



A comparative analysis based on feature selection methods and machine learning algorithms for evaluating heart disease prediction performance

Özlem Akarçay Pervin¹ , Nimet Yapıcı Pehlivan^{2,3*} , Gerhard Wilhelm Weber^{4,5}

¹Department of Computer Programming, Vocational School of Commerce and Industry, KTO Karatay University, Konya, 42050, Türkiye

²Department of Statistics, Science Faculty, Selçuk University, Konya, 42250, Türkiye

³Department of Biostatistics, Graduate School of Health Science, Ankara University, Ankara, 06110, Türkiye

⁴Faculty of Engineering Management, Poznań University of Technology, Poznań, 60-695, Poland

⁵Institute of Applied Mathematics, Middle East Technical University, Ankara, 06800, Türkiye

Highlights:

- Use of machine learning algorithms in the diagnosis of Heart Disease
- The effect of SMOTE resampling method and feature selection methods on classification performance
- Performance comparison of the machine learning algorithms

Keywords:

- Heart disease diagnosis
- SMOTE resampling
- Feature selection
- Machine Learning
- Performance metrics

Article Info:

Research Article

Received: 13.07.2024

Accepted: 11.11.2025

DOI:

10.17341/gazimmfd.1513919

Acknowledgement:

This study is supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) with project number 125F351 and by Selçuk University Scientific Research Projects Coordination Office (BAP) with project number 25401078

Correspondence:

Author: Nimet Yapıcı Pehlivan
e-mail: nimet@selcuk.edu.tr
phone: +90 533 360 0858

Graphical/Tabular Abstract

In this study, performance comparisons were made using Synthetic Minority Oversampling Technique (SMOTE) resampling method for imbalanced data set, three feature selection methods, Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, and Least Angle Regression (LARS), and five machine learning algorithms, K-Nearest Neighbor, Support Vector Machine, Random Forest, Decision Tree, Logistic Regression, for heart disease diagnosis as shown in Figure A.

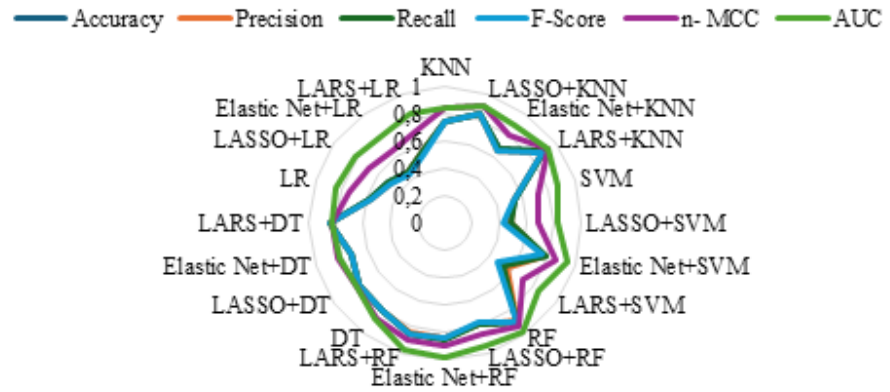


Figure A. Performance metrics for feature selection and machine learning with SMOTE resampling

Purpose: This study aims to diagnose heart disease by considering the SMOTE resampling method, feature selection methods and machine learning algorithms together. A comparative analysis is performed using the UCI Heart Disease Cleveland dataset.

Theory and Methods: Heart disease status is determined by five machine learning algorithms (KNN, SVM, RF, DT, LR) used for classification aim. Performance metrics can be affected by the number of features a dataset contains and whether it is balanced. The SMOTE resampling method that combines existing ones and creates synthetic samples, is applied to solve data imbalance problem. Three feature selection methods (LASSO, Elastic Net, LARS) are used to extract the most relevant features avoiding overfitting and underfitting issues. Performance metrics (accuracy, precision, recall, F-score, n-MCC, and AUC) are calculated by the machine learning algorithms with and without using the SMOTE resampling technique and feature selection methods.

Results: Performance metrics were used to evaluate the performances of all the mentioned algorithms. The RF model with SMOTE and LARS+KNN model with SMOTE achieved the highest accuracy (0.90), precision (0.89), and recall (0.90). The highest F1-score (0.90) is handled by the RF model with SMOTE. On the other hand, the highest n-MCC (0.94) and the highest ROC-AUC (0.98) are obtained by the LARS+KNN model with SMOTE.

Conclusion: In the case of class imbalance problems and irrelevant features, SMOTE resampling and feature selection methods have been increased the classification performances of the machine learning algorithms significantly.



Kalp hastalığı tahmin performansının değerlendirilmesinde özellik seçimi yöntemlerine ve makine öğrenmesi algoritmalarına dayalı karşılaştırmalı bir analiz

Özlem Akarçay Pervin¹ , Nimet Yapıcı Pehlivan^{2,3*} , Gerhard Wilhelm Weber^{4,5}

¹KTO Karatay Üniversitesi, Ticaret ve Sanayi Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, 42050, Konya, Türkiye

²Selçuk Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 42250, Konya, Türkiye

³Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı, 06110, Ankara, Türkiye

⁴Poznań Teknoloji Üniversitesi, Mühendislik Yönetimi Fakültesi, 60-695, Poznań, Polonya

⁵Ortadoğu Teknik Üniversitesi, Uygulamalı Matematik Enstitüsü, 06800, Ankara, Türkiye

Ö N E Ç İ K A N L A R

- Kalp hastalığını teşhis etmede makine öğrenmesi algoritmalarının kullanımı
- Sınıflandırma performansı üzerine SMOTE yeniden örnekleme ve özellik seçimi yöntemlerinin etkisi
- Makine öğrenmesi algoritmalarının performanslarının karşılaştırılması

Makale Bilgileri

Araştırma Makalesi

Geliş: 13.07.2024

Kabul: 11.11.2025

DOI:

10.17341/gazimmfd.1513919

Anahtar Kelimeler:

Kalp hastalığı,
SMOTE yeniden örnekleme,
özellik seçimi,
makine öğrenmesi,
performans metrikleri

ÖZ

Kalp hastalığı, son yıllarda dünya genelinde başlıca ölüm nedenlerinden biri olmuştur. Bu çalışmada, kalp hastalığı durumunu tahmin etmek amacıyla K-en yakın komşuluk (KNN), Destek Vektör Makinesi (SVM), Karar Ağacı (DT), Rastgele Orman (RF) ve Lojistik Regresyon (LR) gibi makine öğrenmesi algoritmaları kullanılmıştır. Makine öğrenmesi (ML) algoritmalarının sınıflandırma performansı, veri kümesinin çok sayıda özellik içermesinden ve sınıf dengesizliğinden etkilenebilmektedir. Bu nedenle, veri kümesini dengelemek amacıyla özellik seçimi yöntemlerinden Sentetik Azınlık Örnekleme Tekniği (SMOTE) uygulanmıştır. İlgili özellikleri belirlemek amacıyla En Küçük Mutlak Büzülme ve Seçim Operatörü (LASSO), ElasticNet ve En Küçük Açılı Regresyon (LARS) kullanılmıştır. Sınıflandırma performansları doğruluk, kesinlik, duyarlılık, F1-skoru, Matthew korelasyon katsayısı (MCC), normalize-MCC (n-MCC) ve ROC-AUC metrikleri ile değerlendirilmiştir. Karşılaştırmalı analizler, gerçek bir veri kümesi üzerinde SMOTE ve özellik seçimi yöntemleri uygulanmadan ve uygulanarak gerçekleştirilmiştir. Sonuçlara göre en yüksek doğruluk (0,90), kesinlik (0,89) ve duyarlılık (0,90) değerleri SMOTE uygulanmış RF ve LARS+KNN'den, en yüksek F1-skoru (0,90) değeri SMOTE uygulanmış RF'den, en yüksek n-MCC (0,94) ve ROC-AUC (0,98) değerleri ise SMOTE uygulanmış LARS+KNN'den elde edilmiştir. Genel olarak, SMOTE ve özellik seçimi yöntemlerinin kullanılması durumunda tüm ML algoritmalarının performansının önemli ölçüde arttığı görülmüştür.

A comparative analysis based on feature selection methods and machine learning algorithms for evaluating heart disease prediction performance

H I G H L I G H T S

- Use of machine learning algorithms in the diagnosis of Heart Disease
- The effect of SMOTE resampling method and feature selection methods on classification performance
- Performance comparison of the machine learning algorithms

Article Info

Research Article

Received: 13.07.2024

Accepted: 11.11.2025

DOI:

10.17341/gazimmfd.1513919

Keywords:

Heart disease,
SMOTE resampling, feature
selection,
Machine learning,
performance metrics

ABSTRACT

Heart disease has become one of the leading causes of death worldwide in recent years. In this study, ML algorithms including KNN, SVM, DT, RF, and LR were employed to predict heart disease status. The classification performance of ML algorithms can be adversely affected by class imbalance and the presence of a large number of features in the dataset. Therefore, the SMOTE was applied to balance the dataset. To identify relevant features, feature selection methods including LASSO, ElasticNet, and LARS were utilized. Classification performance was evaluated using accuracy, precision, recall, F1-score, MCC, n-MCC, and ROC-AUC. Comparative analyses were conducted on a real-world dataset with and without the application of SMOTE and feature selection methods. According to the results, the highest accuracy (0.90), precision (0.89) and recall (0.90) are computed from the RF and LARS+KNN with SMOTE. The highest F1-score (0.90) is handled by the RF model with SMOTE. The highest n-MCC (0.94) and ROC-AUC (0.98) are obtained from the LARS+KNN with SMOTE. It has been observed that the performance of all ML algorithms considered in the study increases significantly when SMOTE is used to address the class imbalance problem and feature selection methods are employed to eliminate irrelevant features.

*Sorumlu Yazar / Yazarlar / *Corresponding Author / Authors: ozlem.akarçay@karatay.edu.tr, *nimet@selcuk.edu.tr, gerhard.weber@put.poznan.pl
/ Tel: +90 533 360 0858

1. Giriş (Introduction)

Günümüzde, eğitim ve gelir düzeyindeki yükselmeye birlikte beslenme alışkanlıklarının değişmesi, genellikle yaşlı nüfusta görülen Bulaşıcı Olmayan Hastalıkların (Non-communicable diseases, NCD) çocuklar ve yetişkinler arasında da yaygınlaşmasına yol açmıştır [1]. Dünya genelinde meydana gelen ölümlerin en önemli nedeni olan bulaşıcı olmayan hastalıklar, küresel olarak halen önemli bir hastalık yükü oluşturmaktadır. Gelişmemiş ve gelişmekte olan ülkelerde ölümlerin %78'nin ve erken ölümlerin %85'nin bulaşıcı olmayan hastalıklardan kaynaklandığı bilinmektedir. NCD, en önemli küresel ölüm nedeni olup özellikle düşük ve orta gelirli ülkeleri etkilemekte ve genellikle önlenabilir risk faktörleri nedeniyle meydana gelmektedir. Kardiyovasküler hastalıklar, kronik solunum sistemi hastalıkları, kanser ve diyabet en önemli NCD arasında yer almaktadır. Bu hastalıklar tütün kullanımı, aşırı alkol kullanımı, fiziksel hareketsizlik ve sağlıksız beslenme gibi davranışsal risk faktörleri ile ilişkilidir. NCD'ye bağlı ölümlerin %44'ü (17,9 milyon) kalp ve damar hastalıklarına, %22'si (9 milyon) kanserlere, %9'u (3,8 milyon) astım ve kronik obstrüktif akciğer hastalıklarını içeren kronik solunum yolu hastalıklarına, %4'ü (1,6 milyon) diyabete bağlıdır. Bu dört ana bulaşıcı olmayan hastalık, NCD'ye bağlı ölümlerin %79'undan sorumludur. Hem gelişmiş hem de gelişmekte olan ülke ekonomilerinde, NCD büyük finansal ve ekonomik risk oluşturmaktadır [2].

Sağlık hizmetlerinde sağlık operasyonları ve ekonomik analizlerin yanı sıra, risk değerlendirmesi, klinik teşhis koyma, karar verme, hastalıkların modellenmesi, tedavilerin tasarlanması gibi klinik uygulamalar sağlık hizmetlerinde oldukça önemli yer tutmaktadır.

Çok fazla özellik içeren karmaşık gerçek dünya problemleri için çeşitli analitik ve matematiksel yöntemler geliştirilmiştir. Bu yöntemlerden biri olan özellik seçimi (Feature selection, FS), bir veri kümesinden ilgisiz ya da gereksiz özellikleri ortadan kaldırarak özelliklere ilişkin bir alt küme oluşturmak amacıyla tüm ilgili özellikleri seçme sürecidir. FS yöntemleri, karma tamsayılı programlama, doğrusal programlama ve karesel programlama gibi optimizasyon yöntemlerine dayalıdır. Özellik alt küme seçimi, özelliklerin tüm olası kombinasyonlarını ele aldığı için bir NP-zor optimizasyon problemidir [3]. FS yöntemleri; filtreleme (filter), sarmalayıcı (wrapper) ve gömülü (embedded) olmak üzere üç ana kategoriye ayrılmaktadır. Relief algoritması ve bilgi kazanımı (Information gain) gibi filtreleme yöntemleri, herhangi bir öğrenme algoritması dahil etmeden özellik seçim sürecini bağımsız olarak gerçekleştirir. Sıralı ileri seçim ve sıralı geri eleme gibi sarmalayıcı yöntemler, özellik alt kümelerini değerlendirmek için önceden belirlenmiş bir öğrenme algoritmasını kullanır. Gömülü yöntemler ise, model oluşturma sırasında özellikleri seçmektedir. Boyut küçültme için kullanılan optimizasyon modellerine dayalı LASSO, Elastic Net ve LARS gibi teknikler ise gömülü yöntemler arasında yer almaktadır [3-5].

Öte yandan, dengesiz veri kümeleri ile yanlış sınıflandırma sorunları ortaya çıkabilmektedir. Bu sorunun üstesinden gelebilmek için sıklıkla veri düzeyinde yaklaşımlar kullanılmaktadır [6]. Bu yaklaşımlardan biri olan SMOTE yeniden örnekleme yöntemi, dengesiz veri kümelerini dengeli hale getirmek için sıklıkla kullanılmaktadır. Makine öğrenmesi (ML) algoritmaları, son yıllarda hastalıkların teşhisinde sıklıkla uygulanmaktadır. Hastalıkların tahmini için sınıflandırmaya yönelik birçok çalışma, makine öğrenmesi teknikleri kullanılarak gerçekleştirilmiştir. Makine öğrenmesi modelleri büyük ölçüde bilgisayar bilimi, istatistik, yöneylem araştırması ve optimizasyon tekniklerine dayanmaktadır [3].

Kalp ve damar hastalıkları (Cardiovascular diseases, CVD), dünya çapında yaygın bir sağlık sorunudur ve Dünya Sağlık Örgütü verilerine göre her yıl tahminen 17,9 milyon cana mal olan en önde gelen ölüm nedeni olarak kabul edilmektedir. CVD, koroner kalp hastalığı, serebrovasküler hastalık, romatizmal kalp hastalığı ve diğer durumları içermektedir. Kalp hastalığı için birincil davranışsal risk faktörleri arasında sağlıksız beslenme, fiziksel aktivite eksikliği, tütün kullanımı ve aşırı alkol tüketimi yer alır. Bunlar bireylerde yüksek tansiyon, kan şekerinin yükselmesi, kan yağlarının yüksek olması, aşırı kilo ve obezite olarak ortaya çıkabilmektedir [1, 2]. Literatürde, kalp hastalığı teşhisine yönelik olarak farklı veri kümeleri üzerinde çeşitli makine öğrenmesi algoritmalarının uygulandığı çok sayıda çalışma mevcuttur. Bu çalışmalarda, genellikle Lojistik Regresyon (LR), Naive Bayes (NB), K-En Yakın Komşu (K-NN), Destek Vektör Makinesi (SVM), Karar Ağaçları (DT), Rastgele Orman (RF), Gradient Boosting (GB) ve Yapay Sinir Ağları (ANN) modelleri kullanılmaktadır.

Literatürde en sık kullanılan veri kümelerinden biri olan UCI-Cleveland Kalp Hastalığı (UCI-HDC) veri kümesi üzerinde, farklı sınıflandırma teknikleri, yeniden örnekleme yöntemleri ve özellik seçimi yöntemleri uygulanmış ve performans karşılaştırması yapılmıştır. Bu çalışmalardan bazıları şunlardır:

Kahramanlı ve Allahverdi [7], kalp hastalığını teşhis etmek için ANN ve bulanık sinir ağı (FNN) içeren hibrit bir sistem geliştirilmiştir. Geliştirilen hibrit sistem, gerçek zamanlı problemlerdeki performans test etmek için Cleveland veri kümesi üzerinde uygulanmış ve %87,4 doğruluk elde edilmiştir. Palaniappan ve Awang [8], çeşitli semptomlara dayalı veri kümelerini kullanarak kalp hastalığını tahmin etmeye yönelik bir hastalık teşhis sistemi olan Akıllı Kalp Hastalığı Tahmin Sistemini (IHDPs) geliştirmiştir. NB, DT ve ANN sınıflandırıcılarını kullanan web tabanlı sistemin uygulanması sonucunda NB için %86,12; DT için %80,4 ve ANN için %85,68 doğruluk değerleri elde edilmiştir. Gudadhe vd. [9], SVM ve ANN kullanarak kalp hastalıklarının sınıflandırılmasına yönelik bir karar destek sistemi oluşturmuşlardır. SVM, kalp hastalığı verilerini %80,41 doğrulukla kalp hastalığının varlığını veya yokluğunu gösteren iki sınıfa ayırmıştır. ANN ise bu verileri %97,5 doğrulukla 5 kalp hastalığı kategorisi içerisine sınıflandırmıştır. İki yöntemin de verileri sınıflandırmada iyi performans gösterdiği, ancak ANN'nin SVM'ye göre daha yüksek doğruluk değeri ile sınıflandırma yaptığı gösterilmiştir.

Panda vd. [10], UCI-HDC veri kümesindeki değişken sayısını en aza indirmek için FS yöntemlerinden LASSO ve Ridge Regresyonu uygulamıştır. Değişken sayısını azaltmanın, sınıflandırma amacı ile kullanılan RF, LR, Extra Trees (ET) ve Gaussian NB (GNB) gibi ML algoritmalarının performansını artırdığı gösterilmiştir. En yüksek doğruluk değeri (%94,92), LASSO ve Ridge regresyonu uygulandıktan sonra GNB algoritmasında elde edilmiştir. Li vd. [11], kalp hastalığını tahmin etmede ilgili özellikleri seçmek için LASSO, Relief, FCMIM ve MRMR dahil olmak üzere FS tekniklerini geliştirmişlerdir. Seçilen özelliklerin çeşitli ML algoritmalarında kullanılması sonucunda, FCMIM ve SVM algoritması ile %92,37 doğruluk elde edildiği gösterilmiştir. Ghosh vd. [12], UCI-HDC veri kümesi üzerinde FS yöntemlerinden Relief ve LASSO uygulanması durumunda ML algoritmalarının daha iyi performans sergilediğini göstermiştir. Relief yöntemi ve RFBM'nin tüm yöntemler içinde %99,05 doğruluk ile en iyi performansa sahip olduğu gösterilmiştir. Waqar vd. [13], UCI-HDC veri kümesinde bağımlı değişkenin kategorik olması ve sınıflar arası dengesizlik durumunu vurgulamıştır. Veri kümesinin SMOTE yöntemi kullanılarak dengelenmesi sonucunda çalışmada ele alınan ML algoritmalarının performansında iyileşme olduğu gözlemlenmiştir. SMOTE tabanlı ANN

algoriması ile en yüksek doğruluk (%96,1), kesinlik (%95,7), duyarlılık (%95,7), F1-skoru (%96) ve AUC (%100) değerlerine ulaşılmıştır. Ullah vd. [14], UCI-HDC veri kümesi üzerinde ilgili özellikleri seçmek için LASSO, ANOVA, MultiSURF, varyans eşiği ve karşılıklı bilgileri uygulamışlardır. Ayrıca, LR, KNN, SGD, RF SVM, NB ve DT gibi çeşitli ML algoritmalarının performansı karşılaştırılmıştır. Özellik seçimi sonucunda modelde yer alan en önemli beş değişkene ML algoritmalarının uygulanması sonucunda, en yüksek doğruluk (%97,5), kesinlik (%97), duyarlılık (%97), F1-skoru (%97) ve MCC (%95) değerlerinin SVM yönteminden elde edildiği gösterilmiştir.

Takci [15], kalp hastalıkları teşhisi için UCI Statlog ve Cleveland veri kümeleri üzerinde KNN algoritmasını uygulamıştır. Statlog veri kümesi için en yüksek doğruluk değerinin komşu sayısı 5, uzaklık yöntemi Euclidian ve ağırlıklandırma Distance alınan KNN algoritması ile %67,90; komşu sayısı 5, uzaklık yöntemi Euclidian ve ağırlıklandırma Distance alınan Genetik Algoritmalar ile %88,88 olarak elde edilmiştir. Cleveland veri kümesi için elde edilen en yüksek doğruluk değeri, optimizasyon öncesi %71,42 iken komşu sayısı 3, uzaklık yöntemi Manhattan ve ağırlıklandırma Uniform alınan KNN ile %90,11'e çıkmıştır. Ghosh ve Islam [16], özellik seçimi ve veri dengeleme yöntemlerinin sınıflandırma algoritmalarının performansı üzerindeki etkisini araştırmıştır. Sınıflar arası dengesizlik durumunun giderilmesi için SMOTE yöntemi, FS yöntemi olarak ise Elastic Net yöntemi kullanılmıştır. Sınıflandırma algoritmalarının performansını değerlendirmek için doğruluk, duyarlılık, seçicilik, AUC, kesinlik, F1-skoru ve MCC gibi çeşitli performans metrikleri kullanılmıştır. RF ve Adaboost algoritmalarına ilişkin doğruluk değerleri, tüm özellikler modelde iken sırasıyla %99 ve %94 iken, Elastic Net uygulandıktan sonra %92 ve %83'e düşmüştür. Ancak, sınıf dengeleme ve FS yöntemlerinin uygulanmasıyla çalışmada ele alınan diğer sınıflandırma algoritmalarının doğruluk değerlerinde artış olmuş ve LR (%82), KNN (%87), SVM (%88), ANN (%88) ve MLP (%86) olarak hesaplanmıştır. Naseer vd. [17], sınıflar arası dengesizlik problemini çözmek için SMOTE yöntemini kullanmışlardır. Ensemble Bagging tekniği (BM) uygulandıktan sonra, ele alınan ML algoritmalarının performansı karşılaştırılmıştır. BM ile birleştirilmiş ML algoritmalarının performanslarının daha yüksek olduğu gösterilmiştir. SVM-BM modeli uygulandığında, doğruluk (%88,52), kesinlik (%91), duyarlılık (%88), F1-skoru (%90) ve ROC-AUC (%89) sonuçlarına ulaşılmıştır. NB-BM modeliyle, doğruluk (%86,88), kesinlik (%88), duyarlılık (%88), F1-skoru (%88) ve ROC-AUC (%87) sonuçları elde edilmiştir. KNN-BM modeli, doğruluk (%86,93), kesinlik (%84), duyarlılık (%94), F1-skoru (%89) ve ROC-AUC (%87) sonuçlarını vermiştir. RF-BM modeli ise, doğruluk (%86,88), kesinlik (%86), duyarlılık (%91), F1-skoru (%89) ve ROC-

AUC (%87) sonuçlarını vermiştir. Son olarak, HLRBM modeliyle, doğruluk (%90,16), kesinlik (%93), duyarlılık (%88), F1-skoru (%90) ve ROC-AUC (%90) sonuçlarına ulaşılmış ve en iyi performans gösteren model olmuştur. Sevlı [18], diyabeti hastalığının teşhisine yönelik olarak Pima Indian Diabetes veri kümesi üzerinde 6 ML algoritması kullanarak sınıflandırma yapmıştır. Sınıflandırıcıların başarısını arttırmak için 14 yeniden örnekleme yöntemi kullanılmıştır. Tüm ML algoritmaları için yeniden örnekleme yapılmadan ve yapılarak sınıflandırma işlemi yapılmış ve 5 farklı performans metriği hesaplanmıştır. En yüksek doğruluk (%96,29), InstanceHardnessThreshold az örnekleme tekniği ile RF modelinden elde edilmiştir. Yeniden örnekleme tekniklerinin genel olarak sınıflandırıcıların başarısını arttırdığı görülmüştür. Jafar vd. [19], GB modelini kullanan HypGB adı verilen otomatik ML sistemini geliştirmişlerdir. Kalp hastalığı olan bireyleri sınıflandırmak için FS yöntemi olarak LASSO kullanılmıştır. Cleveland ve Kaggle veri kümeleri için sırasıyla %97,32 ve %97,72 en yüksek doğruluk değerlerine ulaşılmıştır.

Tablo 1'de verilen Sahid vd. [20], Muntasir Nishat vd. [24], çalışmalarında SMOTE+ML algoritmaları; Oleiwi vd. [23] ve Li vd. [11] çalışmalarında FS+ML algoritmaları ve Ghosh vd. [16] çalışmasında ise SMOTE+FS+ML yöntemleri ele alınmıştır. Fakat, bu çalışmaların hiçbirinde Cleveland kalp hastalığı veri kümesi için veri dengeleme yöntemlerinden SMOTE, özellik seçim yöntemlerinden LASSO, Elastic Net ve LARS'ın makine öğrenmesi algoritmaları ile birlikte ele alındığı ve performans metrikleri üzerinden karşılaştırmalı analiz yapıldığı bir çalışmaya rastlanılmaması, bu çalışmanın özgün değerini göstermektedir.

Bu çalışma, makine öğrenmesi algoritmalarını kullanarak kalp hastalığını teşhis etmeyi amaçlamakta ve üç bölümden oluşmaktadır. İlk olarak, veri kümesindeki dengesizlik sorununu çözmek için SMOTE yeniden örnekleme yöntemi uygulanmış ve veri kümesi dengeli hale getirilmiştir. Daha sonra, aşırı uyum (overfitting) ve yetersiz uyum (underfitting) sorunlarından kaçınmak için veri kümesindeki en ilgili özellikleri belirlemek amacıyla özellik seçimi yöntemlerinden LASSO, Elastic Net ve LARS uygulanmıştır. Son olarak, makine öğrenmesi algoritmalarından KNN, SVM, RF, DT ve LR kullanılarak kalp hastalığı durumunun sınıflandırılmasına yönelik çeşitli performans metrikleri üzerinden karşılaştırmalı bir analiz gerçekleştirilmiştir. Doğruluk, kesinlik, duyarlılık/geri çağırma, F1-skoru, MCC ve ROC-AUC gibi performans metrikleri, veri kümesinin içerdiği özelliklerin sayısından ve dengeli olup olmamasından etkilenmektedir. Sonuçları doğru bir şekilde değerlendirebilmek amacıyla, SMOTE uygulanmadan ve uygulandıktan sonra FS yöntemlerinin kullanıldığı ML algoritmalarına ilişkin performans metrikleri hesaplanmıştır.

Tablo 1. UCI-Cleveland kalp hastalığı veri kümesi üzerinde yapılan bazı güncel çalışmalar
(Some recent works on the UCI-Cleveland heart disease dataset)

Çalışma	Yöntem	Özellik Sayısı	Doğruluk	Kesinlik	Duyarlılık	F1-skoru	MCC	ROC-AUC
Sahid vd. [20]	SMOTETomek+SVM	-	%96	%96	%96	%96	-	%97
Tompra vd. [21]	ANN+SMOTE-ENN	-	%54	%14	%87	%24	-	%69
Oleiwi vd. [22]	Mutual information MI+SVM	11	%96,7	%100	%94,1	%96,9	-	-
Hasanah vd. [23]	RF+SMOTE	-	%90,53	%85,71	%95,35	-	-	%65,10
Li vd. [11]	FCMIM+SVM	10	%92,37	-	%89	-	%90	-
Ghosh vd. [16]	RF+MICE+Elastic Net+SMOTE	7	%92	%92	%94	%93	%84	%98
Muntasir Nishat vd. [24]	SMOTE-ENN+RF	-	%90	%87,8	%97,3	%92,3	-	%91,3
Naseer vd. [17]	SMOTE+LR+BM (HLRBM)	-	%90,16	%93	%88	%90	-	-

2. Materyal ve Metod (Material and Method)

Bu çalışmada, bireylerin kalp hastası olma riskinin yaş, cinsiyet, serum kolesterol düzeyi, açlık kan şekeri düzeyi gibi etkenlere göre değerlendirilmesi amaçlanmıştır. UCI-HDC veri kümesi üzerinde çeşitli makine öğrenmesi algoritmaları kullanılarak hastalığın erken teşhisine yönelik bir çalışma gerçekleştirilmiştir. Bu amaçla, sınıflandırma için ML algoritmalarından KNN, SVM, RF, DT ve LR kullanılmış ve farklı metrikler üzerinden performansları değerlendirilmiştir. Veri kümesinin dengesiz yapıda olması nedeniyle, sınıflandırmanın tahmin başarısını arttırmak için SMOTE yeniden örnekleme yöntemi uygulanmıştır. Sınıflandırma algoritmalarında, modelin tahmin performansını arttırmak, aşırı uyum riskini azaltmak ve ilgisiz değişkenleri veri kümesinden çıkarmak için özellik seçimi yöntemlerinden LASSO, Elastic Net, LARS uygulanmıştır. Çalışmada ele alınan sınıflandırma algoritmalarının performansları, SMOTE ve FS yöntemlerinin uygulandığı ve uygulanmadığı durumlarda doğruluk, kesinlik, duyarlılık, n-MCC ve ROC-AUC metrikleri üzerinden değerlendirilmiştir.

2.1. Veri Yeniden Örnekleme (Data Resampling)

Bir veri kümesi, sınıflandırma kategorilerinin dağılımı eşit olmadığında dengesiz olarak kabul edilir. Bu dengesizlik, “sınıf dengesizliği” sorununa yol açmakta ve gerçek dünya veri kümelerinde sıklıkla karşılaşılmaktadır. Sınıf dengesizliği, bir sınıfın az sayıda gözlemden oluşması (azınlık sınıfı, minority class), diğer sınıfın ise daha fazla sayıda gözlem içermesi (çoğunluk sınıfı, majority class) olarak ifade edilmektedir. Çoğunluk sınıfının sınıflandırma hatası azınlık sınıfınınkinden daha yüksek olduğundan, makine öğrenmesi algoritmaları dengesiz verileri doğru bir şekilde sınıflandırmayabilir. Dengesiz veri kümeleri yanlış sınıflandırma sorunlarına yol açabildiğinden, bu sorunun üstesinden gelebilmek amacıyla çok sayıda yöntem geliştirilmiştir. Bu yöntemler, algoritma düzeyindeki yaklaşımlar, veri düzeyindeki yaklaşımlar, maliyete duyarlı öğrenme yaklaşımları ve topluluk temelli yaklaşımlar olmak üzere dört grupta sınıflandırılmaktadır.

Veri düzeyinde yaklaşımlar, veri ön işleme aşamasında sınıf dengesizliğinin neden olduğu etkilerden kaçınmak için yeniden örnekleme kullanarak sınıf dağılımını dengelemektedir. Yeniden örnekleme prosedürleri; yetersiz örnekleme (undersampling), aşırı örnekleme (oversampling) ve hibrit yöntemler olmak üzere üç grupta ele alınmaktadır. Yetersiz örneklemede Tomek Linkleri ve aşırı örneklemede SMOTE, literatürde en çok bilinen yeniden örnekleme yöntemleridir. Sınıf dengesizliği problemini çözmek için en çok kullanılan aşırı örnekleme tekniklerinden biri olan SMOTE yöntemi, mevcut azınlık örneklerini (minority samples) en yakın azınlık komşularıyla birlikte uygulayarak sentetik azınlık sınıfı gözlemlerini üretmektedir [6, 25].

2.2. Özellik Seçim Yöntemleri (Feature Selection Methods)

Son yıllarda yapay zeka ve makine öğrenmesi alanlarının gelişmesiyle birlikte büyük ilgi gören yüksek boyutlu veriler, genellikle gerçek veriler olup açık erişimli olarak çeşitli veri tabanlarında yer almaktadır. Yüksek boyutlu verilerde, özellikler arasındaki yüksek korelasyon, çok fazla özellik veya gereksiz/ilgisiz özellikler gibi birçok sorunla karşılaşılmaktadır. Veri kümesinin çok sayıda özellik içermesi durumunda, gereksiz ve ilgisiz özellikler ML algoritmalarının gereğinden fazla takılmasına neden olabilmekte ve hata oranını artırabilmektedir. Bu tür sorunları çözmek için, veri ön işlemenin bir parçası olan “Boyut Azaltma (Dimensionality reduction)” teknikleri uygulanmaktadır. Boyut azaltmada en sık kullanılan yaklaşımlar, özellik seçimi (Feature selection, FS) ve özellik çıkarma (Feature Extraction, FE)’dir. Veri ön işleme, ML

algoritmalarının verimli bir şekilde kullanılması için oldukça önemlidir. Veri ön işleme en yaygın ve önemli tekniklerden biri olarak öne çıkan FS yöntemleri, orijinal veri kümesinden ilgili özelliklerin bir alt kümesini elde etmeyi ve böylece ML algoritmalarını hızlandırırken tahmin doğruluğunu da artırmayı amaçlamaktadır. Literatürde, birçok FS yöntemi önerilmiş ve filtreleme (filter), sarmalayıcı (wrapper) ve gömülü (embedded) yöntemler olmak üzere üç gruba ayrılmıştır. Gömülü yöntemler, özellik seçme sürecini ve öğrenme sürecini tek bir optimizasyon problemine birleştirmektedir. En iyi bilinen gömülü yaklaşımlar; LASSO, Elastic Net ve LARS yöntemleridir [26-29].

LASSO yöntemi, Tibshirani [32] tarafından önerilen ve yüksek boyutlu verilerde düzenleme ve özellik seçimi gibi iki ana amaca hizmet eden kullanışlı bir yöntemdir. Gereksiz özellikleri ortadan kaldırarak önemli özellikleri belirleyerek aşırı uyum sorununu azaltan bu yöntem, model parametrelerinin mutlak değerlerinin toplamına bir kısıt ekler. Bu kısıt, toplamın önceden belirlenmiş bir sabit değerden (üst sınır) daha küçük olmasını sağlar ve Eş. 1’de verilen optimizasyon problemi elde edilir.

$$\text{Minimize } L(\beta) = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (1)$$

Eş. 1’de; y_i bağımlı değişken, X_{ij} bağımsız değişken, β_j regresyon katsayısı, β_0 sabit terim olarak tanımlanır ve $\lambda \geq 0$ büzülm miktarını kontrol eden ayar parametresidir. λ ’nın belirlenmesi büyük önem taşımaktadır ve yeterince büyük olduğunda daha fazla katsayı sıfıra zorlanır, böylece boyut azaltma gerçekleştirilir [30-32].

Elastic Net yöntemi, uygulanmasındaki bazı sınırlamalar nedeniyle LASSO’nun bir uzantısı olarak Zou ve Hastie [33] tarafından önerilmiştir. Bu sınırlamalar; i) veri kümesinin gözlem sayısından (n) daha büyük bağımsız değişken sayı (p) içermesi ($p > n$) durumunda, LASSO’nun dişbükey yapısından dolayı maksimum n tane değişken seçme eğiliminde olması, ii) veri kümesi gruplandırılmış değişkenlerden oluştuğunda, yani yüksek korelasyonlu bağımsız değişkenler olduğunda LASSO’nun diğerlerini göz ardı ederek her gruptan bir tane değişken seçme eğiliminde olmasıdır. LASSO’nun bu sınırlamalarını ele alabilmek için LASSO ve Ridge Regression yöntemlerini birleştiren Elastic Net önerilmiş ve Eş. 2’de verilen optimizasyon problemi elde edilmiştir.

$$\text{Minimize } L(\beta) = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p X_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \quad (2)$$

Eş. 2’de; y_i bağımlı değişkeni, X_{ij} bağımsız değişkeni, β_j regresyon katsayısını, β_0 sabit terimi; $\lambda_1 \geq 0$ ve $\lambda_2 \geq 0$ ayarlama parametrelerini göstermektedir [30, 33].

LARS yöntemi, LASSO parametrelerini karesel programlamadan daha hızlı tahmin etmek için etkili bir algoritma üreten, En küçük açı regresyonu (Least angle regression, LAR)’nın LASSO’ya dönüştürülmüş halidir ve ileri adımli regresyona benzerdir. Bununla birlikte, ileri adımli regresyonda olduğu gibi her adımda bağımsız değişkenler gibi özellikleri dahil etmek yerine, tahmin edilen parametreleri, her birinin artık ile korelasyonuna eşit bir açıyla artırır [34].

2.3. Makine Öğrenmesi Algoritmaları (Machine Learning Algorithms)

Makine öğrenmesi, büyük veri çağında hesaplamalı algoritmaların sürekli gelişen bir alanıdır. Bu algoritmalar çevreden öğrenerek insan zekasını taklit etmek üzere tasarlanmıştır. ML algoritmaları; örüntü tanıma, bilgisayarla görme, mühendislik, finans ve tıbbi çalışmalar gibi farklı alanlarda başarıyla uygulanmaktadır ve genel olarak

denetimli (supervised) ve denetimsiz (unsupervised) olmak üzere iki ana kategoriye ayrılmaktadır [35-37]. Denetimli öğrenme, regresyon ve sınıflandırma problemleri olarak gruplandırılabilir ve çıktıların sürekli olması durumunda regresyon, kategorik olması durumunda ise sınıflandırma problemi olarak değerlendirilmektedir. Sınıflandırmaya ilişkin denetimli ML algoritmaları arasında Lojistik Regresyon, Destek Vektör Makinesi, Karar Ağacı, Rastgele Orman ve K-En Yakın Komşu yer almaktadır.

K-En Yakın Komşu (KNN) algoritması, bir gözlemin diğer gözlemlerle benzerliğine dayalı bir sınıflandırma yöntemidir. Yeni bir gözlem verildiğinde veri kümesindeki her bir gözleme olan uzaklığı Öklid uzaklığı kullanılarak hesaplanır ve o gözlem “K” en yakın komşusunun ait olduğu sınıfa atanır. KNN algoritması, basit olması ve herhangi bir model yapısı ya da ek varsayım gerektirmemesi nedeniyle tercih edilmektedir. KNN algoritmasının önemli bir dezavantajı, her yeni gözlem için K tane komşuya olan mesafenin birçok kez yeniden hesaplanmasını gerektirmesinden dolayı hesaplama süresini artırmasıdır [38, 39].

Destek Vektör Makinesi (SVM), yeni gözlemleri sınıflandırmak için iki destek vektörünün oluşturduğu düzlemin simetri çizgisi boyunca uzanan ideal bir hiperdüzlemi tanımlamayı amaçlamaktadır. İki gözlem arasındaki benzerliği ölçen çekirdek (kernel) fonksiyonları, SVM’de daha iyi sınıflandırma elde etmek amacıyla gözlemleri yüksek boyutlu bir uzaya haritalamak amacıyla kullanılmaktadır. Polinom, doğrusal, sigmoid ve radyal tabanlı fonksiyon (RBF) gibi çeşitli çekirdek fonksiyonları mevcuttur. Doğrusal çekirdek fonksiyonu, SVM’nin en basit ve en yaygın kullanılan çekirdek fonksiyonlarından biridir. SVM, yüksek boyutlu verilerde ve veri kümesinin gözlem sayısından daha büyük sayıda değişken içerdiği durumlarda bile etkili bir algoritmadır [40, 41].

Karar Ağacı (DT), kök düğüm, iç düğümler (dallar) ve uç düğümler (yapraklar)’den oluşan ve yaygın olarak kullanılan bir sınıflandırıcıdır. Bir karar ağacında kök düğüm, diğer tüm düğümlerin ebeveyni olan en üst düğümdür. Kök düğümün, dallar olarak bilinen bir alt kümesi vardır. Her bir dalda bir değişken test edilir ve sonuçlar (yani kararlar) yaprak düğümde temsil edilir. Yaprak düğümlerde gösterilen sonuçlar, kategorik veya sürekli değerler olabilir [42-44]. Rastgele Orman (RF), hem regresyon hem de sınıflandırma amacıyla kullanılabilen, güçlü ve çok yönlü bir denetimli makine öğrenmesi algoritmasıdır. Yüksek boyutlu veriler için tek bir karar ağacından çok daha etkin olan RF algoritması, bir orman oluşturmak için birden fazla karar ağacını birleştirir. Orijinal veri kümesinden birden fazla veri alt kümesi rastgele oluşturulur ve her karar ağacı farklı veri alt kümeleri üzerinde eğitilir. Özellikler, ayrıca ağaç inşası sırasında rastgele seçilir. Birden fazla ağaç tarafından oluşturulan tahmin, çoğunluk oyu kullanılarak birleştirilir. Ormandaki her ağacın verilere farklı bir bakış açısı sağlaması nedeniyle RF sınıflandırıcısının performansı ormandaki ağaç sayısı arttıkça artar. Ancak ormana çok fazla ağaç eklenmesi performansın düşmesine neden olabilmektedir [45-47].

Lojistik regresyon (LR), kategorik bir bağımlı değişken ile kategorik veya sürekli bağımsız değişkenler arasındaki model ilişkisini belirlemeyi amaçlayan denetimli sınıflandırma yöntemidir. Lojistik regresyon, kategorik verilerin bir lojistik fonksiyon aracılığıyla sınıflandırılmasını sağlar. İkili LR, çoklu LR ve sıralı LR olmak üzere üç grupta incelenen lojistik regresyon, kategorik bağımlı değişkenlerin tahmin edilmesinde önemli bir rol oynamaktadır [11, 48, 49].

Makine öğrenmesi algoritmaları, kalp hastalığı gibi klinik veriler üzerinde sınıflandırma amacıyla yaygın olarak kullanılmaktadır. ML algoritmalarının performansı; veri kümesinin boyutu, sınıf dengesizliği ve özellik sayısı gibi çeşitli faktörlerden etkilenebilmektedir. Bu

nedenle, SMOTE yeniden örnekleme ve özellik seçimi yöntemlerinin uygulanması, sınıflandırma başarısını artırmak ve klinik açıdan daha anlamlı sonuçlar elde etmek açısından kritik öneme sahiptir.

2.4. Performans Metrikleri (Performance metrics)

Sınıflandırma amacıyla kullanılan ML algoritmalarının tahmin performansının değerlendirilmesinde, karmaşıklık matrisinden yararlanarak hesaplanan çeşitli performans metrikleri kullanılmaktadır. Performans metrikleri, veri kümesinin dengeli ya da dengesiz olması durumundan etkilenmektedir. İkili sınıflandırma problemleri için, gerçek pozitifler (TP) doğru olarak etiketlenen pozitif örnekler, gerçek negatifler (TN) ise doğru olarak etiketlenen negatif örnekler göstermektedir. Yanlış pozitifler (FP), yanlış olarak etiketlenen negatif örnekler ve yanlış negatifler (FN) ise yanlış olarak etiketlenen pozitif örnekler ifade etmektedir. İkili sınıflandırma problemleri için kullanılan birçok performans metriği, çoklu sınıflandırma problemlerine başarılı bir şekilde uyarlanmıştır. Tablo 2’de, çoklu sınıflandırma için bir karmaşıklık matrisi gösterilmiş olup $C_i, i = 1, 2, \dots, K$ sınıfları ifade etmektedir.

Tablo 2. Çoklu sınıflandırma için karmaşıklık matrisi
(Confusion matrix for multiple classification)

		Tahmin Sonucu			
		C ₁	C ₂	...	C _K
Gerçek Durum	C ₁	C ₁₁	C ₁₂	...	C _{1K}
	C ₂	C ₂₁	C ₂₂	...	C _{2K}
	...	...	...	...	...
	C _K	C _{K1}	C _{K2}	...	C _{KK}

Çoklu sınıflandırma için performans metrikleri, Tablo 2’de yer alan karmaşıklık matrisinden yararlanarak aşağıdaki gibi tanımlanmıştır [55-57].

Doğruluk (accuracy), bir sınıflandırma modelinin genel performansını ölçmek için yaygın olarak kullanılır ve doğru tahminlerin toplam örneklerin sayısına oranıdır.

$$\text{Doğruluk} = \frac{\sum_{i=1}^K C_{ii}}{\sum_{i=1}^K \sum_{j=1}^K C_{ij}} \quad i, j = 1, \dots, K \quad (3)$$

Ancak, doğruluk özellikle dengesiz veri kümelerinde her zaman en iyi ölçü olmayabilir. Bu gibi durumlarda, model çoğunluk sınıfına daha fazla odaklanma eğiliminde olup azınlık sınıfında iyi performans gösteremeyebilmektedir. Bu nedenle, veri kümesindeki dengesizliği hesaba katan kesinlik, duyarlılık/geri çağırma, F1-skoru, MCC ve ROC-AUC gibi performans metriklerinin kullanılması önerilmektedir. Çoklu sınıflandırmada; kesinlik, geri çağırma/duyarlılık ve F1-skoru için mikro-ortalama ve makro-ortalama değerleri hesaplanmaktadır. Mikro-ortalama, sınıf büyüklüklerini dikkate alarak örnek bazlı ortalama alırken; makro-ortalama, sınıf büyüklüklerinden bağımsız olarak her sınıfa eşit ağırlık vererek sınıf bazlı bir ortalama alır. Ağırlıklı ortalama ise, her sınıfın metrikleri sınıf büyüklüğüne göre ağırlıklandırılır ve ortalaması alınır, böylece dengesiz veri kümelerinde daha gerçekçi bir performans değerlendirmesi yapılmasını sağlar.

Kesinlik (precision), sınıflandırıcının pozitif olarak tahmin ettiği örneklerin kaç tanesinin gerçekte pozitif olduğunu ölçmektedir.

$$\text{Kesinlik}_{\text{Makro-ort}} = \frac{1}{K} \sum_{i=1}^K \frac{C_{ii}}{C_{i.}}, C_{i.} = \sum_{j=1}^K C_{ji} \quad (4)$$

$$\text{Kesinlik}_{\text{Mikro-ort}} = \frac{\sum_{i=1}^K C_{ii}}{\sum_{i=1}^K \sum_{j=1}^K C_{ij}} = \text{Doğruluk} \quad (5)$$

$$Kesinlik_{AğırlıklıMakro-ort} = \frac{1}{\sum_{i=1}^K w_i} \left(\sum_{i=1}^K w_i \frac{C_{ii}}{C_{i.}} \right) \quad (6)$$

Duyarlılık (sensitivity)/geri çağırma (recall), gerçekte pozitif sınıfa ait örneklerden kaç tanesinin doğru tahmin edildiğini ölçmektedir.

$$Geriçağırma_{Makro-ort} = \frac{1}{K} \sum_{i=1}^K \frac{C_{ii}}{C_{i.}}, C_{i.} = \sum_{j=1}^K C_{ij} \quad (7)$$

$$Geriçağırma_{Mikro-ort} = \frac{\sum_{i=1}^K C_{ii}}{\sum_{i=1}^K C_{i.}} = Doğruluk \quad (8)$$

$$Geriçağırma_{AğırlıklıMakro-ort} = \frac{1}{\sum_{i=1}^K w_i} \left(\sum_{i=1}^K w_i \frac{C_{ii}}{C_{i.}} \right) \quad (9)$$

F1-skoru (F1-score), kesinlik ve duyarlılığın harmonik ortalaması olup özellikle veri kümesi dengesiz olduğunda ve modelin aynı anda yüksek kesinlik ve duyarlılık elde etmesi gerektiğinde kullanışlıdır [50-53].

$$F1 - skoru_{Makro-ort} = 2 \frac{Kesinlik_{Makro-ort} \times Geriçağırma_{Makro-ort}}{Kesinlik_{Makro-ort} + Geriçağırma_{Makro-ort}} \quad (10)$$

$$F1 - skoru_{Mikro-ort} = Kesinlik_{Mikro-ort} = Geriçağırma_{Mikro-ort} = Doğruluk \quad (11)$$

Matthew Korelasyon Katsayısı (MCC), gerçek ve tahmin edilen örnekler arasındaki korelasyonu hesaplamak için karmaşıklık matrisinin tüm unsurlarını dikkate alan bir performans metriğidir. MCC, -1 ile 1 arasında değer alır ve 1 mükemmel sınıflandırmayı, 0 rastgele sınıflandırmayı, -1 tamamen yanlış sınıflandırmayı ifade etmektedir. Normalize edilmiş MCC (n-MCC) değeri, $n - MCC = \frac{MCC+1}{2}$ eşitliğinden hesaplanır [50].

$$MCC = \frac{(\sum_{i=1}^K C_{ii})(\sum_{i=1}^K \sum_{j=1}^K C_{ij}) - \sum_{i=1}^K (C_{i.} C_{.i})}{\sqrt{[(\sum_{i=1}^K \sum_{j=1}^K C_{ij})^2 - \sum_{i=1}^K (C_{i.} C_{.i})^2]}} \quad (12)$$

ROC eğrisi altında kalan alan (ROC-AUC), dengesiz veri durumunda azınlık sınıftaki hatalar daha belirgin olduğundan performans metriği olarak kullanılmaktadır. ROC-AUC, farklı kesme noktalarında TPR (Gerçek Pozitif Oran) ve FPR (Yanlış Pozitif Oran) arasındaki ilişkiyi gösterir. TPR, gerçek pozitiflerin tüm pozitif örnekler içindeki oranını, FPR ise yanlış pozitiflerin tüm negatif örnekler içindeki oranını ifade etmektedir. ROC-AUC değeri, 0 ile 1 arasında değişir ve daha yüksek bir ROC-AUC genellikle daha iyi bir

sınıflandırıcıyı göstermektedir. Çoklu sınıflandırmada, her sınıf için ROC-AUC hesaplanıp ardından makro-ortalama veya ağırlıklı ortalama alındığında OvR (ROC-AUC) hesaplanmış olur. [53, 54, 56].

3. Sonuçlar ve Tartışmalar (Results And Discussion)

Bu çalışmada, sınıf dengesizliği, özellik seçimi ve sınıflandırma modellerinin birlikte kullanıldığı hibrit bir yaklaşım önerilerek sınıflandırma performansına olası etkileri incelenmiştir. Çalışmada ele alınan UC Irvine Machine Learning Repository Cleveland Kalp Hastalığı (UCI-HDC) veri kümesi, 1 bağımlı ve 13 bağımsız olmak üzere toplam 14 değişken içermektedir (<https://archive.ics.uci.edu/dataset/45/heart+disease>). Bu veri kümesi, birçok çalışmada çeşitli yöntemler kullanılarak ele alınmıştır [11, 16, 17; 20-24].

İlk olarak, veri kümesindeki eksik değerler çıkartılmış ve kategorik değişkenlere kukla değişkenler atanarak veri ön işleme tamamlanmıştır. İkinci olarak, Python programında yer alan Imblearn paketi (Sürüm 3.9.7) kullanılarak sınıf dengesizliği sorununu çözmek amacıyla SMOTE yeniden örnekleme yöntemi kullanılmıştır. Üçüncü olarak, veri kümesinin boyutunu azaltmak için özellik seçim yöntemlerinden LASSO, Elastic Net ve LARS uygulanmıştır. Veri kümesindeki verilerin %70'i eğitim verisi, geri kalan %30'u ise test verisi olarak rastgele ayrılmıştır. Son olarak, makine öğrenme algoritmalarından KNN, DT, RF, SVM ve LR uygulanmıştır.

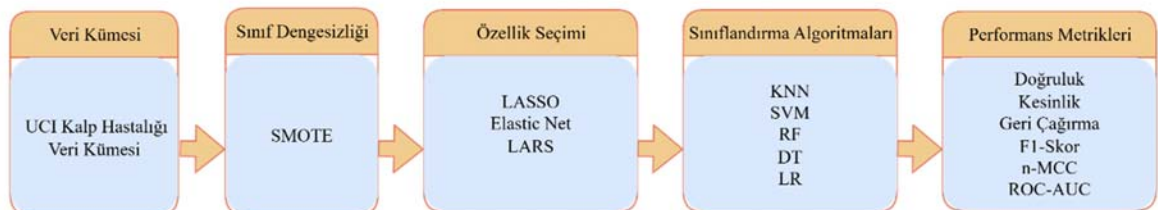
Bu çalışmada hiperparametre optimizasyonu gerçekleştirilmemiş olup, ele alınan makine öğrenmesi algoritmalarına ilişkin parametreler, UCI-HDC veri kümesi üzerinde yapılmış önceki araştırmalar referans alınarak belirlenmiş ve Tablo 3'te ayrıntılı olarak sunulmuştur.

Ele alınan her bir durum için sınıflandırma performansını değerlendirmek amacıyla; doğruluk, kesinlik, duyarlılık/geri çağırma, F1-skoru, n-MCC ve ROC-AUC performans metrikleri hesaplanmıştır. MCC dışındaki performans metrikleri [0,1] aralığında olduğundan, metriklerin karşılaştırılabilirliğini sağlamak amacıyla normalize edilmiş MCC (n-MCC) değerleri dikkate alınmıştır. Önerilen modelin yapısı Şekil 1'de gösterilmektedir.

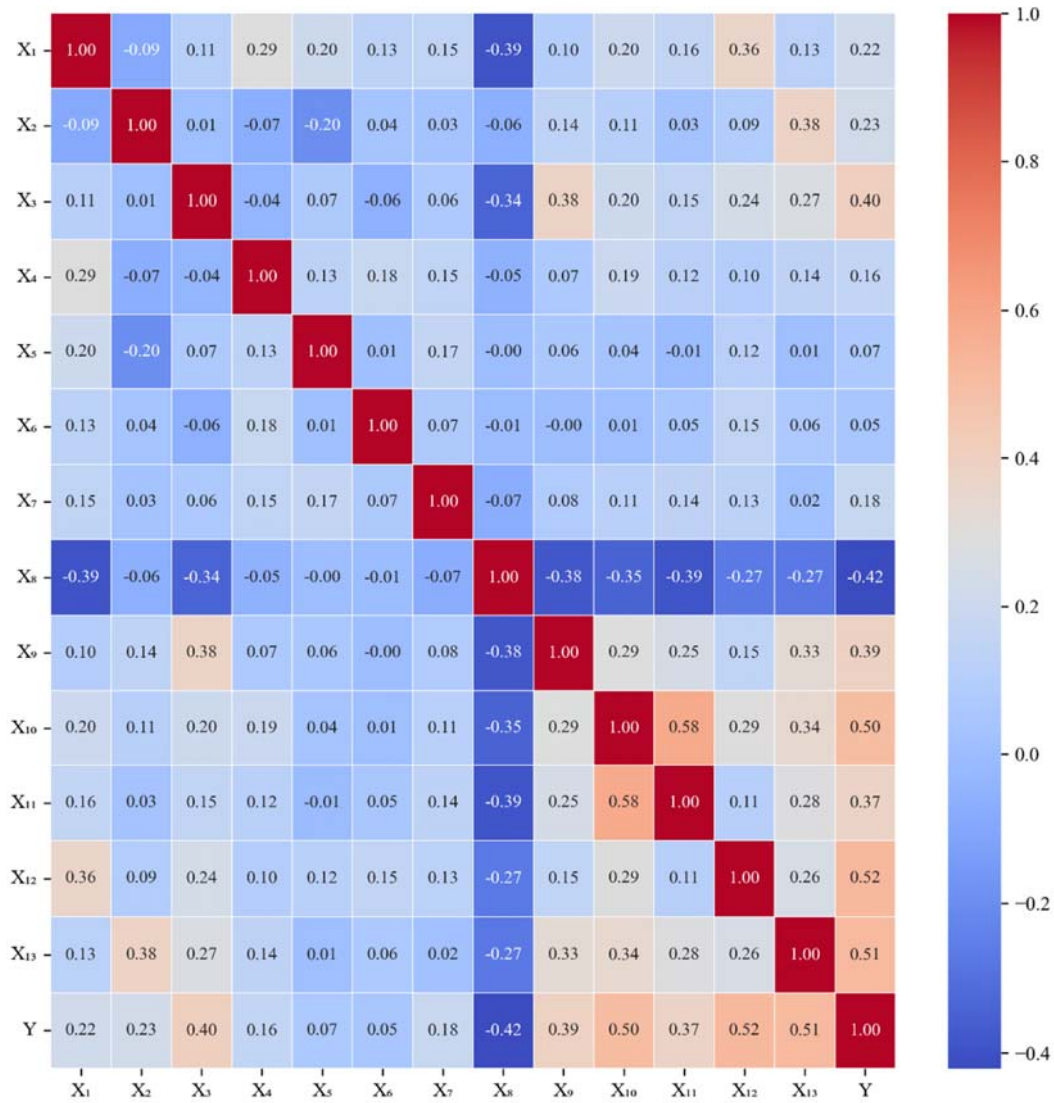
Veri kümelerindeki farklı özellikler arasındaki korelasyon katsayılarını hesaplayan ve bunları görsel olarak temsil eden "ısı haritası (heatmap)", çeşitli özellikler arasındaki ilişkileri görselleştirmektedir.

Tablo 3. Makine öğrenmesi algoritmaları için ele alınan parametreler (Parameters considered for machine learning algorithms)

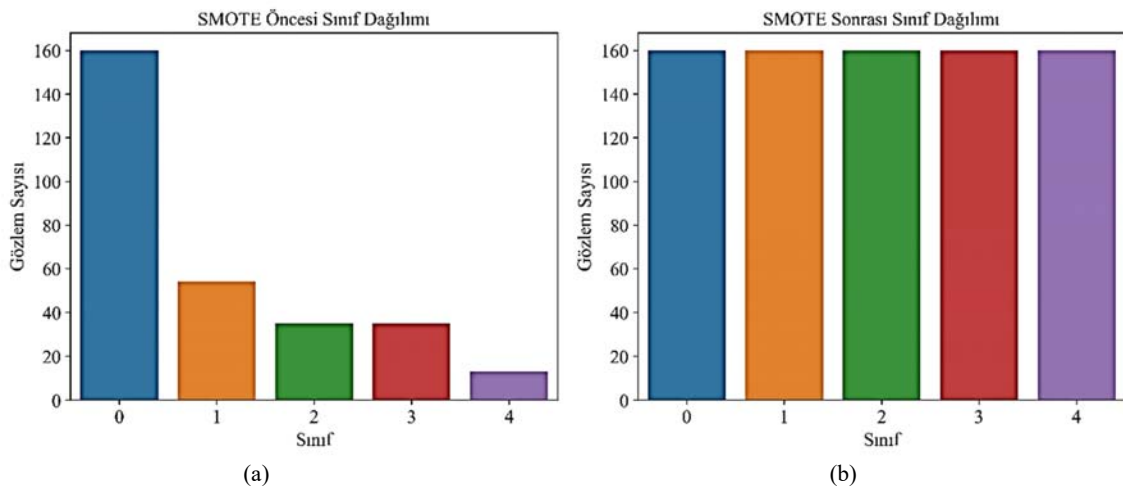
ML Algoritması	Parametre	Kaynak
KNN	k=1	Almustafa (2020) [58]
SVM	kernel='rbf', C=10, gamma=0.001	Li vd. (2020) [11]
RF	gini kriteri, estimators=100 ve max depth =20	Kadhim ve Radhi (2023) [59]
DT	kriter (Criterion) =gini, dallanma (splitter) = best	Ayon vd. (2022) [39]
LR	C=1.0	Lupague vd. (2023) [60]



Şekil 1. Önerilen modelin yapısı (Structure of the proposed model)



Şekil 2. Bağımsız değişkenler (özellikler) arasındaki korelasyon katsayılarına ilişkin ısı haritası
(Heatmap of correlation coefficients between independent variables (features))



Şekil 3. UCI Cleveland kalp hastalığı veri kümesi için (a) Gerçek veri noktaları (b) SMOTE uygulanmış veri noktaları
(UCI Cleveland heart disease data set (a) Real data points (b) Data points after SMOTE)

Tablo 4. SMOTE uygulanmadan / uygulanarak özellik seçim yöntemlerinden hesaplanan R^2 , MSE değerleri ve seçilen özellikler
(Without and with SMOTE, calculated R^2 and MSE values by feature selection methods and selected features)

SMOTE uygulanmadan										SMOTE uygulanarak					
Özellik seçimi yöntemleri										Özellik seçimi yöntemleri					
			LASSO		Elastic Net		LARS		LASSO		Elastic Net		LARS		
			Eğitim	Test	Eğitim	Test	Eğitim	Test	Eğitim	Test	Eğitim	Test	Eğitim	Test	
			R ²	0.54	0.54	0.23	0.28	0.56	0.54	0.53	0.54	0.55	0.55	0.57	0.58
			MSE	0.70	0.65	1.18	1.02	0.67	0.64	0.90	0.92	0.89	0.93	0.84	0.86
Bağımlı değişken															
Y			0=sağlıklı; 1=düşük; 2=orta; 3=yüksek; 4=ciddi												
(Status)															
Bağımsız değişkenler															
X ₁	Hasta yaşı	Tamsayı	-			-0.0073 *		-		-		-		0.0211 *	
(Age)															
X ₂	Hasta cinsiyeti	0=erkek; 1=kadın	1.77e-17 *			0.0112 *		-0.0810 *		1.29e-16 *		-		0.0868 *	
(Gender)															
X ₃	Göğüs ağrısı	0=tipik anjin 1=tipik olmayan anjin 2=anjin olmayan ağrı 3=semptomsuz	1.79e-01 *			0.2553 *		0.1150 *		1.75e-01 *		-0,0638 *		0.1544 *	
(Cp)															
X ₄	Dinlenme kan basıncı	Tamsayı	5.88e-03 *			0.0046 *		0.0389 *		5.27e-02 *		0.0069 *		0.1551 *	
(Trestbps)															
X ₅	Serum kolesterol	Tamsayı	-			0.0003 *		-		-		-0.0015 *		-0.1002 *	
(Chol)															
X ₆	Açlık kan şekeri	0=Normal (<120 mg/dl) 1=Yüksek (> 120 mg/dl)	-2.08e-01 *			-0.2160 *		-0.1969 *		-1.00e-01 *		-0.5137 *		-0.0117 *	
(Fbs)															
X ₇	Dinlenme EKG	0=normal, 1=st-t dalgası anormalliği 2=sol ventrikül hipertrofisi	-			-0.0248 *		-		-		-0.04506 *		0.0606 *	
(Restecg)															
X ₈	Maks. kalp atış hızı	Tamsayı	-1.14e-01 *			-0.0103 *		-0.1304 *		-		-0.0069 *		-	
(Thalach)															
X ₉	Egzersize bağlı anjina	0=Hayır 1=Evet	-4.59e-02 *			0.0085 *		-0.0668 *		-1.52e-01 *		-0.1558 *		-0.1465 *	
(Exang)															
X ₁₀	Dinlenmeye göre egzersizin neden olduğu ST çökmesi	Tamsayı	1.86e-01 *			0.2223 *		0.1984 *		2.48e-01 *		0.3069 *		0.2547 *	
(Oldpeak)															
X ₁₁	Egzersiz sırasındaki eğim	0=yükselen 1=sabit 2=düşen	-			-		-		4.37e-03 *		-		0.0867 *	
(Slope)															
X ₁₂	Büyük damar sayısı	Tamsayı	3.52e-01 *			0.3717 *		0.3662 *		3.46e-01 *		0.4247 *		0.3788 *	
(Ca)															
X ₁₃	Talasemi kusurları	0=normal 1=sabit kusur 2=geri döndürülebilir kusur	1.58e-02 *			0.1929 *		-0.1019 *		1.29e-01 *		-		0.2118 *	
(Thal)															

Pozitif korelasyonlar açık tonlarla gösterilirken, negatif korelasyonlar koyu tonlarla gösterilmektedir. Isı haritasından elde edilen bilgiler sayesinde yüksek ilişkili bağımsız değişkenlerin (özellikler) kaldırılmasıyla, aşırı uyum ve çoklu bağlantı riski azaltılarak model performansı artırılır. UCI-HDC veri kümesindeki özellikler arasındaki korelasyon katsayıları ısı haritasında çizilmiş ve Şekil 2’de gösterilmiştir. Veri kümesindeki çıktı (bağımlı değişken, Y); 0,1,2,3,4 olmak üzere beş kategoriye ayrılmıştır. Şekil 3 (a)’da, her sınıftaki gözlem sayısının oldukça farklı olduğu ve veri kümesinin dengesiz yapıda olduğu görülmektedir. Şekil 3(b)’de, SMOTE yeniden örnekleme yönteminin uygulanmasıyla sınıf dengesizliği sorununun giderildiği görülmektedir.

Kalp hastalığı durumunu tahmin etmede en önemli özellikleri belirlemek amacıyla LASSO, Elastic Net ve LARS özellik seçim yöntemlerini uygulanmıştır. Tablo 4’te, SMOTE uygulanmadan ve uygulanarak özellik seçimi yöntemlerinden hesaplanan R^2 ve MSE değerleri verilmiştir.

Tablo 4’ten, SMOTE uygulanmış veri kümesi üzerinde uygulanan LASSO, Elastic Net ve LARS yöntemlerine ilişkin MSE ve R^2 değerlerinin SMOTE uygulanmamış olanlara göre daha iyi olduğu görülmektedir. SMOTE uygulanmadan (uygulanarak) LASSO yönteminin 9 (9), Elastic Net yönteminin 12 (9) ve LARS yönteminin ise 9 (12) özellik seçtiği görülmektedir. Özellik seçimi yöntemleri ile

belirlenen X_2 (Gender), X_3 (Cp), X_4 (Trestbps), X_6 (Fbs), X_8 (Thalach), X_9 (Exang), X_{10} (Oldpeak), X_{12} (Ca) ve X_{13} (Thal) gibi özelliklerin hem SMOTE uygulanmamış hem SMOTE uygulanmış durumda literatürde bilinen kalp hastalığı risk faktörleriyle uyumlu olduğu görülmektedir.

Özellik seçimi yöntemlerine dayalı ML algoritmaları için SMOTE uygulanmadan ve uygulanarak hesaplanan performans metrikleri sırasıyla Tablo 5 ve Tablo 6'da verilmiştir. Performans metrikleri olarak, dengesiz veri kümeleri için sıklıkla kullanılan doğruluk, kesinlik, duyarlılık, F1-Skoru, n-MCC ve ROC-AUC seçilmiştir. Çalışmada; sınıf sıklığından bağımsız olarak tüm sınıfları eşit öneme sahip olarak alan makro-ortalama değerleri ele alınmış olup, her sınıfın nihai ortalama katkısinin kullanıcı tanımlı bir ağırlık

tarafından kontrol edildiği ağırlıklı makro-ortalama kesinlik, ağırlıklı makro-ortalama duyarlılık/geri çağırma ve ağırlıklı makro-ortalama F1-skoru değerleri hesaplanmıştır.

Tablo 5'te, özellik seçimi yöntemleri uygulanmadan ve uygulanarak ML algoritmalarının (ML/FS+ML) performanslarında meydana gelen değişimler incelenmiştir. Tablo 6'da ise, hem SMOTE yeniden örnekleme hem FS yöntemleri uygulanmamış ve uygulanmış ML algoritmalarının (SMOTE_ML/SMOTE_FS+ML) performanslarında meydana gelen değişimler verilmiştir. Her bir algoritmadan elde edilen en iyi performans metriklerine ilişkin değerler Şekil 4 - Şekil 8'de gösterilmiştir.

KNN algoritmasının performansının, SMOTE ve LARS yöntemlerinin uygulanması sonucunda büyük ölçüde arttığı

Tablo 5. SMOTE uygulanmadan, makine öğrenmesi algoritmalarından hesaplanan performans metrikleri
(Without SMOTE, performance metrics calculated from machine learning algorithms)

ML algoritması	Doğruluk	Kesinlik	Duyarlılık	F1-skoru	n-MCC	ROC-AUC
KNN	0,57	0,32	0,31	0,31	0,65	0,60
LASSO+KNN	0,53	0,30	0,32	0,30	0,64	0,60
Elastic Net+KNN	0,50	0,12	0,18	0,14	0,51	0,52
LARS+KNN	0,56	0,30	0,33	0,31	0,65	0,60
SVM	0,60	0,29	0,29	0,27	0,63	0,81
LASSO+SVM	0,59	0,28	0,29	0,24	0,64	0,83
Elastic Net+SVM	0,44	0,16	0,19	0,17	0,52	0,52
LARS+SVM	0,60	0,19	0,30	0,23	0,65	0,83
RF	0,61	0,27	0,31	0,28	0,67	0,75
LASSO+RF	0,61	0,34	0,35	0,33	0,67	0,75
Elastic Net+RF	0,60	0,25	0,30	0,27	0,66	0,74
LARS+RF	0,63	0,37	0,35	0,34	0,69	0,76
DT	0,52	0,33	0,35	0,33	0,64	0,61
LASSO+DT	0,61	0,34	0,35	0,33	0,57	0,55
Elastic Net+DT	0,51	0,29	0,28	0,28	0,62	0,57
LARS+DT	0,63	0,37	0,35	0,34	0,63	0,54
LR	0,53	0,18	0,21	0,19	0,60	0,74
LASSO+LR	0,57	0,26	0,26	0,25	0,64	0,79
Elastic Net+LR	0,57	0,23	0,27	0,24	0,63	0,80
LARS+LR	0,57	0,26	0,26	0,25	0,57	0,79

ML: Makine öğrenmesi algoritması; LASSO+ML: LASSO uygulanmış ML; Elastic Net+ML: ElasticNet uygulanmış ML; LARS+ML: LARS uygulanmış ML

Tablo 6. SMOTE uygulanarak, makine öğrenmesi algoritmalarından hesaplanan performans metrikleri
(With SMOTE, performance metrics calculated from machine learning algorithms)

SMOTE uygulanmış ML algoritması	Doğruluk	Kesinlik	Duyarlılık	F1-skoru	n-MCC	ROC-AUC
KNN	0,75	0,75	0,75	0,74	0,84	0,84
LASSO+KNN	0,85	0,84	0,85	0,84	0,90	0,90
Elastic Net+KNN	0,68	0,66	0,68	0,66	0,80	0,88
LARS+KNN	0,90	0,89	0,90	0,89	0,94	0,93
SVM	0,54	0,53	0,54	0,52	0,72	0,87
LASSO+SVM	0,47	0,48	0,48	0,43	0,68	0,83
Elastic Net+SVM	0,77	0,76	0,78	0,75	0,86	0,94
LARS+SVM	0,52	0,56	0,53	0,48	0,71	0,85
RF	0,90	0,89	0,90	0,90	0,93	0,98
LASSO+RF	0,78	0,78	0,78	0,77	0,86	0,95
Elastic Net+RF	0,84	0,84	0,85	0,84	0,90	0,98
LARS+RF	0,85	0,84	0,85	0,85	0,90	0,97
DT	0,79	0,79	0,79	0,79	0,87	0,87
LASSO+DT	0,78	0,78	0,78	0,77	0,81	0,80
Elastic Net+DT	0,72	0,72	0,73	0,72	0,83	0,82
LARS+DT	0,85	0,84	0,85	0,85	0,84	0,83
LR	0,58	0,57	0,58	0,57	0,74	0,85
LASSO+LR	0,51	0,49	0,51	0,49	0,70	0,82
Elastic Net+LR	0,46	0,46	0,46	0,45	0,66	0,79
LARS+LR	0,55	0,55	0,54	0,53	0,72	0,84

görülmüştür. En yüksek doğruluk (0,90), kesinlik (0,89), duyarlılık (0,90), F1-Skoru (0,89), n-MCC (0,93) ve ROC-AUC (0,93) değerleri SMOTE uygulanmış LARS+KNN yönteminden hesaplanmıştır. KNN yönteminin performansının SMOTE ve FS yöntemlerinden nasıl etkilendiği, Şekil 4'te verilen performans metrikleri ile gösterilmektedir.

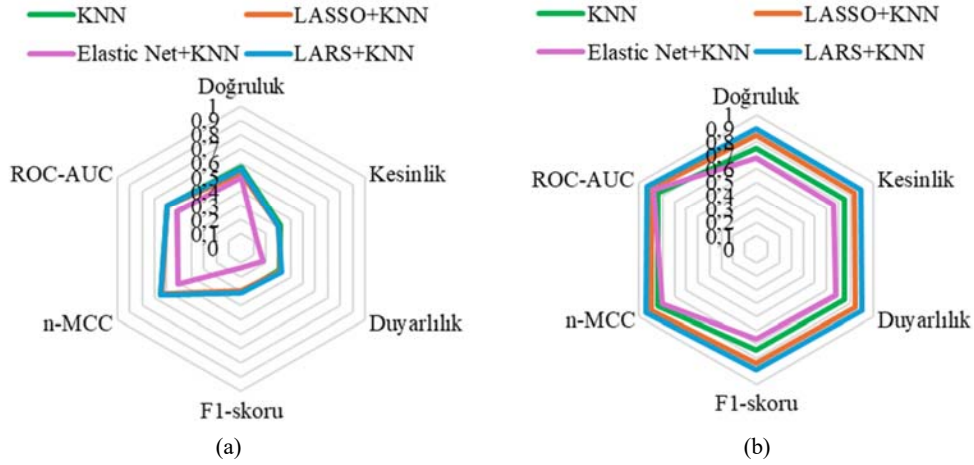
SVM algoritmasının performansının, SMOTE ve Elastic Net yönteminin uygulanması sonucunda büyük ölçüde arttığı görülmüştür. En yüksek doğruluk (0,7), kesinlik (0,76), duyarlılık (0,78), F1-Skoru (0,75), n-MCC (0,85) ve ROC-AUC (0,94) değerleri SMOTE uygulanmış Elastic Net+SVM yönteminden hesaplanmıştır. SVM yönteminin performansının SMOTE ve FS yöntemlerinden nasıl etkilendiği, Şekil 5'te verilen performans metrikleri ile gösterilmektedir.

RF algoritmasının performansının, SMOTE yönteminin uygulanması sonucunda büyük ölçüde arttığı görülmüştür. En yüksek doğruluk (0,90), kesinlik (0,89), duyarlılık (0,90), F1-Skoru (0,90), n-MCC (0,93) ve ROC-AUC (0,98) değerleri SMOTE uygulanmış RF

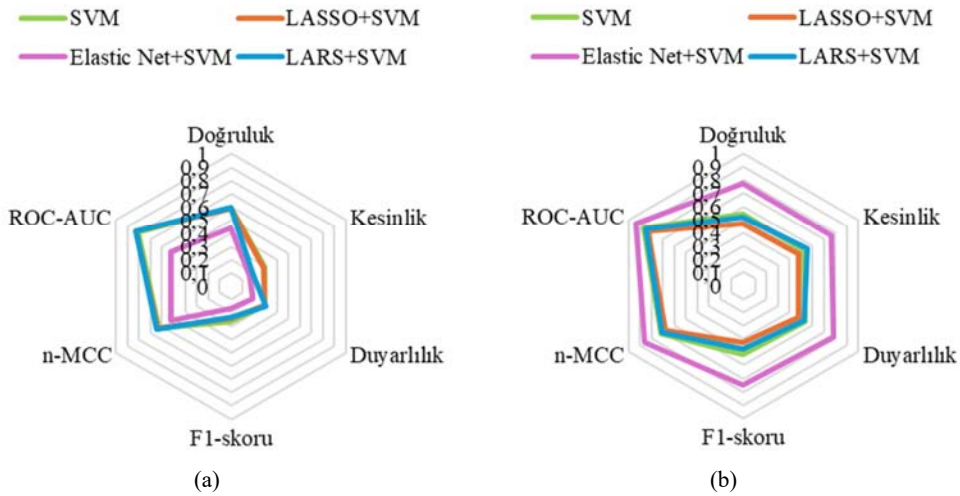
yönteminden hesaplanmıştır. RF yönteminin performansının SMOTE ve FS yöntemlerinden nasıl etkilendiği, Şekil 6'da verilen performans metrikleri ile gösterilmektedir.

DT algoritmasının performansının, SMOTE ve LARS yöntemlerinin uygulanması sonucunda büyük ölçüde arttığı görülmüştür. En yüksek doğruluk (0,90), kesinlik (0,89), duyarlılık (0,90), F1-Skoru (0,90), n-MCC (0,93) ve ROC-AUC (0,98) değerleri SMOTE uygulanmış LARS+DT yönteminden hesaplanmıştır. DT yönteminin performansının SMOTE ve FS yöntemlerinden nasıl etkilendiği, Şekil 7'de verilen performans metrikleri ile gösterilmektedir.

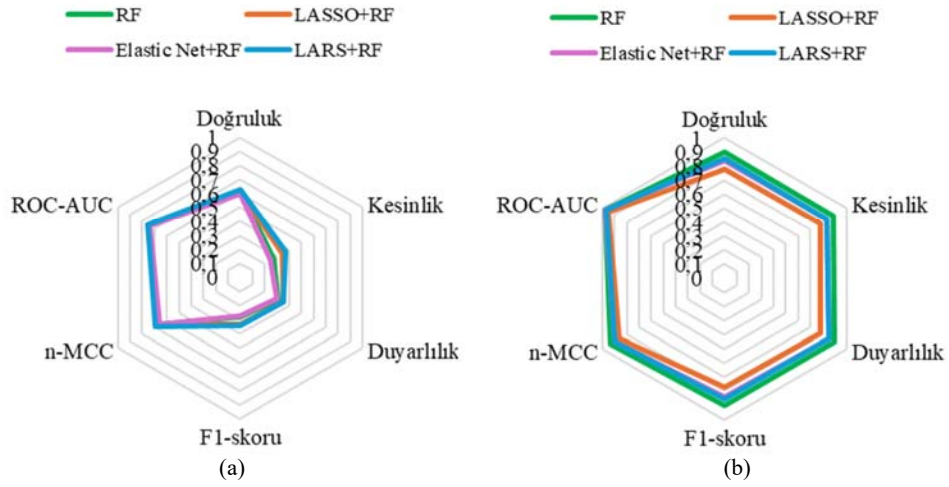
LR algoritmasının performansının, SMOTE yönteminin uygulanması sonucunda büyük ölçüde arttığı görülmüştür. En yüksek doğruluk (0,58), kesinlik (0,57), duyarlılık (0,58), F1-Skoru (0,57), n-MCC (0,73) ve ROC-AUC (0,85) değerleri SMOTE uygulanmış LR yönteminin performansının SMOTE ve FS yöntemlerinden nasıl etkilendiği, Şekil 8'de verilen performans metrikleri ile gösterilmektedir.



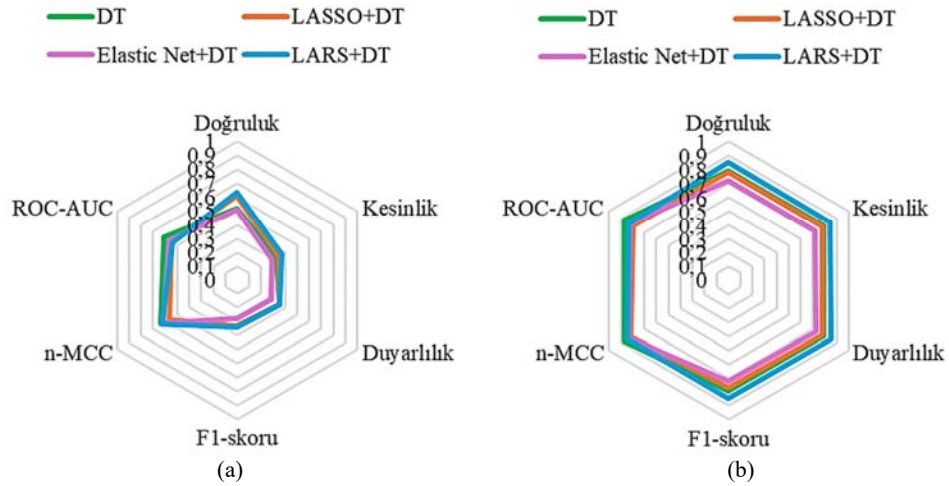
Şekil 4. KNN için performans metrikleri (a) SMOTE uygulanmamış (b) SMOTE uygulanmış (Performance metrics for KNN (a) without SMOTE (b) with SMOTE)



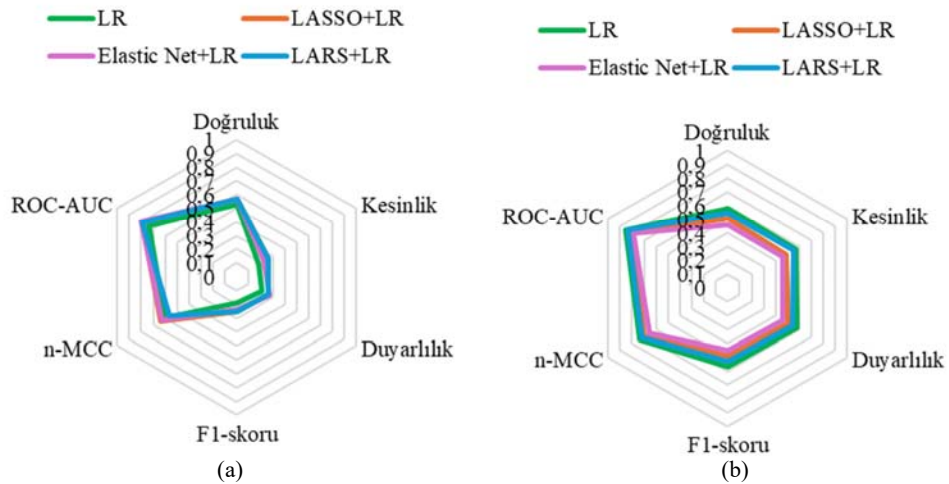
Şekil 5. SVM için performans metrikleri (a) SMOTE uygulanmamış (b) SMOTE uygulanmış (Performance metrics for SVM (a) without SMOTE (b) with SMOTE)



Şekil 6. RF için performans metrikleri (a) SMOTE uygulanmamış (b) SMOTE uygulanmış
(Performance metrics for RF (a) without SMOTE (b) with SMOTE)



Şekil 7. DT için performans metrikleri (a) SMOTE uygulanmamış (b) SMOTE uygulanmış
(Performance metrics for DT (a) without SMOTE (b) with SMOTE)



Şekil 8. LR için performans metrikleri (a) SMOTE uygulanmamış (b) SMOTE uygulanmış
(Performance metrics for LR (a) without SMOTE (b) with SMOTE)

ML algoritmalarından hesaplanan “performans metrikleri” arasında SMOTE uygulanmadan önce ve SMOTE uygulandıktan sonra istatistiksel olarak anlamlı fark olup olmadığını incelemek amacıyla $\alpha = 0.05$ anlamlılık düzeyinde gerçekleştirilen Wilcoxon İşaretli Sıralar Testi (Wilcoxon Signed Rank Test) sonuçları Tablo 7’de verilmiştir. Tablo 7’den; doğruluk, kesinlik, duyarlılık/geri çağırma, F1-skoru, n-MCC ve ROC-AUC performans metriklerinin tümü

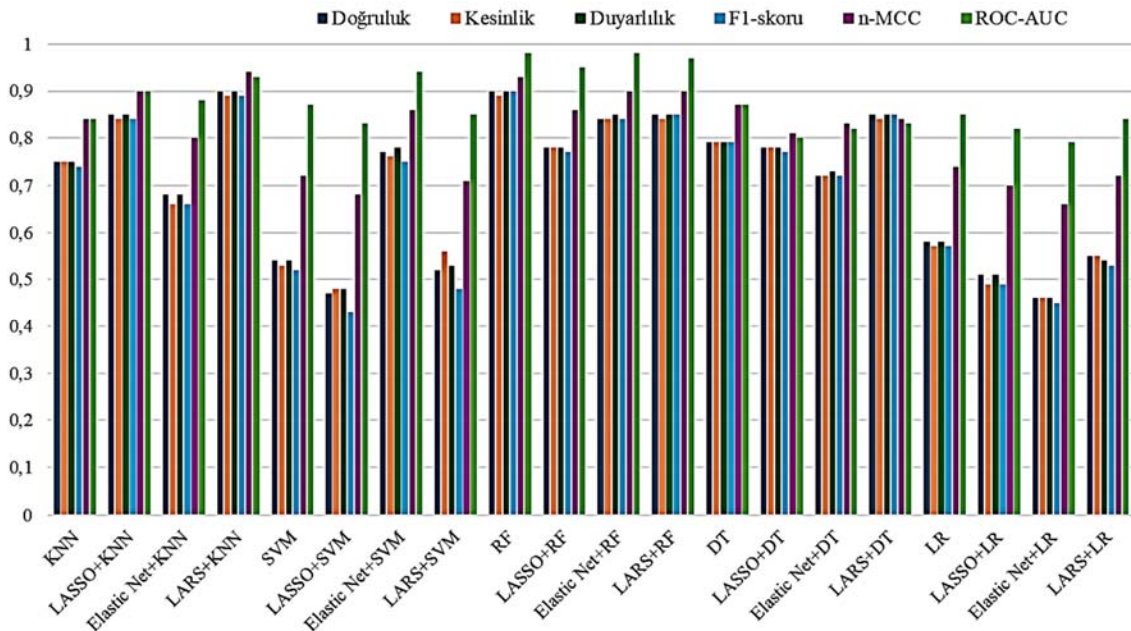
açısından, SMOTE uygulanmadan önce ve SMOTE uygulandıktan sonraki değerler arasında istatistiksel olarak anlamlı fark olduğu ($p < \alpha$) görülmektedir. Bu durum, UCI-HDC veri kümesi üzerinde SMOTE yönteminin uygulanmasının ele alınan ML algoritmalarının sınıflandırma performansını önemli ölçüde artırdığını göstermektedir.

Tablo 7. Performans metrikleri için SMOTE öncesi ve SMOTE sonrası karşılaştırmalara ilişkin sonuçlar
(Wilcoxon Signed Rank test results for pre-SMOTE and post-SMOTE comparison of performance metrics)

SMOTE öncesi ve SMOTE sonrası karşılaştırma	Test istatistiği - Z	p-değeri
Doğruluk vs. SMOTE Doğruluk	-2,950	0,003*
Kesinlik vs. SMOTE Kesinlik	-3,922	0,000*
Duyarlılık vs. SMOTE Duyarlılık	-3,923	0,000*
F1-skoru vs. SMOTE F1-skoru	-3,922	0,000*
n-MCC vs. SMOTE n-MCC	-3,923	0,000*
ROC-AUC vs. SMOTE ROC-AUC	-3,784	0,000*

Tablo 8. ML/FS+ML algoritmaları için SMOTE öncesi ve SMOTE sonrası karşılaştırmalara ilişkin sonuçlar
(Wilcoxon Signed Rank test results for pre-SMOTE and post-SMOTE comparison of ML/FS+ML algorithms)

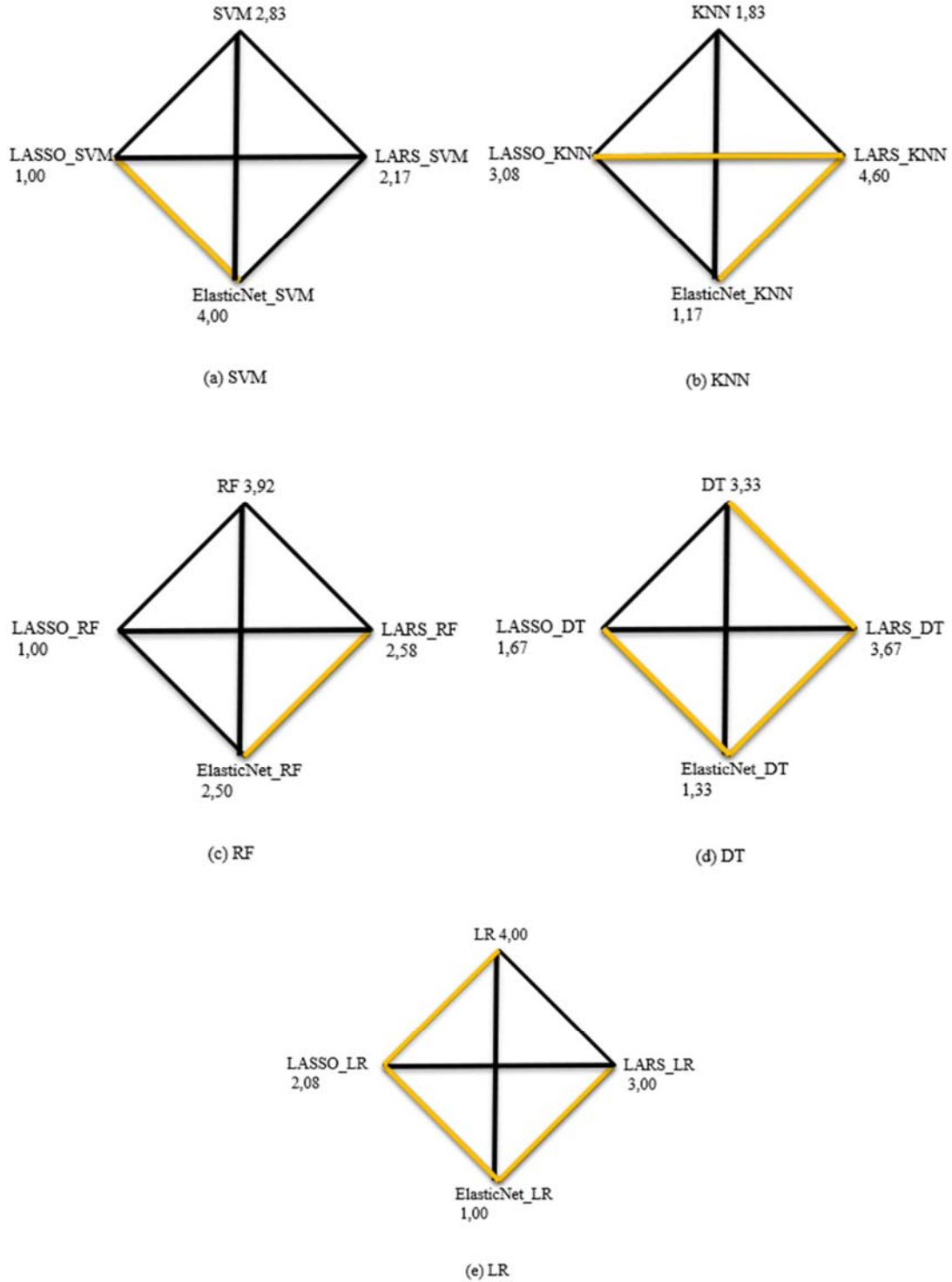
SMOTE öncesi ve SMOTE sonrası karşılaştırma	Test istatistiği - Z	p-değeri
KNN vs. SMOTE_KNN	-2,207	0,027*
SVM vs. SMOTE_SVM	-1,897	0,058
RF vs. SMOTE_RF	-2,207	0,027*
DT vs. SMOTE_DT	-2,207	0,027*
LR vs. SMOTE_LR	-2,201	0,028*
LASSO+KNN vs. SMOTE_LASSO+KNN	-2,207	0,027*
LASSO+SVM vs. SMOTE_LASSO+SVM	-1,490	0,136
LASSO+RF vs. SMOTE_LASSO+RF	-2,207	0,027*
LASSO+DT vs. SMOTE_LASSO+DT	-2,207	0,027*
LASSO+LR vs. SMOTE_LASSO+LR	-1,682	0,093
ElasticNet+KNN vs. SMOTE_ElasticNet+KNN	-2,201	0,028*
ElasticNet+SVM vs. SMOTE_ElasticNet+SVM	-2,201	0,028*
ElasticNet+RF vs. SMOTE_ElasticNet+RF	-2,226	0,026*
ElasticNet+DT vs. SMOTE_ElasticNet+DT	-2,207	0,027*
ElasticNet+LR vs. SMOTE_ElasticNet+LR	-1,363	0,173
LARS+KNN vs. SMOTE_LARS+KNN	-2,207	0,027*
LARS+SVM vs. SMOTE_LARS+SVM	-1,572	0,116
LARS+RF vs. SMOTE_LARS+RF	-2,207	0,027*
LARS+DT vs. SMOTE_LARS+DT	-2,201	0,028*
LARS+LR vs. SMOTE_LARS+LR	-1,997	0,046*



Şekil 9. SMOTE ve FS yöntemlerine dayalı makine öğrenmesi algoritmalarına ilişkin performans metrikleri
(Performance metrics for machine learning algorithms based on SMOTE and FS methods)

Bununla birlikte; SMOTE yeniden örnekleme yönteminin, ele alınan tüm ML ve özellik seçimi yöntemleri uygulanmış ML (FS+ML) algoritmaları üzerinde istatistiksel olarak anlamlı fark yaratıp yaratmadığını incelemek amacıyla $\alpha = 0.05$ anlamlılık düzeyinde gerçekleştirilen Wilcoxon İşaretli Sıralar Testi (Wilcoxon Signed Rank Test) sonuçları Tablo 8’de verilmiştir.

Tablo 8’de verilen sonuçlar, SMOTE yönteminin birçok ML ve FS+ML algoritmasında istatistiksel olarak anlamlı farklılıklar oluşturduğunu göstermektedir. Ancak SVM, LASSO+SVM, LARS+SVM, LASSO+LR ve ElasticNet+LR ile bunların SMOTE uygulanmış durumları arasında istatistiksel olarak anlamlı fark saptanmamıştır ($p > \alpha$).



Şekil 10. SMOTE uygulanmış durumda ML ve FS+ML algoritmalarına ilişkin ikili karşılaştırmalar (Pairwise comparisons of ML and FS+ML algorithms with SMOTE)

Şekil 9'da, dengesiz verilere uygulanan SMOTE yeniden örnekleme ve LASSO, Elastic Net, LARS özellik seçimi yöntemlerine dayalı olarak ML algoritmalarından hesaplanan performans metrikleri gösterilmiştir.

Tablo 6 ve Şekil 9'da verilen SMOTE uygulanmış ML algoritmalarından hesaplanan performans metrikleri incelendiğinde;

- Doğruluk değeri açısından RF ve LARS+KNN algoritmalarının en iyi performansı (0,90) gösterirken, Elastic Net+LR en kötü performansı (0,46) göstermiştir.
- Kesinlik değeri göre, RF ve LARS+KNN algoritmalarının en iyi (0,89), Elastic Net+LR algoritmasının ise en kötü (0,46) performansa sahip olduğu görülmüştür.
- Duyarlılık/geri çağırma değeri bakımından, RF ve LARS+KNN algoritmalarının en yüksek (0,90), Elastic Net+LR'nin en düşük (0,46) seviyeye ulaştığı gösterilmiştir.
- F1-Skoru açısından değerlendirildiğinde, RF algoritmasının en yüksek (0,90), LASSO+SVM'nin ise en kötü (0,43) skora sahip olduğu görülmüştür.
- n-MCC açısından LARS+KNN algoritmasının en iyi performansı (0,94), Elastic Net+LR algoritmasının en kötü performansı (0,66) gösterdiği görülmüştür.
- ROC-AUC açısından değerlendirildiğinde, RF ve Elastic Net+RF algoritmaları en yüksek (0,98) değere, Elastic Net+LR algoritması en düşük (0,79) değere sahiptir.

Çalışmada, LASSO, ElasticNet ve LARS özellik seçimi yöntemlerinin ML algoritmalarının sınıflandırma performansına olan etkisini de değerlendirebilmek amacıyla SMOTE uygulanmış durum için hesaplanan tüm performans metrikleri üzerinden $\alpha = 0,05$ anlamlılık düzeyinde Friedman testi yapılmıştır. Buna göre; KNN ile FS+KNN yöntemleri arasında ($p = 0,001 < \alpha$), SVM ile FS+SVM yöntemleri arasında ($p = 0,001 < \alpha$), RF ile FS+RF yöntemleri arasında ($p = 0,001 < \alpha$), DT ile FS+DT yöntemleri arasında ($p = 0,002 < \alpha$), LR ile FS+LR yöntemleri arasında ($p = 0,000 < \alpha$) istatistiksel olarak anlamlı fark olduğu görülmüştür. Anlamlı farklılık görülen ML ve FS+ML algoritmalarından hangilerinin farklılık yarattığına ilişkin ikili karşılaştırma sonuçları Şekil 10'da verilmiştir. İstatistiksel olarak anlamlı olan farklılıklar, her bir grubun ortalama sıralamasını gösteren turuncu çizgiler ile gösterilmiştir.

Şekil 10'da görsel olarak verilen ikili karşılaştırma test sonuçlarına göre; SMOTE uygulanmış durumda KNN ile LARS+KNN ($adj. p = 0,022 < \alpha$) ve ElasticNet+KNN ile LARS+KNN ($adj. p = 0,001 < \alpha$); LASSO+SVM ile ElasticNet+SVM ($adj. p = 0,000 < \alpha$); RF ile LASSO+RF ($adj. p = 0,001 < \alpha$); DT ile ElasticNet+DT ($adj. p = 0,044 < \alpha$), ElasticNet+DT ile LARS+DT ($adj. p = 0,010 < \alpha$) ve LASSO+DT ile LARS+DT ($adj. p = 0,044 < \alpha$); LR ile ElasticNet+LR ($adj. p = 0,000 < \alpha$), LR ile LASSO+LR ($adj. p = 0,044 < \alpha$) ve ElasticNet+LR ile LARS+LR ($adj. p = 0,044 < \alpha$) arasında istatistiksel olarak anlamlı performans farklılıkları olduğu görülmüştür. Diğer karşılaştırma çiftleri arasında anlamlı bir fark olduğuna dair kanıt bulunamamıştır.

3. Sonuçlar (Conclusions)

Hastalıkların erken tanısı ve doğru tahmin edilmesi, sağlık alanında en çok araştırılan konular arasında yer almaktadır. Kalp hastalıklarının erken tanısı hayati öneme sahip olduğundan, tahmin için makine öğrenmesi kapsamında yer alan çeşitli sınıflandırma algoritmaları sıklıkla kullanılmaktadır. Bu çalışma, kalp hastalığı tahmininde sınıflandırma algoritmalarının uygulanmasını ve karşılaştırmalı bir analiz yapılmasını amaçlamaktadır. Bu amaçla K en yakın komşuluk, destek vektör makinesi, rastgele orman, karar ağacı ve lojistik regresyon teknikleri ele alınmıştır.

Birçok ML algoritması dengesiz veri kümelerine karşı duyarlıdır ve bu veri kümelerinin dengeli hale getirilmesiyle performansları önemli ölçüde artabilmektedir. Çalışmada ele alınan UCI-HDC veri kümesindeki kategorik bağımlı değişkenin beş sınıfı arasındaki dengesizlik sorunu, SMOTE yeniden örnekleme yöntemi kullanılarak giderilmiştir.

Modelde yer alacak değişkenler, kullanılan özellik seçimi yöntemlerine göre değişiklik gösterebilmektedir. LASSO yöntemiyle "Gender, Cp, Trestbps, Fbs, Exang, Oldpeak, Slope, Ca ve Thal" özellikleri, Elastic Net yöntemiyle "Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak ve Ca" özellikleri, LARS yöntemiyle ise "Age, Gender, Cp, Trestbps, Chol, Fbs, Restecg, Exang, Oldpeak, Slope, Ca ve Thal" özellikleri belirlenmiştir. Ele alınan FS yöntemleriyle en yüksek önem katsayısına sahip özellik "Ca (büyük damar sayısı) (LASSO:0,3460, Elastic Net: 0,4247, LARS: 0,3788)" olarak tespit edilmiştir. Seçilen özelliklere, KNN, SVM, RF, DT, LR makine öğrenmesi algoritmaları sınıflandırma amacıyla uygulanmış ve bu algoritmaların performanslarını değerlendirebilmek amacıyla performans metriklerinden doğruluk, kesinlik, duyarlılık, F1-skoru, n-MCC ve ROC-AUC hesaplanmıştır.

SMOTE yönteminin uygulanmasıyla, ele alınan ML ve FS uygulanan ML algoritmalarının sınıflandırma performansının arttığı görülmüştür. SMOTE uygulanmış RF ve SMOTE uygulanmış LARS+KNN ile en yüksek doğruluk (0,90), kesinlik (0,89) ve duyarlılık (0,90) değerleri elde edilmiştir. SMOTE uygulanmış LARS+KNN ile en yüksek n-MCC (0,93) değerine ulaşılmıştır. SMOTE uygulanmış RF ile en yüksek F1-Skoru (0,90) ve en yüksek AUC (0,98) elde edilmiştir. Sonuç olarak, sınıf dengesizliği sorununu çözmek için SMOTE yeniden örnekleme yönteminin ve ilgisiz özellikleri elimine etmek için özellik seçimi yöntemlerinin kullanılması durumunda, çalışmada ele alınan tüm makine öğrenmesi algoritmalarının performansının önemli ölçüde arttığı görülmüştür.

Bu çalışmada geliştirilen SMOTE yeniden örnekleme ve özellik seçimi yöntemlerine dayalı makine öğrenmesi algoritmaları, kardiyovasküler hastalıkların erken tanısı ve risk sınıflandırmasında etkili bir yaklaşım sunmaktadır. Elde edilen yüksek performans metrikleri, erken teşhis, gereksiz tetkiklerin azaltılması ve kişiselleştirilmiş tedavi planlarının desteklenmesi açısından önemli klinik faydalar sağlamaktadır. Ayrıca, önerilen modellerin elektronik sağlık kayıtları ve rutin laboratuvar verilerine uygulanabilir olması, klinik kullanım potansiyelini artırmaktadır.

Gelecek çalışmalarda, bu çalışmada ele alınan yöntem ve algoritmaların lojistik, finans ve bankacılık, çevre ve enerji gibi karmaşık yönetsel konulardaki karar verme problemlerine uygulanması planlanmaktadır. Çalışmadan elde edilen sonuçlar ışığında, dengesiz veriler için Tomek Linkleri, Borderline SMOTE gibi diğer yeniden örnekleme yöntemlerinin ve topluluk tabanlı yaklaşımların kullanılması önerilmektedir. Ayrıca, özellik seçimi için Parçacık Sürü Optimizasyonu, Yapay Arı Kolonisi, Diferansiyel Evrim gibi metasezgisel algoritmaların uygulanması öngörülmektedir.

Teşekkür (Acknowledgement)

Bu makale, birinci yazarın Selçuk Üniversitesi Fen Bilimleri Enstitüsü'nde tamamladığı "Sınıflandırma problemlerinde veri dengeleme ve özellik seçiminin makine öğrenmesi algoritmalarının performansına etkisinin çok kriterli karar verme ile değerlendirilmesi" başlıklı doktora tezinden üretilmiştir. Çalışma, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) tarafından 125F351 numaralı proje ve Selçuk Üniversitesi Bilimsel Araştırma Projeleri Koordinatörlüğü (BAP) tarafından 25401078 numaralı proje ile desteklenmiştir.

Kaynaklar (References)

1. T.C. Sağlık Bakanlığı Halk Sağlığı Genel Müdürlüğü, Türkiye Kalp ve Damar Hastalıkları Önleme ve Kontrol Programı (2021-2026). T.C Sağlık Bakanlığı Yayın No: 1215, ISBN: 978-975-590-832-8, 2021.
2. World Health Organization (WHO), Noncommunicable diseases country profiles, ISBN 978-92-4-151462-0, 2018.
3. Gambella C., Ghaddar B., Naoum-Sawaya J., Optimization Problems for Machine Learning: A Survey, *European Journal of Operational Research*, 290 (3), 807-828, 2021.
4. Chen G., Chen J., A Novel Wrapper Method for Feature Selection and Its Applications, *Neurocomputing*, 159, 219-226, 2015.
5. Jiang H., Luo S., Dong Y., Simultaneous Feature Selection and Clustering Based on Square Root Optimization, *European Journal of Operational Research*, 289 (1), 214-231, 2021.
6. Ramyachitra D., Manikandan P., Imbalanced Dataset Classification and solutions: A Review, *International Journal of Computing and Business Research*, 5(4), 1-29, 2014
7. Kahramanli H., Allahverdi N., Design of a hybrid system for the diabetes and heart diseases, *Expert Systems with Applications*, 35 (1-2), 82-89, 2008.
8. Palaniappan S., Awang R., Intelligent Heart Disease Prediction System Using Data Mining Techniques, *IEEE/ACS International Conference on Computer Systems and Applications*, 108-115, 31 March - 04 April 2008, Doha, Qatar.
9. Gudadhe M., Wankhade K., Dongre S., Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network, *2010 International Conference on Computer and Communication Technology (ICCCCT)*, 741-745, Allahabad, India, 17-19 September 2010.
10. Panda D., Ray R., Abdullah A.A., Dash S.R., Predictive systems: Role of feature selection in prediction of heart disease, *Journal of Physics: Conference Series*, Volume 1372, International Conference on Biomedical Engineering, Penang Island, Malaysia, 26-27 August 2019.
11. Li J.P., Haq A.U., Din S.U., Khan J., Khan A., Saboor A., Heart Disease Identification Method Using Machine Learning Classification in e-Healthcare, *IEEE Access*, 8, 107562-107582, 2020
12. Ghosh P., Azam S., Jonkman M., Karim A., Shamrat F.J.M., Ignatious E., Shultana S., Beeravolu A.R., De Boer F. Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms with Relief and LASSO Feature Selection Techniques, *IEEE Access*, 9, 19304-19326, 2021.
13. Waqar M., Dawood H., Dawood H., Majeed N., Banjar A., Alharbey R., An efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction, *Scientific Programming*, 6621622, 1-12, 2021.
14. Ullah F., Chen X., Rajab K., Al Reshan M.S., Shaikh A., Hassan M.A., Rizwani M., Davidekova M., An Efficient Machine Learning Model Based on Improved Features Selections for Early and Accurate Heart Disease Prediction, *Computational Intelligence and Neuroscience*, 1906466, 2022.
15. Takci, H., Performance-enhanced KNN algorithm-based heart disease prediction with the help of optimum parameters. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 38 (1), 451-460, 2022.
16. Ghosh S., Islam M.A., Performance Evaluation and Comparison of Heart Disease Prediction Using Machine Learning Methods with Elastic Net Feature Selection, *American Journal of Applied Mathematics and Statistics*, 11 (2), 35-49, 2023.
17. Naseer A., Khan M.M., Arif F., Iqbal W., Ahmad A., Ahmad I., An Improved Hybrid Model for Cardiovascular Disease Detection Using Machine Learning in IoT, *Expert Systems*, 42(1), e13520, 2025.
18. Sevlı, O., Diagnosis of diabetes mellitus using various classifiers. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 38 (2), 989-1002, 2022.
19. Jafar A., Lee M., HypGB: High Accuracy GB Classifier for Predicting Heart Disease with Hyperopt HPO Framework and LASSO FS Method, *IEEE Access*, 11, 138201-138214, 2023.
20. Sahid M.A., Hasan M., Akter N., Tareq M.M.R., Effect of Imbalance Data Handling Techniques to Improve The Accuracy of Heart Disease Prediction Using Machine Learning and Deep Learning, *2022 IEEE Region 10 Symposium (TENSYP)*, Mumbai, India, 1-3 July 2022.
21. Tompra K.V., Papageorgiou G., Tjortjis C., Strategic Machine Learning Optimization for Cardiovascular Disease Prediction and High-Risk Patient Identification, *Algorithms*, 17 (5), 178, 2024.
22. Olewi Z.C., AlShemmary E.N., Al-Augby S., Adaptive features selection technique for efficient heart disease prediction, *Journal of Al-Qadisiyah for Computer Science and Mathematics*, 15 (1), 1-18, 2023
23. Hasanah U., Soleh A.M., Sadik K., Effect of Random Under sampling, Oversampling, and SMOTE on the Performance of Cardiovascular Disease Prediction Models, *Jurnal Matematika, Statistika dan Komputasi*, 21 (1), 88-102, 2024.
24. Muntasir Nishat M., Faisal F., Jahan Ratul I., Al-Monsur A., Ar-Rafi A.M., Nasrullah, S. M., Reza, T., Khan, R.H., A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset, *Scientific Programming*, 3649406, 2022.
25. Krawczyk B., Galar M., Jeleń L., Herrera F., Evolutionary Undersampling Boosting for Imbalanced Classification of Breast Cancer Malignancy, *Applied Soft Computing*, 38, 714-726, 2016.
26. Kumar V., Minz S., Feature Selection: A Literature Review, *SmartCR*, 4 (3), 211-229, 2014.
27. Li Y., Li T., Liu H., Recent Advances in Feature Selection and Its Applications, *Knowledge and Information Systems*, 53, 551-577, 2017.
28. Venkatesh B., Anuradha J., A Review of Feature Selection and Its Methods, *Cybernetics and Information Technologies*, 19 (1), 3-26, 2019.
29. Chen Y., Tsionas E.G., Zelenyuk V., LASSO DEA for Small and Big Data, *Omega*, 102, 102419, 2021.
30. Fonti V., Belitser E., Feature Selection Using LASSO, *VU Amsterdam Research Paper in Business Analytics*, 30, 1-25, 2017.
31. Ransam J., Cook J (2018) LASSO Regression, *Journal of British Surgery*, 105 (10), 1348-1348, 2017.
32. Tibshirani R., Regression Shrinkage and Selection via the LASSO, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58 (1), 267-288, 1996.
33. Zou H., Hastie T., Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67 (2), 301-320, 2005.
34. Yu Q., Miche Y., Eirola E., Van Heeswijk M., Séverin E., Lendasse A., Regularized Extreme Learning Machine for Regression with Missing Data, *Neurocomputing*, 102, 45-51, 2013.
35. El Naqa I., Murphy M.J., What is Machine Learning? In: *Machine Learning in Radiation Oncology*, Springer, Berlin, 2015.
36. Mandal J.K., Bhattacharya D., *Emerging Technology in Modelling and Graphics*, Proceedings of IEM Graph 2018, Springer, e-ISBN: 978-981-13-7403-6, 2020
37. Soofi A.A., Awan A., Classification Techniques in Machine Learning: Applications and Issues, *Journal of Basic & Applied Sciences*, 13 (1), 459-465, 2017.
38. Almomany A., Ayyad W.R., Jarrah A., Optimized Implementation of An Improved KNN Classification Algorithm Using Intel FPGA Platform: COVID-19 Case Study, *Journal of King Saud University-Computer and Information Sciences*, 34 (6), 3815-3827, 2022.
39. Ayon S.I., Islam M.M., Hossain M.R. Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques, *IETE Journal of Research*, 68 (4), 2488-2507, 2022.
40. Arora N., Kaur P.D. A Bolasso Based Consistent Feature Selection Enabled Random Forest Classification Algorithm: An Application to Credit Risk Assessment, *Applied Soft Computing*, 86, 105936, 2020.
41. Soman K., Loganathan R., Ajay V., Machine learning with SVM and Other Kernel Methods: PHI Learning Pvt. Ltd, ISBN: 8120334353, 9788120334359, 2009.
42. Ileberi E., Sun Y., Wang Z., A Machine Learning Based Credit Card Fraud Detection Using the GA Algorithm for Feature Selection, *Journal of Big Data*, 9 (1), 1-17, 2022.
43. Kingsford C., Salzberg S.L., What are Decision Trees?, *Nature Biotechnology*, 26 (9), 1011-1013, 2008.
44. Patel H.H., Prajapati P., Study and Analysis of Decision Tree Based Classification Algorithms, *International Journal of Computer Sciences and Engineering*, 6 (10), 74-78, 2018.
45. Rani P., Kumar R., Ahmed N.M.S., Jain A., A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, 7 (3), 263-275, 2021.
46. Izmirlian G., Application of the Random Forest Classification Algorithm to a SELDI-TOF Proteomics Study in the Setting of a Cancer Prevention Trial, *Annals of the New York Academy of Sciences*, 1020 (1), 154-174, 2004.

47. Oshiro T.M., Perez P.S., Baranauskas J.A. How Many Trees in a Random Forest?, 8th International Conference of Machine Learning and Data Mining in Pattern Recognition (MLDM 2012), Berlin, Germany, 13-20 July 2012..
48. Lemon S.C., Roy J., Clark M.A., Friedmann P.D., Rakowski W., Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression, *Annals of Behavioral Medicine*, 26, 172-181, 2003.
49. Shah D., Patel S., Bharti S.K., Heart Disease Prediction Using Machine Learning Techniques, *SN Computer Science*, 1, 1-6, 2020.
50. Gupta A., Kumar R., Arora H.S., Raman B., MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis, *IEEE Access*, 8, 14659-14674, 2019.
51. Kumar P.T., Sriram Praveen V.A., Maheshwari R., Gowda S.D., A Comparative Study of Machine Learning Techniques in Heart Disease Detection, *Perspectives in Communication, Embedded-systems and Signal-processing-PiCES*, 4 (10), 264-272, 2021.
52. Pouyanfar S., Tao Y., Mohan A., Tian H., Kaseb A.S., Gauen K., Dailey R., Aghajanzadeh S., Lu Y.-S., Chen, S.-C., Shyu M.-L., Dynamic Sampling in Convolutional Neural Networks for Imbalanced Data Classification, 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, USA, 10-12 April 2018.
53. Xue J.-H., Titterington, D.M., Do Unbalanced Data Have a Negative Effect on LDA?, *Pattern Recognition*, 41 (5), 1558-1571, 2008.
54. Lee J.-S., AUC4. 5: AUC-Based C4. 5 Decision Tree Algorithm for Imbalanced Data Classification, *IEEE Access*, 7, 106034-106042, 2019.
55. De Diego I.M., Redondo A.R., Fernández R.R., Navarro J., Moguerza J.M., General Performance Score for Classification Problems, *Applied Intelligence*, 52 (10), 12049-12063, 2022.
56. Terven, J., Cordova-Esparza, D.M., Romero-González, J.A., Ramírez-Pedraza, A., Chávez-Urbiola, E.A., A Comprehensive Survey of Loss Functions and Metrics in Deep Learning, *Artificial Intelligence Review*, 58 (195), 2025.
57. Farhadpour, S., Warner, T.A., Maxwell, A.E. Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices. *Remote Sensing*, 16 (533), 2024.
58. Almustafa K.M. Prediction of Heart Disease and Classifiers' Sensitivity Analysis, *BMC Bioinformatics*, 21, 1-18, 2020.
59. Kadhim M.A., Radhi A.M., Heart Disease Classification Using Optimized Machine Learning Algorithms, *Iraqi Journal for Computer Science and Mathematics*, 4 (2), 31-42, 2023.
60. Lupague R.M.J.M., Mabborang R.C., Bansil A.G., Lupague M.M., Integrated Machine Learning Model for Comprehensive Heart Disease Risk Assessment Based on Multi-Dimensional Health Factors, *European Journal of Computer Science and Information Technology*, 11 (3), 44-58, 2023.

