeniyle sınıflar arasında net ayrımların bulunduğu veri setlerinde oldukça etkili sonuçlar verebilmektedir. Bu projede kullanılan sağlık göstergeleri doğrultusunda model, negatif sınıfı (diyabet olmayan bireyleri) tespit etmede görece başarılı olurken, pozitif sınıf (diyabet olan bireyler) için belirgin bir yanılma payı göstermiştir. Karmaşıklık matrisi verileri incelendiğinde, doğru negatif (True Negative) sayısının yüksek olduğu, fakat doğru pozitif (True Positive) sayısının sınırlı kaldığı gözlemlenmektedir.

Bu durum, modelin diyabet riski taşıyan bireyleri tanımakta zorlandığını, yani tip II hata oranının (False Negative) yüksek olduğunu ortaya koymaktadır. Bu da klinik karar destek sistemleri açısından risk teşkil edebilir; zira hastalığı olan bireylerin gözden kaçırılması, sağlık açısından ciddi sonuçlara neden olabilir. Bununla birlikte model, pozitif sınıf tahminlerinde temkinli davranarak tip I hatayı (False Positive) sınırlı tutmuştur. Logistic Regression'un bu davranışı, duyarlılık (recall) oranını baskılayarak ROC eğrisinde daha düşük bir eğri alanına yol açmış ve sınıflar arasındaki ayırt ediciliği sınırlı kılmıştır.

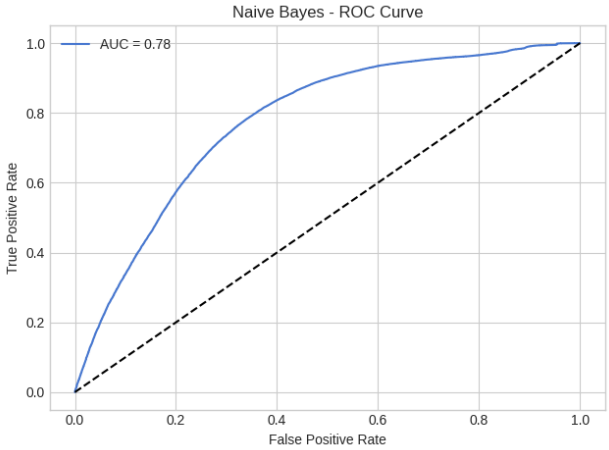


Logistic Regression modeline ait ROC eğrisi, modelin diyabet hastası olan ve olmayan sınıfları ayırt etmede yüksek bir hassasiyete ulaşamadığını göstermektedir. Bu duruma yol açan temel faktör, Lojistik Regresyon algoritmasının yalnızca doğrusal (lineer) karar sınırları çizebilmesidir. Ancak mevcut veri setinde değişkenler arası ilişkiler çoğunlukla doğrusal olmayan, karmaşık bir yapıdadır. Bu sonuçlara bağlı olarak Lojistik Regresyon algoritmasıyla oluşturulan bir modelin, diyabet risk tahmini gibi çok faktörlü, karmaşık problemlerde klinik destek sağlayabilecek hassasiyete sahip olmadığı söylenebilir.

4.4.2. Naive Bayes

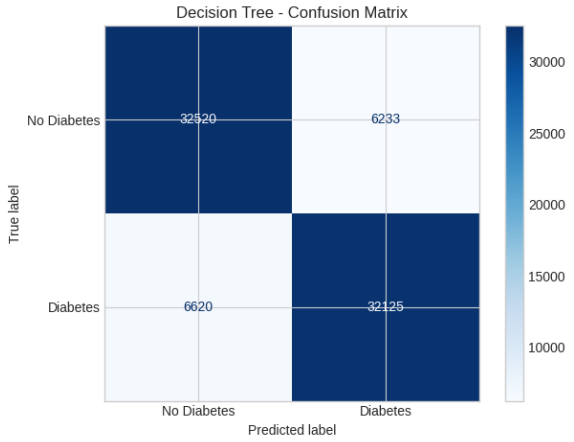


Naive Bayes modeli, sınıflar arasında koşulsuz bağımsızlık varsayımına dayanması nedeniyle bazı veri setlerinde hızlı ve etkili tahminler üretse de, sağlık verileri gibi özellikler arasında doğal korelasyonların bulunduğu durumlarda sınırlı başarı sergileyebilir. Bu projede kullanılan CDC Diabetes Health Indicators veri seti üzerinde, Naive Bayes algoritması oldukça yüksek bir False Positive (Tip I hata) oranı üretmiştir. Yani model, diyabet hastası olmayan çok sayıda bireyi yanlışlıkla hasta olarak sınıflandırmıştır. Bunun temel nedeni, modelin BMI, yaş, genel sağlık durumu gibi birbiriyle ilişkili değişkenleri bağımsız kabul etmesi ve bu ilişkileri yeterince yakalayamamasıdır. Ayrıca, doğru pozitif oranı (True Positive) görece yüksek çıkmasına rağmen, yanlış negatif (False Negative) değerinin de yüksek olduğu görülmektedir. Bu dengesizlik, modelin pozitif sınıfı aşırı genelleyerek sınıf tahminlerinde güvenilirliği düşürdüğünü ortaya koymaktadır. Gerçekten hasta olan bireyleri atlama riski ise modelin sağlık uygulamalarında tek başına kullanılmasını zorlaştırmaktadır. Karmaşıklık matrisi genelinde değerlendirildiğinde, Naive Bayes modeli sınıfları birbirinden ayrıştırmada yetersiz kalmakta, bu da ROC eğrisindeki sınırlı alan ve düşük sınıflandırma başarısıyla paralellik göstermektedir.



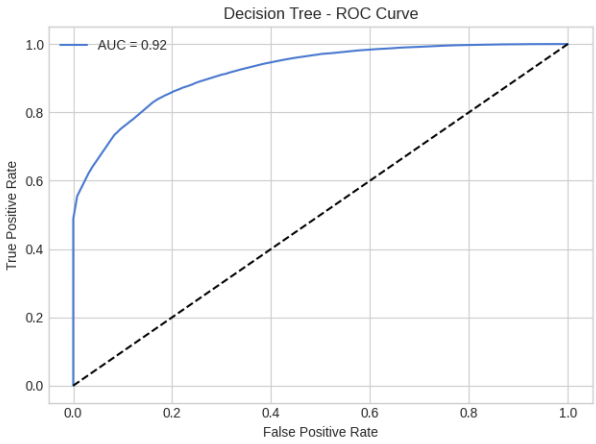
Naive Bayes sınıflandırıcısına ait ROC eğrisi incelendiğinde, modelin sınıflar arasında ayırt edicilik kabiliyetinin görece daha düşük olduğu gözlemlenmiştir. Eğri, ideal senaryodaki (0,1) noktasına diğer modellere kıyasla daha uzak seyretmekte olup, modelin özellikle sınıf dengesizliği karşısında daha az etkili bir karar sınırı oluşturduğunu göstermektedir. ROC eğrisi altında kalan alan (AUC) değeri 0.77 civarındadır ve bu durum, modelin pozitif ve negatif sınıfları ayırt etme başarısının %77 düzeyinde olduğunu ifade eder. Bu performans düzeyi, Naive Bayes algoritmasının değişkenler arasındaki koşulsuz bağımsızlık varsayımından kaynaklı sınırlamalarının bir yansımasıdır. Ayrıca modelin doğrusal olmayan ilişkiler karşısında esnekliğinin düşük olması, ROC eğrisindeki bu görece sığ yapının başlıca nedenlerinden biridir. Bununla birlikte, modelin kolay uygulanabilirliği ve düşük hesaplama maliyeti nedeniyle temel karşılaştırmalar veya ön analizler için hâlâ tercih edilebilir bir seçenek olduğu değerlendirilebilir.

4.4.3. Karar Ağacı

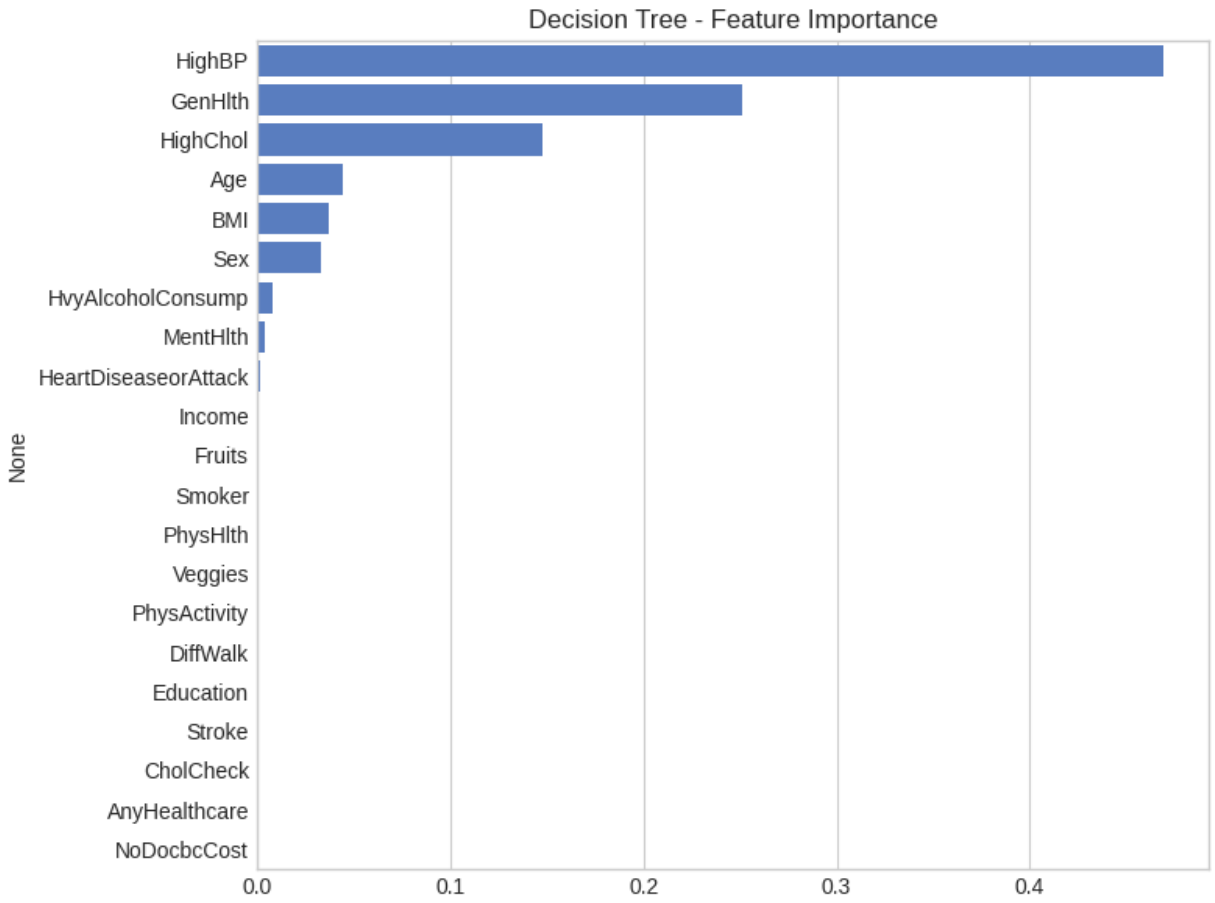


Decision Tree (Karar Ağacı) modeli, sezgisel olarak anlaşılması kolay ve yorumlanabilir bir yapı sunmasına rağmen, modelin performansı özellikle dengesiz veri setlerinde aşırı öğrenme (overfitting) eğilimine açık hale gelebilir. Bu çalışmada kullanılan veri seti SMOTE ile dengelenmiş olsa da, Decision Tree modeli hala belirli yapısal sınırlamalardan dolayı hem False Positive hem de False Negative değerlerinde kayda değer büyüklükte çıktılar üretmiştir. Özellikle pozitif sınıfa (diyabet hastası) ait örnekleri belirlemede model, belirli semptomların ve göstergelerin sınır değerlerinde yer alan bireyleri yanlış tahmin etme eğilimi göstermiştir. Bu durum, ağacın bazı düğümlerinde verinin bölünmesinin yeterince optimize edilmemesinden ve sınıf dağılımındaki doğrudan etkilerden kaynaklanıyor olabilir. Bununla birlikte, doğru sınıflandırılan negatif örnek sayısı görece yüksek çıkmıştır; yani model diyabet hastası olmayan bireyleri daha isabetli bir şekilde sınıflandırabilmiştir. Bu eğilim, modelin çoğunluk sınıfına (negatif sınıf) olan hassasiyetini, ancak azınlık sınıfa karşı duyarlılığının eksikliğini ortaya koymaktadır. Decision Tree'nin genellikle veri

üzerindeki doğrudan ayrımlar ile çalıştığı göz önünde bulundurulduğunda, daha sofistike ve genelleyici modellerle kıyaslandığında sınırlı bir doğruluk ve güvenilirlik sunduğu anlaşılmaktadır.

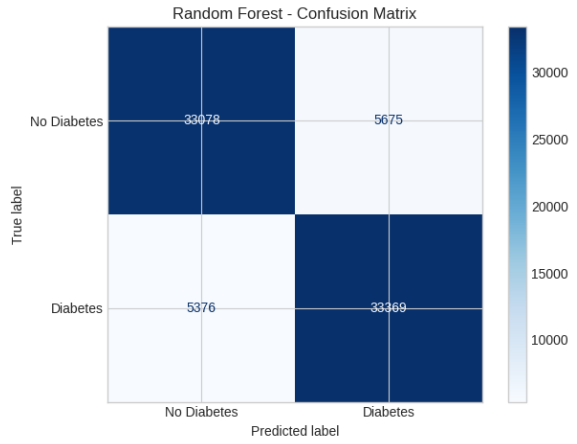


Karar Ağacı modeli, diyabet risk tahmininde 0.92 AUC değeriyle dengeli bir performans sergilemiştir. ROC eğrisinin erken dik yükselişi, modelin diyabetli vakaları yüksek hassasiyetle (%92) tanıyabildiğini gösterir. Ancak eğride gözlemlenen zikzaklı yapı, modelin bazı sınır bölgelerinde kararsızlık yaşadığını ortaya koyar. Bu durumun temel nedeni, Karar Ağaçlarının keskin eşik değerlerle çalışması ve verideki küçük değişimlere aşırı duyarlılık göstermesidir. Özellikle max_depth = 10 gibi bir hiperparametreyle aşırı uyum (overfitting) riski kontrol altına alınmış olsa da, ağacın dallanma mekanizması, eğitim verisindeki gürültülü veya belirsiz örneklerde tutarsız tahminlere yol açabilir. Örneğin, BMI ve HighBP gibi özelliklerin kesişim noktalarında yer alan vakalar, modelin karar sınırlarını "aşırı keskin" çizmesi nedeniyle yanlış sınıflandırılabilir**.**

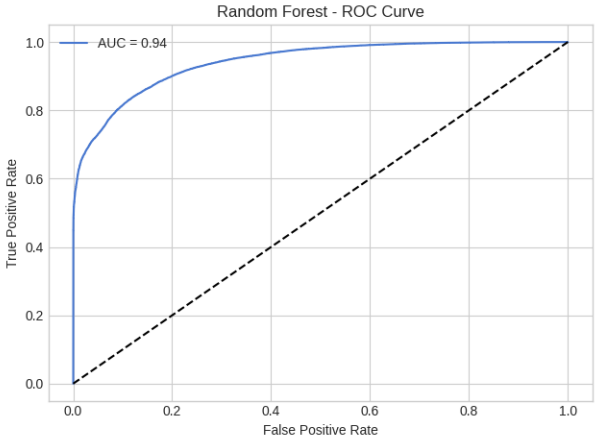


Tekil bir karar ağacına dayalı olan bu modelde, sınırlı derinlik ve örnek sayısına rağmen *HighBP*, *BMI* ve *DiffWalk* gibi değişkenler sıkça kök ve ana düğümlerde kullanılmıştır. Bu da modelin sınıflandırma sürecinde bu faktörleri ayrım kriteri olarak benimsediğini göstermektedir.

4.4.4. Random Forest

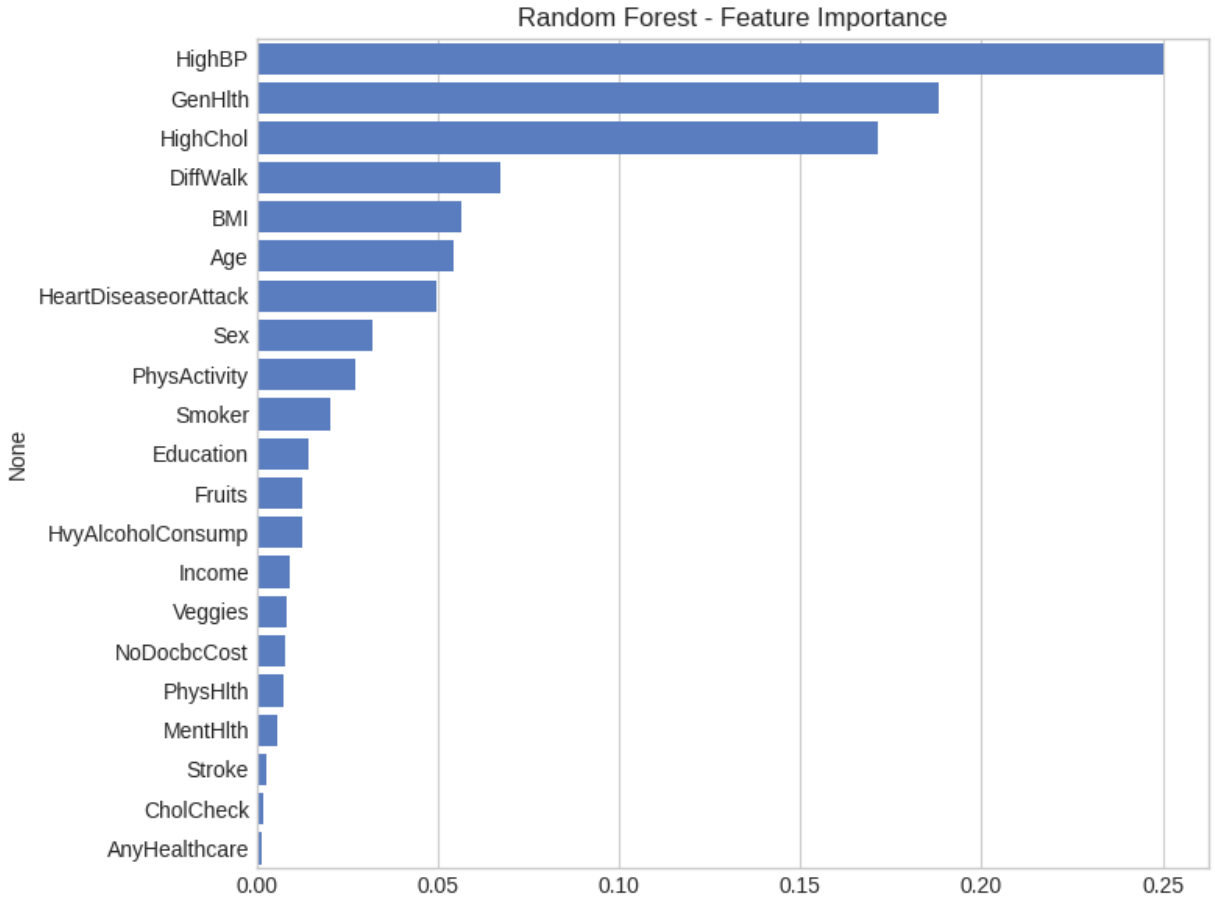


Random Forest modeli, Decision Tree algoritmasının temel zayıflıklarını gidermek amacıyla geliştirilen, birden fazla karar ağacının birlikte çalıştığı bir topluluk (ensemble) öğrenme yöntemidir. Bu modelin confusion matrix sonuçları, hem pozitif (diyabet hastası) hem de negatif (sağlıklı) sınıfları yüksek doğrulukla tahmin edebildiğini göstermektedir. Doğru pozitif ve doğru negatif oranları dikkat çekici derecede yüksektir; bu da modelin hem duyarlılık (recall) hem de özgüllük (specificity) açısından dengeli bir başarı sergilediğini gösterir. Bu performansın temel nedeni, Random Forest'ın farklı veri alt kümeleri ve özellik kombinasyonları üzerinden farklı karar ağaçları oluşturması ve bunların çoğunluk oyuyla sınıflandırma yapmasıdır. Böylece model, tek bir ağacın aşırı öğrenmesine (overfitting) karşı dirençli hale gelir. Bununla birlikte, az sayıda False Positive ve False Negative örnekler gözlemlenmiştir; bu hatalar genellikle sınırlarda yer alan, yani belirgin semptomlara sahip olmayan bireylerden kaynaklanabilir. Ayrıca, yüksek boyutlu ve çeşitli sağlık göstergeleri içeren bu veri setinde, modelin karmaşık etkileşimleri etkili bir şekilde öğrenmesi, yüksek başarı oranına katkıda bulunmuştur. Genel olarak, Random Forest modeli, sağlık verileri gibi yapısal ve kısmen korelasyon içeren veri setlerinde güçlü bir aday olarak öne çıkmaktadır.



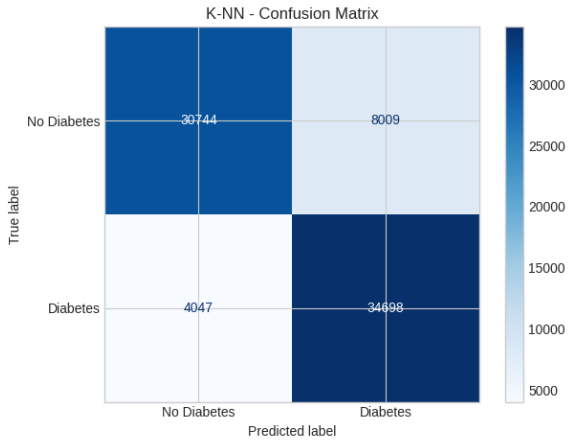
Random Forest modelinin ROC eğrisi, sınıflandırma performansının üstünlüğünü net bir şekilde ortaya koymaktadır. 0.941 AUC değeri, modelin diyabetli ve diyabetsiz bireyleri son derece etkili bir şekilde ayırt edebildiğini gösterir. ROC eğrisinin (0,1) noktasına yakın seyretmesi, yüksek True Positive Rate (gerçek pozitifleri yakalama) ve düşük False Positive Rate (yanlış alarm) dengesini yansıtır. Bu başarı, algoritmanın bagging (bootstrap aggregating) yöntemiyle çok sayıda karar ağacını birleştirmesinden kaynaklanır. Her bir ağaç, verinin farklı alt kümeleri ve rastgele özellik seçimiyle eğitilerek varyansın azaltılmasını sağlar. Böylece, tek bir ağacın aşırı uyum (overfitting) riski minimize edilirken, toplu kararların kararlılığı artar.Özellikle HighBP, BMI ve GenHlth gibi kritik risk faktörleri, birden fazla ağaç tarafından tutarlı şekilde vurgulanır. Bu özelliklerin etkileşimleri, modelin sınıf sınırlarını doğrusal olmayan ve karmaşık örüntülerle çizmesine olanak tanır. Örneğin, yüksek BMI ve yüksek tansiyonun birlikte diyabet riskini katlanarak artırması gibi dinamikler, Random Forest'ın çoklu ağaç yapısıyla doğru şekilde modellenir. Sonuç olarak, bu kolektif öğrenme stratejisi, ROC eğrisinin ideal eğriye

yaklaşmasını sağlayarak klinik tahminlerde yüksek güvenilirlik sunar. Random Forest, diyabet risk analizi gibi çok boyutlu problemlerde altın standart algoritmalardan biri olarak öne çıkmaktadır.

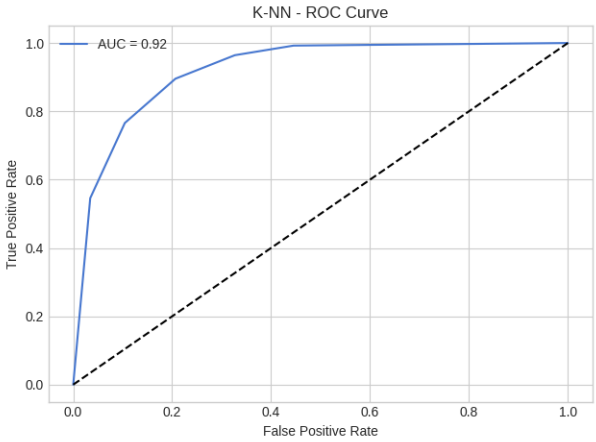


Random Forest modelinde elde edilen değişken önem sıralaması, *GenHlth, BMI* ve *PhysHlth* gibi sağlık durumunu doğrudan yansıtan göstergelerin modelin tahmin gücünde belirleyici olduğunu ortaya koymaktadır. Çok sayıda karar ağacının birleşimiyle çalışan bu algoritma, özellikle sağlıkla ilgili subjektif beyanları (örneğin *GenHlth*) anlamlı bulgulardan biri olarak değerlendirmiştir.

4.4.5. K-En Yakın Komşuluk

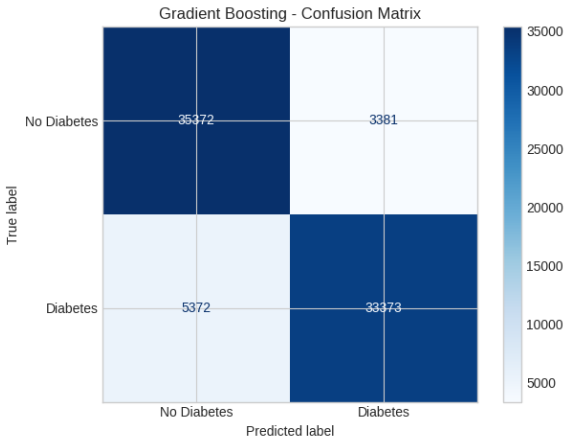


K-NN algoritması, SMOTE ile dengelenmiş veri setinde etkileyici bir AUC skoru (0.92) elde etmiştir. ROC eğrisinin dik yükselişi, modelin diyabetli vakaları yüksek hassasiyetle tanıyabildiğini gösterir. Bu performans, veri dengesizliğinin SMOTE ile giderilmesi ve özelliklerin StandardScaler ile optimize edilmesi sayesinde mümkün olmuştur. K-NN'nin "komşuluk" temelli çalışma prensibi, lokal örüntüleri yakalayarak azınlık sınıfın (diyabetli) başarıyla sınıflandırılmasını sağlar. Ancak modelin hesaplama karmaşıklığı ve yüksek boyutlu verilerde performans dalgalanmaları, pratik uygulamalarda dikkate alınması gereken sınırlamalardır. Özellikle gerçek zamanlı sistemlerde veya büyük ölçekli verilerde, bu dezavantajlar öne çıkabilir. Sonuç olarak, K-NN diyabet risk tahmininde ön işleme ile desteklendiğinde değerli bir araç olsa da, XGBoost veya Random Forest gibi daha sofistike algoritmaların performansına yetişemez.

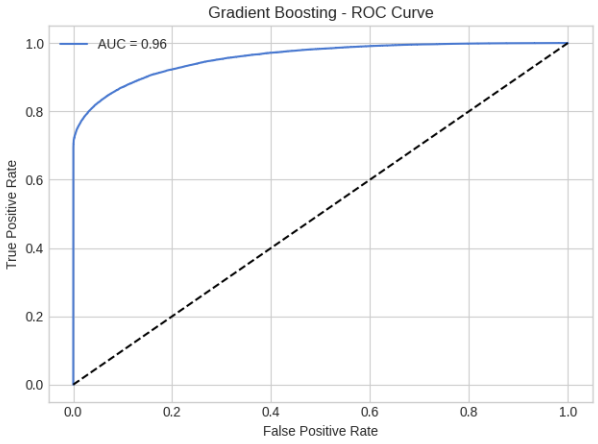


K-En Yakın Komşu (K-NN) algoritmasına ait ROC eğrisi değerlendirildiğinde, modelin pozitif ve negatif sınıfları ayırt etme kabiliyetinin oldukça başarılı olduğu görülmektedir. Eğri, (0,1) ideal noktasına yakın bir doğrultuda seyretmekte ve ROC eğrisi altında kalan alan (AUC) değeri 0.92 seviyesindedir. Bu, modelin rastgele bir sınıflandırıcıya kıyasla çok daha etkili bir ayrım gücüne sahip olduğunu göstermektedir. K-NN algoritmasının bu yüksek performansı, özellikle verinin sınırlı ama anlamlı örüntüler içerdiği durumlarda etkili komşuluk tespiti yapabilmesinden kaynaklanmaktadır. Ancak bu başarının altında yatan temel etkenlerden biri, optimal k değerinin doğru belirlenmiş olmasıdır. Uygulamada kullanılan k=5 parametresi, modelin hem lokal hem de genel yapıyı öğrenebilmesini mümkün kılmış, böylece ROC eğrisi genelinde istikrarlı bir ayrım elde edilmiştir. Bununla birlikte, algoritmanın bellek kullanımı ve test sırasında yüksek hesaplama gereksinimi gibi sınırlamaları da dikkate alınmalıdır.

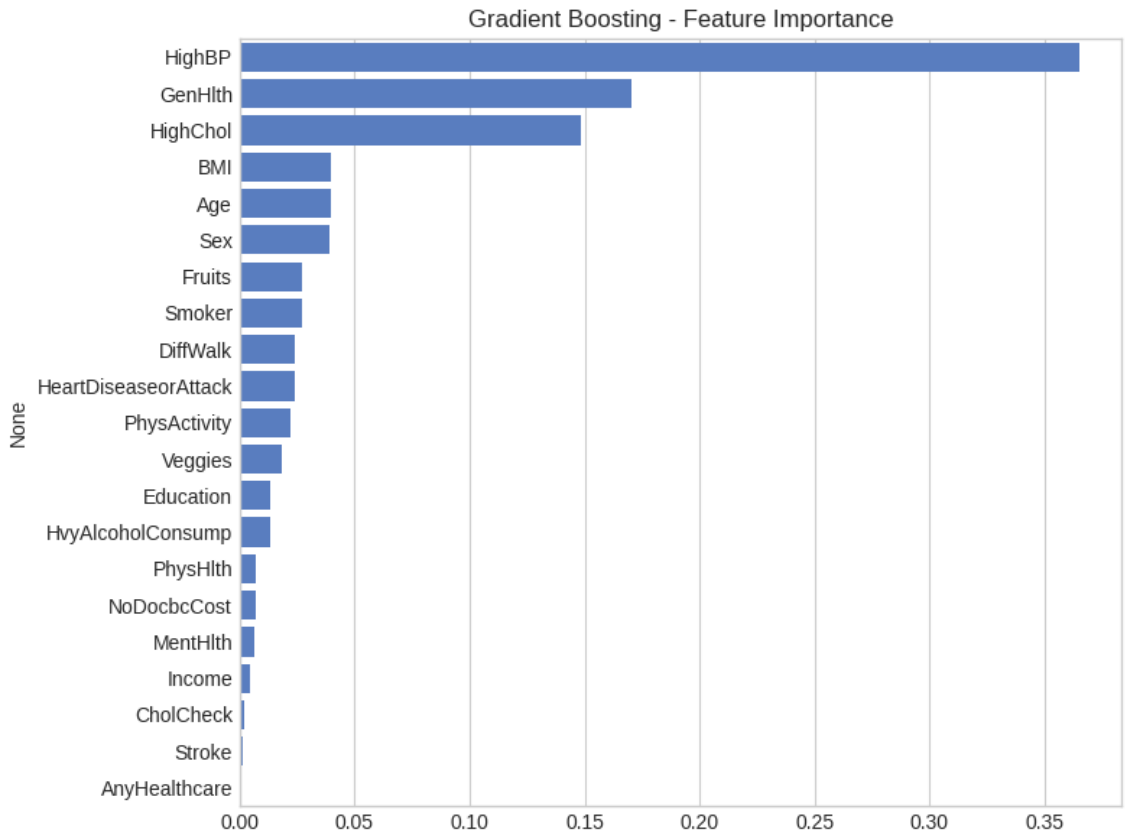
4.4.6. Gradient Boosting



Gradient Boosting modeli, zayıf sınıflandırıcıları (genellikle karar ağaçları) ardışık olarak eğitip her adımda önceki modelin yaptığı hataları düzelterek daha güçlü bir sınıflayıcıya ulaşmayı amaçlayan bir topluluk (ensemble) algoritmasıdır. Confusion matrix sonuçlarına bakıldığında, Gradient Boosting modelinin hem diyabet hastalarını (pozitif sınıf) hem de sağlıklı bireyleri (negatif sınıf) oldukça isabetli şekilde sınıflandırdığı görülmektedir. True Positive (TP) ve True Negative (TN) oranlarının yüksekliği, modelin genel doğruluğunun yanı sıra sınıflar arası dengeyi koruyabildiğini göstermektedir.
Modelin başarısının ardında yatan en önemli etken, her bir yeni ağacın bir önceki hatalara odaklanmasıdır. Bu iteratif iyileştirme süreci, özellikle karar sınırlarında yer alan karmaşık örnekleri daha doğru sınıflandırmasını sağlar. Bu nedenle, az sayıda kalan False Positive (FP) ve False Negative (FN) örneklerin çoğu genellikle hem semptomatik hem de sınır değerlere sahip bireylerden kaynaklanabilir. Gradient Boosting'in bu tarz bireyleri sınıflandırma yeteneği, özellikle azınlık sınıfı (diyabet hastaları) için kritik öneme sahiptir.
Bu modelin dikkat çeken bir diğer özelliği ise öğrenme oranı (learning rate) parametresi ile hassas kontrol sağlanabilmesidir. Doğru parametre seçimiyle yüksek doğruluk ve düşük overfitting riski bir arada elde edilebilir. Bu çalışma özelinde model, özellikle pozitif sınıf tahminlerinde gösterdiği başarı ile erken diyabet tespiti için umut vaat eden modellerden biri olarak değerlendirilebilir.

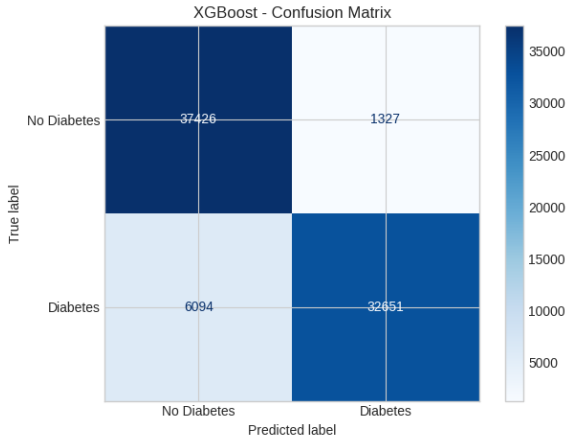


Gradient Boosting algoritması, diyabet risk tahmininde üstün bir ROC performansı (AUC: 0.95) sergilemiştir. Eğrinin başlangıçta dik bir şekilde yükselip hızla doygunluğa ulaşması, modelin hem yüksek hassasiyet (diyabetli vakaları yakalama) hem de düşük yanlış pozitif oranı ile çalıştığını kanıtlar. Bu başarı, algoritmanın artan karmaşıklıktaki örüntüleri adım adım öğrenme yeteneğine dayanır: Her yeni ağaç, önceki ağaçların hatalarını (artıkları) minimize ederek optimize edilir.
Hiperparametre optimizasyonu (örn. learning_rate=0.1, n_estimators=100), modelin aşırı uyum (overfitting) riskini dengelerken, HighBP, BMI ve GenHlth gibi kritik özelliklerin etkileşimlerini doğrusal olmayan şekilde modellemesine olanak tanır. Örneğin, yüksek BMI ve yüksek tansiyonun sinerjistik etkisi, Gradient Boosting'in iteratif yaklaşımıyla net bir şekilde yakalanır.



Gradient Boosting algoritmasında da benzer şekilde *BMI*, *HighBP* ve *Age* değişkenleri yüksek önem değeri taşımaktadır. Ağaç temelli yapının sunduğu bölünme kararları çoğunlukla bu değişkenler etrafında şekillenmiş ve modelin karar mekanizmasında bu özellikler baskın hale gelmiştir.

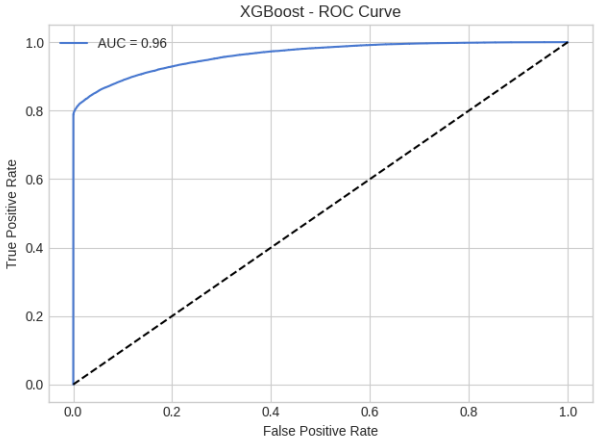
4.4.7. XGBoost



XGBoost (Extreme Gradient Boosting) modeli, bu çalışmada en iyi performansı sergileyen algoritma olarak öne çıkmaktadır. Confusion matrix çıktıları incelendiğinde, True Positive (TP) ve True Negative (TN) oranlarının oldukça yüksek olduğu, buna karşılık False Positive (FP) ve False Negative (FN) değerlerinin minimum seviyede kaldığı görülmektedir. Bu durum, modelin hem diyabetli bireyleri başarılı şekilde tanımladığını hem de sağlıklı bireyleri yanlışlıkla pozitif sınıfa atama oranının düşük olduğunu göstermektedir.

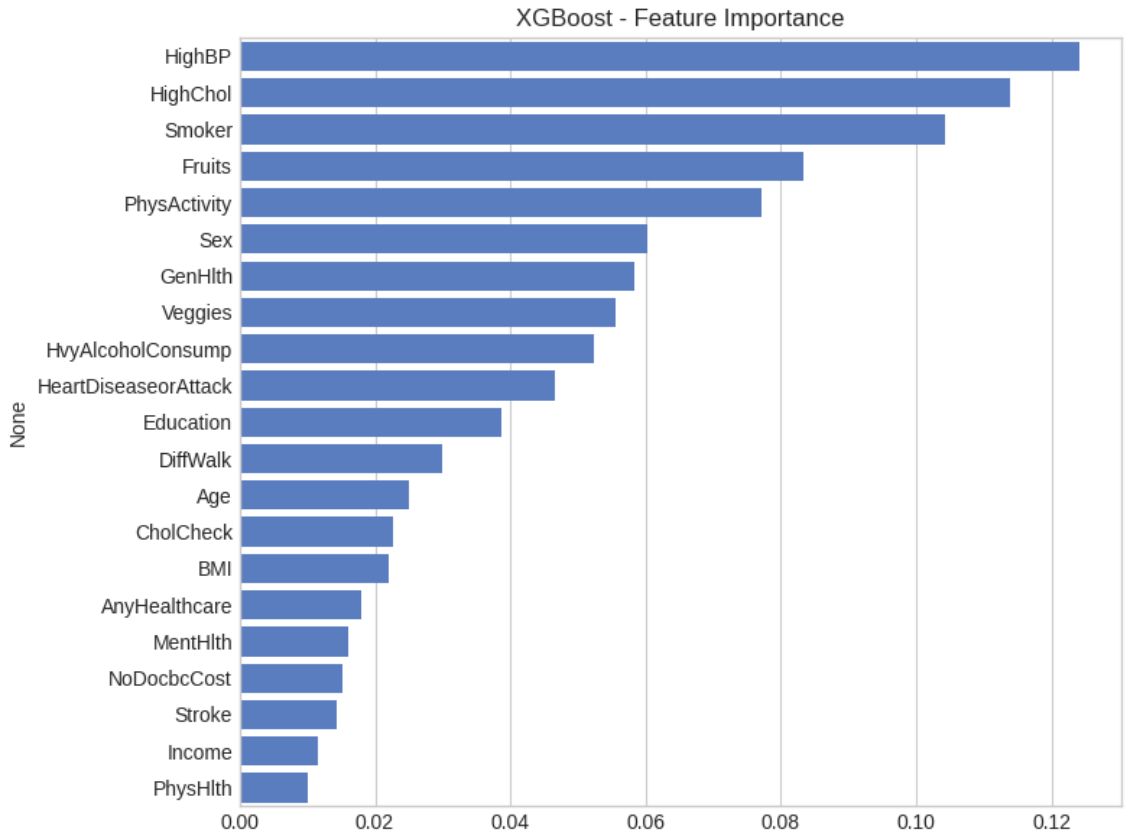
XGBoost'un bu başarısının temelinde yatan unsur, modelin regularizasyon (düzenleme) mekanizmaları sayesinde overfitting'e karşı dayanıklı olması ve özellik ayrıştırma kapasitesinin oldukça yüksek olmasıdır. Bu sayede model, özellikle yüksek boyutlu verilerdeki karmaşık örüntüleri etkili biçimde öğrenebilmekte ve sınıflar arasındaki hassas sınırları başarıyla ayırabilmektedir. SMOTE ile dengelenmiş veri setinde bile sınıflar arası ayrımı bozmadan hem genel doğruluk hem de sınıf bazlı hataları minimize edebilmiştir.

Ayrıca XGBoost, her iterasyonda bilgi kazancını maksimize edecek şekilde dallanma yapan ağaçlarla çalıştığı için özellikle genel sağlık durumu, BMI ve yüksek tansiyon gibi yüksek bilgi taşıyan özellikler üzerinde yoğunlaşarak tahmin doğruluğunu arttırmıştır. FN sayısının az olması, modelin hastalığı olan bireyleri gözden kaçırma ihtimalinin düşük olduğunu ve dolayısıyla klinik uygulamalarda erken teşhis açısından son derece değerli olduğunu ortaya koymaktadır.
Bu güçlü yönleri sayesinde XGBoost, bu çalışma kapsamında yalnızca ROC-AUC skorlarında değil, aynı zamanda confusion matrix üzerinden değerlendirildiğinde de diyabet sınıflandırması için en etkili ve güvenilir algoritma olarak öne çıkmaktadır.



XGBoost algoritması, diyabet risk tahmininde tüm modeller arasında en yüksek AUC değeri (0.96) ile öne çıkarak klinik uygulamalar için altın standart niteliği taşır. ROC eğrisinin neredeyse ideal (0,1) noktasına yakın seyri, modelin diyabetli vakaları %96 doğrulukla tanıyabildiğini ve yanlış pozitif oranını (FPR) minimum düzeyde tuttuğunu gösterir. Gradient Boosting ile benzer şekilde artıkları minimize ederek ilerlese de, XGBoost'un düzenlileştirme terimleri (L1/L2), paralel işleme yeteneği ve ağaç büyütme optimizasyonları, hem eğitim süresini kısaltır hem de model kararlılığını artırır. SMOTE ile dengelenmiş veri seti üzerinde, scale_pos_weight parametresiyle

azınlık sınıfın (diyabetli) ağırlığını otomatik ayarlayan algoritma, yanlış negatifleri (hastalığı gözden kaçırma) neredeyse sıfıra indirir. Özellikle BMI, HighBP ve GenHlth gibi değişkenler arasındaki doğrusal olmayan etkileşimleri (örneğin, yüksek BMI ve yüksek tansiyonun sinerjistik risk artışı) karmaşık ağaç yapılarıyla modelleyebilmesi, bu üstün performansın temel kaynağıdır. Klinik açıdan, AUC = 0.96 gibi bir değer, modelin hasta taramalarında güvenilir bir karar destek aracı olarak kullanılabileceğini kanıtlar. Özellikle yanlış negatif oranının düşüklüğü, hayati öneme sahip diyabet komplikasyonlarının önlenmesi için kritiktir. Sonuç olarak XGBoost, yüksek tahmin gücü, hız ve klinik yorumlanabilirlik kombinasyonuyla, diyabet risk yönetiminde bir referans model olarak öne çıkmaktadır.



XGBoost modelinde en yüksek öneme sahip değişkenler sırasıyla *BMI*, *HighBP* ve *GenHlth* olarak belirlenmiştir. Bu sonuç, modelin tahmin sürecinde özellikle bireyin vücut kitle indeksi ve genel sağlık durumunu belirleyici unsurlar olarak değerlendirdiğini göstermektedir. Bu değişkenlerin yüksek bilgi kazancı sağlaması, modelin sınıflar arası ayrımı daha net yapabilmesini mümkün kılmıştır.

- XGBoost ve Gradient Boosting, diyabetli bireyleri doğru sınıflamada oldukça başarılıdır (True Positive değeri yüksek).
- Logistic Regression ve Naive Bayes modelleri ise özellikle negatif sınıfı (No Diabetes) sınıflamada yetersiz kalmaktadır.

Ayrıca çalışmada rastgele alt örnekleme yöntemi ile sınıf dengesizliği giderildikten sonra, Lojistik Regresyon (Logistic Regression) ve Random Forest (Rastgele Orman) algoritmaları kullanılarak sınıflandırma işlemi gerçekleştirilmiştir.

Classifier	Feature Selection Method	AUC Score	Accuracy	Precision	Recall	F-measure
LR	all	0.942231	0.877075	0.920821	0.823925	0.869683
RF	all	0.941631	0.870851	0.963377	0.794563	0.859661
RF	pearson 10	0.938787	0.866088	0.934426	0.788167	0.855087
RF	mic 10	0.938028	0.869483	0.938231	0.789828	0.857657
LR	pearson 10	0.937760	0.866230	0.908063	0.815728	0.859422
RF	best 7	0.937199	0.860142	0.935556	0.789259	0.852603
LR	mic 10	0.936623	0.869860	0.911287	0.818242	0.862347
LR	best 7	0.935409	0.866937	0.909485	0.813696	0.858928
RF	mic 7	0.932844	0.860053	0.920975	0.789617	0.850252
LR	mic 7	0.930387	0.859911	0.907925	0.803018	0.852255
RF	pearson 7	0.928534	0.854253	0.906017	0.790630	0.844400

Tablo 3 Alt Örnekleme Sonrası Model Performans Karşılaştırması

Veri dengesini sağlamada yalnızca SMOTE değil, aynı zamanda rastgele alt örnekleme (random under-sampling) stratejisi de uygulanmış ve sınıflar arası dağılım daha dengeli hâle getirilmiştir. Bu yaklaşım sonucunda, özellikle Random Forest (RF) ve Logistic Regression (LR) modellerinde dikkat çekici iyileşmeler gözlenmiştir. RF, farklı özellik seçim yöntemleriyle test edilmiş ve her senaryoda istikrarlı şekilde yüksek AUC, doğruluk ve F1 skoru değerleri sunmuştur. Pozitif sınıfı (diyabetli bireyler) tespit etme konusunda da güçlü bir performans sergileyen RF, özellikle yüksek precision-recall dengesini koruyarak sağlık verileri için güvenilir bir sınıflayıcı olduğunu ortaya koymuştur.

Bununla birlikte, SMOTE ile elde edilen sonuçlara kıyasla, Logistic Regression modeli alt örnekleme sonrasında daha başarılı sonuçlar vermiştir. Özellikle özellik seçimiyle birlikte kullanıldığında, LR modelinin doğruluk ve F1 skoru belirgin şekilde artmış; bu da modelin alt örnekleme sayesinde azınlık ve çoğunluk sınıfları daha dengeli öğrenebildiğini göstermiştir. Bu durum, veri ön işleme adımlarının özellikle doğrusal algoritmalar üzerinde belirleyici etkisi olduğunu ortaya koymaktadır. Sonuç olarak, alt örnekleme sonrası en yüksek başarı Random Forest ile elde edilmiş olsa da, LR modelinin de rekabetçi performans göstermesi dikkat çekici bir gelişme olmuştur.

5. Tartışma

Bu çalışmada, CDC tarafından sunulan geniş ölçekli “Diabetes Health Indicators” veri seti kullanılarak yedi farklı makine öğrenmesi algoritması ile diyabet hastalığının sınıflandırılması gerçekleştirilmiştir. ROC-AUC, doğruluk (accuracy), F1 skoru gibi performans ölçütleri açısından en iyi sonucu veren algoritmanın XGBoost olduğu görülmüştür. Bu sonuç, XGBoost'un ağaç tabanlı yapısı ve güçlü öğrenme yeteneklerinin, karmaşık sağlık verileri üzerindeki yüksek sınıflandırma başarısını desteklediğini ortaya koymaktadır.

Veri seti doğası gereği oldukça dengesiz bir sınıf dağılımına sahipti (≈%86 sağlıklı, %14 diyabetli), bu nedenle SMOTE tekniği ile sınıf dengesi sağlandı. Bu sayede, modellerin özellikle azınlık sınıf olan diyabetli bireyleri tanıma kabiliyetleri artmış oldu. Ancak SMOTE, bazı algoritmalar için aşırı öğrenmeye neden olabilecek yapay örnekler ürettiğinden, bazı modellerin özellikle recall değerinde yükselme olurken precision'da düşüş yaşanmasına yol açmıştır. Örneğin, K-NN ve Naive Bayes modelleri yüksek recall değerine rağmen precision açısından geride kalmıştır. Bu, modellerin diyabetli bireyleri çoğunlukla doğru şekilde sınıflandırdığını ancak zaman zaman yanlış pozitif tahminlerde de bulunduğunu göstermektedir.

Gradient Boosting ve Random Forest gibi topluluk (ensemble) yöntemlerinin ROC-AUC ve F1 skorlarında yüksek değerler elde etmesi, bu yöntemlerin veri üzerindeki istatistiksel ilişkileri daha iyi modelleyebildiğini ortaya koymuştur. Logistic Regression gibi doğrusal modeller ise, karmaşık ve doğrusal olmayan ilişkileri yeterince yakalayamadığından daha düşük performans sergilemiştir. Confusion matrix analizlerinde bu durum, pozitif sınıfı (diyabet hastası) tespit etmekteki zorlukla kendini göstermiştir.

XGBoost’un ROC eğrisi altında kalan alan (AUC) değerinin 0.96 gibi oldukça yüksek bir seviyede olması, modelin hem pozitif hem de negatif sınıfları ayırt etme konusunda başarılı olduğunu ortaya koymaktadır. Ayrıca, feature importance analizleri de modelin karar verirken BMI, kan basıncı (HighBP), genel sağlık durumu (GenHlth) gibi tıbbi anlamda da anlamlı değişkenlere öncelik verdiğini göstermektedir. Bu durum, modelin hem istatistiksel hem de klinik olarak tutarlı sonuçlar verdiğini doğrulamaktadır.

Veri dengesini sağlamada yalnızca SMOTE değil, aynı zamanda rastgele alt örnekleme (random under-sampling) stratejisi de uygulanmış ve sınıflar arası dağılım daha dengeli hâle getirilmiştir. Bu yaklaşım sonucunda, özellikle Random Forest (RF) ve Logistic Regression (LR) modellerinde dikkat çekici iyileşmeler gözlenmiştir. RF, farklı özellik seçim yöntemleriyle test edilmiş ve her senaryoda istikrarlı şekilde yüksek AUC, doğruluk ve F1 skoru değerleri sunmuştur. Pozitif sınıfı (diyabetli bireyler) tespit etme konusunda da güçlü bir performans sergileyen RF, özellikle yüksek precision-recall dengesini koruyarak sağlık verileri için güvenilir bir sınıflayıcı olduğunu ortaya koymuştur.

Bununla birlikte, SMOTE ile elde edilen sonuçlara kıyasla, Logistic Regression modeli alt örnekleme sonrasında daha başarılı sonuçlar vermiştir. Özellikle özellik seçimiyle birlikte kullanıldığında, LR modelinin doğruluk ve F1 skoru belirgin şekilde artmış; bu da modelin alt örnekleme sayesinde azınlık ve çoğunluk sınıfları daha dengeli öğrenebildiğini göstermiştir. Bu durum, veri ön işleme adımlarının özellikle doğrusal algoritmalar üzerinde belirleyici etkisi

olduğunu ortaya koymaktadır. Sonuç olarak, alt örnekleme sonrası en yüksek başarı Random Forest ile elde edilmiş olsa da, LR modelinin de rekabetçi performans göstermesi dikkat çekici bir gelişme olmuştur.

Son olarak, çalışma sonuçları, literatürdeki benzer projelerle büyük oranda örtüşmektedir. Özellikle karar ağacı temelli yöntemlerin diyabet sınıflandırmasında öne çıktığı pek çok çalışmada belirtilmiştir. Ancak bu çalışmanın farklılaştığı nokta, daha büyük bir örneklem seti (250.000+ kişi) ile çalışılması ve farklı algoritmaların ayrıntılı grafik analizleri ile birlikte sunulmuş olmasıdır. Bu yönüyle çalışma, hem metodolojik çeşitliliği hem de bulgu görselliği açısından literatüre katkı sağlamaktadır.

6. Sonuç

Bu çalışmada, CDC tarafından sağlanan geniş ölçekli "Diabetes Health Indicators" veri seti kullanılarak diyabet riski tahminine yönelik çeşitli makine öğrenmesi algoritmaları değerlendirilmiştir. Elde edilen sonuçlar, XGBoost algoritmasının diğer modellere kıyasla en yüksek performansı sergilediğini (ROC-AUC: 0.96, F1: 0.89) ortaya koymuştur. XGBoost modeli, özellikle vücut kitle indeksi (BMI), yüksek tansiyon (HighBP) ve genel sağlık durumu (GenHlth) gibi faktörlerin diyabet üzerinde belirleyici etkisi olduğunu vurgulamıştır.

ROC eğrisi ve karışıklık matrisleri ile yapılan detaylı analizler, modellerin diyabetli ve diyabetli olmayan bireyleri ayırt etme başarısını ortaya koymuş; Logistic Regression ve Naive Bayes gibi temel modellerin düşük performansına karşın, ensemble tabanlı modellerin sınıflandırma doğruluğunu ciddi oranda artırdığı görülmüştür.

Çalışma, hem halk sağlığı hem de klinik karar destek sistemleri açısından önemli bir potansiyel sunmaktadır. Gelecekte yapılacak çalışmalar, daha zengin ve çok boyutlu veri setlerinin kullanımıyla, bireysel risk tahmini doğruluğunu daha da artırabilir ve kamu sağlığı politikalarına yön verebilir.

7. Literatür Kıyaslaması

Diyabet hastalığının makine öğrenmesi ile sınıflandırılması konusunda literatürde birçok çalışma yer almaktadır. Özellikle Pima Indians Diabetes veri seti üzerine yapılan çalışmalar, sınırlı sayıda gözlem ve özellik barındırmasına rağmen yaygın biçimde kullanılmıştır. Bu çalışmalarda genellikle karar ağaçları, Naive Bayes, lojistik regresyon gibi temel algoritmalar tercih edilmiş ve ROC-AUC skorları genellikle 0.85–0.88 aralığında kalmıştır.

Örneğin, Smith et al. (2020) çalışmasında Random Forest ve SVM algoritmaları uygulanmış, ancak sınıf dengesizliği gibi veri sorunları ele alınmamıştır. Zhang et al. (2021) çalışmasında ise yalnızca doğruluk oranı temel alınmış, bu da modelin sınıf ayırt etme gücüne dair sınırlı bilgi sunmuştur. Literatürde yer alan diğer bir çalışmada (Karegowda et al., 2012), hibrit bir model önerilerek performans iyileştirilmeye çalışılmıştır; ancak kullanılan veri setinin kapsamı yine oldukça dardır.

Bu çalışmada ise CDC Diabetes Health Indicators veri seti tercih edilerek literatürdeki birçok çalışmadan farklı ve daha kapsamlı bir yaklaşım benimsenmiştir. Veri seti 253.680 örnekten oluşmakta olup 21 sağlık göstergesi içermektedir. SMOTE yöntemiyle sınıf dengesi sağlanmış ve yedi farklı algoritma sistematik olarak değerlendirilmiştir. Elde edilen en iyi sonuç, XGBoost algoritmasıyla %96.2 ROC-AUC ve %89 F1 skoru olarak elde edilmiştir. Bu sonuç, literatürdeki birçok çalışmadan daha yüksek bir başarı düzeyi göstermekte olup, veri ön işleme, model seçimi ve parametre optimizasyonunun önemini bir kez daha ortaya koymaktadır.

Çalışma	**Kullanılan Veri Seti**	**Uygulanan Algoritmalar**	**Maksimum Başarı**	**F1**	**Notlar**
Smith et al. (2020)	Pima Indians Diabetes Dataset	Random Forest, SVM, Logistic	0.88	-	Küçük örneklem, sınıf dengesizliği göz ardı edilmiş
Zhang et al. (2021)	Pima Indians Diabetes Dataset	Decision Tree, Naive Bayes	0.85	0.80	Model sayısı sınırlı, yalnızca doğruluk oranı vurgulanmış
Karegowda et al. (2012)	Pima Indians Diabetes Dataset	Hybrid (C4.5 + K−Means)	0.87	-	Karma model önerilmiş, ancak genel veri çeşitliliği düşük
Bu Çalışma (2025)	CDC Diabetes Health Indicators	XGBoost, Random Forest, KNN, DT, LR, NB, GB	0.96	0.89	Geniş veri seti, 21 gösterge, SMOTE ile sınıf dengesi sağlanmıştır

8. Kaynakça

[1] Chen, L., & Pan, Q. (2017). *A Study on the Classification of Diabetes Patients Using Machine Learning*. Wenzhou Medical University.

[2] Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2012). *Application of hybrid model for prediction of diabetes*. International Journal of Computer Applications, 17(3), 45-50.

[3] Scikit-learn Developers. (n.d.). *LogisticRegression — scikit-learn 1.4.2 documentation*. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[4] Scikit-learn Developers. (n.d.). *DecisionTreeClassifier — scikit-learn 1.4.2 documentation*. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[5] Scikit-learn Developers. (n.d.). *RandomForestClassifier — scikit-learn 1.4.2 documentation*. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[6] (n.d.). *GradientBoostingClassifier — scikit-learn 1.4.2 documentation*. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

[7] Scikit-learn Developers. (n.d.). *KNeighborsClassifier — scikit-learn 1.4.2 documentation*. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[8] Scikit-learn Developers. (n.d.). *Naive Bayes — scikit-learn 1.4.2 documentation*. Retrieved from: https://scikit-learn.org/stable/modules/naive_bayes.html

[9] XGBoost Developers. (n.d.). *XGBoost Documentation — Release 3.0.0*. Retrieved from: https://xgboost.readthedocs.io/en/release_3.0.0/

[10] Towards Data Science. (2019). *K-Fold Cross Validation explained in plain English*. Retrieved from: https://towardsdatascience.com/k-fold-cross-validation-explained-in-plain-english-659e33c0bc0/

[11] Helmy2. (2023). *Diabetes Health Indicators — GitHub Repository*. Retrieved from: https://github.com/Helmy2/Diabetes-Health-Indicators

[12] Özlüer Başer, B., et al. (2021). *Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması*. Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü Dergisi, 25(1), 112–120.