T.C. KOCAELİ SAĞLIK VE TEKNOLOJİ ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

SINIFLANDIRMA ALGORİTMALARINDA ÖZNİTELİK SEÇİMİ YÖNTEMLERİNİN UYGULAMALI KARŞILAŞTIRMASI

DAMLA KEKLİK

BİTİRME PROJESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
KOCAELİ, 2025

T.C. KOCAELİ SAĞLIK VE TEKNOLOJİ ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

SINIFLANDIRMA ALGORİTMALARINDA ÖZNİTELİK SEÇİMİ YÖNTEMLERİNİN UYGULAMALI KARŞILAŞTIRMASI

DAMLA KEKLİK

BİTİRME PROJESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ
KOCAELİ, 2025

ABSTRACT

Applied Comparison of Feature Selection Methods in Classification Algorithms

KEKLİK, Damla

Bachelor's Thesis, Computer Engineering
Supervisor: Dr. Vildan YAZICI

Working with high-dimensional data sets presents significant challenges in machine learning projects, both in terms of computational cost and model performance. In particular, the presence of irrelevant or noisy variables can reduce the accuracy of classification algorithms, lead to overfitting, and weaken the model's generalizability. In this context, the feature selection process, which aims to create simpler and more effective models by selecting the most informative features, is a critical preprocessing step.

This thesis proposes a fully automated feature selection framework based on multi-objective optimization principles, replacing traditional manual methods based on fixed threshold values. The study uses the Diabetes Health Indicators Dataset (22 features, 70,692 data points) obtained from Kaggle, which contains health indicators related to individuals' risk of diabetes. Three different feature selection methods were compared: Spearman correlation (ordered relationships), Mutual Information (non-linear dependencies), and Random Forest Feature Importance (model-based selection). The feature subsets selected using these methods were tested with Lojistik Regresyon and Random Forest classifiers.

The main contribution of the proposed system is the development of an adaptive optimization mechanism that automatically determines the optimal threshold value for each method. This process evaluates model performance using a three-component multi-objective scoring function: Classification Performance (60%), Feature Efficiency (25%), and Performance-Efficiency Balance (15%).

According to experimental results, the highest performance was achieved with the Spearman correlation + Lojistik Regresyon configuration, which selected 20 out of 34 features (F1-score: 0.7339, Accuracy: 0.7328). This study demonstrates that simple models can outperform complex algorithms with proper feature selection; it also contributes to the literature by providing an automatic, objective, and reproducible feature selection methodology.

Keywords: Feature Selection, Multi-Objective Optimization, Classification Performance, Dimension Reduction Diabetes Prediction

ÖZET

Sınıflandırma Algoritmalarında Öznitelik Seçimi Yöntemlerinin Uygulamalı Karşılaştırması

KEKLİK, Damla

Lisans Bitirme Projesi, Bilgisayar Mühendisliği

Tez Danışmanı: Dr. Öğretim Üyesi Vildan YAZICI

Makine öğrenimi projelerinde, yüksek boyutlu veri setiyle çalışmak hem hesaplama maliyeti açısından hem de model başarısı açısından önemli zorluklar içermektedir. Veri setinde veri ile ilgisi olmayan değişkenlerin var olması işleri zorlaştırabilir. İlgisiz değişkenlerin varlığı sınıflandırma algoritmalarının doğruluğunu düşürür, aşırı öğrenmeye yol açar ve modelin genellenebilir olmasını zayıflatır. Bu tarz sorunların modelde olmaması ya da daha aza indirmek için öznitelik seçimi kullanılır. Daha etkili model oluşturulması için öznitelik seçimi kritik ön işleme adımıdır.

Bu tez, geleneksel sabit eşik değerlerine dayalı manuel yöntemlerin yerini alan, çok amaçlı optimizasyon prensiplerine dayalı, tamamen otomatik bir öznitelik seçim çerçevesi önermektedir. Çalışmada, Kaggle'dan temin edilen ve bireylerin diyabet riskiyle ilişkili sağlık verileri bulunan Diabetes Health Indicators Dataset kullanılmıştır. Üç farklı öznitelik seçimi yöntemi kullanılarak karşılaştırılmıştır: Spearman korelasyonu (sıralı ilişkiler), Mutual Information (doğrusal olmayan bağımlılıklar) ve Random Forest Öznitelik Önemi (model-tabanlı seçim) yöntemleri kullanılmıştır. Bu yöntemlerle seçilen öznitelik alt kümeleri, Lojistik Regresyon ve Random Forest sınıflandırıcılarıyla test edilmiştir.

Önerilen sistemin temel katkısı, her yöntem için en uygun eşik değerini otomatik olarak belirleyen bir adaptif optimizasyon mekanizması geliştirmiş olmasıdır. Bu süreç, model başarısını üç bileşenli çok amaçlı bir skorlama fonksiyonu ile değerlendirmiştir: Sınıflandırma Performansı (%60), Öznitelik Verimliliği (%25) ve Performans-Verimlilik Dengesi (%15).

Deneysel sonuçlara göre en yüksek performans, 34 özelliğin 20'sini seçen Spearman korelasyonu + Lojistik Regresyon konfigürasyonuyla elde edilmiştir (F1-skor: 0.7339, Doğruluk: 0.7328). Bu çalışma, basit modellerin doğru öznitelik seçimiyle karmaşık algoritmalardan daha başarılı olabileceğini göstermekte; aynı zamanda otomatik, nesnel ve tekrarlanabilir bir öznitelik seçimi metodolojisi sunarak literatüre katkı sağlamaktadır.

Anahtar kelimeler: Öznitelik Seçimi, Çok Amaçlı Optimizasyon, Sınıflandırma Başarımı, Boyut Azaltma Diyabet Tahmini.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

1. GİRİŞ

Gerçek dünyadaki veri analizi süreçlerinde sıklıkla karşılaşılan temel zorluklardan biri, veri setlerinin yüksek boyutlu ve karmaşık yapıda olmasıdır. Özellikle büyük veri kümeleriyle çalışırken, her bir gözleme ait onlarca hatta yüzlerce öznitelik bulunabilir. Bu değişkenlerin tamamı her zaman problemle doğrudan ilişkili olmayabilir. Problemle ilgisi bulunmayan ya da gürültü içeren değişkenlerin analizde yer alması, yalnızca modelin hesaplama maliyetini artırmakla kalmaz; aynı zamanda sınıflandırma algoritmalarının doğruluğunu düşürebilir, aşırı uyuma (overfitting) neden olabilir ve modelin genellenebilir olmasını azaltabilir.

Bu nedenle, veri madenciliği ve makine öğrenmesi uygulamalarında öznitelik seçimi süreci, veri ön işleme aşamasının en kritik bileşenlerinden biri olarak ön plana çıkmaktadır. Öznitelik seçimi, veri setinde yer alan çok sayıda değişken içerisinden, modele en fazla bilgi katkısı sağlayan veya hedef değişkenle en güçlü ilişkili olan değişkenlerin seçilmesini; diğer yandan ilgisiz ya da yenilenen (redundant) değişkenlerin elenmesini hedefler. Böylece hem modelin başarımı artar hem de daha az sayıda değişkenle daha sade, daha yorumlanabilir ve daha hızlı çalışan bir sistem elde edilmiş olur. Aynı zamanda, boyut indirgeme sayesinde modelin eğitimi ve test süreci hızlanır; özellikle yüksek boyutlu veri setlerinde bu durum belirgin bir avantaj sağlar.

Bu çalışmada, sınıflandırma problemlerinde üç farklı matematiksel öznitelik seçim metotları etkinliği incelenecektir. Bu öznitelik seçimleri şu şekildedir; Spearman korelasyonu, Random Forest'ın sunduğu öznitelik önem değeri ve Mutual Information olarak belirlenmiştir. Bu yöntemler, elimizdeki veri seti üzerinde daha iyi sonuçlar verdikleri için tercih edilmiştir. Bu yöntemlerin sınıflandırma algoritmalarının performansına olan etkileri değerlendirilmiştir.

Çalışma kapsamında kullanılmak için diyabet veri seti seçilmiştir. Diyabetin sebebi insülin üretiminin yetersiz olması ya da vücut hücrelerinin insüline uygun şekilde yanıt vermemesi olabilir. Veya bu ikisi nedeniyle kişinin yüksek kan şekerine sahip olduğu bir grup metabolik hastalığı tanımlar. Dolayısıyla günümüzde diyabet tespitine duyulan ihtiyaç artmaktadır. Çalışma için Kaggle platformundan veri seti alınmıştır [1]. Bu platformdan alınan veri seti bireylerin diyabet riskiyle ilişkili çeşitli sağlık verilerini içeren Diabetes Health Indicators Datasettir. Bu veri seti, çok sayıda sağlık temelli değişken içerdiğinden, öznitelik seçimi süreçlerinin etkisini gözlemlemek açısından daha uygundur.

Öznitelik seçimi yöntemlerinin seçtiği değişken alt kümeleriyle Random Forest ve Lojistik Regresyon gibi iki farklı sınıflandırma algoritması eğitilecek ve bu modellerin başarımı, ROC eğrisi, karmaşıklık matrisi, AUC, doğruluk (accuracy) ve F1 skoru performans metrikleriyle karşılaştırılacaktır. Ayrıca, literatürdeki benzer çalışmalardan farklı olarak, bu projede otomatik eşik değeri optimizasyonu ve çok amaçlı skor fonksiyonu ilave değerlendirme stratejileri kullanılmıştır. Bu kullanım sayesinde, seçilen yöntemlerin etkisi daha nesnel ve sistematik biçimde analiz edilmiştir.

Sonuç olarak, bu çalışma, yüksek boyutlu veri kümelerinde öznitelik seçimi uygulamalarının sınıflandırma başarımı üzerindeki etkisini detaylı biçimde ortaya koymayı; veri boyutunun azaltılmasıyla sağlanan avantajları, model kalitesiyle birlikte değerlendirmeyi amaçlamaktadır. Elde edilen sonuçlar hem akademik araştırmalar hem de gerçek hayattaki sağlık verisi analizi gibi uygulamalı alanlar için rehber niteliği taşıyacaktır.

2. LİTERATÜR TARAMASI

Son yıllarda yapılan çalışmalarda öznitelik seçim yöntemleri ve sınıflandırma algoritmalarının birlikte kullanıldığı göze çarpmaktadır. Bu yöntemin performansta başarılı sonuçlar verdiği tespit edilmiş. Birçok çalışma da sabit eşik değerleriyle seçimler yapıldığı fark edilerek, literatürde bulunan önemli araştırmaların amaçları ve metodolojileri incelenmiştir.

2.1. İlgili Çalışmalar

Abdelhafez ve Amer tarafından yürütülen çalışmada, diyabet hastalığının tahmini için sekiz farklı makine öğrenimi algoritmasının (**Naive Bayes,Decision Trees ,Lojistik Regresyon, SVM , Random Forest , ANN, LogitBoost** ve **Voting Classifier)** performansı karşılaştırılmıştır. Araştırmanın temel odak noktası, özellikle Korelasyon temelli CfsSubsetEval ve sarmalayıcı temelli WrapperSubsetEval olmak üzere iki farklı öznitelik seçimi yönteminin sınıflandırma başarısı üzerindeki etkisini detaylı bir şekilde incelemektir. Yazarlar, bu yöntemlerle oluşturulan 4 ve 7 öznitelikle daha küçük veri kümeleri kullanarak hem model karmaşıklığını azaltmayı hem de doğruluk ve F1 skoru gibi performans metriklerinde iyileşmeler sağlamayı hedeflemiştir. Çalışmanın bulguları, seçilmiş öznitelik kümeleriyle karar ağacı ve Random Forest algoritmalarının en yüksek başarı oranlarına ulaştığını ve aykırı değer temizliğinin SVM gibi algoritmaların performansını olumlu yönde etkilediğini göstermiştir [2].

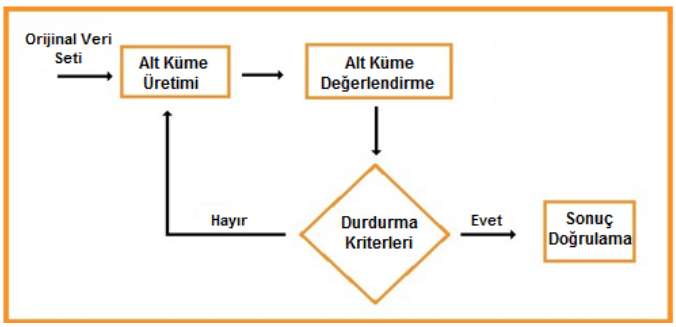
Koren ve arkadaşları, öznitelik seçimi için öznitelik önem eşiklerini otomatik olarak öğrenen bir çerçeve önermektedir. Çalışmalarında, üç farklı etiketli veri seti üzerinde birden fazla öznitelik seçim tekniğini karşılaştırmışlar ve sınıflandırma doğruluğunu maksimize etmek amacıyla eşik değerleri üzerinde sistematik bir arama (örneğin, 0,01'lik adımlarla 0 ila 0,2 arasında) gerçekleştirmişlerdir. Geliştirdikleri bu yöntemle, öznitelik seçiminden sonra sınıflandırma doğruluğunda %20'ye varan artışlar ve daha iyi hassasiyet/geri çağırma (precision/recall) oranları elde ettiklerini raporlamışlardır. Bu araştırma, özellikle boyutluluğu azaltmada otomatik eşik ayarlamasının değerini vurgulamaktadır. Ancak, yazarların belirttiği üzere, çalışma performansı ağırlıklı olarak doğruluk metriği üzerinden değerlendirmekte, çok kriterli puanlama yaklaşımlarını veya sabit bir yöntem setiyle yapılan karşılaştırmaları tam anlamıyla kapsamamaktadır [3].

Pechprasarn ve arkadaşları, farklı filtreleme yöntemlerini (MRMR, Chi², ReliefF, ANOVA) kullanarak temel klinik değişkenleri belirlemiş ve çeşitli makine öğrenmesi modellerinin performanslarını karşılaştırmıştır. En yüksek doğruluğun, seçilmiş 9 öznitelik ile eğitilmiş Kuadratik SVM modelinde elde edildiği rapor edilmiştir. Ancak, burada da öznitelik sayısının seçimi sabit tutulmuş ve eşik değerlerine dayalı dinamik bir optimizasyon yapılmamıştır [4].

2.2. Öznitelik Seçimi

Öznitelik seçimi, modelin başarısını artırmak ve hesaplama maliyetini azaltmak amacıyla yaygın olarak kullanılan bir tekniktir. Bu yöntem, orijinal veri kümesindeki en anlamlı ve en iyi temsil yeteneğine sahip alt kümenin belirlenmesini amaçlamaktadır. "N" adet öznitelik arasından en iyi "k" adet özelliği seçerek öğrenme süresini optimize etmektir. Bu kapsamda, öznitelik seçiminin temel amacı, tüm verileri kullanmadan en yüksek veri bütünlüğünü koruyarak optimal performansı elde etmektir. Bu şekilde modelin hesaplama süresi kısalır, aşırı öğrenme riski azaltılır ve yorumlanabilirliği artırılmış olur. Öznitelik seçimin genel akışı Şekil 1'de bulunmaktadır [5].

Veri kümesini oluşturmak için gerekli olan veri toplama sürecinde, özellik seçimi kaynak tasarrufu sağlar. Bu süreçte yalnızca en anlamlı ve etkili özellikler belirlendiğinden, gereksiz veri toplama ihtiyacı azalır. Ayrıca, veri kümesini daha basit bir şekilde tanımlanabilir, görselleştirilebilir ve anlaşılabilir hale getirir. Bu da analiz sürecini kolaylaştırır ve yorumlamayı daha etkili kılar. Öte yandan, ilgili olmayan yani gürültülü verilerin ortadan kaldırılması sayesinde hem veri kalitesi artar hem de algoritmaların daha hızlı ve verimli çalışması sağlanır.



Şekil 1:Öznitelik seçimi genel akış şeması [6]

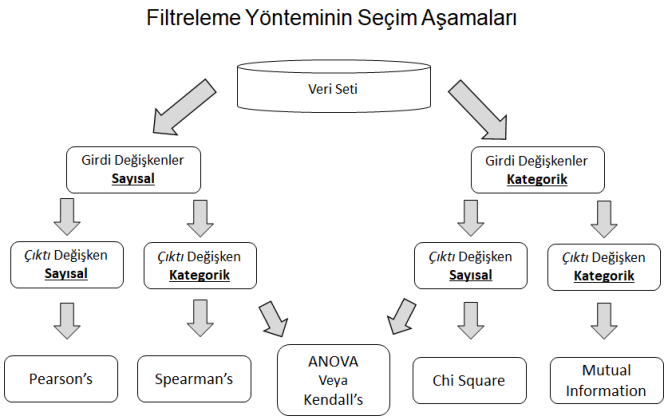
Üç çeşit öznitelik seçim yöntemi bulunur. Bunlar; birincisi sadece istatiksel bilgilere dayalı olan filtreleme (filter) yöntemler, öznitelikler üzerinde arama işlemleri gerçekleştiren sarmal (wrapper) yöntemler ve en iyi bölen ölçütünü bulmaya dayalı olan gömülü (embedded) yöntemler olmak üzere üç grupta incelenir [7].

2.2.1. Filtre Yöntemleri

Değişkenlerin hedef değişkenle ile olan ilişkisini değerlendirerek önemli olanları belirler ve düşük bilgi katsayısı bulunan değişkenleri eler. Modelden bağımsız çalışır. Herhangi makine öğrenme algoritmasına bağlı değildir. Hesaplama maliyetleri düşüktür. Hızlı ve verimli olması büyük bir avantajıdır. Veri setine uygun filtreleme yöntemini seçerken Şekil 2'yi kullanabilir [8].

Dezavantajları olarak, değişkenler bireysel olarak değerlendirilir, bu da değişkenler arası etkileşimleri göz ardı edebilir. Karmaşık veri yapılarında, değişkenlerin birlikte sağladığı bilgi kaybolabilir ve modelin doğruluğunu etkileyebilir. Başlıca filtreleme yöntemleri:

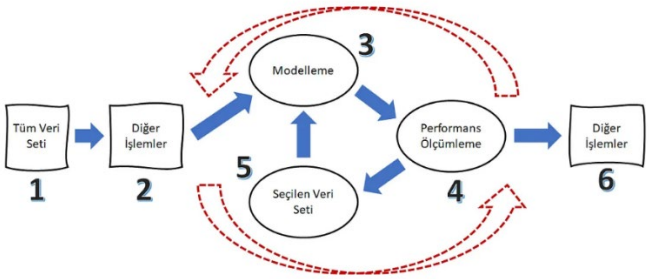
- **Spearman Korelasyonu:** iki değişken arasındaki sıralı (monoton) ilişkiyi ölçen, verilerin sıralarına dayalı non-parametrik bir korelasyon katsayısıdır.
- **Mutual Information:** Değişkenler arasındaki bilgi paylaşımını değerlendirerek, hedef değişkenle daha fazla ortak bilgiye sahip olan değişkenleri seçer.



Şekil 2: Filtreleme yönteminin seçim aşamaları [8]

2.2.2. Sarmal Yöntem

Belirli bir öznitelik alt kümesinin performansını değerlendirmek için bir sınıflandırma algoritması kullanır. Bu yöntemler, modelin kendisini bir "kara kutu" olarak değerlendirir ve farklı öznitelik kombinasyonlarıyla modelin performansını ölçer. Şekil 3 de yöntemin modelleme adımları gözükmektedir [9].



Şekil 3: Sarmal yöntem modelleme adımları. [9]

Hesaplama maliyeti yüksek olduğundan dolayı bu yöntemin kullanılması tercih edilmedi.

2.2.3. Gömülü Yöntem

Bu yöntemler, filtreleme ve sarmalayıcı yöntemlerinin avantajlarını birleştirir hem hesaplama verimliliği hem de model performansı açısından dengeli bir yaklaşım sunar. En büyük avantajı, modelin eğitim sürecine entegre olmalarıdır. Bu, hem daha az hesaplama gücü gerektirdiği hem de doğrudan modelin performansına odaklandığı için verimlidir [10].

2.3. Sınıflandırma Algoritmaları

Sınıflandırma algoritmaları, denetimli öğrenme kapsamında kullanılan ve verileri belirli kategorilere veya sınıflara ayırmayı amaçlayan makine öğrenmesi yöntemleridir. Bu algoritmalar, etkili eğitim verisinden öğrenerek yeni gelen verileri doğru sınıflara atamaya çalışır. Bu algoritmaların performansı genellikle doğruluk, kesinlik, duyarlılık ve F1 skoru gibi metriklerle değerlendirilir. Sınıflandırma problemleri genellikle iki temel kategoriye ayrılır bu iki kategori; ikili sınıflandırma veriler sadece iki sınıfa ayrılır, çok sınıflı sınıflandırma veriler ikiden fazla sınıfa ayrılır [11,12].

- **Lojistik Regresyon:** Doğrusal bir modelin lojistik fonksiyonu ile dönüştürülmesine dayanan bir sınıflandırma algoritmasıdır. Özellikle ikili sınıflandırma problemlerinde kullanılır, ancak çok sınıflı problemlere de uyarlanabilir.
- **Random Forest:** Birden fazla karar ağacının tahminlerini birleştiren bir topluluk öğrenme yöntemidir. Her ağaç, rastgele seçilen öznitelikler ve veri örnekleri kullanılarak eğitilir.

Sınıflandırma algoritmalarının başarısını değerlendirmek için kullanılan temel metrikler:

- **Accuracy (Doğruluk):** Doğru tahmin edilen örneklerin toplam örnek sayısına oranı.
- **F1-Skoru:** Kesinlik ve duyarlılığın harmonik ortalaması.
- **ROC Eğrisi ve AUC:** Alıcı işletim karakteristiği eğrisi ve bu eğrinin altında kalan alan, sınıflandırıcının ayrım gücünü gösterir.
- **Karmaşıklık Matrisi (Confusion Matrix):** Tahminlerin gerçek sınıflarla karşılaştırıldığı bir tablo.

3. YÖNTEMLER

3.1. Veri

Kaggle platformundan “Diabetes Health Indicators Dataset” adlı veri seti üzerinde çalışmalar gerçekleştirilmiştir. Diyabet ve diyabet olmayan iki değişkenden oluşuyor. Bu iki değişkenin veri setinde dengeli dağıldığı gözlemlenmektedir. Dengeli dağılım nedeniyle sınıf dengesizliği problemi ortadan kalkmıştır. Veri setinde toplam 70692 satır ve 22 değişkenden bulunmaktadır. Veri setindeki bağımlı değişken bireyin diyabet hastası olup olmadığını belirtir:

- 1: Birey diyabet hastasıdır.
- 0: Birey diyabet hastası değildir.

Bağımsız değişkenler ise kişilerin yaşam tarzı, kültür seviyesi ve sağlık durumları gibi değişkenlerden oluşmaktadır. Değişkenlerin büyük bir kısmı sağlık göstergelerini modellemek amacıyla özetlenmiştir:

- Yaşam Tarzı:
    - Smoker: Sigara kullanıp kullanmadığı hakkında bilgi verir.
    - Stroke: Felç geçmişi.
    - HeartDiseaseorAttack: Kalp hastalığı geçmişi.
    - PhysActivity: Fiziksel aktivite yapıp yapmadığı.
    - Fruts ve Veggies: Sağlıklı beslenme durumu.
    - HvAlcoholConsump: Alkol tüketimi.
    - AnyHealthcare, NoDocbcCost: Hastaneye erişimi ve ekonomik durumu.
- Yaşam Kalitesi:
    - HighBP: Yüksek tansiyon.
    - HighChol: Yüksek kolesterol.
    - CholCheck: Kolesterol kontrol testi.
    - BMI: Beden kitle indeksi.
- Fiziksel ve Psikolojik Sağlık:
    - MenHlth ve PhysHlth: Bir ay içindeki ruhsal ve fiziksel durum.
    - DifWalk: Yürüyüş bozukluğu.

- Diğer Faktörler:
    - Sex: Kişinin cinsiyeti.
    - Age: Kişinin yaşı.
    - Education: Kişinin eğitim seviyesi.
    - Income: Kişinin gelir düzeyi.

Veri setindeki kategorik değişkenler sayısal değerler şeklinde kodlanmış olup, modelleme sürecinde yeniden ölçeklendirilmesi gerekebilir. Veri, analiz için doğrudan kullanılabilir bulunmaktadır. Çalışmanın temel amacı, bu göstergeleri kullanarak bireylerin diyabet riski tahmininde bulunmak ve model performanslarını karşılaştırmalı olarak değerlendirmektir.

Çalışmada, Spearman korelasyonu, Random Forest öznitelik önem değerleri ve Mutual Information gibi matematiksel tabanlı öznitelik seçimi yöntemlerinin etkinliği değerlendirilmiştir. Seçilen özniteliklerin, Random Forest ve Lojistik Regresyon sınıflandırma algoritmalarının performansına etkisi sistematik olarak analiz edilmiş; diyabet teşhisinde en etkili öznitelik kombinasyonlarının belirlenmesi amaçlanmıştır.

3.2. Kullanılan Öznitelik Seçim Yöntemleri

3.2.1. Spearman Korelasyonu

Öznitelik seçim sürecinde sıralama tabanlı bir yöntem olan Spearman korelasyon katsayısı kullanılmıştır. Spearman korelasyonu, değişkenler ile hedef değişken arasındaki monoton ilişkileri değerlendirir; yani bir değişken artarken diğerinin artıp artmadığını ya da azalıp azalmadığını ölçer. Bu özelliği sayesinde, yalnızca doğrusal değil, tüm artan ya da azalan ilişki türlerini tespit edebilir. Spearman korelasyonu tercih edilme nedenleri: Sayısal değişkenler ile hedef değişken arasındaki sıralı ilişkileri doğru şekilde yakalayabilme kapasitesi, aykırı değerlere karşı daha dayanıklı olması, filtreleme temelli, hızlı ve yorumlanabilir bir ön seçim yöntemi sunmasıdır. Bu yöntem, her değişkenin gözlem değerlerini küçükten büyüğe sıralayarak sıralama değerleri üretir. İki değişkenin sıralama değerleri arasındaki farklar (d) hesaplanır, ardından bu farkların karelerinin toplamı alınır [13].

$$p = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

d : İki değişkenin sıra değerleri arasındaki fark.

n : Gözlem sayısı.

3.2.2. Karşılıklı Bilgi

Bir değişkenin diğeri hakkında sağladığı bilgi miktarı ölçülür bu şekilde değişken ile hedef değişkene olan bağımsızlığı azsa Mutual Information yüksektir denilir. Mutual Information seçilme sebebi, değişkenler arasındaki lineer olmayan bağımlılıkları da yakalayabilen bilgi kuramı tabanlı bir yöntem olduğu için tercih edilmiştir. Sağlık verileri doğası gereği lineer olmayan ilişkilere sahip olabileceğinden, bu yaklaşım daha doğru öznitelik değerlendirmesi sağlar [14].

$$MI(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(X,Y) log_2 \left(\frac{p(X,Y)}{p(X)p(Y)} \right)$$

P(x, y): X ve Y'nin ortak olasılık dağılımı.

P(x) ve p(y): X ve Y'nin marjinal olasılık dağılımlarıdır.

3.2.3. Rastgele Orman Öznitelik Önemi

Random Forest öznitelik önemi kullanımının avantajı doğrusal olmayan yani karmaşık ilişkileri belirlemektedir. Diyabet gibi hastalarda, risk faktörleri tek tek değil birbirleriyle birlikte etkileşim içinde çalışmaktadır. Çok sayıda öznitelik arasında bile seçim yaparken güvenilir sonuçlar elde etmemizi sağlar. Birden fazla karar ağaçlarını birleştirerek sonuçlar üretir. Her bir özelliğe atanan puan, karar ağacı uygulamalarındaki başarısının değerlendirilmesinden kaynaklanır [15].

Karar ağaçlarındaki düğümler, her bölmeden elde edilen veri saflığı sonuçlarını azaltan öznitelik etkinliğini hesaplamak için kullanılan puanlama metriklerini belirler.

$$I_j = \frac{1}{N_T} \sum_{T} \sum_{t \in T: v(s_t)=j} p;(t) \Delta i(s_t, t)$$

N_T: Ağaç sayısı.

T: Random Forest'taki tüm ağaçlar.

t : Bir düğüm.

$v(s_t)$: t düğümünde bölünme için kullanılan değişken.

p(t): t düğümüne ulaşan örneklerin alanı.

$\Delta i(s_t,t)$: t düğümündeki saflık artışı

3.3. Performans Ölçüm Metrikleri

Makine öğrenmesi model taleplerinin sağlamasını yapmak ve özellikle sınıflandırma problemlerinde farkında olmak için çeşitli metriklere yer verilmiştir. Bu metrikler ile model başarısını ölçme işlemimiz sağlanmaktadır [16]. Ölçümler sonucunda karşılaştırma yapabilme avantajı vardır. Çalışmada kullanılan temel sınıflandırma metrikleri:

- **Doğruluk (Accuracy):** Modeli doğru sınıflandırdığı örnek sayısının toplam örnek sayısına oranıdır. Genel performansı ölçmek için sıkça kullanılır ama dengesiz veri setlerini ölçerken hatalı sonuçlar verebilmektedir.

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN}$$

	1	0
1	TP	FP
0	FN	TN

TP: Modelin pozitif olduğunu düşündüğü ve gerçekte pozitif olan örnekler.

TN: Modelin negatif olduğunu düşündüğü ve gerçekte negatif olan örnekler.

FP: Modelin pozitif olduğunu düşündüğü ve gerçekte negatif olan örnekler.

FN: Modelin negatif olduğunu düşündüğü ve gerçekte pozitif olan örnekler.

- **F1- Skoru (F1-Score):** Kesinlik ve duyarlılık arasındaki dengeyi ölçen ortalamadır. Dengesiz veri setlerinde de güvenilirdir.

$$F1 - Skoru = 2 \times \frac{Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık}$$

- **ROC Eğrisi ve AUC (Area Under Curve):** ROC eğrisi farklı eşik değerleri için doğru pozitif oranına karşı yanlış pozitif oranını gösterir. AUC, ROC eğrisinin altında kalan alanı ifade eder ve modelin ayrıştırma yeteneğini ölçer.
- **Karşılaştırma matrisi:** Modelin tahminlerinin gerçek sınıflara göre dağılımını gösteren matris.

3.4. Kullanılan Sınıflandırma Algoritmaları

3.4.1. Rastgele Orman

Çok sayıda karar ağacının birleşimi ile oluşmaktadır. Ağaçların tahminlerinin birleşmesine dayanan algoritmadır. Random Forest bireysel karar ağaçlarının zayıf yönlerini azaltmanın yanında modelin genelleme kapasitesini çoğaltmak amacı da vardır.

İlk adımı veri setinden rastgele ve tekrarlı örnekler çekilir. Her bir karar ağacı bu örnekleri kullanarak eğitilir. Düğüm kısımlarında tüm öznitelikler yerine rastgele seçilmiş bir alt küme kullanarak ağaçlar arasındaki korelasyon azaltılıp çeşitlilik arttırılmış olur. Bu şekilde aşırı öğrenme önlenir. Karar ağaçları seçilen örneklerle ve rastgele öznitelik alt kümeleri ile budama yapılmadan büyütülür ve her ağaç kendi öğrenme yolunu geliştirmiş olur [12].

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

T: Bir düğümdeki örneklem.

C: Sınıf sayısı.

P_i: Düğümdeki i.sınıfa ait örneklerin oranı

3.4.2. Lojistik Regresyon

İstatistiksel bir modeldir. Bağımsız değişkenler ve bağımlı değişken (sınıf etiketi) arasındaki ilişkiyi lojistik fonksiyonu kullanarak modeller. Temel amacı bir olayın gerçekleşme olasılığını tahmin eder.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}}$$

P(Y=1|X): Bağımsız değişkenler X verildiğinde, bağımlı değişkenin 1 sınıfına ait olma olasılığı.

$\beta_0, \beta_1, \ldots, \beta_n$: Modelin öğrenmeye çalıştığı katsayılardır.

$X_1, X_2, \ldots, X_n$: Bağımsız değişkenler.

Lojistik Regresyon'un öğrenme süreci, genellikle **maksimum olabilirlik (likelihood) tahmini** yöntemine dayanır. Amaç; gözlemlenen veriye en uygun model parametrelerini (katsayıları) bulmaktır. Bu süreçte kullanılan maliyet fonksiyonu, **logaritmik kayıp** olarak adlandırılır.

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} \log\left(h_\beta\left(X^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\beta\left(X^{(i)}\right)\right)\right]$$

m: Toplam örnek sayısı.

$y^{(i)}$: i. Gözlemin gerçek sınıf etiketi.

$h_\beta(x^{(i)})$: Modelin. gözlem için tahmin ettiği olasılık.

β: Modelin parametre vektörüdür.

3.5. Modelleme Aşamaları

3.5.1. Veri Ön İşleme Katmanı

Bu kısım, "çöp içeri, çöp dışarı" (garbage in, garbage out) prensibini önlemeyi amaçlanmaktadır. Yöntem sayesinde ham veri, modelleme ile temiz ve yapılandırılmış bir formata dönüştürülür.

- **Otomatik Değişken Tipi Tespiti:** "***grab_col_name()***" fonksiyonu aracılığı ile veri setindeki sütunları otomatik olarak kategorize eder. Veri kategorik ve sayısal olmak üzere iki gruba ayrılmaktadır. İşlem her bir değişkene göre doğru bir şekilde uygulamayı garanti altına almış olur. (Örn: sayısal için ölçeklendirme, kategorik için kodlama.)
- **Aykırı Değer Yöntemi:** Aykırı değerler, modelin performansını ve genelleme kabiliyetini olumsuz etkileyebilir. Bu sebeple gerekli kontroller yapılmalıdır. Projede, aykırı değerlere karşı istatistiksel olarak dirençli olan **çeyrekler arası genişlik (IQR)** yöntemi kullanıldı. Yöntem, verinin birinci çeyreklik (Q1) ve üçüncü çeyreklik (Q3) değerlerini, verilen yüzdeliklere (Projede kullandığımız: Q1=0.10, Q3=0.90) göre hesaplar.

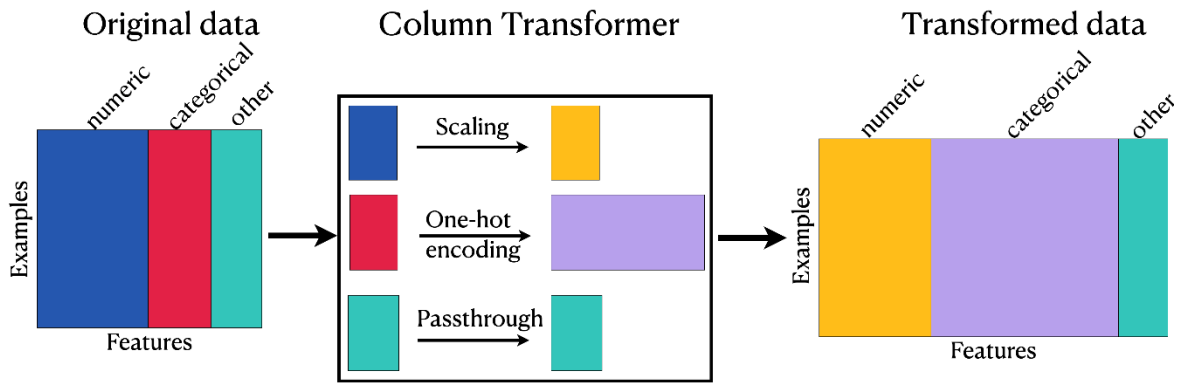
$$AltSınır = Q1 - 1.5 * IQR$$

$$ÜstSınır = Q3 + 1.5 * IQR$$

  Bu sınırın dışında kalan değerler, aykırı kabul edilir ve sınır değerleri ile değiştirilir. Bu yaklaşım, silme yöntemine göre daha koruyucudur.

- **Öznitelik Ölçeklendirmesi:** Sayısal öznitelikler genellikle farklı birimlere ve ölçeklere sahiptir (örn.: yaş ve kan basıncı). Bu durum, özellikle mesafe tabanlı inişi kullanan algoritmaların performansını olumsuz etkiler. Bu sorunun çözümü için "StandartScaler" kullanıldı
    - **StandardScaler:** Öznitelikler ortalamasını 0 ve standart sapma 1 olacak şekilde dönüştürür. Algoritmaların verinin normal dağıldığı varsayımına dayandığı durumlarda etkilidir.
- **Kategorik Değişken Kodlaması:** Makine öğrenmesi algoritmaları matematiksel denklemlerle çalıştığı için metin tabanlı kategorik verileri işleyemez. Bu modül, "one-hot encoding" tekniğini kullanarak her bir kategoriyi ikili (binary) bir sütuna dönüştürür. Özellikle, doğrusal modellerde (örn.: Lojistik Regresyon) mükemmel çoklu doğrusallık (multicollinearity) problemine yol açan "dummy variable trap" tuzağını önlemek için "*drop_first=True"* parametresini destekler.

Bu veri ön işleme aşağıdaki Şekil 4'de bu süreç görsel olarak özetlenmiştir.



Şekil 4: Veri ön işleme adımları [17]

3.5.2. Öznitelik Seçim Algoritması

- **Spearman Sıra Korelasyonu Tabanlı Seçim**
  - **Teorik Arka Plan:** Spearman katsayısı (ρ), iki değişken arasındaki ilişkinin doğrusal olması gerekmez; yalnızca monotonik (sürekli artan veya sürekli azalan) olma durumunu ölçer ve daha geniş bir ilişki türü yelpazesini yakalamada daha esnek kılar. Hesaplama, ham değerler yerine bu değerlerin sıralamaları üzerinden yapılır.
  - **Uygulama Detayları:** Aynı değere sahip gözlemler için ortalama sıralama atayan bir mekanizma içerir, bu da sonuçların doğruluğunu artırır. Seçim aşamasında; her bir özelliğin hedef değişkenle olan Spearman korelasyonunun mutlak değeri hesaplanır ve bu değer belirlenen eşikten (threshold) büyük veya eşitse o öznitelik seçilir. İlişkinin yönü (pozitif veya negatif) değil, gücü önemli olduğu için mutlak değer kullanılır.
- **Karşılıklı Bilgi (Mutual Information) Tabanlı Seçim**
  - **Teorik Arka Plan:** Bir değişken hakkındaki bilginin, diğer bir değişken hakkındaki belirsizliği ne kadar azalttığını ölçer. Karşılıklı bilginin en büyük gücü doğrusal, karmaşık, periyodik veya parçalı gibi her türlü doğrusal olmayan ilişkiyi tespit edebilmesidir. Mutual Information, Shannon entropisi kavramına dayanır.
  - **Uygulama Detayları:** Değişkenlerin ortak ve marjinal olasılık dağılımlarını oluşturur ve base-2 logaritma kullanarak Shannon entropisini hesaplar. Yüksek Mutual Information skoruna sahip öznitelikler, hedef değişken hakkında daha fazla bilgi taşıdığı varsayılarak seçilir.

- **Random Forest Tabanlı Bilgi Kazancı Seçimi**
    - **Teorik Arka Plan:** Bu yöntem, "gömülü" (embedded) bir yöntemdir, yani öznitelik seçimi modelin kendi eğitim süreci içinde gerçekleşir. Bir karar ağacı oluşturulurken, her düğümde veriyi en iyi şekilde bölecek öznitelik ve bölünme noktası aranır. "En iyi" bölünme, genellikle bilgi kazancı ile ölçülür. Bölünme sonrası alt düğümlerdeki entropinin ne kadar azaldığını gösterir.
    - **Uygulama Detayları:** Bir Random Forest modelinde birçok karar ağacı bulunur. Özelliğin önem puanı, o özelliğin tüm ağaçlardaki bölünmelerde sağladığı ortalama bilgi kazancı olarak hesaplanır. Daha yüksek önem puanına sahip öznitelikler, modelin genel tahmin gücüne daha fazla katkıda bulunduğu için seçilir. Uygulama, sürekli değişkenler için olası tüm bölünme noktalarını değerlendirir ve bilgi kazancını maksimize eden eşiği arar.

3.5.3. Performans Değerlendirme ve Otomatik Optimizasyon Çerçevesi

- **Çok Amaçlı Optimizasyon Stratejisi:** Bu çerçevenin temel yeniliği, tek bir metriğe odaklanmak yerine, birbiriyle çelişebilen üç hedefi akıllıca dengeleyen çok amaçlı bir yaklaşım benimsemesidir. Bu, gerçek dünya problemlerinde "en iyi" modelin genellikle bir ödünleşim (trade-off) sonucu olduğu gerçeğini yansıtır.
- **Ağırlıklandırılmış Skorlama Fonksiyonu:** Bu stratejiyi hayata geçiren matematiksel araç, her bir aday öznitelik alt kümesini değerlendiren ağırlıklandırılmış skorlama fonksiyonudur.

$$Toplam\ Skor = (0.60 * Performans) + \left(0.25 * (1 - Özellik\ Oranı)\right) + 0.15 * \frac{Performans}{Özellik\ Oranı}$$

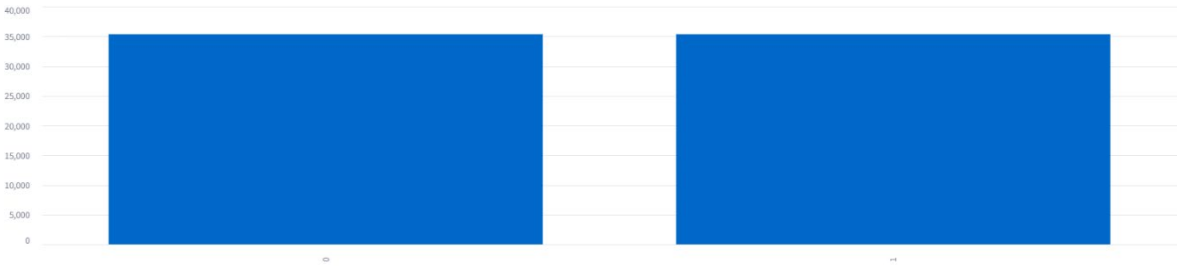
- Bu formül; her bir adayın performansını, verimliliğini (Öznitelik Oranı = seçilen öznitelik / toplam öznitelik) ve denge skorunu (Performans / Öznitelik Oranı) birleştirerek bütünsel bir "iyilik" puanı üretir.
- **Otomatik Arama ve Hata Yönetimi:** "***auto_optimize_threshold_selector***" fonksiyonu, her bir yöntem için belirlenmiş eşik aralıkları üzerinde iterasyon yapar. Bu süreç, istisna yönetimi (try-except blokları) ve sınır koşulu işleme (örn.: hiç öznitelik seçilmemesi durumu) gibi kapsamlı hata yönetimi mekanizmalarıyla donatılarak üretim ortamına hazır bir sağlamlıkta tasarlanmıştır.

3.6. Görsel Sunum Modülü

3.6.1. Keşifçi Veri Analizi (EDA) Modülü

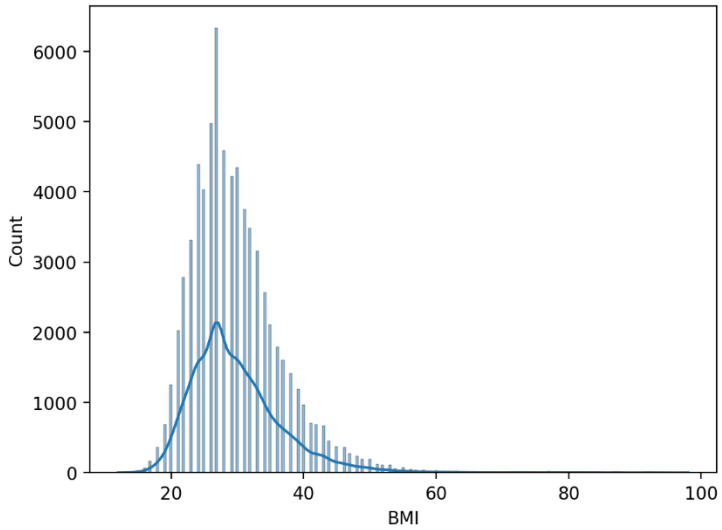
Kullanıcıya veri seti hakkında önemli bilgileri edinmesini sağlayan arayüzdür. Veri setinde bulunan özniteliklere ve veri setinin ilk beş satırını gözlemlenmesi sağlandı. Eksik veri analizi yapıldı. Bu analiz sonucunda eksik veri bulunamadı. Hedef değişkenin dağılımı kontrol edildi. Daha sonrasında öznitelikler kategorik ve sayısal olarak ayrıldı. Bu ayrım sayesinde her bir değişkenin dağılımını grafikler ile gösterimi yapıldı. Bu değişkenlerin dağılım grafikleri ve anlamaları:

- **Hedef Değişken Dağılımı:** Hedef değişkenin dağılımı kontrol edildi. Şekil 5'de görüldüğü gibi dengeli (%50,%50) dağılmıştır.



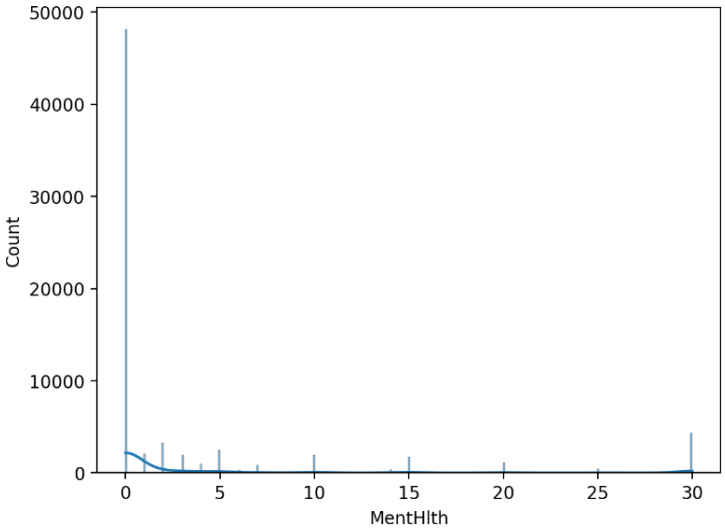
Şekil 5: Hedef değişkenin dağılım grafiği

- **BMI (Vücut Kitle İndeksi) Sayısal Değişkeninin Dağılımı:** Şekil 6'daki histogram grafiği veri setinde bulunan bireylerin vücut kitle indeks dağılımını göstermektedir. Histogram çubukları, belirli BMI aralıklarında kaç birey olduğunu gösterir. Üzerindeki mavi çizgi yani kde eğrisi dağılımın yoğunluk tahminini sunar. Grafikte görüldüğü gibi 20 ile 45 arasında yoğunluk bulunuyor. En yüksek frekansın yaklaşık 28 civarında olduğu söylenebilir. Bu durum, veri stindeki bireylerin çoğunluğunun normal kilolu (18,5-24,9) ile fazla kilolu (25-29,9) [9] kategorileri arasında yer aldığını gösterir. Grafiğin devamında az sayıda da olsa obez bireylerin olduğunu belirtebilirz. Yüksek BMI değerine sahip bireylerin az olması modelin bu grubu doğru tahmin etmesini zorlaştırabilir [18].



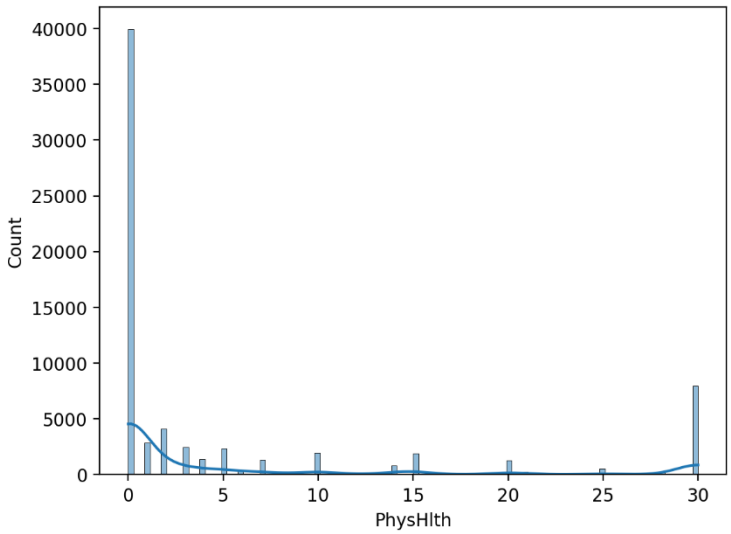
Şekil 6: BMI dağılımı

- **MentHlth (Ruh sağlığı Durumu) Dağılım Grafiği:** Şekil 7'deki histogram grafiği bireylerin son 30 gün içinde ruhsal sağlık sorunları yaşadıkları gün sayısını gösterir. Histograma göre veri setindeki kişilerin çoğunluğu son 30 gün içinde hiç ruh sağlığı problemi yaşamadıklarını gösterir. Az sayıda kişinin 30 gün boyunca sürekli bir ruh sağlığı problemi yaşadığını göstermektedir.



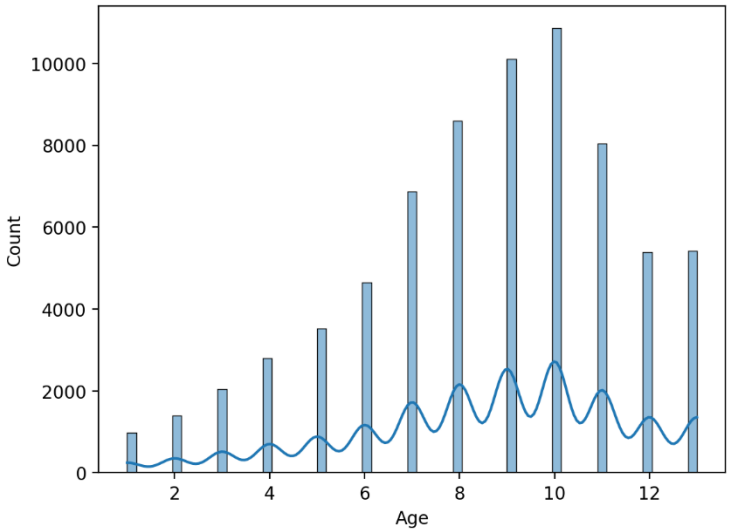
Şekil 7: MenHlth dağılımı

- **PhysHlth (Fiziksel Sağlık) Dağılım Grafiği:** Şekil 8'deki histogram grafiği bireylerin fiziksel sağlık sorunu hakkında bilgi vermektedir. Bireylerin büyük bir çoğunluğunun fiziksel sağlık sorunları yaşamadığı anlaşılmaktadır. 0'dan sonra değerler hızla azalıyor ve 1-5 arasında orta seviyede frekanslar görülüyor (2,000-5,000 arası). 5'ten sonra genel olarak düşük frekanslar var, ancak 30 değerinde tekrar bir yükseliş görülüyor (yaklaşık 8,000). Bu muhtemelen "30 gün" üst sınırını temsil ediyor.



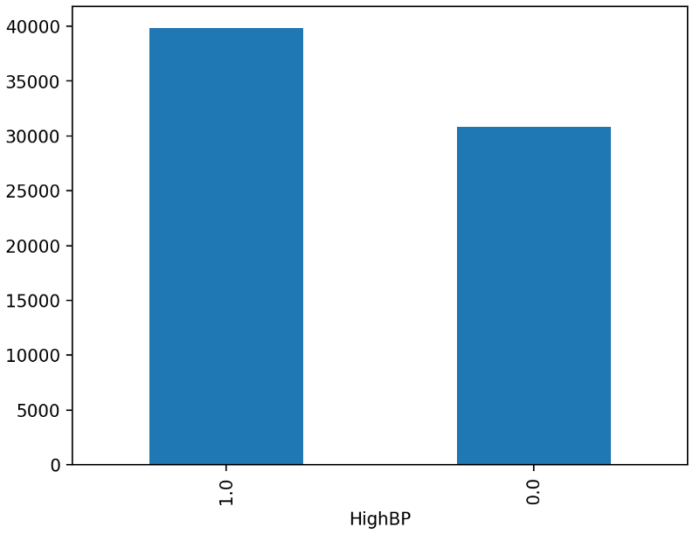
Şekil 8: PhysHlth dağılımı

- **Age (Yaş) Dağılım Grafiği:** Şekil 9'daki histogram grafiği kişilerin yaşları hakkında bilgi sunuyor. Bu değişken için belitrilken değerlerden bazılar; Kategori 1= 18-24, Kategori 9 = 60-64 ve Katagori 13 = 80 yaş ve üzeri. frekanslar 9-10 kategorilerinde (60-69 yaş grubu) yoğunlaşmış durumda. Bu, örneklemin büyük çoğunluğunun yaşlı nüfustan oluştuğunu ortaya koyuyor. Genç yaş grupları (kategori 1-5, yani 18-44 yaş arası) nispeten düşük temsil edilirken, orta yaş (kategori 6-8, 45-59 yaş) ve özellikle 60 yaş üzeri gruplar (kategori 9-13) örneklemin ana kısmını oluşturuyor. 80 yaş üzeri grupta (kategori 13) bile yaklaşık 5,500 kişi bulunuyor.



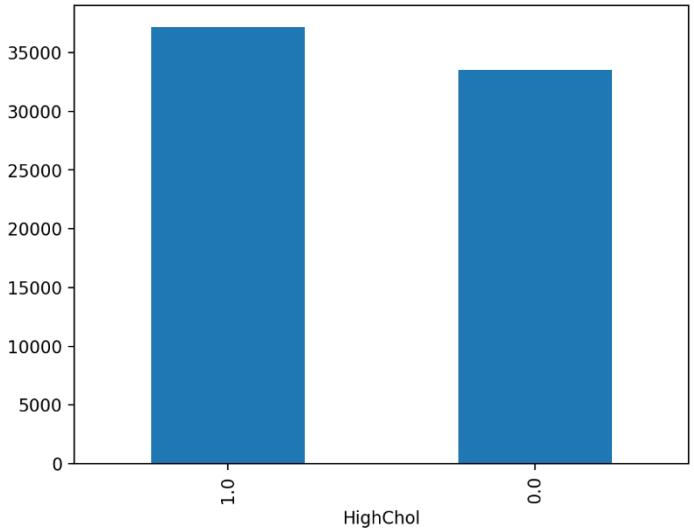
Şekil 9: Age dağılımı

- **HighBP (Yüksek Kan Basıncı) Dağılım Grafiği:** Şekil 10'daki grafik ile kategorik değişkenlere geçildi. Grafik 1 ve 0 dan oluşmaktadır. 1 ile yüksek kan basıncı bulunan bireylerin sayısını ve 0 ile yüksek kan basıncı bulunmayan bireylerin sayısını göstermektedir.



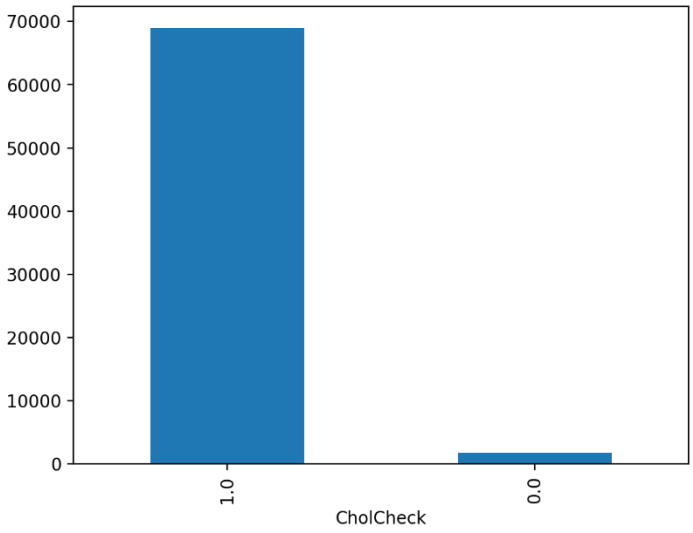
Şekil 10: HighBP dağılımı

- **HighChol (Yüksek Kolesterol) Dağılım Grafiği:** Şekil 11'deki grafikte yüksek kolesterol dağılımı görünmektedir. 1 değeri ile ifade edilmek istenen yüksek kolesterole sahip kişilerin sayısını ve 0 değeri ile yüksek kolesterolü olmayan kişilerin sayısını göstermektedir.



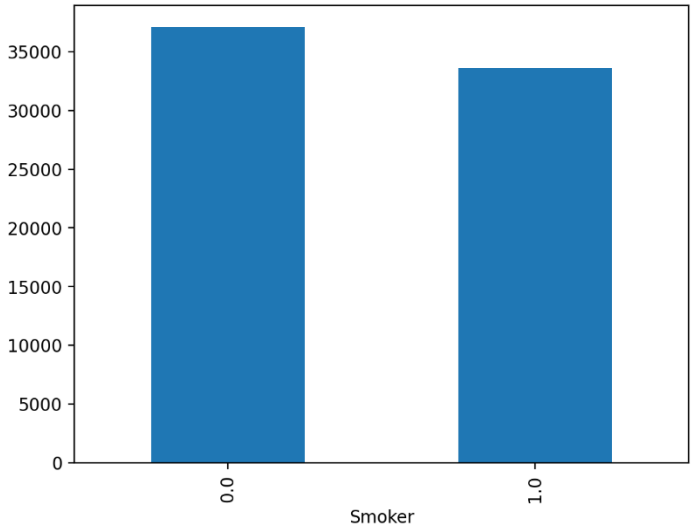
Şekil 11: HighChol dağılımı

- **CholCheck (Kolesterol Kontrolü) Dağılım Grafiği:** Şekil 12'deki grafik kolesterol kontrolü yaptırıp yaptırmadığı dağılımını ifade ediyor. Veri setindeki kişilerin çoğunluğu kolesterol kontrolü yaptırmıştır.1 ile kolesterol kontrolü yaptıranlar ve 0 ile yaptırmayanlar gözükmektedir.



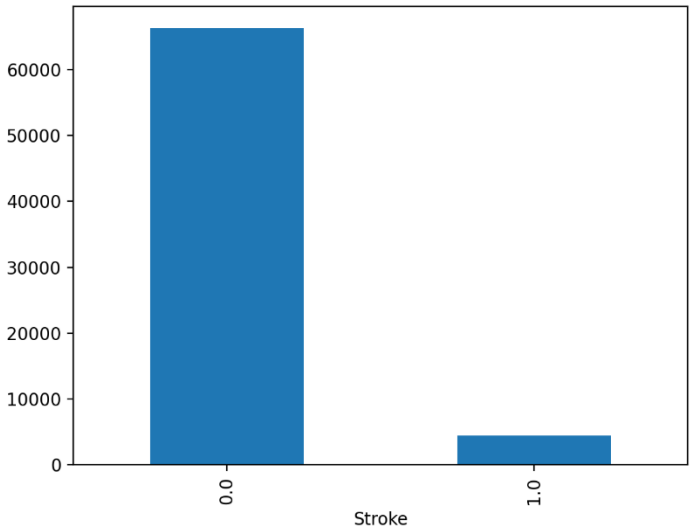
Şekil 12: CholCheck dağılımı

- **Smoker (Sigara Kulanım) Dağılım Grafiği:** Şekil 13'deki grafik sigara kullanıp kullanmayanların dağılımını göstermektedir. Dağılıma göre 1 ile sigara kulanan kişilerin sayısını ve 0 ile sigara kullanmayan kişilerin sayısını göstermektedir.



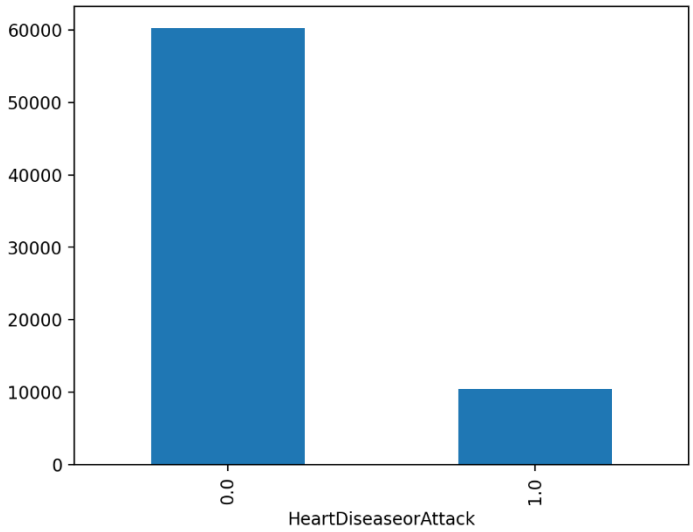
Şekil 13: Smoker dağılımı

- **Stroke (İnme Geçirme Durumu) Dağılım Grafiği:** Şekil 14'deki grafik inme geçirme durumuna göre dağılımıdır. 1 inme geçirmiş kişlerin sayısını gösterir ve 0 inme geçirmemiş kişilerdir.



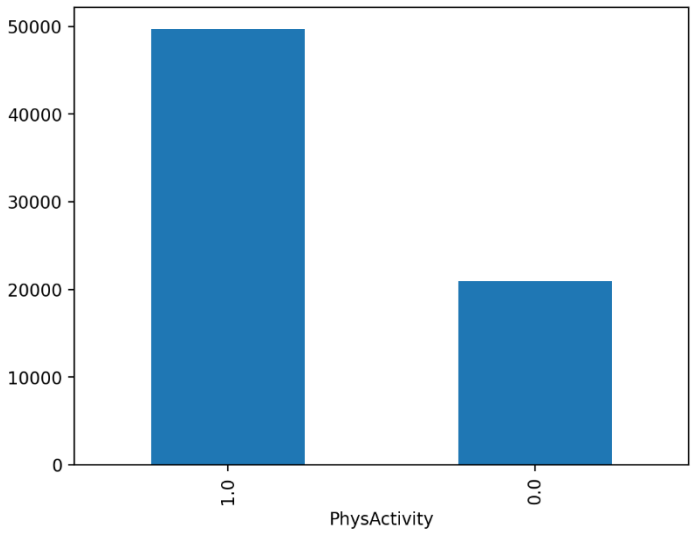
Şekil 14: Stroke dağılımı

- **HeartDiseasorAttack (Kalp Krizi Geçirme Durumu) Dağılım Grafiği:** Şekil 15'deki grafik kalp krizi geçirme durumunu gösterir. 1 ile kalp krizi geçirmiş kişi sayısını belirtir. 0 ile kalp krizi geçirmeyenleri gösterir.



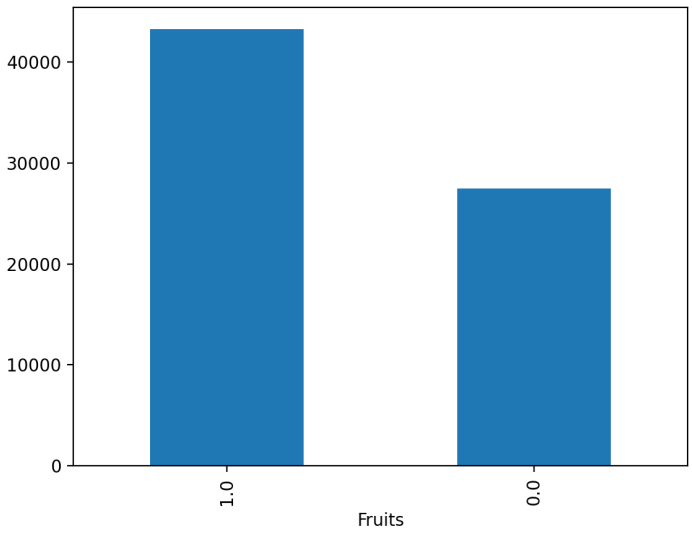
Şekil 15: HeartDiseasorAttack dağılımı

- **PhysActivaty (Fiziksel Aktivite Durum) Dağılım Grafiği:** Şekil 16'daki grafik fiziksel aktivite yapıp yapmadığının dağılımını gösterir. 1 ile fiziksel aktivite yapan kişilerin sayısını ve 0 ile fiziksel aktivite yapmayanların sayısını ifade eder.



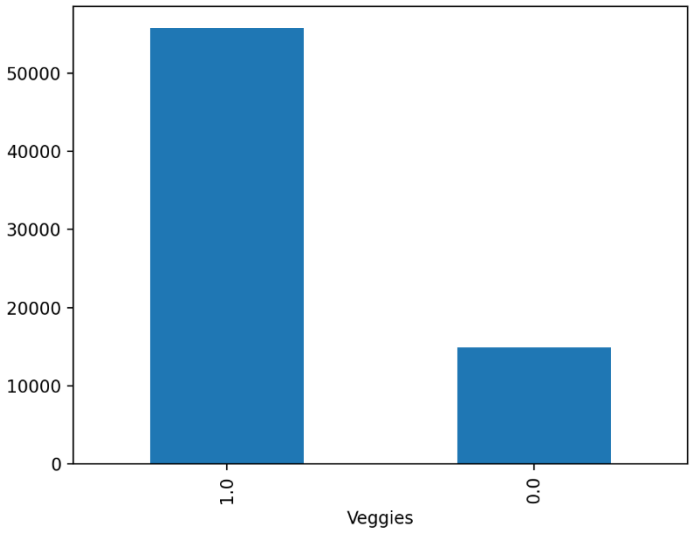
Şekil 16: PhysActivity dağılımı

- **Fruits (Meyve Tüketimi) Dağılım Grafiği:** Şekil 17'deki grafik meyve tüketimine göre dağılımını gösterir. 1 ile meyve tüketenlerin sayısını gösterir. 0 ile meyve tüketemeyenlerin sayısını ifade eder.



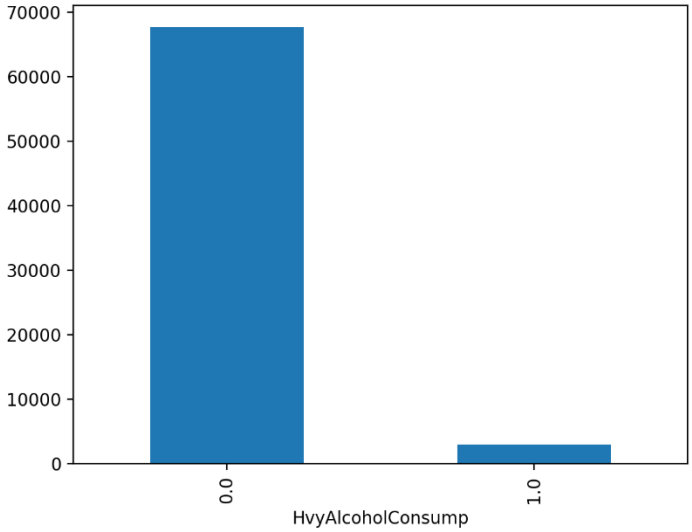
Şekil 17: Fruits dağılımı

- **Veggies (Sebze Tüketimi) Dağılım Grafiği:** Şekil 18'deki grafik bireylerin sebze tüketim alışkanlıklarını gösterir. 1 ile sebze tüketen bireyleri ve 0 ile sebze tüketmeyen bireyleri temsil etmektedir.



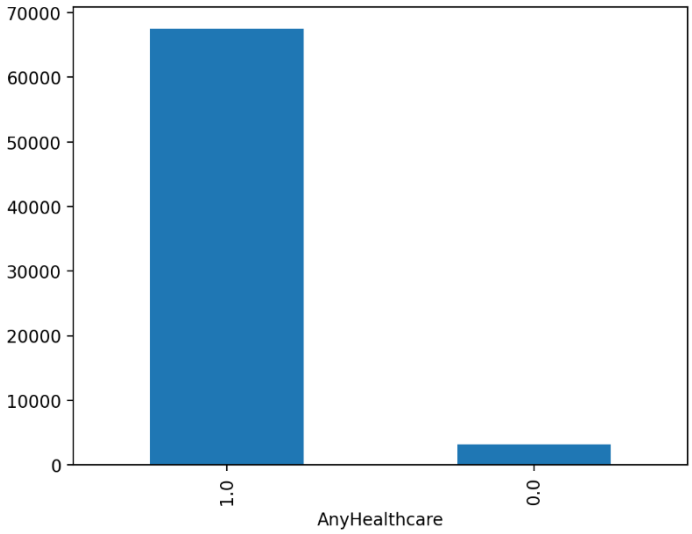
Şekil 18: Veggies dağılımı

- **HvyAlcoholConsump (Yoğun Alkol Tüketimi):** Şekil 19'daki grafik bireylerin yoğun alkol tüketim durumunu göstermektedir. 1 değeri yoğun alkol kullanan bireyleri temsil eder ve 0 değeri yoğun alkol tüketmeyen bireylerin sayısıdır.



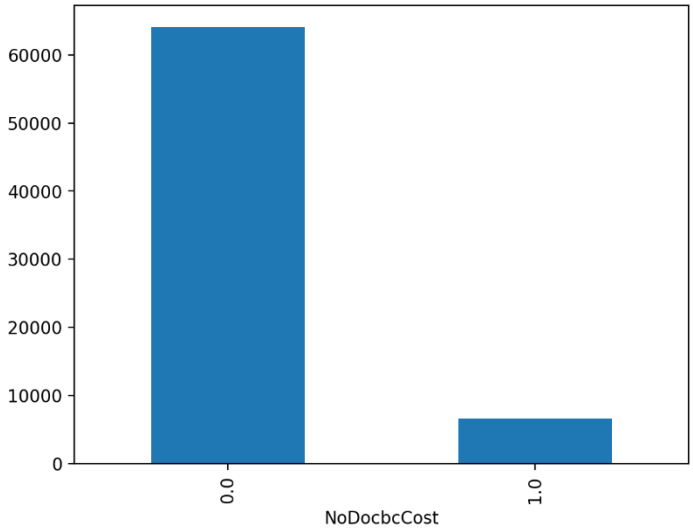
Şekil 19: HvyAlcoholConsump dağılımı

- **AnyHealthcare (Sağlık Hizmeti Alma Durumu) Dağılılm Grafiği:** Şekil 20'deki grafik bireylerin sağlık hizmetine erişim durumunu göstermektedir. 1 değeri ile sağlık hizmeti alan bireyleri ve 0 değeri ile sağlık hizmeti almayan bireylerin sayısı ifade edilmiştir.



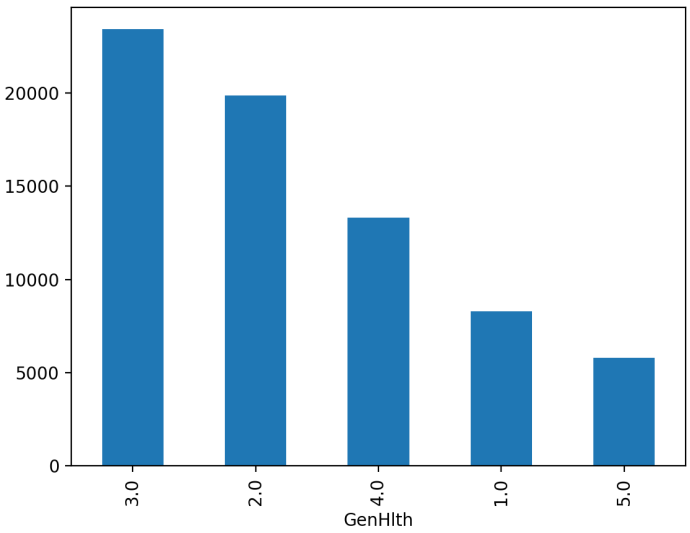
Şekil 20: AnyHealthcare dağılımı

- **NoDocbcCost (Maliyet Nedeniyle Doktora Gidememe) Dağılım Grafiği:** Şekil 21'deki grafikte doktora gidememe durumunu gösterir. 1 değeri doktora gidemediğini gösterir. 0 değeri ise doktora gitme durumunu ifade eder.



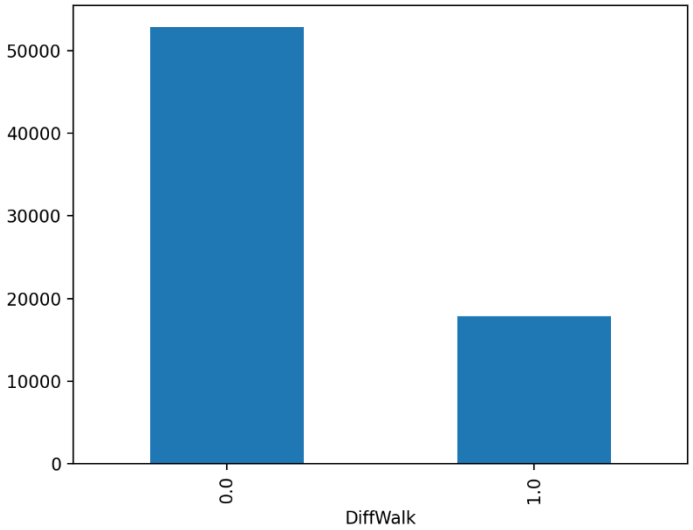
Şekil 21: NoDocbcCost dağılımı

- **GenHlth (Katılımcıların sağlık Durum) Dağılım Grafiği:** Şekil 22'deki grafik çalışma kapsamında incelenen örneklemdeki bireylerin genel sağlık durumlarına (GenHlth) ilişkin öz bildirimlerinin frekans dağılımı gözükmektedir. "Genel olarak sağlığınızın nasıl olduğunu söylersiniz?" sorusuna verilen yanıtlar, 1'den 5'e kadar bir Likert ölçeği kullanılarak değerlendirilmiş olup; burada 1=Mükemmel, 2=Çok İyi, 3=İyi, 4=Orta ve 5=Kötü sağlık durumunu ifade etmektedir.



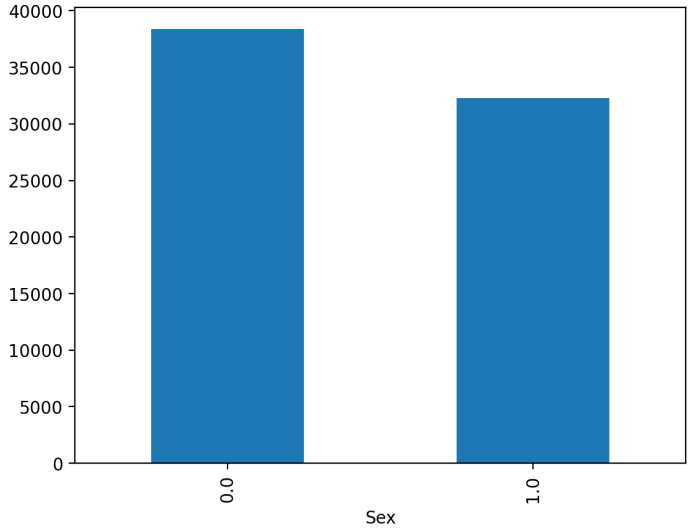
Şekil 22: GenHlth dağılımı

- **DiffWalk (Yürüme Güçlüğü) Dağılım Grafiği:** Şekil 23'deki grafikte bireylerin yürüme güçlüğü çekip çekmediği durumunu gösterir. 1 değeri ile yürüme güçlüğü olduğu ve 0 değeri ise yürüme güçlüğü olmadığını ifade eder.



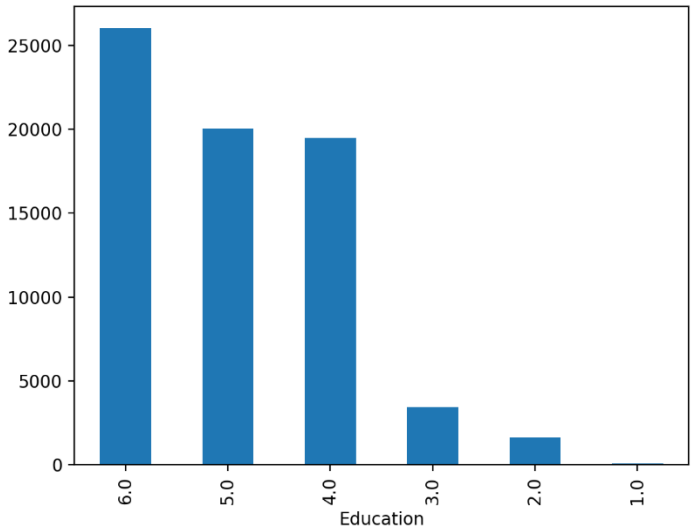
Şekil 23: DiffWalk dağılımı

- **Sex (Cinsiyet) Dağılım Grafiği:** Şekil 24'deki grafik bireylerin cinsiyet durumunu gösterir. 1 değeri ile erkeklerin sayısını ve 0 değeri ile kadınların sayısı gösterilmektedir.



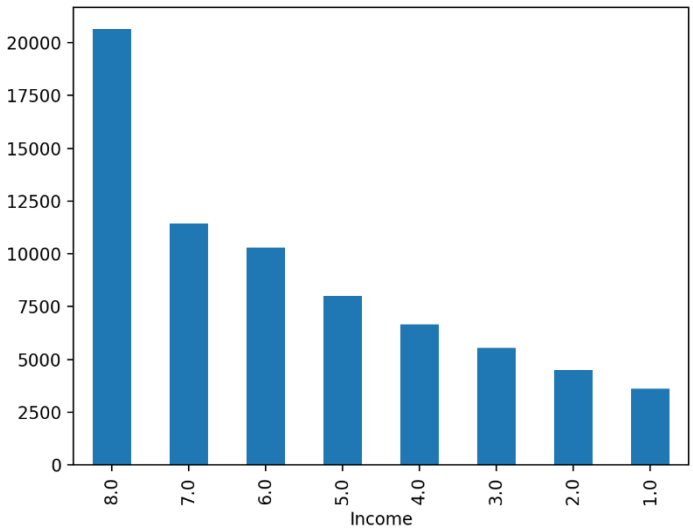
Şekil 24: Sex dağılımı

- **Education (Eğitim Düzeyi) Dağılım Grafiği:** Şekil 25'deki grafik bireylerin eğitim düzeyi hakkında bilgi vermektedir. 1'den 6'ya kadar bir ölçekte değerlendirilmiş olup; burada 1değeri ile okula hiç gitmemiş veya sadece anaokulu, 2 değeri ile ilkokul mezunu anlamına gelmektedir. Diğer değerler için veri setinde açıklama yapılmamıştır. En yüksek değer yani 6 için üniveriste veya daha üst bir eğitim diyebiliriz. En fazla birey 6 değerinde bulunmaktadır.



Şekil 25: Education dağılımı

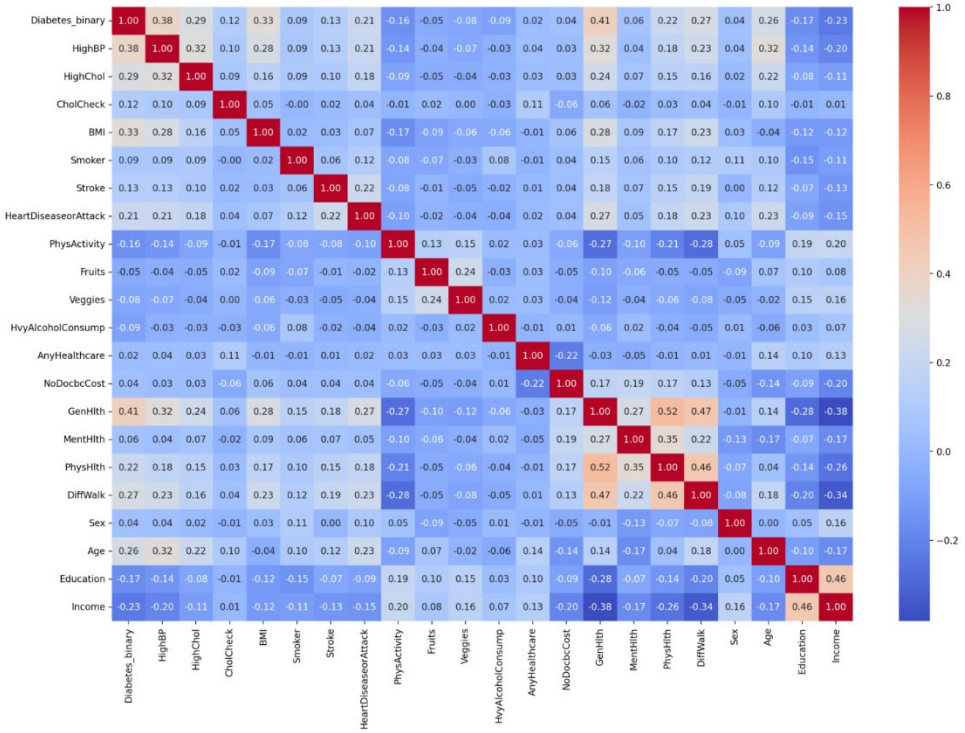
- **Income (Yıllık Gelir) Dağılım Grafiği:** Şekil 26'daki grafikte bireylerin yıllık gelir dağılımı bulunmakta. Gelir düzeyleri 1'den 8'e kadar kategorize edilmiştir. Veri seti açıklamasında 1 değeri ile 10.000'den daha az, 5 değeri ile 35.000'den az ve 8 değeri ile 75.000 veya daha fazla yıllık gelire sahip olunduğu söyleniyor.



Şekil 26: Income dağılımı

Şekil 27'de, keşifçi veri analizi sayfasında gözüken değişkenler arasında Spearman Korelasyon katsayıları ısı haritası olarak gösterilmiştir. Isı haritası, değişken çiftleri arasındaki doğrusal olmayan sıralı ilişkileri ölçen korelasyon değerlerini içerir. Korelasyon katsayıları -1 ile +1 arasında değerler alır. Kırmızı tonları pozitif korelasyon değerlerini. Mavinin tonları ise negatif korelasyon değerlerini göstermektedir. Renklerin yoğunluğu ilişkinin gücünü belirtmektedir [19]. Diabetes_binary değişkeni ile en yüksek pozitif korelasyona sahip değişkenler:

- GenHlth = 0,41,
- HighBp = 0,38,
- BMI = 0,33,
- HighChol = 0,29 ve
- DiffWalk =0,27 şeklidir

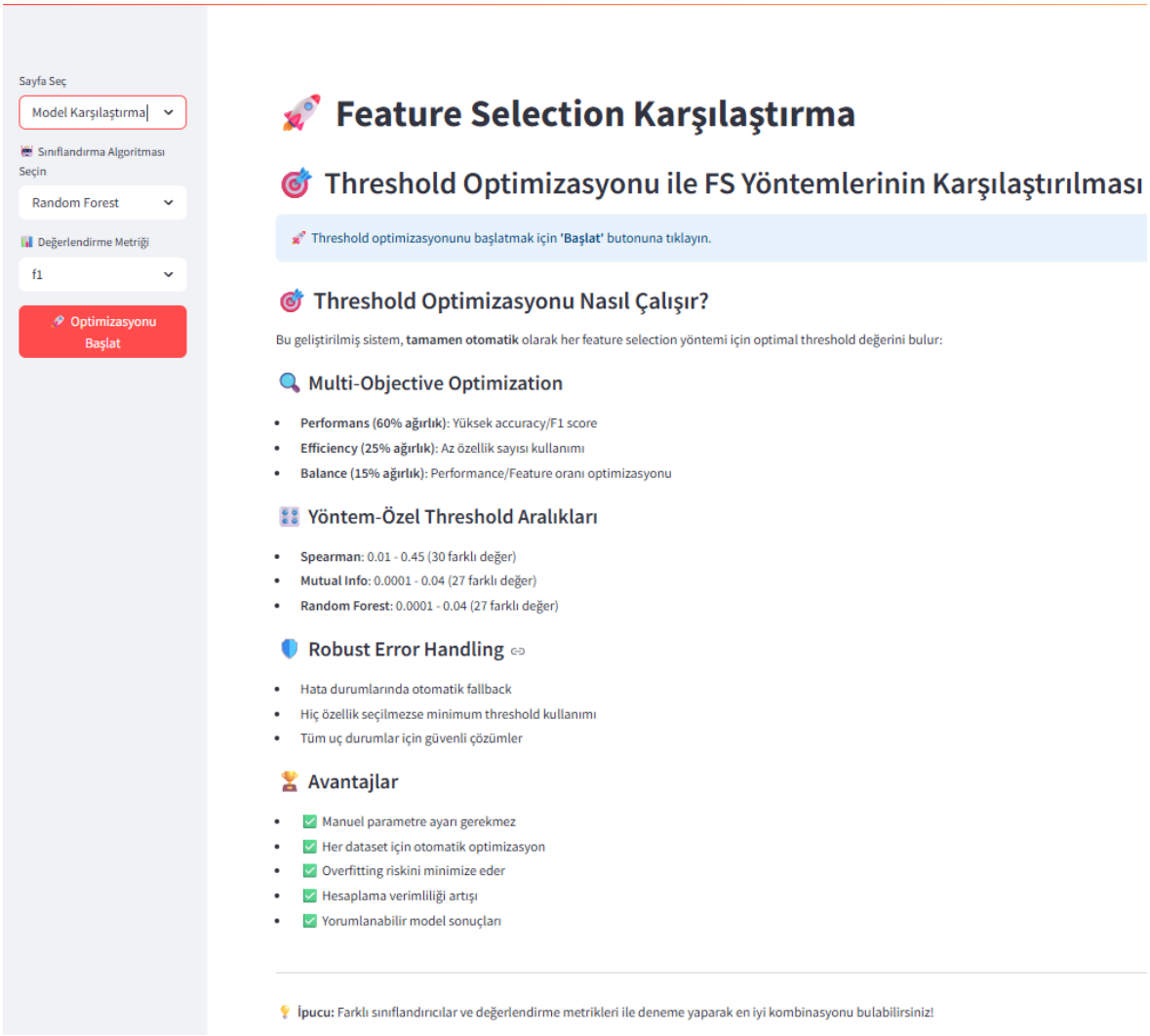


Şekil 27: Isı haritası

3.6.2. Model Karşılaştırma Sayfası

Sayfada kuullanıcıdan sınıflandırma algoritması ve değerlendirme metriği seçmesi istenilir. Sınıflandırma olarak Random Forest veya Lojistik Regresyon arasından seçim yapabilir, başarı metriği olarak F1-Skoru veya doğruluğu tercih edebilir. Analiz başlatıldığında, optimizasyon süreci bir ilerleme çubuğu ile gerçek zamanlı olarak takip

edilebilir. Sonuçlar, farklı yöntemlerin performansını doğrudan karşılaştıran görseller sunar. Şekil 27'de arayüz bulunmaktadır. Kullanılan analiz yöntemleri.



Şekil 28: Arayüz

- **ROC Eğrisi Analizi:** Farklı modellerin true positive rate ve false positive rate arasındaki ödünleşimini karşılaştırır ve AUC (Area Under Curve) değeri ile genel performanslarını özetler.
- **Karmaşıklık Matrisi (Confusion Matrix):** Modellerin hangi sınıflarda ne tür hatalar yaptığı detaylı olarak gösterir.
- **Öznitelik Önem Grafikleri:** Her bir yöntemin seçtiği en iyi öznitelik alt kümesini ve bu özniteliklerin önem skorlarını gösteren bar grafikleri sunar.

3.7. Projenin İşleyişi

Çalışma akışı, veri setinin yüklenmesinden, en uygun ve verimli modeli bulunmasına kadar geçen süreci kapsayan mantıksal bir sırayı takip eder. Bu aşamalar aşağıda bulunmaktadır.

- **Aşama 1:** Verinin yüklenmesi ve otomatik hazırlığı denilmesi daha doğrudur. Veri setinin dosya yolu koda tanımlanarak aktarılır. Bu işlem sayesinde ham veri alınır; sütun tipleri belirlenir, eğer veri setinde aykırı değer varsa sınır içine alınır, sayısal değişkenler standartlaştırılır ve kategorik öznitelikleri makine öğrenmesine uygun ikili formata dönüştürülür. Bu şekilde otomatikleştirerek veriyi daha hızlı temizleme işlemini gerçekleştirmemizi sağlamıştır.
- **Aşama 2:** Ön işlenmiş veri; keşifçi veri analizi modülünde görselleştirmeye hazır hale getirir. Burada veri setinin temel dinamiklerini inceler. İkinci sayfa olan model karşılaştırma sayfasında asıl hedefimiz olan sınıflandırma algoritmalarında öznitelik seçimi yöntemlerinin uygulamalı karşılaştırması kısmıdır. Bu sayfada "Hangi sınıflandırma modeli (Random Forest, Lojistik Regresyon) kullanılacak" ve "Başarı hangi metrikle (F1-skoru, doğruluk) ölçülecek" sorularının yanıtlarını arayüz üzerinden seçer.
- **Aşama 3:** Başlat butonuna basıldığında "***auto_optimize_threshold_selector***" devreye girer. Bu fonksiyon, her bir öznitelik seçim yöntemine göre önceden belirlemiş olduğumuz eşik listesi üzerinde bir döngü başlatılır. Her bir eşik değeri için şu adımlar sırasıyla ve otomatik olarak gerçekleşir:
    - "***spearman_selector_auto(X, y, threshold=thresh)***" fonksiyonu çağrılır ve o eşiğe göre bir öznitelik alt kümesi *(X_sel)* ve önem skorları *(scores_dict)* elde edilir.
    - Sistem, seçilen öznitelik sayısının sıfır olup olmadığını kontrol eder. Eğer sıfırsa, bu iterasyon atlanır.
    - Elde edilen (*X_sel*) ve "y" verisi, "***random_state=42***" kullanılarak %70 eğitim ve %30 test verisi olarak deterministik bir şekilde bölünür. Bu sabit "***random_state***" kullanımı, deneylerin tekrarlanabilirliğini sağlar.
    - Kullanıcının seçtiği sınıflandırma modeli, yalnızca eğitim verisinin seçilmiş öznitelikleri *(X_train)* üzerinde eğitilir.
    - Model, test verisi *(X_test)* üzerinde tahminler *(y_pred)* yapar ve performans metriği hesaplanır.

- Son olarak, bu performans değeri, kullanılan öznitelik oranı ve denge oranı, çok amaçlı skorlama fonksiyonuna beslenerek eşik değeri için nihai *(total_score)* hesaplanır.
- Bu döngüdeki her bir sonucun *(threshold, performance, feature_count total_score gibi)* bir sözlük (dictionary) yapısında geçici olarak saklanması, sürecin sonunda en iyi sonucun kolayca bulunmasını sağlar.

- **Aşama 4:** Sonuçların karşılaştırmalı analizi. Tüm eşik değerlerini denedikten sonra, sistem en yüksek *total_score* değerine sahip olan iterasyonu "en iyi" olarak belirler. Arayüzde en iyi sonucun detaylarını ve diğer yöntemlerin en iyi sonuçlarıyla olan karşılaştırmasını sunar. Kullanıcı, farklı yöntemlerin ROC eğrilerini tek bir grafikte görerek, hangi modelin farklı eşiklerde daha istikrarlı bir performans sergilediğini analiz edebilir. Bu bütünsel sunum, en iyi teknik çözümü bulmanın yanı sıra, farklı öznitelik seçimi felsefelerinin problem üzerindeki etkilerini anlamak için de güçlü bir analitik araçtır.

4. BULGULAR

Seçilen öznitelik seçim yöntemleri sırasıyla faklı threshold değerleri ile otomatik test edildi. Kullanılan sınıflama algoritması ve değerlendirme metriğine göre Spearman korelasyonu 30 farklı eşik değeri ile test edildi. Her algoritma ve metriğe göre en iyi 5 threshold değer tabloları aşağıdaki gibidir.

Tablo 1: Random Forest F1-skoru Spearman korelasyonu en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.4	0.7045	20	0.7053
0.35	0.7276	24	0.6647
0.3	0.7417	28	0.6242
0.25	0.7456	29	0.6153
0.2	0.749	31	0.5947

Tablo 2: Random Forest accuracy Spearman korelasyonu en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.4	0.7058	20	0.7064
0.35	0.7194	24	0.658
0.3	0.732	28	0.6167
0.25	0.7355	29	0.6074
0.2	0.7391	31	0.5871

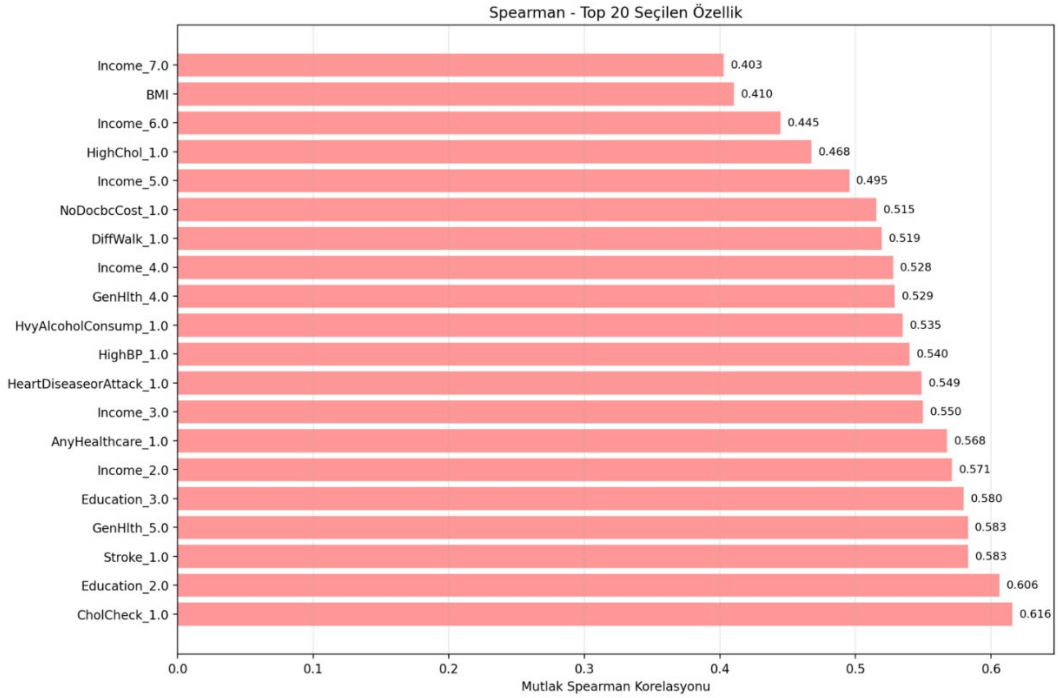
Tablo 3: Lojistik Regresyon F1-skor Spearman korelasyon en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.4	0.7339	20	0.7304
0.35	0.7553	24	0.6872
0.3	0.754	28	0.6339
0.25	0.7544	29	0.6221
0.2	0.7554	31	0.5996

Tablo 4: Lojistik Regresyon accuracy Spearman korelasyon en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.4	0.7328	20	0.7295
0.35	0.7502	24	0.6831
0.3	0.7491	28	0.63
0.25	0.7498	29	0.6185
0.2	0.7503	31	0.5957

Bulunan threshold değerleri ile 34 öznitelikten 20 öznitelik seçilmiştir. Seçilen öznitelikleri Şekil 29'da görebilirsiniz.



Şekil 29: Spearman korelasyonu ile seçilen özniteliklerin listesi

Random Forest öznitelik önemini 27 farklı threshold değeri ile test ederek en az özniteliklerle en iyi performansı vermesi hedeflendi. Her algoritma ve metriğe göre en iyi 5 threshold değer tabloları aşağıdaki gibidir.

Tablo 5: Random Forest F1-skoru Random Forest en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.04	0.7242	5	1.3865
0.035	0.7181	7	1.1526
0.03	0.7064	9	1.008
0.025	0.7052	10	0.9592
0.02	0.7062	11	0.9203

Tablo 6: Random Forest accuracy Random Forest en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.04	0.7147	5	1.371
0.035	0.7138	7	1.1468
0.03	0.7045	9	1.0057
0.025	0.7017	10	0.9553
0.02	0.7027	11	0.9165

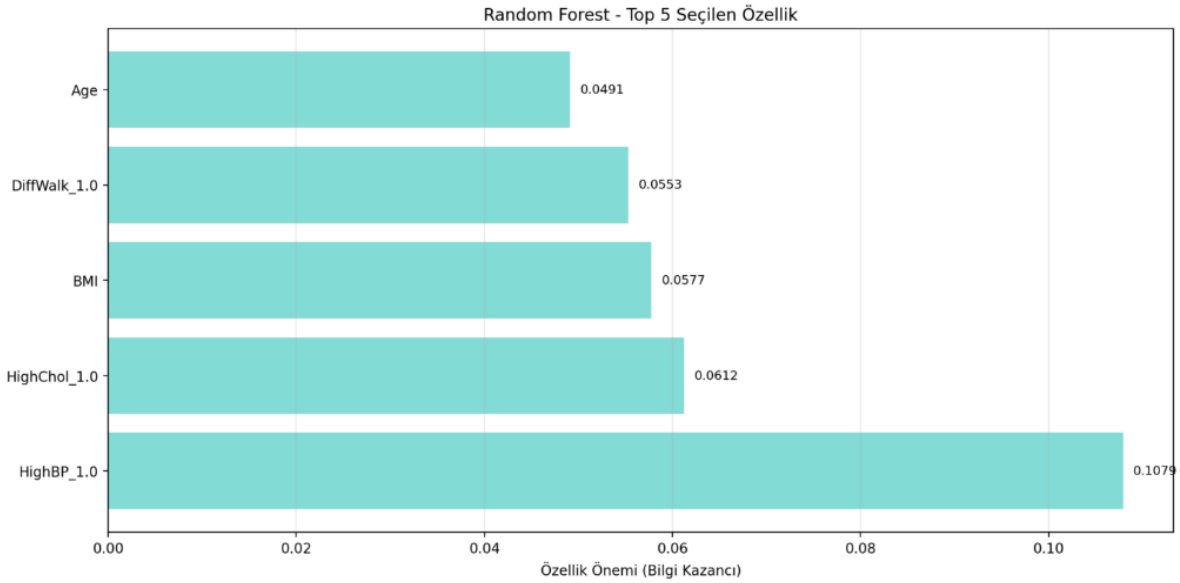
Tablo 7: Lojistik Regresyon F1-skor Random Forest en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.04	0.7292	5	1.3945
0.035	0.7399	7	1.1815
0.03	0.7426	9	1.0501
0.025	0.7407	10	0.9987
0.02	0.7432	11	0.9596

Tablo 8: Lojistik Regresyon accuracy Random Forest en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.04	0.7233	5	1.3849
0.035	0.735	7	1.175
0.03	0.7384	9	1.0453
0.025	0.7367	10	0.9942
0.02	0.7398	11	0.956

Belirlenen threshold değerleri ile 34 öznitelikten 5 öznitelik seçilmiştir. Seçilen öznitelikleri Şekil 30'da görebilirsiniz.



Şekil 30: Random Froest ile seçilen özniteliklerin listesi

Mutual Information öznitelik önemini 27 farklı threshold değeri ile test ederek en az özniteliklerle en iyi performansı vermesi hedeflendi. Her algoritma ve metriğe göre en iyi 5 threshold değer tabloları aşağıdaki gibidir.

Tablo 9: Random Forest F1-skoru Mutural Information en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.04	0.7242	5	1.3865
0.035	0.7049	8	1.0635
0.03	0.7064	9	1.008
0.025	0.7052	10	0.9592
0.02	0.7062	11	0.9203

Tablo 10: Random Forest accuracy Mutual Information en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.04	0.7147	5	1.371
0.035	0.702	8	1.0599
0.03	0.7045	9	1.0057
0.025	0.7017	10	0.9553
0.02	0.7027	11	0.9165

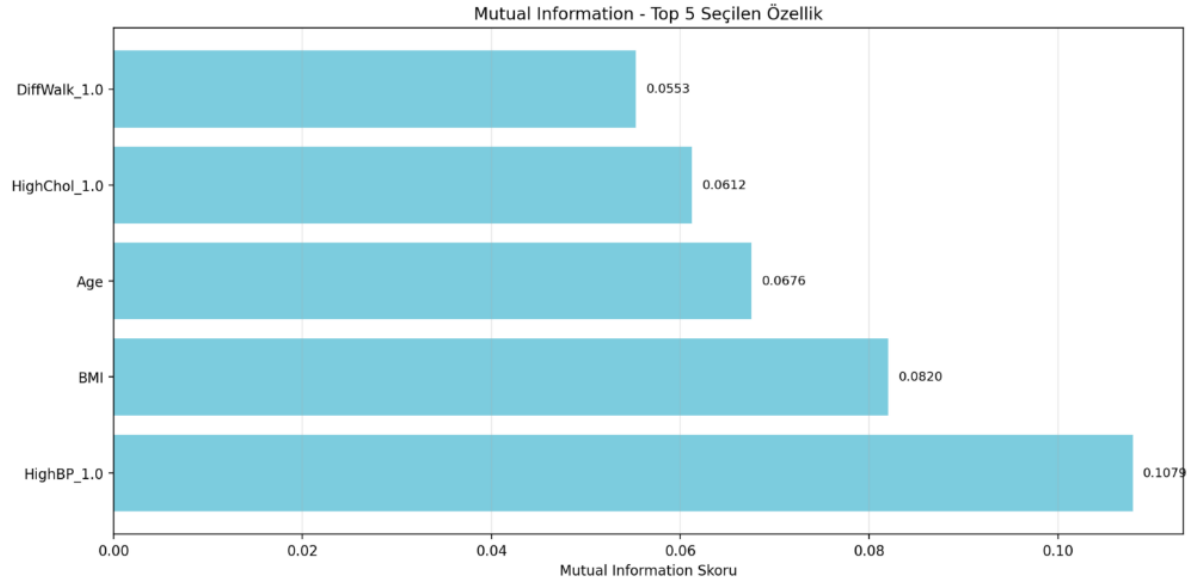
Tablo 11: Lojistik Regresyon F1-skor Mutual Information en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.04	0.7292	5	1.3945
0.035	0.7389	8	1.1056
0.03	0.7426	9	1.0501
0.025	0.7407	10	0.9987
0.02	0.7432	11	0.9596

Tablo 12: Lojistik Regresyon accuracy Mutual Information en iyi 5 threshold

Threshold	Performans	Öznitelik Sayısı	Toplam Skor
0.04	0.7233	5	1.3849
0.035	0.7342	8	1.0997
0.03	0.7384	9	1.0453
0.025	0.7367	10	0.9942
0.02	0.7398	11	0.956

Belirlenen threshold değerleri ile 34 öznitelikten 5 öznitelik seçilmiştir. Seçilen öznitelikleri Şekil 30'da görebilirsiniz.



Şekil 31: Mutual Information ile seçilen özniteliklerin listesi

En etkili öznitelik seçimi yöntemi, yapılan karşılaştırmalar sonucunda Spearman korelasyonu olmuştur. Spearman korelasyonu, iki değişken arasındaki monotonik ilişkiyi değerlendirerek sıralama temelli güçlü bağıntılar bulur. Bu yaklaşım, doğrusal olmayan fakat sıralı korelasyonları da yakalayabildiği için, sağlık verileri gibi karmaşık ve kısmen kategorik yapılarda daha isabetli sonuçlar vermektedir. Özellikle öznitelikler ile hedef değişken arasında doğrudan sıralı ilişki olduğunda, Spearman korelasyonu etkili bir filtreleme yöntemi olarak öne çıkar.

Spearman korelasyonu ile Lojistik Regresyonun birlikte yüksek performans göstermesinin temel nedeni, her iki yöntemin de verideki sıralı (monotonik) ilişkileri etkili bir şekilde değerlendirebilmesidir. Spearman korelasyonu, öznitelikler ile hedef değişken arasındaki sıralı ilişkileri yakaladığı için, Lojistik Regresyon'un ihtiyaç duyduğu bağımsız ve anlamlı özniteliklerin ön plana çıkmasını sağlar. Böylece modele yalnızca hedef değişkenle güçlü ve düzenli ilişki gösteren değişkenler sunulur.

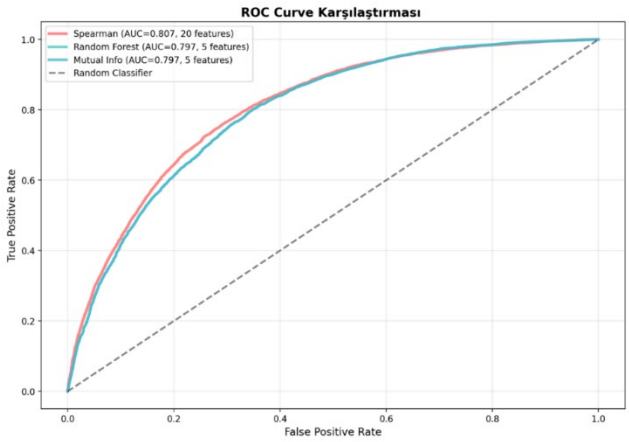
Ayrıca Lojistik Regresyon modeli, doğrusal karar sınırları oluşturur ve ilgisiz değişkenlere karşı duyarlıdır. Spearman ile yapılan öznitelik seçimi, bu tür değişkenleri filtrelediği için modelin overfitting yapmadan daha iyi genelleme yapmasını sağlar.

- Lojistik Regresyon, en yüksek F1-skoru (0.7339) ve doğruluk (0.7328) değerlerini bulmuştur.
- Random Forest sınıflandırıcısı, teorik üstünlüğüne rağmen uygulama çalıştırıldığında daha düşük performans göstermiştir.
- En etkili öznitelik seçimi yöntemi Spearman korelasyonu olmuştur; bu yöntemle 34 öznitelikten 20 tanesi seçilerek en iyi performans elde edilmiştir.
- Seçilen özniteliklerin neler olduğunu gösteren grafikler ve modellerin ROC eğrileri ile karmaşıklık matrisleri ekler bölümünde bulunmaktadır.

5. SONUÇ

Bu çalışmada, ROC eğrisi analizleri kapsamında en başarılı sonuç Lojistik Regresyon modeli ile elde edilmiştir. Modelin doğruluk ve F1 skorları arasında az fark bulunmuş, her iki metrik açısından da yüksek performans sağlanmıştır. Şekil 32'de ROC eğrisini görebilirsiniz. Diğer modellerin sınıflandırma ROC eğrilerini ek Şekil 1-3 arasında, karmaşıklık matrislerini ekte bulunan Tablo 1-6 arasında yer almaktadır.

En karmaşık algoritmaların her zaman en iyisi olmadığı prensibini güçlü bir şekilde desteklemiştir. Lojistik Regresyon gibi klasik modellerin, doğru koşullar altında modern algoritmalara karşı hala güçlü alternatifler olabildiğini göstermiştir. Geliştirilen otomatik optimizasyon çerçevesi önemlidir. Sonuçlar, makine öğrenmesi projelerinde başarının sadece algoritma seçimiyle değil, veri ön işleme, öznitelik seçimi ve optimizasyon hedeflerinin uyumlu bir şekilde tasarlanmasıyla elde edildiğini ortaya koymaktadır



Şekil 32: Lojistik Regresyon ROC eğrisi

5.1. Teorik ve Pratik Katkılar

- **Otomatik Eşik Optimizasyonu:** Geliştirilen çerçeve, manuel eşik belirleme ihtiyacını ortadan kaldırarak tekrarlanabilir sonuçlar sağladı.
- **Çok Amaçlı Değerlendirme:** Performans ve verimliliği dengeleyen skorlama sistemi, iki faklı metrik yani F1-skoru ve doğruluk ile "tek metrik tuzağı" önlendi.
- **Model-Yöntem Etkileşimi:** Model performansının sadece seçilen sınıflandırma algoritmasına değil, aynı zamanda uygulanan öznitelik seçimi yöntemine de bağlı olduğunu ve bu iki kararın birbirinden bağımsız alınamayacağını göstermektedir.

6. TARTIŞMA

6.1. Bulguların Değerlendirilmesi ve Tartışılması

6.1.1. Lojistik Regresyon'un Üstünlüğünün Nedenleri

Lojistik Regresyon sınıflandırma algoritması, özellikle filtre tabanlı öznitelik seçim yöntemleri ile birlikte kullanıldığında daha uyumlu sonuçlar vermektedir. Bu projede, iki farklı filtre tabanlı öznitelik seçim yöntemi olan Spearman korelasyonu ve Mutual Information kullanılmıştır. Bu yöntemler, veri setindeki her bir özniteliği bağımsız olarak değerlendirerek güçlü ve anlamlı olanları seçmekte etkilidir. Lojistik Regresyon'un doğrusal yapısı ise, bu şekilde seçilmiş öznitelikleri verimli bir biçimde kullanma avantajına sahiptir. Bununla birlikte, projede geliştirilen çok amaçlı skorlama fonksiyonu, daha az sayıda öznitelikle yüksek sınıflandırma performansı gösteren modelleri ödüllendirmiştir. Bu durum, Lojistik Regresyon'un doğası gereği sade ve açıklanabilir yapısını öne çıkararak onu daha avantajlı bir hale getirmiştir. Ayrıca, Lojistik Regresyon'un parametrik yapısı ve regülarizasyon mekanizmaları sayesinde, seçilmiş öznitelik alt kümeleriyle çalışıldığında aşırı öğrenmeye (overfitting) karşı daha dirençli bir yapı sergilediği ve daha stabil bir genelleme performansı gösterdiği düşünülmektedir.

6.1.2. Rastgele Orman'ın Sınırlı Performansı

Random Forest'in neden beklenenin aksine daha düşük bir performans sergilemesinin temel nedenleri olabilecek etkenler: Filtre tabanlı öznitelik seçme yöntemleri, modelin performansı için önemli olan bazı öznitelik etkileşimlerini göz ardı edebilir veya tamamen eleyebilir. Çok amaçlı optimizasyonun, daha fazla öznitelik kullanan konfigürasyonları cezalandırması, öznitelik seçimi sonrası veri yapısının Random Forest'in karmaşık modelleme kapasitesini gerektirmemesi, düşük performansın sebeplerinden biri olabilir.

6.2. Karşılaşılan Zorluklar ve Çözüm Stratejileri

Projenin geliştirilmesi sırasında sistemin sağlamlığını, verimliliğini ve güvenilirliğini sağlamak amacıyla bir dizi teknik zorlukla karşılaşılmış ve bunlara karşı etkili mühendislik çözümleri geliştirilmiştir.

- **Sınır Koşulu Yönetimi:**
    - **Problem:** Optimizasyon döngüsü sırasında denenen bir eşik değeri çok kısıtlayıcı olabilir ve sonuç olarak hiçbir özelliğin seçilmemesine yol açabilir. Bu "boş küme" durumu, modelleme aşamasının çökmesine (*ValueError*) ve tüm analiz sürecinin başarısız olmasına neden olur.
    - **Çözüm:** Optimizasyon döngüsünün her adımında, öznitelik seçiminden sonra *(X_sel.shape[1])* kontrolü yapılır. Eğer öznitelik sayısı sıfır ise, continue ifadesiyle o iterasyon atlanır ve bir sonraki eşik değerine geçilir. Bu basit ama etkili kontrol, sistemin istisnai durumlarda bile çalışmaya devam etmesini sağlayan bir "fail-safe" mekanizmasıdır.
- **Tek Metrik Tuzağı:**
    - **Problem:** Bir modelin kalitesini sadece "doğruluk" gibi tek bir metrikle değerlendirmek, yanıltıcı sonuçlara yol açabilir. Örneğin, "%91 doğrulukla 5 öznitelikle bir model mi, yoksa %92 doğrulukla 40 öznitelikle bir model mi daha iyidir?" Bu soruya cevap vermek, bağlama göre değişir ve tek bir metrik bu ödünleşimi yakalayamaz.
    - **Çözümü:** Projenin en temel yeniliklerinden biri olan çok amaçlı optimizasyon stratejisi bu soruna doğrudan bir çözümdür. Performansı, verimliliği (model basitliği) ve dengeyi birleştiren ağırlıklı skorlama fonksiyonu, "en iyi" modelin daha bütünsel ve pratik bir tanımını yapar. Bu sayede sistem, sadece isabetli değil, aynı zamanda verimli ve zarif modelleri de ödüllendirir.
- **Veri Setinin Dengesiz Olması:**
    - **Problem:** Veri seti dağılımı %90 diyabet değil, %10 diyabetli şeklindeydi. Bu dağılım sonucunda model çalıştığında her veriye diyabetli değil olarak kabul ediyordu. Bu şekilde kullanılması aşırı uyum (overfitting) sorunu ile karşılaşılmasına neden oldu.
    - **Çözüm:** %50 %50 yani dengeli bir veri seti ile değiştirilerek aşırı öğrenme sorunun önüne geçilmiş oldu. Bu şekilde kontrol ettiği verilere daha doğru şekilde tahminlerde bulunması sağlandı.

Bu çalışmanın temelini oluşturan otomatik ve çok amaçlı optimizasyon çerçevesi, yalnızca mevcut proje kapsamında değil, gelecekte yapılacak araştırmalar açısından da önemli bir potansiyel sunmaktadır.

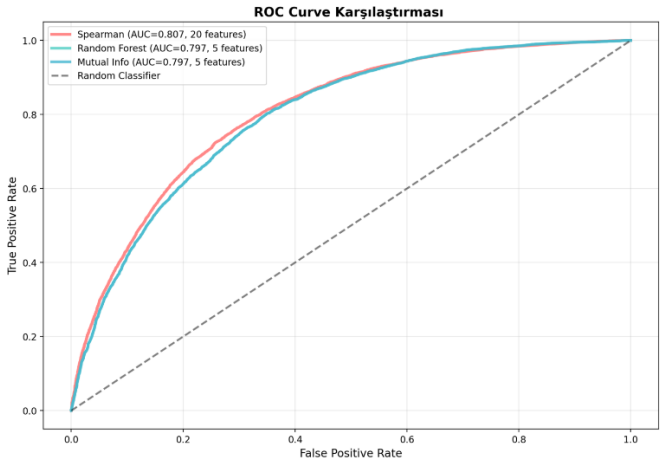
Proje; iki filtre tabanlı öznitelik seçim yöntemi ve bir gömülü yöntem ile değerlendirilmiştir. İlerleyen çalışmalarda sarmal (*wrapper*) tabanlı yöntemler ve başka gömülü (embedded) yöntemlerle farklı stratejilerin entegrasyonu ile daha kapsamlı karşılaştırmalar yapılabilir. Ayrıca, geliştirilen çerçevenin genelleştirilmesi ve farklı veri yapılarıyla ne kadar uyumlu çalıştığını test etmek amacıyla, çeşitli alanlardan elde edilmiş ve farklı dağılım özellikleri taşıyan veri setleri üzerinde benzer analizlerin gerçekleştirilmesi faydalı olacaktır.

KAYNAKÇA

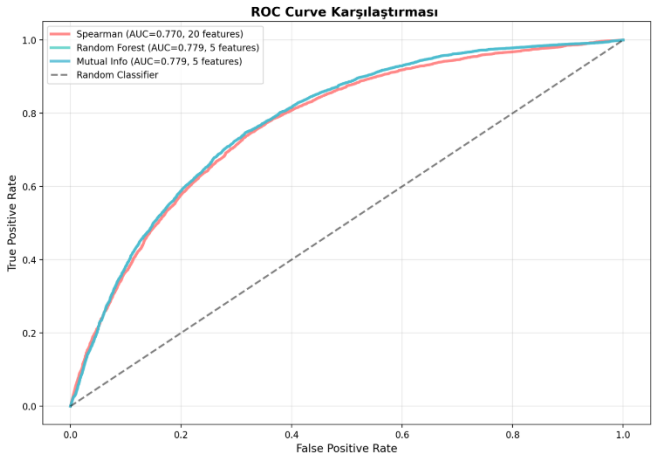
1. Teboul, A. (2022). Diabetes Health Indicators Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
2. Abdelhafez, H. A., & Amer, A. A. (2024). Machine learning techniques for diabetes prediction: A comparative analysis. *Journal of Applied Data Sciences*, 5(2), 792–807. https://doi.org/10.47738/jads.v5i2.219
3. Koren, M., Peretz, O., & Koren, O. (2023). Automated threshold learning for feature selection optimization. SSRN. https://doi.org/10.2139/ssrn.4350765
4. Pechprasarn, S., Srisaranon, N., & Yimluean, P. (2025). Optimizing diabetes prediction: An evaluation of machine learning models through strategic feature selection. Journal of Current Science and Technology, 15(1), Article 75. https://doi.org/10.59796/jcst.V15N1.2025.75
5. Forman, G. 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research, 3, 1289–1305.
6. Kılınç, D., & Kılıç, B. (2018). Makine öğrenmesi algoritmaları ile diyabet hastalığı sınıflandırılması. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 6(2), 1183–1193. https://doi.org/10.29130/dubited.433502
7. Saeys, Y., Inza, I., Larranaga, P. 2007. A review of feature selection techniques in bioinformatics, Bioinformatics, 23(19), 2507-2517
8. Şener, Y. (2023, Ağustos 8). *Makine öğrenmesinde değişken seçimi (feature selection) yazı serisi: Filtreleme yöntemleri*. Medium. https://yigitsener.medium.com/makine-%C3%B6%C4%9Frenmesinde-de%C4%9Fi%C5%9Fken-se%C3%A7imi-feature-selection-yaz%C4%B1-serisi-filtreleme-y%C3%B6ntemleri-ve-415a894d5b93
9. Şener, Y. (2023, Ağustos 7). *Makine öğrenmesinde değişken seçimi (feature selection) yazı serisi: Sarmal, wrapper yöntemler ve embedded yöntemler*. Medium. https://yigitsener.medium.com/makine-%C3%B6%C4%9Frenmesinde-de%C4%9Fi%C5%9Fken-se%C3%A7imi-feature-selection-yaz%C4%B1-serisi-sarmal-wrapper-y%C3%B6ntemler-ve-dd3b99c6c372

10. **Şener, Y. (2023, Ağustos 7).** *Makine öğrenmesinde değişken seçimi (feature selection) yazı serisi: Gömülü (embedded) yöntemler.* Medium. https://yigitsener.medium.com/makine-%C3%B6%C4%9Frenmesinde-de%C4%9Fi%C5%9Fken-se%C3%A7imi-feature-selection-yaz%C4%B1-serisi-g%C3%B6m%C3%BCl%C3%BC-embedded-y%C3%B6ntemler-c23293915b39
11. Steinwart, I., & Christmann, A. (2008). Support Vector Machines. Springer.
12. Kayademir, S. (2023, Mart 15). Makine öğrenmesi: Temel kavramlar ve uygulamalar. Medium. https://kayademirs.medium.com/makine-%C3%B6%C4%9Frenmesi-acc2b18d875a
13. BilgisayarKavramlari (2015,Temmuz 9) Spearsman's Rank Correlation Değerinin Hesaplanması https://www.youtube.com/watch?v=KBXJCcibFLs
14. Hall, M. 1999. Correlation-based Feature Selection for Machine Learning, The University of Waikato, PhD Thesis,Hamilton.
15. Novakavic, J., Strbac, P., Bulatovic, D. 2011. Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms, Yugoslav Journal of Operations Research, 21(1), 119-135.
16. Devreyakan. Performans metrikleri nedir? *Devreyakan*. https://devreyakan.com/performans-metrikleri/
17. Nguyen, Q. (2022). Preprocessing Categorical Features and Column Transformer. In BAIT 509: Business Applications of Machine Learning. University of British Columbia. Erişim tarihi 7 Haziran 2025, https://bait509-ubc.github.io/BAIT509/lectures/lecture5.html
18. BLK Pediatric Practice. (2021, June 17). Body Mass Index (BMI). https://blk-pediatric-practice.com/2021/06/17/body-mass-index-bmi/
19. QuantHub. (t.y.). How to read a correlation heatmap. *QuantHub*. https://www.quanthub.com/how-to-read-a-correlation-heatmap/
20. Budak, H. (2018). Özellik seçim yöntemleri ve yeni bir yaklaşım. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *22*(Özel sayı), 21–31. https://dergipark.org.tr/tr/download/article-file/552933

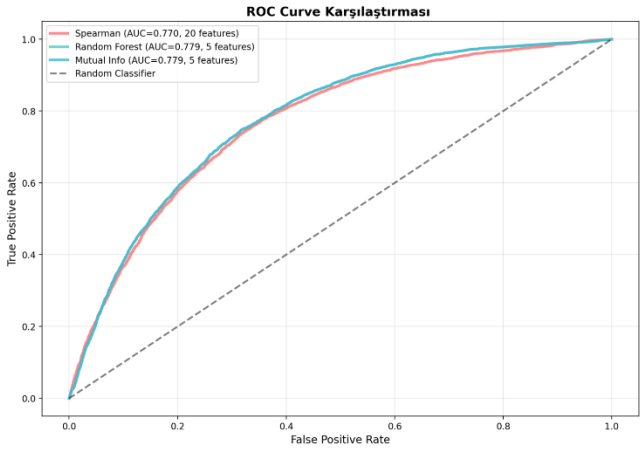
EK



Ek şekil 1: Lojistik Regresyon'un accuracy göre ROC eğrisi



Ek şekil 2: Random Forest'ın accuracy göre ROC eğrisi



Ek şekil 3: Random Forest'ın F1-skoruna göre ROC eğrisi

Tablo 1: Random Forest karmaşıklık matrisi

	0	1
0	7531	3070
1	3169	7438

Tablo 2: Lojistik Regresyon karmaşıklık matrisi

	0	1
0	7728	2873
1	2793	7814

Tablo 3: Random Forest karmaşıklık matrisi

	0	1
0	7211	3390
1	2661	7946

Tablo 4: Lojistik Regresyon karmaşıklık matrisi

	0	1
0	7439	3162
1	2707	7900

Tablo 5: Random Forest karmaşıklık matrisi

	0	1
0	7211	3390
1	2661	7946

Tablo 6: Lojistik Regresyon karmaşıklık matrisi

	0	1
0	7439	3162
1	2707	7900

Tablo 7: Veri setinde bulunan değişkenlerin tanımları

Değişken Adı	Açıklama	Veri Tipi
Diabetes_binary	Diyabet durumu: 0 = diyabet yok, 1 = diyabet	Kategorik (Binary)
HighBP	Yüksek tansiyon: 0 = yüksek tansiyon yok, 1 = yüksek tansiyon var	Kategorik (Binary)
HighChol	Yüksek kolesterol: 0 = yüksek kolesterol yok, 1 = yüksek kolesterol var	Kategorik (Binary)
CholCheck	Son 5 yılda kolesterol kontrolü: 0 = kontrol yapılmamış, 1 = kontrol yapılmış	Kategorik (Binary)
BMI	Vücut Kitle İndeksi	Sayısal (Sürekli)
Smoker	Sigara kullanımı (hayat boyu en az 100 sigara içmiş olma): 0 = hayır, 1 = evet	Kategorik (Binary)
Stroke	İnme geçmişi: 0 = hayır, 1 = evet	Kategorik (Binary)
HeartDiseaseorAttack	Koroner kalp hastalığı veya miyokard enfarktüsü geçmişi: 0 = hayır, 1 = evet	Kategorik (Binary)
PhysActivity	Son 30 günde iş dışı fiziksel aktivite: 0 = hayır, 1 = evet	Kategorik (Binary)
Fruits	Günde 1 veya daha fazla meyve tüketimi: 0 = hayır, 1 = evet	Kategorik (Binary)
Veggies	Günde 1 veya daha fazla sebze tüketimi: 0 = hayır, 1 = evet	Kategorik (Binary)
HvyAlcoholConsump	Ağır alkol tüketimi: 0 = hayır, 1 = evet	Kategorik (Binary)
AnyHealthcare	Herhangi bir sağlık sigortası: 0 = hayır, 1 = evet	Kategorik (Binary)
NoDocbcCost	Son 12 ayda maliyet nedeniyle doktora gidememe durumu: 0 = hayır, 1 = evet	Kategorik (Binary)
GenHlth	Genel sağlık durumu: 1-5 ölçeği (1 = mükemmel, 5 = kötü)	Kategorik (Ordinal)
MentHlth	Son 30 gündeki kötü ruh sağlığı günleri: 1-30 gün ölçeği	Sayısal (Diskret)

PhysHlth	Son 30 gündeki fiziksel rahatsızlık günleri: 1-30 gün ölçeği	Sayısal (Diskret)
DiffWalk	Yürümede veya merdiven çıkmada ciddi zorluk: 0 = hayır, 1 = evet	Kategorik (Binary)
Sex	Cinsiyet: 0 = kadın, 1 = erkek	Kategorik (Binary)
Age	13 seviyeli yaş kategorisi: 1 = 18-24, 9 = 60-64, 13 = 80 veya üzeri	Kategorik (Ordinal)
Education	Eğitim seviyesi: 1-6 ölçeği	Kategorik (Ordinal)
Income	Gelir ölçeği: 1-8 (1 = $10,000'dan az, 8 = $75,000 veya daha fazla)	Kategorik (Ordinal)