ISSN 2757-5543 GÜFFD 7. Cilt (1): 39-56 (2026) DOI: 10.63716/guffd.1911737

Gazi Üniversitesi Fen Fakültesi Dergisi
https://dergipark.org.tr/tr/pub/guffd

Özellik Seçimi Yöntemlerinin Sınıflandırma Performansı ve Model Yorumlanabilirliği Üzerindeki Etkileri: Deneysel Bir Analiz

Derya Turfan Uğurlu*

Hacettepe Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 06500, Ankara, Türkiye

Öne Çıkanlar

- Farklı özellik seçimi yöntemlerinin sınıflandırma performansı üzerindeki etkileri karşılaştırılmıştır.
- Özellik seçimi yöntemlerinin performans üzerindeki etkileri iki farklı veri seti üzerinde incelenmiştir.
- Model performansı tekrarlı çapraz doğrulama yaklaşımı ile değerlendirilmiştir.
- Seçilen modellerin karar mekanizmaları SHAP analizi ile yorumlanmıştır.
- Daha az değişken kullanılarak benzer performansın korunabildiği gösterilmiştir.

Makale Bilgileri

Geliş: 17/03/2026
Kabul: 24/04/2026

Anahtar Kelimeler

Özellik seçimi,
Makine öğrenmesi,
Sınıflandırma,
SHAP,
Model yorumlanabilirliği

Öz

Bu çalışmada farklı özellik seçimi yöntemlerinin sınıflandırma performansı üzerindeki etkileri iki farklı veri seti üzerinde incelenmiştir. Bu kapsamda, filtre yaklaşımını temsil eden karşılıklı bilgi yöntemi ile gömülü yöntemler arasında yer alan L1 düzenlileştirme ve ağaç tabanlı değişken önem ölçüleri kullanılarak farklı özellik alt kümeleri oluşturulmuştur. Elde edilen özellik alt kümeleri, L2 düzenlileştirmeli lojistik regresyon ve Rastgele Orman sınıflandırma modelleri ile değerlendirilmiştir. Deneysel süreçte veri sızıntısını önlemek amacıyla, ön işleme ve özellik seçimi adımları tekrarlı çapraz doğrulama süreci içerisinde yalnızca eğitim verisi üzerinde öğrenilmiş ve aynı dönüşümler test verisine uygulanmıştır. Model performansı doğruluk, F1-skoru ve ROC-AUC ölçütleri kullanılarak değerlendirilmiştir. Elde edilen sonuçlar incelendiğinde, birinci veri setinde (Adult) değişken sayısının azaltılmasının performans üzerinde belirgin bir değişim oluşturmadığı; ikinci veri setinde (Heart) ise seçilen özellik alt kümelerine bağlı olarak performans farklılıklarının daha belirgin hâle gelebildiği görülmüştür. Bu bulgular, özellik seçimi yöntemlerinin etkisinin veri setinin özelliklerine ve kullanılan modelleme yaklaşımına bağlı olarak farklılaşabileceğine işaret etmektedir. Ayrıca her iki veri setinde en iyi performans gösteren modeller için gerçekleştirilen SHAP analizi sayesinde model tahminlerinde etkili olan değişkenler yorumlanmıştır.

The Effects of Feature Selection Methods on Classification Performance and Model Interpretability: An Experimental Analysis

Highlights

- The effects of different feature selection methods on classification performance were compared.
- The effects of feature selection methods on performance were examined using two different datasets.
- Model performance was evaluated using a repeated cross-validation approach.
- The decision mechanisms of the selected models were interpreted using SHAP analysis.
- Comparable performance was maintained with fewer variables.

Article Info

Received: 17/03/2026
Accepted: 24/04/2026

Keywords

Feature selection,
Machine learning,
Classification,
SHAP,
Model interpretability

Abstract

In this study, the effects of different feature selection methods on classification performance were examined using two distinct datasets. In this context, different feature subsets were created using the mutual information method—which represents a filter-based approach—as well as L1 regularization and tree-based variable importance measures, which represent embedded approaches. The resulting feature subsets were evaluated using L2-regularized logistic regression and Random Forest classification models. To prevent data leakage during the experimental process, the preprocessing and feature selection steps were learned only from the training data within a repeated cross-validation framework, and the same transformations were subsequently applied to the test data. Model performance was evaluated using accuracy, F1-score, and ROC-AUC metrics. The results showed that reducing the number of variables in the first dataset (Adult) did not lead to a substantial change in performance, whereas in the second dataset (Heart), performance differences became more pronounced depending on the selected feature subsets. These findings indicate that the impact of feature selection methods may vary depending on dataset characteristics and the modeling approach used. In addition, SHAP analysis conducted on the best-performing models for both datasets enabled interpretation of the variables that influence model predictions.

* Sorumlu Yazar/Corresponding Author: Derya Turfan Uğurlu, deryaturfan@hacettepe.edu.tr

GÜFFD GAZİ ÜNİVERSİTESİ FEN FAKÜLTESİ DERGİSİ

Araştırma Makalesi

1. GİRİŞ

Makine öğrenmesi yöntemleri son yıllarda birçok alanda sınıflandırma ve tahmin problemlerinin çözümünde yaygın olarak kullanılmaktadır. Bununla birlikte gerçek dünya veri setleri çoğu zaman yüksek boyutlu olmakta ve veri içindeki tüm değişkenler model performansına aynı ölçüde katkı sağlamamaktadır. Bu durum model karmaşıklığının artmasına, hesaplama maliyetinin yükselmesine ve aşırı öğrenme (overfitting) gibi problemlerin ortaya çıkmasına yol açabilmektedir. Bu nedenle makine öğrenmesi uygulamalarında model eğitimi öncesinde en bilgilendirici değişkenlerin belirlenmesi önemli bir araştırma konusu hâline gelmiştir. Bu amaçla geliştirilen özellik seçimi (feature selection) yöntemleri, veri içindeki gereksiz veya düşük bilgi içeren değişkenlerin elenmesine yardımcı olarak daha sade ve daha iyi genelleme performansı gösterebilen modellerin kurulmasına katkıda bulunur [1].

Özellik seçimi yöntemleri genel olarak filtre (filter), sarmal (wrapper) ve gömülü (embedded) yöntemler olmak üzere üç ana grupta incelenmektedir [1]. Filtre yöntemleri değişkenler ile hedef değişken arasındaki istatistiksel ilişkilere dayanarak modelden bağımsız biçimde özellik değerlendirmesi yaparken, gömülü yöntemler değişken seçimini doğrudan model öğrenme sürecinin bir parçası olarak gerçekleştirmektedir. Filtre yöntemlerine örnek olarak karşılıklı bilgi (mutual information) tabanlı yaklaşımlar sıklıkla kullanılmaktadır. Karşılıklı bilgi ölçütü, bir değişkenin hedef değişken hakkında taşıdığı bilgi miktarını ölçerek değişkenlerin önem sıralamasını belirlemeye imkân sağlamaktadır [2, 3]. Buna karşılık gömülü yöntemler arasında L1 düzenlileştirmeye dayalı yaklaşımlar ve ağaç tabanlı modeller tarafından üretilen değişken önem ölçüleri yaygın olarak kullanılmaktadır. L1 düzenlileştirme bazı model katsayılarını sıfıra indirerek değişken seçimini doğrudan model eğitimi sırasında gerçekleştirebilmektedir [4]. Benzer şekilde Random Forest gibi ağaç tabanlı yöntemler de değişkenlerin model tahminlerine katkısını ölçen değişken önem değerleri üretmekte ve bu ölçüler özellik seçimi amacıyla kullanılabilmektedir [5, 6].

Özellik seçimi yöntemleri çoğu zaman farklı sınıflandırma algoritmaları ile birlikte değerlendirilmekte ve elde edilen değişken alt kümelerinin model performansı üzerindeki etkileri incelenmektedir. Lojistik regresyon yorumlanabilirliği ve istatistiksel temeli nedeniyle sınıflandırma problemlerinde yaygın olarak kullanılan bir yöntemdir [7]. Buna karşılık Random Forest gibi topluluk öğrenme (ensemble learning) yöntemleri doğrusal olmayan ilişkileri modelleyebilme kapasitesi sayesinde birçok veri madenciliği uygulamasında yüksek performans gösterebilmektedir [5]. Bu nedenle farklı özellik seçimi yaklaşımlarının hem doğrusal hem de ağaç tabanlı sınıflandırma yöntemleri ile birlikte değerlendirilmesi model performansının daha kapsamlı biçimde incelenmesine yardımcı olmaktadır.

Literatürde özellik seçimi ve sınıflandırma yöntemlerinin birlikte değerlendirildiği çeşitli çalışmalar bulunmaktadır. Li ve arkadaşları [8], özellik seçimi yöntemlerini veri perspektifinden ele alarak farklı veri yapılarında çeşitli yöntemlerin değişken seçimi ve model performansı üzerindeki etkilerini kapsamlı biçimde incelemiştir. Ağaç tabanlı yöntemlere dayalı özellik seçimi yaklaşımlarına ilişkin olarak Díaz-Uriarte ve Alvarez de Andrés [6], mikroarray veri setleri üzerinde Random Forest tabanlı değişken önem ölçülerini kullanarak hem değişken seçimi hem de sınıflandırma performansını değerlendirmiştir. Benzer şekilde Genuer ve arkadaşları [9], Random Forest tabanlı değişken seçimi yöntemlerinin yüksek boyutlu veri setlerinde etkili sonuçlar üretebildiğini göstermiştir. Fernández-Delgado ve arkadaşları [10] ise çok sayıda sınıflandırma algoritmasını farklı veri setleri üzerinde değerlendirerek özellikle ensemble tabanlı yöntemlerin birçok problemde güçlü performans gösterebildiğini ortaya koymuştur.

Makine öğrenmesi modellerinin yalnızca performans ölçütleri açısından değil, aynı zamanda karar mekanizmalarının anlaşılabilirliği açısından da incelenmesi giderek önem kazanmaktadır. Ribeiro ve arkadaşları [11] tarafından geliştirilen LIME yöntemi model tahminlerinin yerel düzeyde açıklanmasını amaçlarken, Lundberg ve Lee [12] tarafından önerilen SHAP yöntemi Shapley değerlerine dayalı olarak model tahminlerinde her bir değişkenin katkısını hesaplayan güçlü bir açıklanabilirlik yaklaşımı sunmaktadır.

Makine öğrenmesi yöntemlerinin farklı uygulama alanlarında kullanıldığı çalışmalarda da özellik seçimi ve değişken önem analizleri önemli bir rol oynamaktadır. Chicco ve Jurman [13], klinik veriler üzerinde makine öğrenmesi yöntemlerini kullanarak hasta sağkalımını tahmin etmiş ve özellik önem analizleri sonucunda belirli değişkenlerin model performansı açısından kritik rol oynadığını göstermiştir. Zhou ve arkadaşları [14], karşılıklı bilgi ile korelasyon katsayısını birleştiren bir özellik seçimi yöntemi önermiş ve farklı veri setleri üzerinde yaptıkları deneylerde önerilen yöntemin sınıflandırma doğruluğunu artırabildiğini göstermiştir. Benzer şekilde Gomez-Martinez ve arkadaşları [15], ensemble özellik seçimi yöntemlerini veri artırma teknikleri ile birlikte kullanarak sınıflandırma performansını artırmayı hedeflemiş ve elde edilen modellerin yorumlanabilirliğini SHAP yöntemi ile incelemiştir.

Bu çalışmada farklı özellik seçimi yöntemlerinin sınıflandırma performansı üzerindeki etkileri iki veri seti üzerinde incelenmiştir. Bu amaçla filtre yaklaşımını temsil eden karşılıklı bilgi yöntemi ile gömülü yöntemler arasında yer alan L1 düzenlileştirme ve ağaç tabanlı değişken önem ölçüsü kullanılarak farklı özellik alt kümeleri oluşturulmuştur. Elde edilen özellik alt kümeleri L2 düzenlileştirmeli lojistik regresyon ve Rastgele Orman sınıflandırma modelleri ile değerlendirilmiştir. Bu çalışma, farklı özellik seçimi yaklaşımlarının sınırlı sayıda veri seti üzerinde kontrollü bir deneysel yapı içerisinde karşılaştırılmasını amaçlamaktadır. Ayrıca, model seçim kriteri doğrultusunda belirlenen nihai modeller için SHAP analizi gerçekleştirilmiş ve model tahminlerinde etkili olan değişkenler yorumlanmıştır. Bu kapsamda çalışma, yalnızca model performansını değil, aynı zamanda model tahminlerinde etkili olan değişkenleri de birlikte incelemektedir. Bu yönüyle çalışma, geniş ölçekli genelleme yapmayı amaçlayan bir karşılaştırma çalışmasından ziyade, özellik seçimi yöntemlerinin etkilerini veri seti bağlamında inceleyen bir deneysel analiz sunmaktadır.

2. MATERYAL VE YÖNTEM

Bu bölümde çalışmada kullanılan veri setleri, ön işleme süreci, uygulanan özellik seçimi yaklaşımları ve sınıflandırma modelleri açıklanmaktadır. Ayrıca deneysel tasarım, performans değerlendirme ölçütleri ve model yorumlanabilirliğinin incelenmesinde kullanılan SHAP analizi sunulmaktadır. Tüm analizler R istatistiksel programlama dili kullanılarak gerçekleştirilmiş olup kullanılan paketler ilgili alt başlıklarda belirtilmiştir.

2.1. Veri Setleri

Bu çalışmada, farklı özellik seçimi yöntemlerinin sınıflandırma performansı üzerindeki etkileri, iki farklı yapısal özelliğe sahip veri seti üzerinde incelenmiştir. Veri setleri örneklem büyüklüğü, değişken türleri ve uygulama alanı bakımından farklılık göstermektedir.

Veri setleri OpenML platformu üzerinden temin edilmiş olup, Adult veri seti OpenML ID: 1590 ve Heart veri seti OpenML ID: 53 kimlik numaraları kullanılarak elde edilmiştir. Bu veri setleri UCI makine öğrenmesi veri tabanında yaygın olarak kullanılan veri setlerinin OpenML üzerindeki temsilleridir.

Adult veri seti (OpenML ID: 1590), bireylerin yıllık gelirlerinin 50.000 ABD doları üzerinde olup olmadığını gösteren bir sınıflandırma problemine dayanmaktadır. Veri seti 48842 gözlem, 14 özellik ve bir hedef değişkenden oluşmaktadır. Özellikler hem sayısal hem kategorik yapıdadır. Hedef değişken dağılımı ≤50K sınıfı için 37155, >50K sınıfı için 11687 gözlem şeklindedir.

Heart veri seti (Statlog, OpenML ID: 53), kalp hastalığının varlığı (“present”) veya yokluğu (“absent”) durumunu ifade eden bir sınıflandırma problemine dayanmaktadır. Veri seti 270 gözlem, 13 özellik ve bir hedef değişkenden oluşmaktadır. Özelliklerin büyük bölümü sayısal yapıdadır. Hedef değişken dağılımı hastalık yok sınıfı için 150, hastalık var sınıfı için 120 gözlemdir.

Bu veri setleri, yapısal olarak tabular veri özellikleri göstermekte olup, özellik seçimi yöntemlerinin sistematik olarak değerlendirilmesi açısından uygun bir deneysel ortam sunmaktadır. Bunun yanı sıra, bu veri setleri makine öğrenmesi literatüründe farklı sınıflandırma ve özellik seçimi yaklaşımlarının değerlendirildiği çalışmalarda sıklıkla tercih edilmektedir [16-18].

2.2. Ön İşleme Süreci

Tüm ön işleme adımları çapraz doğrulama katmanları içerisinde yalnızca eğitim verisi üzerinde öğrenilmiş, test verisine ise eğitim verisinden elde edilen parametreler uygulanmıştır. Böylece veri sızıntısı (data leakage) önlenmiştir.

Ön işleme sürecinde sayısal değişkenlerde eksik değerler medyan ile doldurulmuş, kategorik değişkenlerde eksik değerler en sık gözlenen kategori ile tamamlanmıştır. Medyan ile tamamlama yöntemi, ortalama gibi alternatif yaklaşımlara kıyasla uç değerlerden daha az etkilenmesi nedeniyle tercih edilmiştir. Kategorik değişkenler, modele yapay bir sıralama ilişkisi yüklememek amacıyla one-hot kodlama ile sayısal forma dönüştürülmüştür. Varyansı sıfır olan değişkenler, ayırt edici bilgi içermemeleri nedeniyle veri setinden çıkarılmıştır. Son olarak sayısal değişkenler z-skor yöntemi kullanılarak standartlaştırılmıştır; bu yaklaşım özellikle düzenlileştirme içeren modellerde değişkenlerin karşılaştırılabilir ölçeklere getirilmesini sağlaması nedeniyle tercih edilmiştir. Özellik seçimi işlemleri bu dönüşümler sonrasında elde edilen özellik uzayı üzerinde gerçekleştirilmiştir.

Adult veri setinde workclass değişkeninde 2799, occupation değişkeninde 2809 ve native.country değişkeninde 857 olmak üzere bazı kategorik değişkenlerde eksik gözlemler bulunduğu tespit edilmiştir. Heart veri setinde ise eksik gözlem bulunmamaktadır. Uygulanan ön işleme adımları sonrasında Adult veri setinde özellik sayısı 14'ten 106'ya yükselirken, Heart veri setinde özellik sayısında değişim gözlenmemiştir. Ayrıca her iki veri setinde de varyansı sıfır olan bir değişkene rastlanmamıştır.

Ön işleme sürecinin veri setleri üzerindeki etkisi Çizelge 1'de özetlenmiştir. Eğitim verisi üzerinden öğrenilen eksik değer tamamlama ve dönüşüm parametreleri test verisine de aynı şekilde uygulanmıştır. Bu süreç sonunda, eğitim ve test veri setleri aynı özellik uzayında temsil edilmekte ve değişken yapısı tutarlı biçimde korunmaktadır.

Çizelge 1. *Ön işleme öncesi ve sonrası veri seti özet bilgileri*

Veri Seti	Gözlem Sayısı	Özellik Sayısı (öncesi)	Özellik Sayısı (sonrası)	Hedef Sınıf Sayısı	Sınıf Dağılımı
Adult	48842	14	106	2	37155/11687
Heart	270	13	13	2	150/120

Not: Özellik sayıları hedef değişken hariç değişken sayısını göstermektedir.

2.3. Özellik Seçimi Yaklaşımları

Bu çalışmada farklı özellik seçimi yaklaşımlarını temsil eden üç yöntem uygulanmıştır. Seçilen yöntemler, literatürde yaygın olarak kullanılan ve farklı özellik seçimi paradigmasını temsil eden filtre ve gömülü yaklaşımları karşılaştırmalı olarak incelemek amacıyla tercih edilmiştir. Bu kapsamda karşılıklı bilgi (Mutual Information) filtre tabanlı yaklaşımı, L1 düzenlileştirme ve ağaç tabanlı değişken önem ölçüleri ise gömülü özellik seçimi yaklaşımını temsil etmektedir. Her bir yöntem için değişkenler elde edilen önem skorlarına göre sıralanmış ve toplam özellik sayısının belirli oranlarına (%25, %50, %75 ve %100) karşılık gelen sayıda en yüksek skorlu değişkenler seçilerek özellik alt kümeleri oluşturulmuştur. Ayrıca herhangi bir özellik seçimi uygulanmayan tüm özelliklerin kullanıldığı durum referans senaryo olarak ele alınmıştır.

2.3.1. Karşılıklı bilgi (Mutual information)

Karşılıklı bilgi, bir özellik ile hedef değişken arasındaki bilgi paylaşımını ölçen bilgi kuramına dayalı bir filtre yöntemidir [19].

Bu çalışmada her bir özellik için hedef değişken ile karşılıklı bilgi değeri hesaplanmış, elde edilen skorlar azalan sırada düzenlenmiş ve belirlenen oranlara göre en yüksek skorlu özellikler seçilerek özellik alt kümeleri oluşturulmuştur. Karşılıklı bilgi hesaplamaları "infotheo" paketi ile gerçekleştirilmiştir.

2.3.2. L1 düzenlileştirme

L1 düzenlileştirme yöntemi, lojistik regresyon modeline L1 düzenlileştirme terimi eklenerek katsayıların seyrekleştirilmesini sağlar. Bu yaklaşımda küçük katsayılar sıfıra zorlanmakta, böylece gömülü (embedded) bir özellik seçimi mekanizması oluşmaktadır [4]. Elde edilen katsayıların mutlak değerleri kullanılarak değişkenler sıralanmış ve belirlenen oranlara göre en yüksek skorlu özellikler seçilmiştir. Model tahminleri ve katsayı hesaplamaları "glmnet" paketi kullanılarak gerçekleştirilmiştir [17].

2.3.3. Ağaç tabanlı değişken önemi (Tree-based variable importance)

Bu yöntemde değişkenler, Rastgele Orman (Random Forest) modeli üzerinden elde edilen değişken önem (variable importance) skorlarına göre sıralanmıştır [5]. Değişken önem değerleri temel alınarak yapılan sıralama sonucunda, belirlenen oranlara göre en yüksek öneme sahip özellikler seçilmiştir. Rastgele Orman tabanlı değişken önem ölçüleri özellik seçimi amacıyla yaygın olarak kullanılmaktadır [5, 6, 9]. Hesaplamalar "ranger" paketi kullanılarak gerçekleştirilmiştir.

Her özellik seçimi yöntemi için belirlenen oranlara göre oluşturulan özellik alt kümeleri kullanılarak sınıflandırma modelleri eğitilmiş ve performans değerlendirmeleri gerçekleştirilmiştir.

2.4. Sınıflandırma Modelleri

Bu çalışmada L2 düzenlileştirmeli lojistik regresyon ve Rastgele Orman olmak üzere iki farklı sınıflandırma modeli kullanılmıştır. Bu modeller, doğrusal ve doğrusal olmayan modelleme yaklaşımlarını temsil etmeleri ve makine öğrenmesi literatüründe yaygın olarak kullanılan temel yöntemler arasında yer almaları nedeniyle tercih edilmiştir. Bu sayede, farklı özellik seçimi yaklaşımlarının hem doğrusal hem de doğrusal olmayan ilişkileri modelleyebilen yapılar üzerindeki etkileri karşılaştırmalı olarak incelenmiştir. Ayrıca, çalışmanın odağı özellik seçimi yöntemlerinin etkisini değerlendirmek olduğundan, bu doğrultuda yöntemler arası karşılaştırmanın daha kontrollü bir deneysel çerçevede gerçekleştirilmesi hedeflenmiştir.

2.4.1. L2 düzenlileştirmeli lojistik regresyon (L2-regularized logistic regression)

Lojistik regresyon, sınıf olasılıklarını doğrusal bir karar sınırı üzerinden modelleyen ve sınıflandırma problemlerinde yaygın olarak kullanılan bir yöntemdir. Çalışmada model, L2 düzenlileştirme (ridge cezası) ile uygulanmıştır. L2 düzenlileştirme katsayı büyüklüklerini sınırlandırarak aşırı uyumu azaltmayı ve özellikle çoklu bağlantı durumlarında daha kararlı tahminler elde etmeyi amaçlamaktadır.

Düzenlileştirme parametresi (λ), her çapraz doğrulama katında yalnızca eğitim verisi kullanılarak iç çapraz doğrulama ile belirlenmiştir. Model tahminleri "glmnet" paketi aracılığıyla gerçekleştirilmiştir [20].

2.4.2. Rastgele orman (Random forest)

Rastgele Orman, birden fazla karar ağacının rastgele alt örnekleme ve rastgele özellik seçimi mekanizmaları ile birleştirilmesine dayanan doğrusal olmayan bir sınıflandırma yöntemidir [5]. Bu yapı, değişkenler arası etkileşimlerin bulunduğu veri yapılarında esnek bir modelleme sağlar.

Çalışmada Rastgele Orman modeli "ranger" paketi kullanılarak uygulanmıştır [21]. Modelde ağaç sayısı 300 olarak sabitlenmiş ve diğer parametreler paket varsayılanları ile kullanılmıştır. Ağaç sayısı, model performansı ile hesaplama maliyeti arasında denge sağlayan yaygın bir tercih olarak belirlenmiştir.

Rastgele Orman yöntemi bu çalışmada iki farklı amaçla kullanılmıştır: özellik seçimi aşamasında değişken önem sıralaması üretmek amacıyla ve sınıflandırma modeli olarak performans karşılaştırmalarında. Bu iki kullanım, veri sızıntısını önlemek amacıyla deneysel sürecin farklı aşamalarında bağımsız olarak gerçekleştirilmiştir.

Her iki sınıflandırma modeli de özellik seçimi sonrasında elde edilen alt kümeler üzerinde eğitilmiş ve performans değerlendirmeleri test verisi üzerinde gerçekleştirilmiştir.

2.5. Deneysel Tasarım

Bu çalışmada özellik seçimi ve sınıflandırma performansının güvenilir biçimde değerlendirilmesi amacıyla tekrarlı ve tabakalı çapraz doğrulama yaklaşımı uygulanmıştır. Deneysel süreç, veri sızıntısını önleyecek ve model karşılaştırmalarını tutarlı biçimde gerçekleştirecek şekilde yapılandırılmıştır.

2.5.1. Çapraz doğrulama

Performans değerlendirmesi için 5 katlı ve 3 tekrar içeren tekrarlı çapraz doğrulama (5×3) kullanılmıştır. Her tekrar içerisinde veri beş alt kümeye ayrılmış, her alt küme bir kez test verisi olarak kullanılmıştır. Katmanlar sınıf dağılımını koruyacak şekilde tabakalı olarak oluşturulmuştur.

Her katmanda ön işleme ve özellik seçimi adımları yalnızca eğitim verisi üzerinde gerçekleştirilmiş, test verisine ise yalnızca eğitim verisinden elde edilen dönüşüm parametreleri uygulanmıştır. Böylece veri sızıntısı engellenmiştir. Bu doğrultuda her bir model ve özellik seçimi kombinasyonu için performans ölçütleri, çapraz doğrulama süreci sonucunda elde edilen toplam 15 test değerlendirmesi üzerinden hesaplanmıştır.

Seçilen değişken alt kümeleri test verisine de aynı şekilde uygulanmış ve model değerlendirmesi yalnızca bu değişkenler üzerinden gerçekleştirilmiştir. Böylece eğitim ve test setleri aynı özellik alt kümesi ve dönüşüm yapısı ile tutarlı biçimde değerlendirilmiştir.

2.5.2. Oran bazlı değişken azaltma

Her özellik seçimi yöntemi için değişkenler önem sıralamasına göre düzenlenmiş ve toplam özellik sayısının belirli oranlarına karşılık gelen alt kümeler oluşturulmuştur. Çalışmada %25, %50, %75 ve tüm özellikler (%100, referans senaryo) olmak üzere dört farklı oran değerlendirilmiştir. Özellik sayısının bu oranlara tam olarak karşılık gelmediği durumlarda, seçilecek değişken sayısı toplam özellik sayısı ile ilgili oranın çarpımı sonucu elde edilen değerin en yakın tam sayıya yuvarlanması ile belirlenmiştir.

Bu yaklaşım sayesinde değişken sayısının sistematik biçimde azaltılmasının model performansı üzerindeki etkisi incelenmiştir.

2.5.3. Model seçim kriteri

Her özellik seçimi yöntemi, değişken oranı ve sınıflandırma modeli kombinasyonu için elde edilen performans değerleri kullanılarak ortalama ± standart sapma hesaplanmıştır. Ön işleme adımlarına bağlı olarak bazı katmanlarda toplam özellik sayısının değişebilmesi nedeniyle, belirli oranlara karşılık gelen seçilen özellik sayısı küçük farklılıklar gösterebilmiştir. Bu nedenle performans ölçütleri, ilgili kombinasyonlar için çapraz doğrulama süreci boyunca elde edilen tüm değerlendirme sonuçları dikkate alınarak hesaplanmıştır.

Performans değerlendirmesinde birden fazla ölçüt birlikte dikkate alınmıştır (bkz. Bölüm 2.6). Ortalama performans değerleri birbirine oldukça yakın olan modeller arasında daha az sayıda değişken içeren yapı tercih edilmiştir. Bu yaklaşım, benzer performans düzeyini korurken daha sade ve yorumlanabilir modeller elde etmeyi amaçlamaktadır.

Yukarıda açıklanan deneysel süreç, özellik seçimi, model eğitimi ve performans değerlendirme adımlarının bir bütün olarak yürütülmesini içermektedir. Bu sürecin tüm adımlarını özetleyen algoritmik yapı Algoritma 1'de sunulmuştur.

2.6. Performans Ölçütleri

Model performansının değerlendirilmesinde doğruluk (Accuracy), F1-skoru ve ROC-AUC ölçütleri kullanılmıştır. Farklı ölçütlerin birlikte incelenmesi, özellikle sınıf dağılımının tam dengeli olmadığı veri yapılarında daha güvenilir bir değerlendirme sunmaktadır. Bu ölçütler aşağıda ayrıntılı olarak açıklanmaktadır.

2.6.1. Doğruluk (Accuracy)

Doğruluk, doğru sınıflandırılan gözlemlerin toplam gözlem sayısına oranı olarak tanımlanır:

$$Doğruluk = \frac{DP+DN}{DP+DN+YP+YN} \ . \quad (1)$$

Bu ifadede,
DP (Doğru Pozitif): Pozitif sınıfa ait olup doğru sınıflandırılan gözlemler,
DN (Doğru Negatif): Negatif sınıfa ait olup doğru sınıflandırılan gözlemler,
YP (Yanlış Pozitif): Negatif sınıfa ait olup pozitif olarak sınıflandırılan gözlemler,
YN (Yanlış Negatif): Pozitif sınıfa ait olup negatif olarak sınıflandırılan gözlemlerdir.

Doğruluk ölçütü genel başarı oranını yansıtmakla birlikte, sınıf dağılımının dengesiz olduğu durumlarda değerlendirme için tek başına yeterli olmayabilir.

2.6.2. F1-Skoru

F1-skoru, kesinlik (precision) ve duyarlılığın (recall) harmonik ortalamasıdır:

$$F1 = \frac{2*(Kesinlik*Duyarlılık)}{Kesinlik+Duyarlılık} \ . \quad (2)$$

Kesinlik ve duyarlılık aşağıdaki şekilde tanımlanır:

$$Kesinlik = \frac{DP}{DP+YP} \ , \quad (3)$$

$$Duyarlılık = \frac{DP}{DP+YN} \ . \quad (4)$$

F1-skoru, özellikle sınıflardan birinin daha az temsil edildiği veri yapılarında modelin performansını daha dengeli biçimde yansıtmaktadır [22].

2.6.3. ROC-AUC

ROC (Receiver Operating Characteristic) eğrisi, doğru pozitif oranı ile yanlış pozitif oranı arasındaki ilişkiyi farklı eşik değerleri altında gösterir. ROC eğrisi altında kalan alan (AUC), modelin sınıfları ayırt etme yeteneğini özetleyen eşik-bağımsız bir performans ölçütüdür [23].

AUC değeri 0,5 ile 1 arasında değişir; daha yüksek değerler modelin sınıfları ayırt etme gücünün arttığını gösterir.

Tüm performans ölçütleri çapraz doğrulama tekrarları boyunca hesaplanmış ve sonuçlar ortalama ± standart sapma biçiminde raporlanmıştır.

2.7. Açıklanabilirlik Analizi (SHAP)

Bu çalışmada model performansı yalnızca sayısal ölçütlerle değil, seçilen özelliklerin tahminlere katkısı açısından da değerlendirilmiştir. SHAP (Shapley Additive ExPlanations), işbirlikçi oyun teorisindeki Shapley değerine dayalı olarak, her bir özelliğin model tahmini üzerindeki katkısını hesaplayan bir açıklanabilirlik yöntemidir [12, 24].

Bu çalışmada SHAP analizi, model seçim kriterine göre belirlenen nihai sınıflandırma modeli üzerinde uygulanmış ve değişkenlerin küresel önem düzeyi ortalama mutlak SHAP değerleri kullanılarak değerlendirilmiştir. Analiz, model performansını karşılaştırmak amacıyla değil, seçilen modelin tahmin sürecinde hangi değişkenlerin ne ölçüde etkili olduğunu incelemek amacıyla gerçekleştirilmiştir.

Değişken katkılarının görselleştirilmesinde SHAP özet grafiği (beeswarm) tercih edilmiş olup elde edilen görseller ve yorumlar Bulgular bölümünde sunulmuştur.

Algoritma 1. *Özellik seçimi ve sınıflandırma için deneysel süreç*

Girdi:
- Veri seti $D = (X, y)$
- Özellik oranları $R = \{0.25, 0.50, 0.75, 1.00\}$
- Özellik seçimi yaklaşımları
  $M = \{\text{Karşılıklı Bilgi, L1 düzenlileştirme, Ağaç tabanlı önem}\}$
- Sınıflandırıcılar $C = \{\text{L2 düzenlileştirmeli Lojistik Regresyon, Rastgele Orman}\}$
- Çapraz doğrulama parametreleri $T\ (tekrar\ sayısı), K\ (kat\ sayısı)$

Çıktı:
- Her (m, r, c) için performans ölçütleri $(\text{ortalama} \pm \text{standart sapma})$
- Seçilen model için SHAP görselleri

Adımlar
1. Sonuç kümesini $S \leftarrow \emptyset$ olarak başlat.
2. Tekrarlı ve tabakalı K-katlı çapraz doğrulama bölünmelerini oluştur
   $\{(Eğitim_{t,k}, Test_{t,k})\}, \quad t = 1, \dots, T; \quad k = 1, \dots, K$
3. Her tekrar $t = 1, \dots, T$ ve her kat $k = 1, \dots, K$ için:
   3.1 Veriyi eğitim ve test olarak ayır.
   3.2 Eğitim verisi üzerinde ön işlem uygula (eksik değer doldurma, one-hot kodlama, sıfır varyanslı değişkenlerin çıkarılması ve standartlaştırma)
   3.3 Eğitim verisinden öğrenilen parametreleri test verisine uygula.
   3.4 $p \leftarrow$ eğitim verisi üzerinde dönüşüm sonrası toplam özellik sayısı
   3.5 Her özellik seçimi yöntemi $m \in M$ ve her oran $r \in R$ için:
      a) $k_r \leftarrow \max(1, yuvarla(r\ x\ p))$
      b) Eğitim verisi üzerinde m yöntemi ile en yüksek skorlu k_r özelliği seç.
      c) Seçilen özellikler kullanılarak eğitim ve test alt veri kümelerini oluştur.
      d) Her sınıflandırıcı için $c \in C$:
         i. Modeli eğitim verisinde eğit.
         ii. Test verisinde tahmin üret.
         iii. Doğruluk, F1 ve ROC-AUC hesapla.
         iv. Sonuçları S'ye ekle.
4. Tüm tekrarlar tamamlandıktan sonra her (m, r, c) için performans ölçütlerini
   $Ortalama \pm Standart\ Sapma$ biçiminde hesapla.
5. Model seçim kriterine göre belirlenen en iyi model ve ilgili özellik alt kümesi üzerinde SHAP analizi gerçekleştir ve değişken katkılarını görselleştir

3. BULGULAR

Bu bölümde Adult ve Heart veri setleri üzerinde gerçekleştirilen deneysel analizlerin sonuçları sunulmaktadır. Performans değerlendirmeleri doğruluk, F1-skoru ve ROC-AUC ölçütleri kullanılarak gerçekleştirilmiş ve sonuçlar tekrarlı çapraz doğrulama sürecinden elde edilen ortalama ve standart sapma değerleri ile raporlanmıştır.

3.1. Adult Veri Seti Sonuçları

Adult veri seti için farklı özellik seçimi yöntemleri, değişken oranları ve sınıflandırma modelleri ile elde edilen performans sonuçları Çizelge 2'de sunulmuştur.

Çizelge 2. *Adult veri seti için farklı özellik seçimi yöntemleri ve değişken oranlarına göre sınıflandırma performansı*

Özellik seçimi yöntemi	Oran	Sınıflandırma modeli	Doğruluk	F1-skoru	ROC-AUC
Tüm değişkenler	1	L2	0,847 (0,003)	0,635 (0,008)	0,900 (0,003)
L1	0,25	L2	0,845 (0,004)	0,627 (0,011)	0,900 (0,003)
L1	0,5	L2	0,846 (0,004)	0,630 (0,011)	0,900 (0,003)
L1	0,75	L2	0,846 (0,003)	0,629 (0,007)	0,900 (0,003)
Karşılıklı bilgi	0,25	L2	0,842 (0,004)	0,616 (0,011)	0,896 (0,004)
Karşılıklı bilgi	0,5	L2	0,846 (0,003)	0,630 (0,010)	0,900 (0,003)
Karşılıklı bilgi	0,75	L2	0,847 (0,003)	0,635 (0,008)	0,900 (0,003)
Ağaç tabanlı önem	0,25	L2	0,843 (0,004)	0,619 (0,011)	0,897 (0,004)
Ağaç tabanlı önem	0,5	L2	0,846 (0,004)	0,632 (0,012)	0,900 (0,003)
Ağaç tabanlı önem	0,75	L2	0,847 (0,003)	0,635 (0,008)	0,900 (0,003)
Tüm değişkenler	1	RO	0,866 (0,002)	0,687 (0,005)	0,918 (0,003)
L1	0,25	RO	0,867 (0,003)	0,692 (0,009)	0,919 (0,003)
L1	0,5	RO	0,867 (0,003)	0,692 (0,009)	0,920 (0,003)
L1	0,75	RO	0,867 (0,002)	0,687 (0,007)	0,919 (0,003)
Karşılıklı bilgi	0,25	RO	0,862 (0,003)	0,671 (0,008)	0,917 (0,003)
Karşılıklı bilgi	0,5	RO	0,865 (0,003)	0,685 (0,007)	0,918 (0,003)
Karşılıklı bilgi	0,75	RO	0,866 (0,002)	0,688 (0,005)	0,918 (0,003)
Ağaç tabanlı önem	0,25	RO	0,864 (0,003)	0,679 (0,008)	0,917 (0,003)
Ağaç tabanlı önem	0,5	RO	0,866 (0,003)	0,690 (0,008)	0,918 (0,003)
Ağaç tabanlı önem	0,75	RO	0,866 (0,002)	0,688 (0,005)	0,918 (0,003)

L1: L1 düzenlileştirme; L2: L2 Düzenlileştirmeli Lojistik Regresyon; RO: Rastgele Orman
Not: Parantez içindeki değerler tekrarlı çapraz doğrulama sonuçlarına ait standart sapmaları göstermektedir.

Sonuçlar incelendiğinde Rastgele Orman modelinin çoğu senaryoda L2 düzenlileştirmeli lojistik regresyon modeline kıyasla daha yüksek performans değerleri ürettiği görülmektedir. Referans senaryoda (tüm değişkenlerin kullanıldığı durum) Rastgele Orman modeli için doğruluk, F1-skoru ve ROC-AUC değerleri sırasıyla 0,866; 0,687 ve 0,918 olarak elde edilmiştir. L2 düzenlileştirmeli lojistik regresyon modeli için ise aynı ölçütler sırasıyla 0,847; 0,635 ve 0,900 olarak hesaplanmıştır.

Özellik seçimi yöntemlerinin performans üzerindeki etkisi değerlendirildiğinde, farklı yöntemler arasında performans açısından belirgin ve tutarlı bir üstünlük gözlenmemektedir. L1 düzenlileştirme, karşılıklı bilgi ve ağaç tabanlı önem yaklaşımları ile elde edilen performans değerleri genel olarak birbirine oldukça yakın düzeydedir.

Değişken sayısının azaltılması açısından değerlendirildiğinde, %25, %50 ve %75 oranlarında oluşturulan özellik alt kümeleri ile elde edilen performans değerlerinin referans senaryoya oldukça yakın olduğu görülmektedir. Bu durum Adult veri setinde model performansının daha küçük özellik alt kümeleri kullanılarak da büyük ölçüde korunabildiğini göstermektedir.

Bununla birlikte, gözlenen performans farklarının standart sapma değerleri dikkate alındığında sınırlı düzeyde olduğu görülmektedir. Bu durum, yöntemler arasında istatistiksel olarak anlamlı bir fark bulunmayabileceğine işaret etmektedir.

Bu nedenle, Adult veri setinde benzer performans değerleri elde edildiğinden, model seçimi performanstan ziyade değişken sayısını minimize etme kriterine dayalı olarak gerçekleştirilmiştir.

3.2. Heart Veri Seti Sonuçları

Heart veri seti için farklı özellik seçimi yöntemleri, değişken oranları ve sınıflandırma modelleri ile elde edilen performans sonuçları Çizelge 3'te sunulmuştur.

Çizelge 3. *Heart veri seti için farklı özellik seçimi yöntemleri ve değişken oranlarına göre sınıflandırma performansı*

Özellik seçimi yöntemi	Oran	Sınıflandırma modeli	Doğruluk	F1-skoru	ROC-AUC
Tüm değişkenler	1	L2	0,846 (0,027)	0,820 (0,032)	0,908 (0,034)
L1	0,25	L2	0,795 (0,042)	0,764 (0,050)	0,869 (0,045)
L1	0,5	L2	0,828 (0,042)	0,802 (0,050)	0,888 (0,040)
L1	0,75	L2	0,844 (0,030)	0,818 (0,035)	0,899 (0,035)
Karşılıklı bilgi	0,25	L2	0,804 (0,039)	0,772 (0,050)	0,868 (0,039)
Karşılıklı bilgi	0,5	L2	0,847 (0,035)	0,821 (0,043)	0,899 (0,039)
Karşılıklı bilgi	0,75	L2	0,849 (0,036)	0,827 (0,042)	0,906 (0,034)
Ağaç tabanlı önem	0,25	L2	0,802 (0,042)	0,768 (0,054)	0,860 (0,043)
Ağaç tabanlı önem	0,5	L2	0,838 (0,037)	0,811 (0,049)	0,895 (0,036)
Ağaç tabanlı önem	0,75	L2	0,840 (0,035)	0,811 (0,039)	0,891 (0,036)
Tüm değişkenler	1	RO	0,842 (0,035)	0,815 (0,039)	0,905 (0,034)
L1	0,25	RO	0,841 (0,057)	0,820 (0,057)	0,871 (0,042)
L1	0,5	RO	0,822 (0,037)	0,792 (0,045)	0,890 (0,038)
L1	0,75	RO	0,838 (0,032)	0,813 (0,038)	0,897 (0,036)
Karşılıklı bilgi	0,25	RO	0,825 (0,067)	0,795 (0,077)	0,868 (0,037)
Karşılıklı bilgi	0,5	RO	0,822 (0,036)	0,793 (0,044)	0,900 (0,038)
Karşılıklı bilgi	0,75	RO	0,837 (0,037)	0,812 (0,043)	0,906 (0,033)
Ağaç tabanlı önem	0,25	RO	0,807 (0,052)	0,771 (0,060)	0,856 (0,035)
Ağaç tabanlı önem	0,5	RO	0,827 (0,043)	0,798 (0,052)	0,893 (0,039)
Ağaç tabanlı önem	0,75	RO	0,831 (0,036)	0,802 (0,044)	0,895 (0,034)

L1: L1 düzenlileştirme; L2: L2 Düzenlileştirmeli Lojistik Regresyon; RO: Rastgele Orman
Not: Parantez içindeki değerler tekrarlı çapraz doğrulama sonuçlarına ait standart sapmaları göstermektedir.

Adult veri setinden farklı olarak bu veri setinde, L2 düzenlileştirmeli lojistik regresyon modelinin birçok senaryoda Rastgele Orman modeli ile benzer veya yer yer daha yüksek performans değerleri ürettiği görülmektedir. Referans senaryoda (tüm değişkenlerin kullanıldığı durum) lojistik regresyon modeli için doğruluk, F1-skoru ve ROC-AUC değerleri sırasıyla 0,846; 0,820 ve 0,908 olarak elde edilirken, Rastgele Orman modeli için bu değerler sırasıyla 0,842; 0,815 ve 0,905 olarak hesaplanmıştır.

Özellik seçimi yöntemlerinin performans üzerindeki etkisi değerlendirildiğinde, değişken sayısının önemli ölçüde azaltıldığı bazı senaryolarda performans değerlerinde düşüş gözlendiği dikkat çekmektedir. Özellikle değişkenlerin %25 oranında seçildiği durumlarda doğruluk ve F1-skoru değerlerinin diğer oranlara kıyasla daha düşük olduğu görülmektedir.

Buna karşılık değişkenlerin yaklaşık %50–%75 oranında korunduğu özellik alt kümelerinde elde edilen performans değerleri referans senaryoya oldukça yakın düzeydedir. Özellikle karşılıklı bilgi yöntemi ile değişkenlerin %75'inin seçildiği durumda lojistik regresyon modeli en yüksek doğruluk değerine ulaşmıştır. Bu nedenle, Heart veri setinde model seçimi, doğrudan performans ölçütlerine dayalı olarak gerçekleştirilmiştir.

Bu sonuçlar, Heart veri setinde çok küçük özellik alt kümelerinin model performansını olumsuz etkileyebileceğini, ancak değişkenlerin büyük bölümünün korunduğu senaryolarda benzer performans değerlerinin elde edilebildiğini göstermektedir.

Standart sapma değerlerinin Adult veri setine kıyasla daha yüksek olması, Heart veri setinin daha küçük örneklem büyüklüğüne sahip olmasından kaynaklanan varyans artışına işaret etmektedir.

3.3. SHAP Analizi

Adult ve Heart veri setleri için elde edilen sonuçlar birlikte değerlendirildiğinde, Bölüm 2.5.3'te açıklanan model seçim kriteri dikkate alınarak SHAP analizinde kullanılacak modeller belirlenmiştir. Performans değerlerinin birbirine oldukça yakın olduğu senaryolarda, model karmaşıklığını azaltmak amacıyla daha az sayıda değişken içeren yapı tercih edilmiştir. Performans açısından belirgin biçimde öne çıkan durumlarda ise en yüksek performans değerini üreten model seçilmiştir. Bu doğrultuda analiz için seçilen modeller Çizelge 4'te gösterilmiştir. SHAP analizi, her veri seti için seçilen nihai modelin eğitildiği özellik alt kümesi üzerinde gerçekleştirilmiştir.

Çizelge 4. *SHAP analizi için seçilen modeller*

Veri seti	Özellik seçimi yöntemi	Oran	Sınıflandırma modeli	Seçim gerekçesi
Adult	L1 düzenlileştirme	0,25	Rastgele Orman	Benzer performans, daha az değişken
Heart	Karşılıklı bilgi	0,75	L2 düzenlileştirmeli lojistik regresyon	En yüksek performans

Adult veri seti için benzer performans değerleri üreten modeller arasında daha az sayıda değişken içeren Rastgele Orman + L1 (%25) kombinasyonu tercih edilmiş, Heart veri seti için ise en yüksek performans değerini sağlayan L2 düzenlileştirmeli lojistik regresyon + Karşılıklı bilgi (%75) modeli seçilmiştir. Seçilen modeller için elde edilen SHAP sonuçları, değişkenlerin model tahminlerine katkısını incelemek amacıyla değerlendirilmiştir. Bu kapsamda değişkenlerin önem sıralarını ve tahmin üzerindeki yönsel etkilerini birlikte göstermesi nedeniyle SHAP özet grafiği olarak arı sürüsü grafiği (beeswarm) kullanılmıştır.

Bu kapsamda, SHAP analizi sonucunda Adult veri setinde model 24 değişken, Heart veri setinde ise 10 değişken üzerinden eğitilmiş ve değerlendirilmiştir. SHAP değerleri bu seçilen değişkenlerin tamamı için hesaplanmıştır. Ancak görsel sunum aşamasında, yorumlanabilirliği artırmak amacıyla değişkenlerin tamamı yerine en yüksek katkıya sahip olanlar gösterilmiştir. Bu doğrultuda, Çizelge 5'te her iki veri seti için karşılaştırılabilirliği sağlamak amacıyla ortalama mutlak SHAP değerine göre en yüksek katkıya sahip ilk 10 değişken ile bunların Türkçe karşılıkları sunulmuştur.

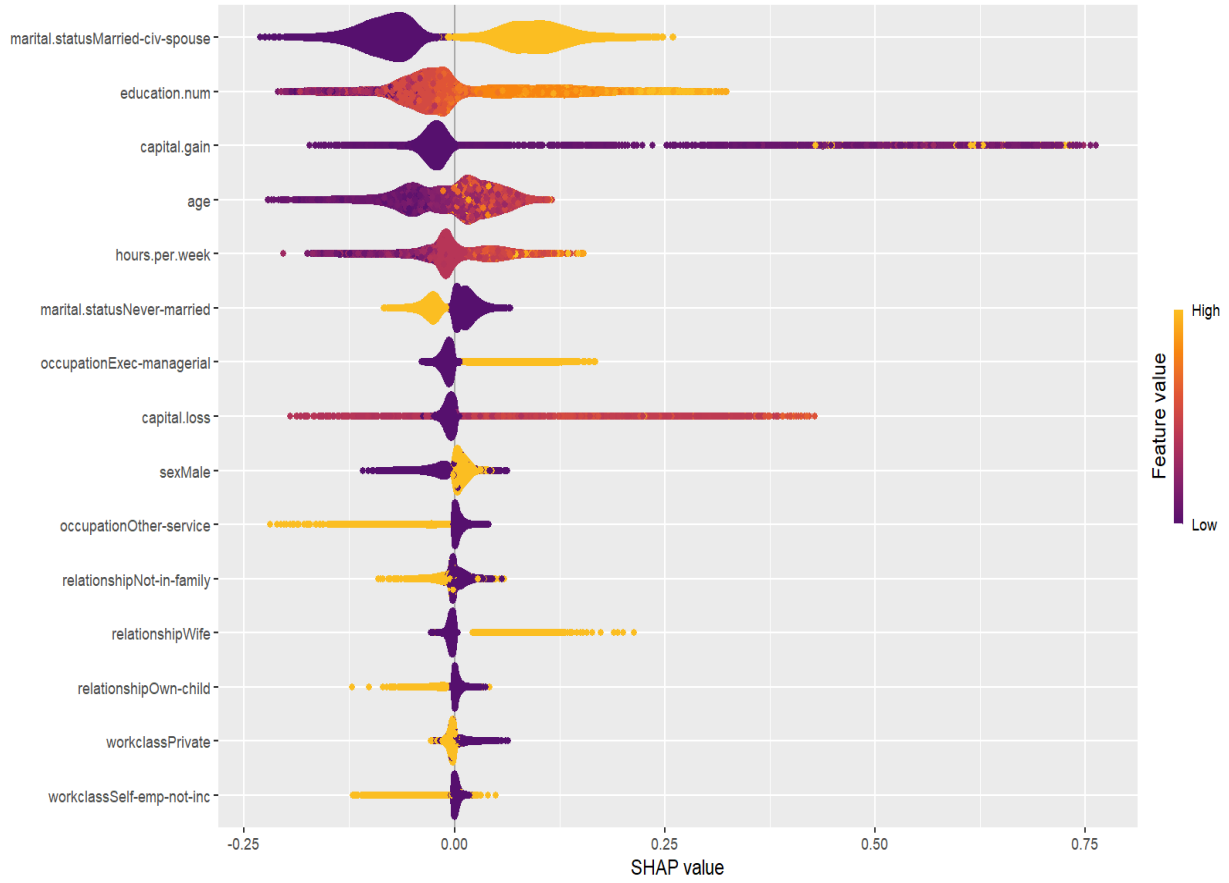
Arı sürüsü grafiklerinde ise Adult veri seti için daha yüksek sayıda değişken arasından görsel olarak ayırt edilebilirliği koruyacak optimal sayıda değişken gösterilirken, Heart veri setinde değişken sayısının sınırlı olması nedeniyle seçilen değişkenlerin tamamı grafikte yer almaktadır.

Çizelge 5. *SHAP analizine göre seçilen en yüksek katkıya sahip ilk 10 değişken ve Türkçe karşılıkları*

Sıra	Adult Veri Seti Değişkeni	Türkçe karşılığı	Heart Veri Seti Değişkeni	Türkçe karşılığı
1	marital.statusMarried-civ-spouse	Evli (eş ile birlikte)	number_of_major_vessels	Majör damar sayısı
2	education.num	Eğitim düzeyi	thal	Talasemi sonucu
3	capital.gain	Sermaye kazancı	chest	Göğüs ağrısı tipi
4	age	Yaş	sex	Cinsiyet
5	hours.per.week	Haftalık çalışma süresi	exercise_induced_angina	Egzersize bağlı anjina
6	marital.statusNever-married	Hiç evlenmemiş	oldpeak	ST depresyonu
7	occupationExec-managerial	Yönetici/üst düzey meslek	maximum_heart_rate_ achieved	Maksimum kalp atış hızı
8	capital.loss	Sermaye kaybı	serum_cholesterol	Serum kolesterol
9	sexMale	Erkek	slope	ST segment eğimi
10	occupationOther-service	Diğer hizmet meslekleri	age	Yaş

Şekil 1'de Adult veri seti için elde edilen SHAP arı sürüsü grafiği sunulmaktadır. SHAP özet grafiğinde x ekseni, ilgili değişkenin model tahmini üzerindeki katkı yönünü ve büyüklüğünü göstermektedir. Renkler ise değişken değerinin düşük (mor) ve yüksek (sarı) olduğu durumları temsil ederek, değişken değerinin artışı ile tahmin üzerindeki etkinin yönünü ortaya koymaktadır. Grafik incelendiğinde model tahminlerine en yüksek katkıyı sağlayan değişkenlerin başında medeni durum (marital.status), eğitim düzeyi (education.num), sermaye kazancı (capital.gain) ve yaş (age) değişkenlerinin geldiği görülmektedir. Özellikle medeni durum değişkenine ait evli (Married-civ-spouse) kategorisinde yer alan gözlemlerin büyük ölçüde pozitif SHAP değerleri ile ilişkili olduğu dikkat çekmektedir. Bu durum, ilgili kategoride yer alan gözlemlerin model tarafından yüksek gelir sınıfı ile daha güçlü biçimde ilişkilendirildiğini göstermektedir. Benzer şekilde eğitim düzeyini temsil eden education.num değişkeninde daha yüksek değerlerin çoğunlukla pozitif SHAP katkıları ile ilişkili olduğu görülmektedir. Ayrıca, bazı değişkenlerde düşük değerlerin negatif SHAP katkıları ile ilişkilendiği görülmekte olup, bu durum modelin karar mekanizmasının değişken değerine duyarlı olduğunu göstermektedir. Bunun yanı sıra sermaye kazancı (capital.gain) ve haftalık çalışma süresi (hours.per.week) gibi ekonomik göstergelerin de model tahminleri üzerinde belirgin bir etkisi olduğu gözlenmektedir. Bu durum, ilgili değişkenlerin yalnızca önemli olmadığını, aynı zamanda değer düzeylerine bağlı olarak model tahminleri üzerinde yönsel bir etki oluşturabildiğini göstermektedir.

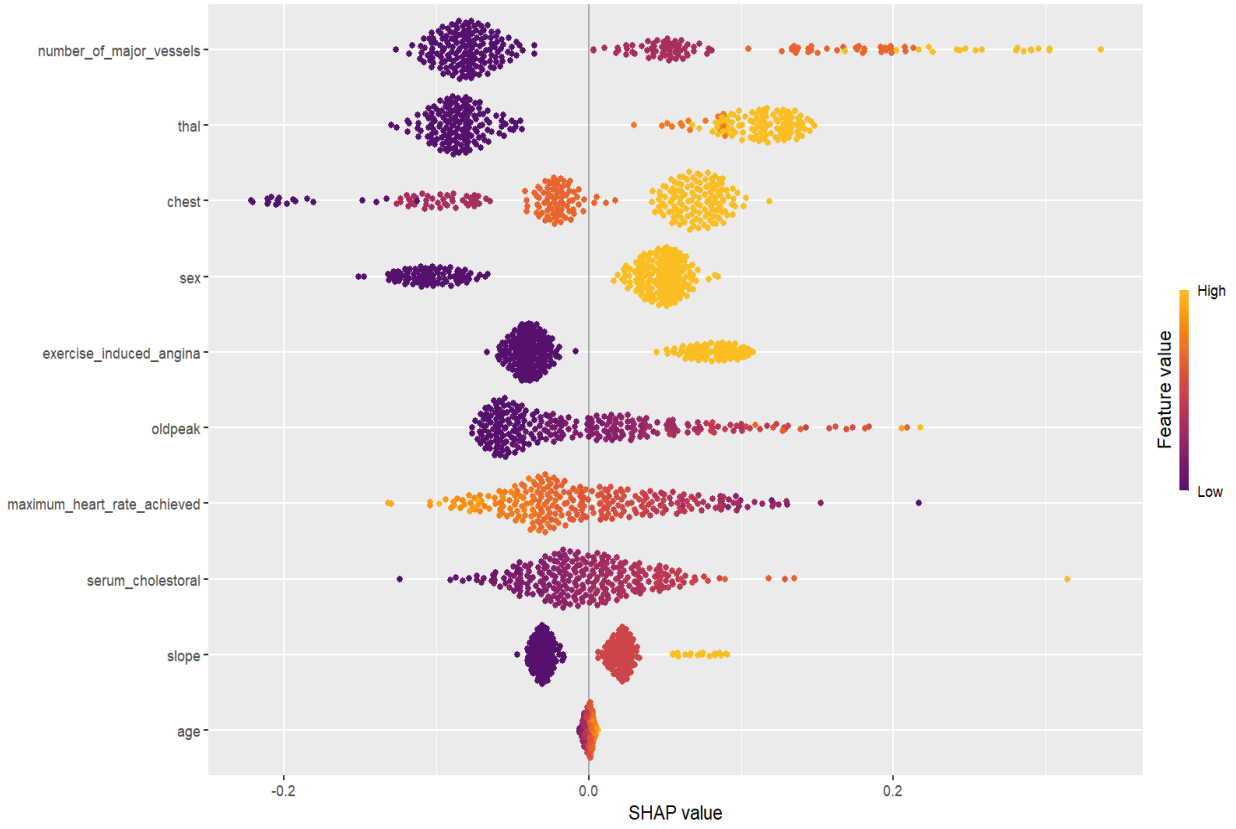
Genel olarak değerlendirildiğinde, Adult veri setinde yüksek gelir tahminlerinin demografik ve sosyoekonomik özelliklerle ilişkilendirildiği görülmektedir. Bu bulgular, Adult veri seti üzerinde gerçekleştirilen önceki makine öğrenmesi çalışmalarında eğitim düzeyi, sermaye kazancı ve demografik değişkenlerin gelir tahmini açısından ayırt edici özellikler arasında yer aldığını gösteren sonuçlarla genel olarak uyumludur [25]. Ayrıca SHAP tabanlı yorumlanabilirlik yaklaşımlarının tabular veri setlerinde değişken katkılarının analiz edilmesinde yaygın olarak kullanıldığı bilinmektedir [12].



Şekil 1. *Adult veri seti için SHAP arı sürüsü grafiği*

Şekil 2'de Heart veri seti için elde edilen SHAP arı sürüsü grafiği sunulmaktadır. Grafik incelendiğinde model tahminlerine en yüksek katkıyı sağlayan değişkenlerin başında majör damar sayısı (number_of_major_vessels), talasemi sonucu, göğüs ağrısı tipi (chest) ve cinsiyet (sex) değişkenlerinin geldiği görülmektedir. Özellikle majör damar sayısı değişkeninin geniş bir SHAP değer aralığına sahip olduğu, hem pozitif hem negatif yönde güçlü katkılar ürettiği ve model tahminleri üzerinde belirgin bir etkiye sahip olduğu dikkat çekmektedir. Bu durum, ilgili değişkenin model karar sürecinde yüksek ayırt ediciliğe sahip olduğunu göstermektedir. Benzer şekilde talasemi ve göğüs ağrısı tipi değişkenlerinin de model çıktısı üzerinde önemli katkılar sağladığı gözlenmektedir. Bunun yanı sıra egzersize bağlı anjina (exercise_induced_angina), ST depresyonu (oldpeak) ve maksimum kalp atış hızı (maximum_heart_rate_achieved) gibi değişkenlerin de model tahminlerine anlamlı katkılar sunduğu görülmektedir. Geniş SHAP değer aralığına sahip değişkenler, modelin karar sürecinde daha ayırt edici bir rol üstlenmektedir.

Genel olarak değerlendirildiğinde, Heart veri setinde model tahminlerinin büyük ölçüde kardiyovasküler durumla ilişkili klinik göstergeler tarafından şekillendiği anlaşılmaktadır. Bu bulgular, kalp hastalığı tanısına yönelik klasik klinik çalışmalarda kullanılan değişkenlerle genel olarak uyumludur. Detrano ve arkadaşları [26] tarafından geliştirilen tanı modelinde damar sayısı, göğüs ağrısı tipi ve egzersize bağlı anjina gibi değişkenler koroner arter hastalığının değerlendirilmesinde kullanılan önemli klinik göstergeler arasında yer almaktadır.



Şekil 2. *Heart veri seti için SHAP arı sürüsü grafiği*

4. TARTIŞMA

Bu çalışmada farklı özellik seçimi yöntemlerinin sınıflandırma performansı üzerindeki etkileri iki farklı veri yapısına sahip veri seti üzerinde incelenmiş ve seçilen modellerin karar mekanizmaları SHAP analizi kullanılarak yorumlanmıştır. Elde edilen bulgular, incelenen veri setleri bağlamında, özellik seçimi yöntemlerinin model performansı üzerindeki etkisinin veri setinin yapısına bağlı olarak farklılaşabildiğine işaret etmektedir. Bu durum, özellik seçimi yöntemlerinin veri dağılımı ve örneklem yapısındaki değişimlere duyarlı sonuçlar üretebildiğini vurgulayan çalışmalarla da örtüşmektedir [27, 28].

Bu çalışmada kullanılan özellik seçimi yöntemleri, filtre ve gömülü yaklaşımları temsil edecek şekilde seçilmiş olup farklı özellik seçimi paradigmasının sınıflandırma performansı üzerindeki etkilerinin karşılaştırmalı olarak incelenmesi amaçlanmıştır. Model performansı değerlendirilirken yalnızca doğruluk ölçütüne bağlı kalınmamış, sınıf dengesizliğinin olası etkisini daha dengeli biçimde değerlendirebilmek amacıyla F1-skoru ve ROC-AUC ölçütleri de birlikte raporlanmıştır.

Bu yaklaşım, özellikle Adult veri seti gibi literatürde yaygın biçimde kullanılan veri yapılarında model performansının daha dengeli ve çok boyutlu biçimde değerlendirilmesine katkı sağlamaktadır [17, 29].

Adult veri seti için elde edilen sonuçlar incelendiğinde, farklı özellik seçimi yöntemleri ve değişken oranları kullanıldığında model performansında belirgin bir değişim oluşmadığı görülmektedir. Özellikle Rastgele Orman modelinin tüm senaryolarda benzer doğruluk, F1 ve ROC-AUC değerleri ürettiği gözlenmiştir. Bu durum Adult veri setinde yüksek gelir tahmininin büyük ölçüde kullanılan sınıflandırma algoritmasının öğrenme kapasitesine bağlı olduğunu ve değişken sayısının azaltılmasının performansı önemli ölçüde etkilemediğini göstermektedir.

Bu bulgu, Adult veri seti üzerinde yapılan güncel çalışmalarla da benzer şekilde rapor edilmiştir. Özellikle Sena ve Machado [17] çalışmasında, modele dahil edilen hassas değişkenlerin çıkarılması veya eklenmesinin performans ölçütleri üzerinde yalnızca sınırlı değişiklikler oluşturduğu, buna karşın model çıktılarında daha çok adalet (fairness) metriklerinin etkilendiği gösterilmiştir. Bu durum, Adult veri setinde tahmin gücünün belirli bir değişken alt kümesine yoğunlaşmadığını, aksine çok sayıda değişken arasında dağıldığını göstermektedir.

Adult veri setinin literatürde yaygın ve tekrar tekrar kullanılan bir karşılaştırma veri seti olması, burada gözlenen görece stabil performans yapısının veri setinin doğası ile ilişkili olabileceğini düşündürmektedir [17, 29]. Bu bağlamda, Adult veri setinin yüksek düzeyde değişken fazlalığı (redundancy) içeren bir yapı sergilediği ve bu nedenle farklı özellik seçimi yöntemlerinin benzer performans sonuçlarına ulaşabildiği söylenebilir. Bu tür veri yapılarında farklı özellik seçimi yöntemlerinin benzer performans üretmesine rağmen seçilen alt değişken kümelerinin farklılaşabilmesi, özellik seçimi süreçlerinin stabilite boyutunun da dikkate alınması gerektiğini göstermektedir [27].

Bununla birlikte özellik seçimi sayesinde daha az sayıda değişken kullanılarak benzer performans değerlerinin elde edilebilmesi, modelin daha sade ve yorumlanabilir hâle getirilebilmesi açısından önemli bir avantaj sağlamaktadır. Bu durum, performans-yorumlanabilirlik arasındaki dengenin sağlanması açısından özellik seçimi sürecinin pratik önemini de ortaya koymaktadır.

Bu bulgular, literatürde farklı özellik seçimi yöntemlerinin benzer performans düzeyleri sunabildiğini ancak seçilen değişken alt kümelerinin farklılaşabildiğini gösteren çalışmalarla da paralellik göstermektedir [27, 28, 30]. Dolayısıyla, özellik seçimi yöntemleri arasındaki temel farkın çoğu zaman performanstan ziyade seçilen değişkenlerin yorumsal anlamı ve modelin açıklanabilirliği üzerinde ortaya çıktığı söylenebilir.

Heart veri setinde ise farklı bir yapı gözlenmiştir. Daha küçük örneklem büyüklüğüne sahip olan bu veri setinde özellik seçimi yöntemlerinin model performansı üzerinde daha belirgin etkiler oluşturabildiği görülmüştür. Özellikle karşılıklı bilgi yönteminin %75 değişken oranı ile birlikte kullanıldığı durumda lojistik regresyon modelinin en yüksek performans değerlerine ulaştığı belirlenmiştir.

Bu durum, daha küçük ve daha düşük boyutlu veri setlerinde gereksiz veya gürültülü değişkenlerin model performansı üzerinde daha belirgin bir olumsuz etki oluşturabildiğini, dolayısıyla özellik seçimi sürecinin bu tür veri yapılarında daha kritik hâle geldiğini göstermektedir. Bu tür veri setlerinde, eğitim verisindeki küçük değişimlerin bile seçilen değişken alt kümelerini etkileyebilmesi, model performansındaki dalgalanmaların daha belirgin hâle gelmesine neden olabilmektedir [28].

Bu bulgu, Heart veri seti üzerinde gerçekleştirilen çalışmalarla model performansının hem kullanılan algoritmaya hem de seçilen özellik alt kümesine duyarlı olabildiğini gösteren sonuçlarla uyumludur [16, 18].

Özellikle Shah ve ark. [18] çalışmasında, toplam değişken sayısının azaltılmasıyla model performansının iyileşebildiği ve daha kompakt modellerin daha başarılı sonuçlar üretebildiği rapor edilmiştir.

Benzer şekilde, Heart veri seti üzerinde gerçekleştirilen çeşitli çalışmalarda farklı model ve özellik seçimi kombinasyonlarının performans üzerinde belirgin farklılıklar oluşturabildiği rapor edilmiştir [31, 32].

Bu sonuçlar birlikte değerlendirildiğinde, özellikle daha küçük örneklem büyüklüğüne sahip veri setlerinde özellik seçim sürecinin model performansı açısından daha kritik bir rol oynayabileceği anlaşılmaktadır.

SHAP analizi sonuçları, seçilen modellerin karar mekanizmalarının daha ayrıntılı biçimde incelenmesine olanak sağlamıştır. Adult veri setinde en yüksek katkıya sahip değişkenlerin medeni durum, eğitim düzeyi, sermaye kazancı ve yaş gibi demografik ve sosyoekonomik özellikler olduğu görülmüştür. Bu bulgular, model tahminlerinin yalnızca istatistiksel ilişkilerden değil, aynı zamanda anlamlı ve yorumlanabilir değişkenlerden beslendiğini göstermektedir. Bu durum, açıklanabilir yapay zekâ yaklaşımlarının model kararlarının insan tarafından anlaşılabilir değişkenler üzerinden yorumlanmasını mümkün kıldığını vurgulayan çalışmalarla da örtüşmektedir [12].

Heart veri setinde ise model tahminlerinin majör damar sayısı, göğüs ağrısı tipi ve egzersize bağlı anjina gibi klinik değişkenlerle güçlü biçimde ilişkilendirildiği görülmüştür.

Bu bulgular, kalp hastalığı tahminine yönelik çalışmalarda bu tür klinik göstergelerin belirleyici rol oynadığını gösteren önceki araştırmalarla paralellik göstermektedir. Özellikle damar sayısı, göğüs ağrısı tipi ve egzersize bağlı göstergelerin kalp hastalığı teşhisinde kritik belirleyiciler olduğu farklı çalışmalarda da vurgulanmaktadır [18, 31]. Bu sonuçlar, klinik veri setlerinde modelin karar sürecinin genellikle tıbbi açıdan anlamlı değişkenler etrafında şekillendiğini ortaya koyan literatür ile de uyumludur [16, 18].

Bununla birlikte, SHAP analizinde yalnızca değişkenlerin önem sıraları değil, aynı zamanda katkı yönü ve büyüklüğü de dikkate alınarak modelin karar mekanizması daha sistematik biçimde yorumlanmıştır. Özellikle SHAP değer aralığı geniş olan değişkenlerin modelin karar sürecinde daha ayırt edici bir rol üstlendiği görülmüştür. Bu çerçevede, SHAP değerlerinin işareti değişkenin tahmin üzerindeki yönünü, mutlak büyüklüğü ise katkı düzeyini yansıtmaktadır. Bu durum, modelin yalnızca değişkenlerin varlığına değil, aynı zamanda değer düzeylerine duyarlı bir karar yapısı geliştirdiğini göstermektedir.

Elde edilen bulgular, farklı özellik seçimi yöntemlerinin benzer performans düzeyleri sunabildiğini, ancak hesaplama maliyetleri açısından farklılık gösterebileceğini düşündürmektedir. Bu çalışmada deneysel süreç tekrarlı çapraz doğrulama ve çoklu model–özellik seçimi kombinasyonlarını içerecek şekilde kapsamlı biçimde kurgulandığından, hesaplama süresi doğrudan karşılaştırmalı bir ölçüt olarak raporlanmamıştır. Bununla birlikte, özellikle büyük veri setleri ve gerçek zamanlı uygulamalar açısından yöntem seçiminde yalnızca performans değil, hesaplama verimliliğinin de dikkate alınması gerektiği açıktır.

Genel olarak değerlendirildiğinde bu çalışma, farklı özellik seçimi yöntemlerinin sınıflandırma performansı üzerindeki etkilerinin veri setinin yapısına bağlı olarak değişebileceğini göstermektedir. Ayrıca SHAP analizi kullanılarak seçilen modellerin karar mekanizmaları incelenmiş ve model tahminlerinin hangi değişkenlerden ve hangi yönde etkilendiği ortaya konulmuştur. Bu yaklaşım, yalnızca model performansını değerlendirmekle kalmayıp aynı zamanda modelin yorumlanabilirliğini de artırmaktadır.

Bu çalışma, geniş veri seti çeşitliliği sunmayı amaçlayan bir benchmark çalışması olmaktan ziyade, özellik seçimi yöntemlerinin etkilerini kontrollü bir deneysel çerçevede incelemeyi hedeflemektedir. Bu nedenle elde edilen bulguların kullanılan veri setleri bağlamında değerlendirilmesi gerektiği vurgulanmalıdır. Bununla birlikte, çalışma iki veri seti ile sınırlıdır ve farklı veri yapılarının analize dahil edilmesi gelecekteki çalışmalar açısından önemli bir araştırma yönü olarak görülmektedir. Gelecek çalışmalarda, daha çeşitli veri setleri ve farklı özellik seçimi stratejileri kullanılarak, model performansı ile yorumlanabilirlik arasındaki ilişkinin daha kapsamlı ve genellenebilir biçimde incelenmesi önerilmektedir.

ÇIKAR ÇATIŞMASI/ÇAKIŞMASI BİLDİRİMİ

Yazarlar arasında çıkar çatışması/çakışması bulunmamaktadır.

YAZAR KATKI ORANLARI

Derya Turfan Uğurlu: Kavramlaştırma, Metodoloji, Yazılım, İçerik analizi, Araştırma, Makalenin yazımı – Orijinal taslak, Makalenin yazımı – İnceleme ve düzenleme.

KAYNAKLAR

[1] Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.
[2] Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4), 537-550.
[3] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226-1238.
[4] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267-288.
[5] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
[6] Díaz-Uriarte, R., and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1), 3.
[7] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). New York: John Wiley & Sons.
[8] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys*, 50(6), 1-45.
[9] Genuer, R., Poggi, J. M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225-2236.
[10] Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1), 3133-3181.
[11] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.
[12] Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4768-4777.
[13] Chicco, D., and Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20(1), 16.
[14] Zhou, H., Wang, X., and Zhu, R. (2022). Feature selection based on mutual information with correlation coefficient. *Applied Intelligence*, 52(5), 5457-5474.
[15] Gomez-Martinez, V., Chushig-Muzo, D., Veierød, M. B., Granja, C., and Soguero-Ruiz, C. (2024). Ensemble feature selection and tabular data augmentation with generative adversarial networks to enhance cutaneous melanoma identification and interpretability. *BioData Mining*, 17(1), 46.
[16] Mohan, S., Thirumalai, C., and Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7, 81542-81554.
[17]Sena, L., & Machado, J. (2024). *Evaluation of fairness in machine learning models using the UCI Adult Dataset*. In *Proceedings of the 39th Brazilian Symposium on Data Bases (SBBD)* (pp. 743–749)
[18] Shah, D., Patel, S., and Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, 1(6), 345.
[19] Cover, T. M. (1999). *Elements of Information Theory*. New York: John Wiley & Sons.
[20] Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 1-22.
[21] Wright, M. N., and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77, 1-17.
[22] Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.
[23] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.
[24] Molnar, C. (2020). *Interpretable Machine Learning*. Lulu Press.
[25] Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 202-207.
[26] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., and Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5), 304-310.

[27] Sen, R., Mandal, A. K., Goswami, S., & Chakraborty, B. (2019). *A comparative study of the stability of filter based feature selection algorithms*. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)* (pp. 1–6). Institute of Electrical and Electronics Engineers.
[28] Nogueira, S., Sechidis, K., and Brown, G. (2018). On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174), 1-54.
[29] Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 6478-6490.
[30] Barbieri, M. C., Grisci, B. I., and Dorn, M. (2024). Analysis and comparison of feature selection methods towards performance and stability. *Expert Systems with Applications*, 249, 123667.
[31] Patel, J., Upadhyay, T., and Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *International Journal of Computer Science and Information Security*, 7(1), 129-137.
[32] Alfadli, K., and Almagrabi, A. (2023). Feature-limited prediction on the UCI heart disease dataset. *Computers, Materials, & Continua*, 74(3), 5871-5883.