Hiper Parametre Ayarlama ve Veri Dengelemenin Kalp Hastalığı Tahmini İçin Kullanılan Makine Öğrenimi Algoritmaları Üzerindeki Etkilerinin İncelenmesi

Araştırma Makalesi/Research Article

Fuat Sungur [1]*, Halit Bakır [2]
[1]Savunma Teknolojileri ABD, Sivas Bilim ve Teknoloji Üniversitesi, Sivas, Türkiye
[2]Bilgisayar Mühendislik Bölümü, Sivas Bilim ve Teknoloji Üniversitesi, Sivas, Türkiye
sungurbey58@hotmail.com, halit.bakir@sivas.edu.tr
(Geliş/Received:18.07.2023; Kabul/Accepted:31.10.2023)
DOI: 10.17671/gazibtd.1399813

Özet— Kalp hastalığı belirtilerinin ihmal edilmesi ciddi rahatsızlıklarla hatta ölümle sonuçlanabilir. Makine öğrenme teknikleri ile ön tanı için bu belirtiler kullanılarak kişide kalp hastalığı olup olmadığına dair tahmin yapılabilmektedir. Bu çalışmada Logistic Regression, Decision Trees, Random Forest, K Nearest Neighbors, Naive Bayes, Gradient Boosting, XGBoost ve Bagging algoritmaları ile kalp hastalığı tahmini yapılmıştır. SMOTE, SMOTETomek, Oversample Minority Class, Undersample Majority Class veri dengeleme yöntemleri ile dört ayrı veri seti oluşturulmuştur. Seçilen tüm makine öğrenme algoritmalarına Random Search ve Bayesian Optimizasyon teknikleriyle hiper parametre optimizasyonu yapılarak sonuçlar karşılaştırılmıştır. Veri dengeleme ve hiper parametre optimizasyonunun kalp hastalığının tahmininde kullanılan makine öğrenme teknikleri performansına etkisi karşılaştırılarak literatüre özgün bir çalışma kazandırılmıştır. Çalışmada Amerika Birleşik Devletleri'nde 319.795 kişi ile yapılan 20 öz nitelikli bir anket olan veri seti kullanılmıştır. Random Forest algoritması SMOTETomek veri dengeleme tekniği kullanılarak ve Bayesian hiper parametre optimizasyonu yapılarak oluşturulan modelde %94 tahmin başarısı elde edilmiştir. Ayrıca, Random Forest algoritması ile Oversample Minority Class veri dengeleme tekniği kullanılarak ve Bayesian hiper parametre optimizasyonu yapılarak %97 sınıflandırma doğruluğu elde edilmiştir.

Anahtar Kelimeler— kalp Hastalıkları, rasgele orman, makine öğrenmesi, smote, smotetomek, bayes optimizasyon

Evaluating The Effects of Hyperparameter Tuning and Data Balancing on Machine Learning Algorithms Used for Heart Disease Prediction

Abstract— Neglecting the symptoms of heart disease can result in serious conditions and even death. Machine learning techniques can be used to make predictions about whether a person has heart disease based on these symptoms. In this study, heart disease prediction was performed using Logistic Regression, Decision Trees, Random Forest, K Nearest Neighbors, Naive Bayes, Gradient Boosting, XGBoost, and Bagging machine learning algorithms. Four separate datasets were created using data balancing methods such as SMOTE, SMOTETomek, Oversample Minority Class, and Undersample Majority Class. Hyperparameter optimization was conducted for all selected machine learning algorithms using Random Search and Bayesian Optimization techniques, and the results were compared. By comparing the impact of data balancing and hyperparameter optimization on the performance of machine learning techniques used in predicting heart disease, this study contributes to the literature with an original approach. The study utilized a dataset from a survey of 319,795 individuals in the United States, which included 20 relevant features. The Random Forest algorithm achieved a prediction accuracy of 94% in the model created using the SMOTETomek data balancing technique and Bayesian hyperparameter optimization. Additionally, the Random Forest algorithm, with the Oversample Minority Class data balancing technique and Bayesian hyperparameter optimization, achieved a classification accuracy of 97%.

Keywords— heart Diseases, random forest, machine learning, smote, smotetomek, bayesian optimization

1. GİRİŞ (INTRODUCTION)

Kalp hastalıkları ölüm nedenleri arasında üst sıralarda yer almaktadır. Birçok insan kalp krizi ve kalp sağlığına bağlı hastalıklar nedeniyle hayatını kaybetmektedir. Dünya Sağlık Örgütü'nün 2019 yılı verilerine göre yaklaşık 17,9 milyon insan Kalp sağlığına bağlı hastalıklardan dolayı hayatını kaybetmiştir. Bu sayı dünya üzerindeki ölümlerin yaklaşık %32'sine denk gelmektedir. Kalp hastalıklarına bağlı ölümlerin büyük bir kısmını (%85) ani ölümler oluşturmuştur [1].

Kalp hastalıklarının çeşitli nedenleri olabilir. Bu nedenler genetik faktörler ve çevresel faktörler olarak iki ana başlıkta toplanabilir. Genetik faktörler önceki kuşaklardan aldığımız hastalığa yatkınlık gösteren genlerle alakalı olup henüz bilim bu konunun çözümünde yeterince ilerleme gösterememiştir. Çevresel faktörler ise bebeğin anne rahmine düştüğü andan itibaren, annenin hamilelik sürecindeki alışkanlıkları, beslenmesi ve hatta duygu durumları ile doğumdan sonra bireyin yaşam tarzını kapsamaktadır. Kalp hastalıklarını tetikleyen bireyin yaşam tarzı ile ilgili başlıca göstergeler beslenme alışkanlıkları, zararlı alışkanlıklar ve yaşanılan çevre gelmektedir. Kalabalık, stresli büyük şehirlerde yaşam, sağlıksız beslenme, obezite, hareketsiz, spordan uzak yaşam tarzı başta kalp hastalıkları olmak üzere birçok hastalığa davetiye çıkarmaktadır. Bununla birlikte alkol, sigara ve uyuşturucu madde kullanımı kalp hastalıklarına önemli ölçüde neden olmaktadır.

Bu çalışmada kalp hastalıklarının tahmin edilmesinde makine öğrenme algoritmalarının karşılaştırmalı performans analizinin yapılması amaçlanmıştır. Kalp hastalıklarının erken teşhisi büyük ölçüde hayat kurtarıcı olduğu için bu çalışmanın sağlık sisteminde bir farkındalık yaratması beklenmektedir.

Günümüzde kalp hastalıklarının nedenleri ve kalp hastalıkları ile ilgili birçok çalışma yapılmakta olup bu konuda birçok kaynak bulunmaktadır. Bu kaynaklardan biriside veri setleridir. Veri seti belirli bir konu ile alakalı bilgilerin sistematik ve anlamlı bir şekilde depolandığı yapılardır. Bu çalışmada Amerika Birleşik Devletleri Hastalık Kontrol ve Önleme Merkezleri tarafından her yıl rutin olarak yapılan Davranışsal Risk Faktörü Gözetleme Sisteminin bir parçası olan veri setinden yararlanacağız. Veri seti 2020 yılında yaklaşık 320bin kişi ile yapılan anketin 20 özniteliğine dair örnekleri içermektedir.

Veri setlerini işlemenin birçok yolu vardır. Bunlardan birisi de makine öğrenmesi algoritmaları ile veri setlerinin analiz edilmesi ve çıkan anlamlı sonuçların değerlendirilmesidir. Bu çalışmada, makine öğrenmesi sınıflandırma algoritmalarından Logistic Regression (LR), Decision Trees (DT), Random Forest (RF), K Nearest Neighbors (KNN), Naive Bayes (NB), Gradient Boosting (GB), XGBoost (XGB) ve Bagging (BAG) algoritmalarının her biri ayrı ayrı çalışılıp en uygun yöntem ya da yöntemlerin seçilmesi planlanmıştır. En uygun algoritmalar belirlendikten sonra hiper parametre optimizasyonu (HPO) teknikleri ile en iyi parametreler belirlenerek bu hiper parametreler ile modeller daha verimli hale getirilecektir.

2. KAYNAK ARAŞTIRMASI (LITERATURE SURVEY)

Bu kısımda çalışmada yararlanılan kaynakların amaçları, teknikleri, elde edilen sonuçları değerlendirilerek literatüre katkıları ve eksiklikleri gözden geçirilecektir.

Yadav ve ark. [3] kalp hastalıklarını tahmin etmek üzere makine öğrenme tekniklerine dayanan bir arayüz oluşturmayı hedeflediler. Rasgele orman ve karar ağaçlarında %97,08 tespit başarısı, Lojistik Regresyonda %80,52 tespit başarısı elde ettiler. Başka bir çalışmada [4] Bhatt ve ark. kalp hastalıklarının tespit edilmesinde RF, XGB, Multi-Layer Perceptron (MLP), DT makine öğrenme algoritmalarını kullandılar. MLP tekniğinde %87,23 tespit başarısı elde ettiler. Ayrıca bu çalışmada veri temizleme, özellik seçimi ve kümeleme gibi veri düzenleme işlemleri yaptılar. Özdemir [5] kalp hastalıklarının teşhisinde EKG verilerini kullandı. DT, RF, Extra Tree, GB ve Support Vector Machine (SVM) teknikleri kullanılarak yapılan çalışmada Extra Tree algoritmasında %96,14 tespit doğruluğu elde edilmiştir. Başka bir çalışmada [6] kalp hastalıklarının tahmin edilmesinde makine öğrenme algoritmaları kullanıldı. Çalışmada GridSearchCV tekniği ile hiper parametre optimizasyonu yapılarak makine öğrenme algoritmalarının performansını artırmak hedeflenmiştir. % 3-5 aralığında performans artışı elde edilen çalışma neticesinde AdaBoost algoritması ile %95 tespit başarısı elde edilmiştir. Anitha ve ark. [7] kalp hastalıklarının tahmin edilmesine yönelik yaptığı çalışmada KNN, NB ve SVM makine öğrenme algoritmalarını karşılaştırmıştır. Çalışmada kullandığı veri setinde NB algoritmasının %86,6 doğruluk oranı ile en iyi sonuçları verdiği çıkarımında bulunulmuştur. Çifçi M. [8] yapay sinir ağlarını birleştirerek oluşturduğu sistemle kalp hastalıkları tahminini amaçlamıştır. Sınıf Nitelik Bağımlılık Maksimizasyonu, Temel Bileşenler Analizi ve SVM yöntemlerini birleştirilerek kalp hastalıkları ön teşhisinde kullanılabilecek bir sistem oluşturmuştur. 270 hastanın 13 öz nitelikli veri setini kullanmıştır. Hibrit sistemin kullanılmadığı sınıflayıcılarda %75,27 doğruluk oranı bulunurken, kendi hibrit sisteminde %89,17 başarı elde etmiştir. Kamat P. [9] çalışmasında J48, NB ve SVM makine öğrenme algoritmalarını k-kat çapraz doğrulama ile birlikte kullanmıştır. Kullanılan yöntemlerin kalp hastalığı tahmininde kullanılabileceği sonucuna ulaşmış ancak model sonuçları ile ilgili paylaşım yapılmamıştır. Çalışmanın bu yönü ile eksik kaldığı söylenebilir. Rajdhan ve ark. [10] kullandıkları veri seti ile kalp hastalıklarını tahmin etmeyi amaçladılar. NB, DT, LR, RF tekniklerini karşılaştırıp RF algoritmasının %90,16 ile diğerlerine kıyasla en yüksek doğruluk oranına sahip olduğunu elde ettiler. Görgün M. [11] 165'i kalp hastası olan 303 kişinin bilgilerinden oluşan veri setini kullanmıştır. Cinsiyet, yaş, kolesterol, diyabet, göğüs ağrısı gibi belirtileri kalp hastalığı tahmininde kullanmıştır. Makine öğrenmesi algoritmalarından KNN, SVM, NB, LR, DT, RF, XGB, LightGBM ve BAG algoritmalarını karşılaştırmış, RF

kullanarak %90,16 doğruluk oranı elde etmiştir. Çalışmada %81,97 tahmin değeri ile en düşük doğruluk oranına sahip LightGBM algoritması göze çarpmaktadır. Repaka ve ark. [12] mobil uygulama yolu ile kişiden aldıkları bilgileri veri setinde işleyerek kişinin kalp hastalığı olup olmadığını tahmin etmek istediler. NB algoritmasını kullandıkları uygulamada %89,77 doğruluk oranına ulaştılar. Köse O. [13] kullandığı veri setinde kalp hastalığına etki eden 38 adet faktör üzerinde çalışmıştır. DT algoritmalarından Chi-Square Automatic Interaction Detection ve CART tekniklerini kullanarak karşılaştırmalı sonuçlar elde etmiştir. Veri setinde %70'i eğitim örneklemi, %30 'u test örneklemi aldığı çalışmanın daha başarılı olduğu sonucuna ulaşmıştır. CART ve Chi-Square Automatic Interaction Detection algoritmalarının farklı ağaç oluşturma yöntemleri benimsediği, Chi-Square Automatic Interaction Detection algoritması ile detaylı sonuçlar elde edilirken CART algoritması ile genel sonuçlar elde edildiğini sonucuna ulaşmıştır. Ramalingam ve ark. [14] 2019 yılında makine öğrenme algoritmalarını kullanarak kalp hastalıklarını tahmin etmeyi amaçlamışlar. DT modellerinin Temel Bileşen Analizi ile kullandığında performansın iyi olmasına rağmen uyum problemleri ile karşılaşıldığı sonucuna ulaşmıştır. RF algoritmasında birden çok DT kullanarak uyum problemlerini çözebildikleri çıkarımında bulundular. NB modellerinin hızlı olduğu ve iyi sonuçlar (%83,49) gösterdiğini belirtiler. SVM yönteminin ise doğruluk oranının çok yüksek olduğu (%92,1) sonucuna ulaştılar. Sharma ve ark. [15] makine öğrenme tekniklerinin karşılaştırmalı tahlilini yapmıştır. KNN modelin SVM modelden daha verimli çalıştığı çıkarımında bulunmuştur. J48 modelinin aşırı uyum konusunda DT modelinden daha iyi olduğu, yapay sinir ağlarının çerim için eğitimde daha iyi sonuçlar verdiği çıkarımında bulunmuştur. Kâmil [16] araştırmasında RF ve Kmean kavramlarına dayalı Parçacık Sürü Optimizasyonuna tekniklerini sentezleyerek en iyi K ortalama tekniğini ve RF'deki ağaç sayısını bulmayı amaçlamıştır. Kullandığı yöntem ile geleneksel yöntemlerden daha kullanışlı işlevsel ve hızlı sonuçlar elde etmiştir. Salman I. [17] 787 hasta ve 24 değişken içeren veri setini kullandığı çalışmasında NB, DT ile güçlendirilmiş Naive Bayes (TAN) ve TAN ve Chow-Liu (TANI) karşılaştırması yaparak kalp hastalıkları tahmini yapmayı amaçlamıştır. Sonuçta TANI metodunun diğerlerine kıyasla en yüksek doğruluk oranına sahip olduğu sonucuna ulaşmıştır. Konda ve ark. [18] kullandıkları veri setindeki 15 ayrı niteliği kullanarak kalp hastalığı ön tanısı için makine öğrenme modellemesi yapmayı amaçladılar. DT, NB ve Yapay Sinir Ağları algoritmalarını kullandıkları çalışmada %89 doğruluk oranı ile DT algoritmasının en uygun yöntem olduğu çıkarımında bulundular. Tarawneh ve ark. [19] 303 örneklem 14 öznitelik içeren araştırmalarında SVM, KNN, RF, J48 algoritma modellerini kullanarak %89,2 doğruluk oranını ile kalp hastalıklarının tahmini ile ilgili çalışma gerçekleştirdiler. Venkatesh ve ark. [2] çalışmalarında Naive Bayes algoritma modelini kullanmıştır. Kalp hastalıklarının tahmininde NB modelinin %97,1 doğruluk oranına sahip olduğunu ortaya koydular. Çil E. [20] 2022 yılında yaptığı çalışmasında Yapay Sinir Ağları, SVM ve KNN, LR, DT, RF, NB algoritmalarını karşılaştırmalı bir şekilde kullanarak kalp hastalıklarını tespit etmeyi amaçlamıştır. LR (%90,77), SVM (%90,52) ve Yapay Sinir Ağları (%90,54) algoritmalarında en anlamlı değerlere ulaşılmıştır. Kesinlik skorlarında LR ve Yapay Sinir Ağları, Duyarlılık skorunda NB, F1 skorlarında ise DT ve NB'nin en başarılı modeller olduğu sonucuna ulaşılmıştır. Sonuç olarak çalışmada LR ve Yapay Sinir Ağları modellerinin en uygun modeller olduğuna karar verilmiştir.

Literatürde kalp hastalıklarının tespit edilmesi ile ilgili çalışmaların daha çok tahmin modelleri oluşturmaya yönelik olduğu görülmüştür. Araştırılan çalışmaların hiç birisinde veri dengelemenin model performansı üzerindeki etkisi incelenmemiştir. HPO yapılan çalışmalarda HPO tekniklerinin karşılaştırması yapılmamıştır. Araştırılan kaynaklar içerisinde en verimli model % 97,1 tahmin başarısı ile NB algoritmasına [2] aittir. Veri dengeleme ve hiper parametre optimizasyonunun kalp hastalığının tahmin edilmesinde kullanılan makine öğrenme algoritmalarının performansı üzerindeki etkisinin literatürdeki çalışmalardan farklı ve araştırılması gereken bir alan olduğu düşüncesi bizi bu çalışmaya motive etmiştir. Bizim çalışmamız kalp hastalıklarının tahmin edilmesinde %97 tahmin başarısının yanında veri dengeleme ve HPO'yu içeren detaylı uygulamaları ile benzer çalışmaların önüne geçmektedir.

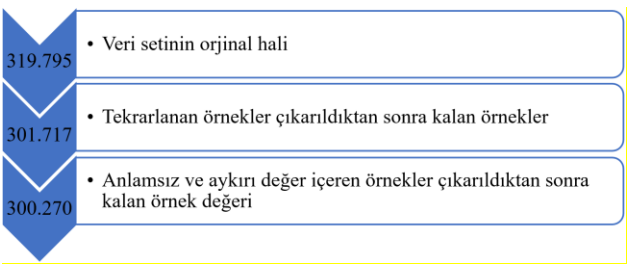
3. MATERYAL ve YÖNTEM (MATERIAL and METHOD)

3.1. Veri Seti (Dataset)

Çalışmada kullanılan en önemli materyal veri setidir. Bu çalışmada kullanılan veri seti, Amerika Birleşik Devletleri Hastalık Kontrol ve Önleme Merkezleri tarafından her yıl rutin olarak yapılan Davranışsal Risk Faktörü Gözetleme Sisteminin bir parçasıdır. 2020 yılında ait veriler kalp rahatsızlıkları ile ilgili yaklaşık 320000 kişi ile yapılan anket araştırması sonuçlarını kapsamaktadır.

3.2. Veri Setinin Hazırlanması (Preparation of the Data Set)

Çalışma sırasında karışıklığı önlemek için öznitelik isimleri daha anlaşılabilir isim kısaltmaları ile değiştirilmiştir. Veri setinde ilk durumda 319.795 olan örnek, tekrarlayan satırlar, anlamsız aykırı değerler çıkarıldıktan yani veri seti temizlendikten sonra 300,270 örnek olarak güncellenmiştir. Temizleme işlemi aşamaları Şekil 1'de gösterilmiştir.



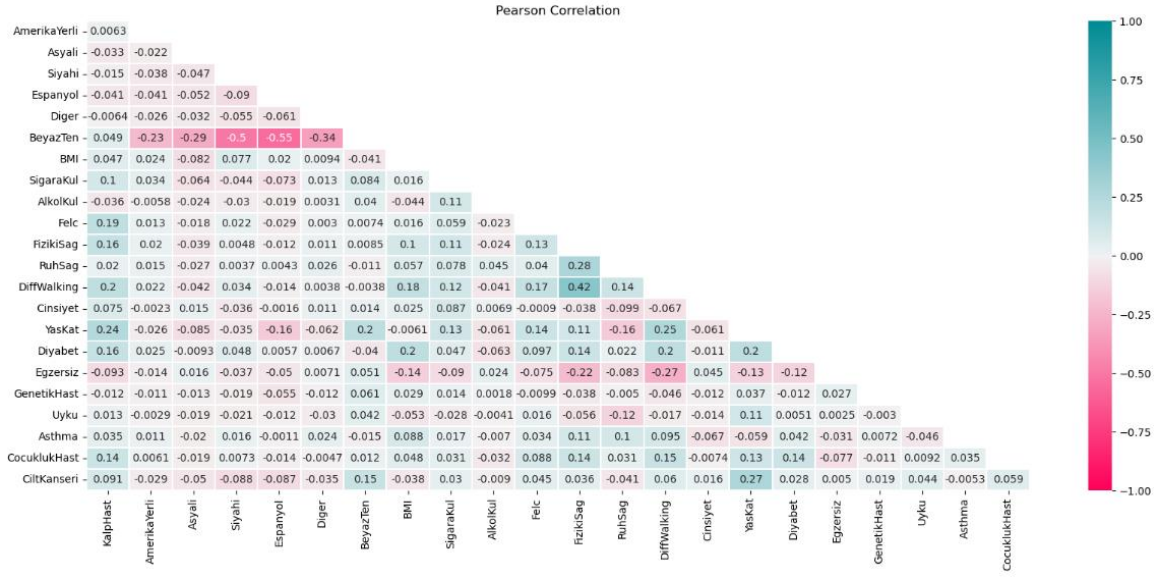
Şekil 1. Veri setinin temizlenme aşamaları (Stages of cleaning the data set)

Veri setinin orijinal halinde 319.795 örnek, 18 öznitelik bulunmaktadır. “Race” özniteliğinin her bir değeri On-Hot-Encoding tekniği kullanılarak birer özniteliğe dönüştürülmüştür. Bu işlemden sonra öznitelik sayısı 23 olmuştur. Öznitelik değerleri üzerinde istatistiksel işlemler yapılabilmesi önce numerik dönüşüm yapılmıştır. Daha sonra her değer kategorik olarak [0,1] arasına ölçeklendirilmiştir. Bu işlem sonrasında korelasyon analizi ile özniteliklerin çıktı ile ilişkisi incelenmiştir. Kaggle sitesinden alınan veri seti 279 soruluk anketin 18 soruya indirgenmiş halidir [21]. Öznitelikler ışığında kişide kalp hastalığı olup olmadığını değerlendirmeye yönelik işlem yapılması amaçlanmaktadır. Bu nedenle kalp hastalığı olup olmadığı ile ilgili öznitelik “HeartDisease” çıktı indeksi olarak belirlenmiştir. Veri setinin %20’si test verisi, %80’i ise eğitim verisi olarak kullanılmıştır. Veri setinde bulunan öznitelikler ve açıklamaları Tablo 1 de verilmiştir.

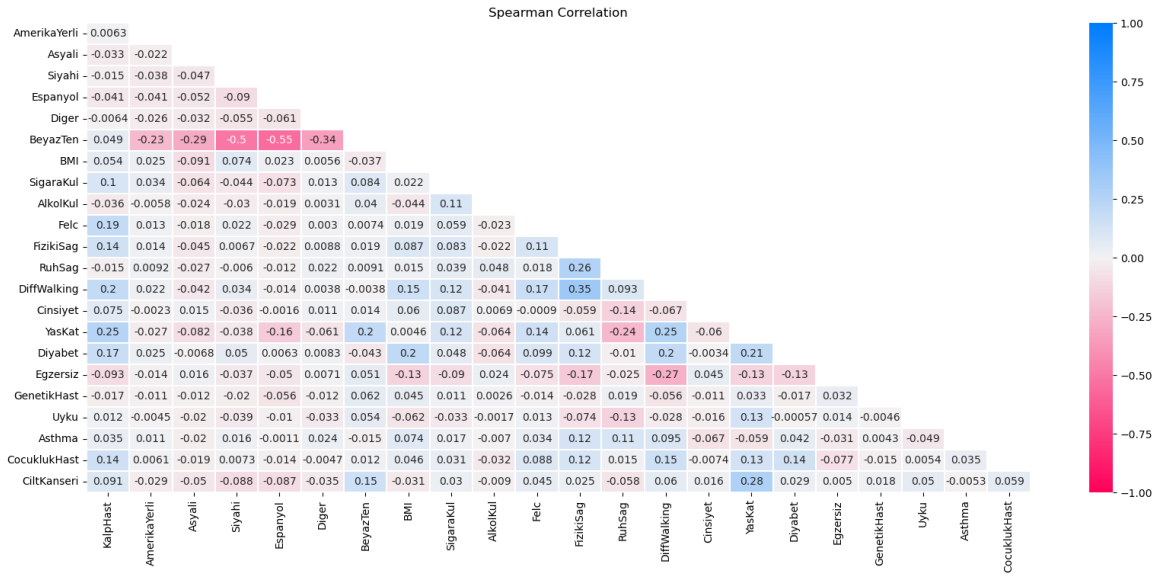
Tablo 1. Veri seti öznitelik bilgileri (Dataset variable information)

Öznitelik Adı	Türü	Aldığı Değerler	Açıklama
HeartDisease	Object	(Yes, No)	Kişide kalp hastalığı olup olmadığını gösterir.
BMI	Float	(12,02<<94,85)	Vücut kitle endeksini (VKİ) gösterir.
Smoking	Object	(Yes, No)	Kişinin sigara kullanıp kullanmadığını gösterir.
AlcoholDrinking	Object	(Yes, No)	Kişinin alkol kullanıp kullanmadığını gösterir.
Stroke	Object	(Yes, No)	Kişinin felç geçirmediğini gösterir
PhysicalHealth	Float	(0<<30)	Kişinin fiziksel sağlığının son 30 günde kaç gün iyi olmadığını gösterir.
MentalHealth	Float	(0<<30)	Kişinin ruhsal sağlığının son 30 günde kaç gün iyi olmadığını gösterir.
DiffWalking	Object	(Yes, No)	Yürürken veya merdiven çıkarken zorluk yaşayıp yaşamadığını gösterir.
Sex	Object	(Male, Female)	Kişinin cinsiyetini gösterir.
AgeCategory	Object	18-24,25-29, …,75-79,80+	13 Yaş kategorisi
Race	Object	(White, Hispanic, Black, Other, Asian, Native)	Kişinin ırkını gösterir. 6 ayrı kategoride ırk bulunur.
Diabetic	Object	No, Yes, No borderline diabetes, Yes (during pregnancy)	Kişinin diyabet hastası olup olmadığını gösterir. 4 ayrı kategoride diyabet verisi bulunur.
PhysicalActivity	Object	(Yes, No)	Kişinin egzersiz yapıp yapmadığını gösterir.
GenHealth	Object	(Yes, No)	Kişinin genetik hastalığı olup olmadığını gösterir.
SleepTime	Float	7<<21	Kişinin günlük uyuduğu süreyi gösterir.
Asthma	Object	(Yes, No)	Kişinin astım hastası olup olmadığını gösterir.
KidneyDisease	Object	(Yes, No)	Kişinin böbrek hastalıkları olup olmadığını gösterir.
SkinCancer	Object	(Yes, No)	Kişinin cilt kanseri olup olmadığını gösterir.

Korelasyon ilişkisi [-1,1] arasında herhangi bir değer alabilir. Korelasyon -1’e yaklaşıyorsa negatif ilişki vardır. Korelasyon +1’e yaklaşıyorsa pozitif ilişki vardır. 0 değeri ise değişkenler arasında ilişki olmadığını gösterir. Pearson korelasyon katsayısı iki değişken arasındaki lineer ilişkiyi incelerken, Spearman korelasyon katsayısı monoton ilişkiye odaklanır. Şekil 2’te Pearson korelasyon analizi ve Şekil 3’te Spearman korelasyon analizi sonuçları gösterilmiştir. Buna göre “Amerikan yerli ırkı”, “Uyku süresi” ve “Genetik hastalıklar” özniteliklerinin çıktımız olan “HeartDisease” ile en düşük ilişkiye sahip olduğu anlaşılmıştır. Sırasıyla “Yaş kategorisi”, “Yürüyüş ve merdiven çıkmada zorluk” ve “Felç geçirme” özniteliklerinin “HeartDisease” ile en yüksek ilişkiye sahip olduğu anlaşılmıştır.

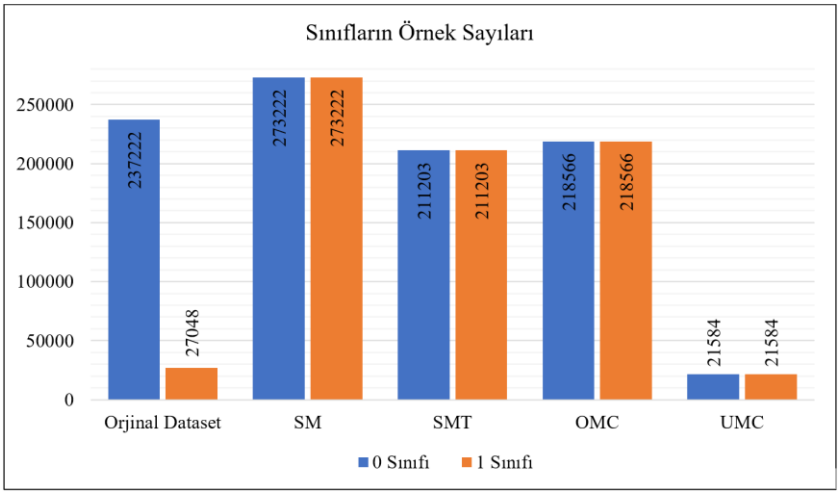


Şekil 1. Özniteliklerin Pearson korelasyon ilişkisi (Pearson correlation relationship of attributes)



Şekil 2.Özniteliklerin Sperman korelasyon ilişkisi (Sperman correlation relationship of attributes)

Korelasyon ilişkisine bakıldıktan sonra veri setinin dengelenme işlemine geçildi. Oversample Minority Class (OMC), Undersample Majority Class (UMC), SMOTE (SM), SMOTETomek (SMT) veri dengeleme yöntemleri ile oluşturulan dört ayrı veri seti veri dengelemenin model performansına etkisini görmek üzere ayrı ayrı kaydedilmiştir. Dengeleme işlemleri sonucunda veri setindeki örnek sayıları dengeleme yöntemine bağlı olarak değişmiştir. Kalp hastalığı sınıfları 0 ve 1 değerlerinin dengelenme yöntemlerine bağlı değişimi Şekil 4'te gösterilmiştir. SM veri dengeleme yöntemi ile oluşturulan veri seti 546444 örneklidir. UMC veri dengeleme yöntemi ile oluşturulan veri seti 43168 örneklidir.



Şekil 3. Veri dengeleme uygulamaları (Data balancing applications)

3.3. Veri Setinin Dengelenmesi (Balancing the Dataset)

3.3.1. SMOTE

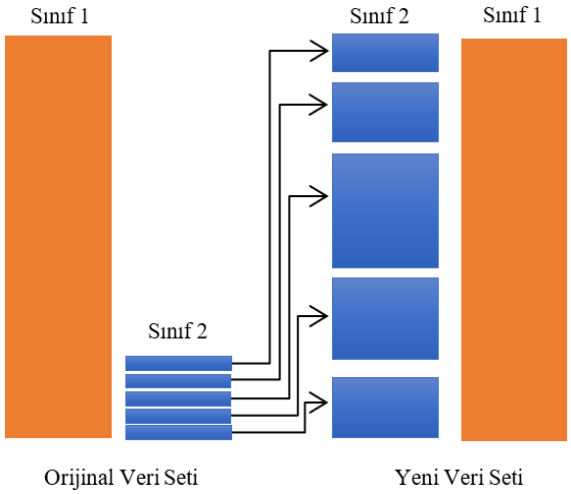
SMOTE yönteminde azınlık sınıfı oluşturulurken her azınlık örneğinden bir numune alınarak KNN sınıfının komşusundan rasgele seçilen parçaları birleştirilir ve bir düzlem boyunca numuneler tanıtılarak örneklem oluşturulur. Gereken örneklem büyüklüğüne bağlı olarak K en yakın komşudan alınacak örneklemler de değişir [22].

3.3.2. SMOTETomek

SMOTETomek tekniğinde ilk adımda azınlık örneklerini çoğaltmak için SMOTE yöntemi kullanılır. İkinci adımda ise çoğaltılmış azınlıklar ile oluşturulan yeni veri setinde Tomek Link eşleri Tomek Link tekniği kullanılarak kaldırılır. Tomek Link yönteminde çoğunluk sınıfı verilerinden azınlık sınıfı verilerine en yakın veriler kaldırılır [23].

3.3.3. Azınlık Sınıfını Yüksek Örnekleme (Oversample Minority Class)

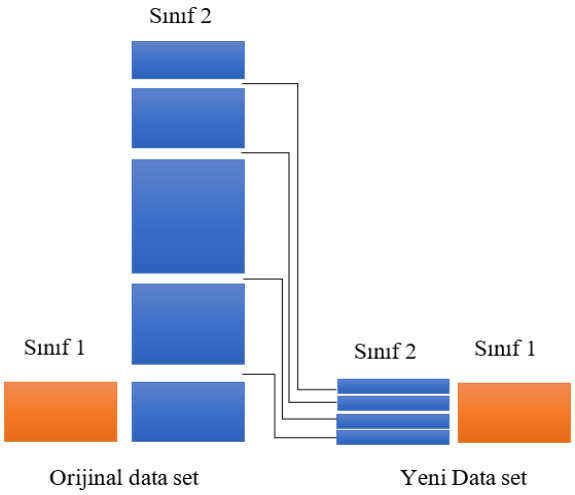
Bu aşırı örnekleme tekniği veri setinin dengesizliğini gidermek için azınlık sınıfı çoğunluk sınıfına eşitlenene kadar azınlık sınıfına yeni örnekler eklenir. Rastgele çoğaltım işleminde Şekil 5'de görüldüğü gibi azınlık sınıfından rastgele örnekler alınarak azınlık sınıfına eklenir. SMOTE gibi sentetik aşırı örnekleme yönteminde ise yapay örnekler oluşturularak azınlık sınıfı çoğaltılır [24].



Şekil 4. Azınlık sınıfını yüksek örnekleme (Oversample minority class)

3.3.5. Çoğunluk Sınıfı Azaltma (Undersample Majority Class)

Çoğunluk sınıfını azınlık sınıfına indirgeme yönteminde çoğunluk sınıfına ait verilerden rasgele veya istatiksel örnekler alınarak azınlık sınıfına Şekil 6'daki gibi eşitlenir. Çoğunluk ve azınlık sınıfı farkı fazla olduğunda veri kaybı fazla olacağı için zayıf örneklem sunar. Ancak bazı durumlarda çoğunluk sınıfı verilerini temizlediği için iyi örneklem sunabilir [24].



Şekil 5. Çoğunluk sınıfı azaltma (Undersample majority class)

3.4. Makine Öğrenme Algoritmaları (Machine Learning Algorithms)

3.4.1. Lojistik regresyon (Logistic regression)

Lojistik Regresyon ikili bağımsız olmayan değişkenlerin olasılıksal tahmininde kullanılan bir yöntemdir. Amaç çıktı değişkenleri ile bağımsız değişkenler arasında anlamlı bir olasılığın varlığını sorgulamaktır. Temel olarak 0,5 değerinin altındaki olasılıklara 0, 0,5 değerinden fazla olasılıklara 1 değeri atanır [20].

3.4.2. Karar Ağaçları (Decision Trees)

Karar ağaçları algoritması akış şeması benzeri hiyerarşik bir düzenle çalışan tahmin algoritmasıdır. DT yönteminde tahmin sınıf etiketi dallanmanın başladığı yerden yani ağaç kökünden beslenir. Değer, öznitelik ile karşılaştırılır ve o değere ait dala kadar takip edilerek sonraki bağlantı noktasına sıçrama yapılır. DT algoritmasında her bağlantı noktası bir sınıf etiketine sahiptir. Düğümlere bağlanan dallar öznitelikleri etkileyen tüm verilerin sınıflandırılması bitene kadar devam eder. Sınıflama, şekillendirme ve analizi diğer yöntemlere göre daha kolay olduğu söylenebilir [10].

3.4.3. Rasgele Orman (Random Forest)

Rasgele orman algoritması küçük boyutlu ve büyük boyutlu problemlerinde başarılı ve hızlı çalışabilen bir tekniktir. Rasgele orman algoritmasının en büyük avantajı regresyon ve sınıflandırma problemlerinde az parametre ile kullanılabilmesidir. Rasgele orman algoritmasında ağaç sayısının attırılmasından optimal ağaç sayısının belirlenmesi önemlidir. Sınıflandırma yaklaşımı açısından rasgele orman algoritması oluşturmak için düğümü durduracak bir karar gelene kadar her düğüm en az iki dala ayrılır. Düğümler bir eşik tarafından veri varyansı minimum olacak şekilde seçilir. Tüm bu dallar ve düğümler içinde ilerleyerek bir yerde tahmin yapılır [25].

3.4.4. K En Yakın Komşu (K Nearest Neighbors)

K En Yakın Komşu algoritması regresyon ve sınıflandırma problemlerinde de kullanılabilmektedir. Seçilen örneklem verilerinin en yakın komşularını etiketlemek suretiyle analiz yapar. Etiketleme yapıldıktan sonra girdilerin etiket değerlerine göre tahminde bulunur. KNN yönteminde girdinin eğitim verilerine uzaklığı çeşitli hesaplama yöntemleri ile yapılır. Ayrıca kullanıcı tarafından komşu verilerin sayısı da seçilebilmektedir. Henüz sınıflandırılmamış veri, veri seti içerisinde kendi konumunun diğer veriler ile mesafesine göre sınıflandırma işleminin yapılması fikrine dayanır. Yani test verisi hangi veri veya veriler yakın ise o veri veya veri gruplarının en uygun sınıfına yerleştirilir. KNN algoritmasında test verileri bulunur ancak eğitim verileri bulunmaz. Eğitim verileri sınıflandırıcı kabul edilerek değerlendirilir. Bu durum tüm yeni verilerin tekrar tekrar taranmasına tarama sürecinin uzamasına neden olur [26].

3.4.5. Naive Bayes

Naive Bayes yönteminde veriler olasılık değerlerine göre sınıflandırılır. Yöntem, verilerin olasılıksal değerlerini tek tek hesaplar ve bu değerin en yüksek olduğu duruma göre sınıflara atamalar yapar. Eğitim verisi çoğaldıkça isabetli sonuçlar alınma oranı da atmaktadır. Bu durum küçük veri setlerinde çalışmasına engel değildir. Dengesiz veri setlerinde rahat çalışabilmektedir [11]. NB sınıflandırmasında verilerin bir sınıfa dahil olduğu olasılığı ile hareket edilir. Bu olasılık hesaplandığında verinin hangi sınıfa dahil olduğu veya olabileceği anlaşılır. NB yöntemi pratik ve kolay anlaşılabilir yöntemdir. Uygulamada verilerin tek bir defa taranması yeterli olacaktır. Bu nedenle hızlı bir yöntem olduğu da söylenebilir [20].

3.4.6. Gradyan Artırıcı Sınıflandırıcı (Gradient Boosting)

Regresyon ve sınıflandırma problemlerinde kullanılabilen Gradyan artırıcı zayıf tahmin kalıplarının birleştirilerek karar ağaçlarının oluşturulduğu bir yöntemdir. Zayıf modeller hesaplanarak genel modelin hatasını azaltır [27]. Gradyan artırıcı genel olarak karar ağaçları gibi tahmin modellerinin geliştirilmesi üzerine kuruludur. Karar ağaçları düğümlerindeki tahminlere göre çıktılar veren düğüm ve dallardan oluşur. Daha stabil sonuçlar veren düğümleri dallandırarak sonuca ulaşır [28]. Karar düğümlerine dayalı olarak tahmine dayalı sonuçlar veren düğümlerden ve yapraklardan oluşur. Regresyon ağaçları, bireysel olarak zayıf modellerdir, ancak bir bütün olarak bakıldığında doğrulukları çok gelişmiştir. Bu nedenle, topluluklar kademeli olarak artan bir şekilde inşa edilir, öyle ki her topluluk önceki topluluktaki hatayı matematiksel olarak düzeltir.

3.4.7. XGBoost

XGBoost neredeyse tüm yapay zekâ problemlerinde kullanılabilen güçlü bir algoritmadır. Teknik aynı koşullarda hali hazırda kullanılan tüm algoritmalardan daha hızlı çalışır ve çok fazla örneği ölçekleyebilir. XGBoost algoritmasının birçok avantajı vardır. Sık olmayan verileri analiz etmek için yeni bir algoritma mantığı geliştirmiştir. Paralel ve birbirinden ayrıksı bilgi sentezi yolu ile daha hızlı tahminlere yol açar. En önemli avantajlarından sayılan harici çekirdek hesaplamalarını kullanmasıdır. Bu sayede bir tek algoritma ile milyonlarca örneğin işlenmesinin yolu açılmış olmaktadır [29].

3.4.8. Torbalama (Bagging)

Bagging algoritması regresyon ve sınıflandırma problemlerinde kullanılabilir bir tekniktir. Aslen kullanılan yöntemlerin kararlılığını ve tahminlerin doğruluğunu artırmak için geliştirilmiştir. Sonuca ulaşmak için rastgele oluşturulan set sınıflarını birleştirir. Bagging algoritması basit mantığı ve yüksek doğru tahminleri ile öne çıkmaktadır. Bu yöntemin temel ilkesi zayıf dalların bir araya gelerek güçlü ağaçlar oluşturmasıdır. Eğitim setindeki tüm ağaçların her birine tekrar tekrar gidilir ve her bir ağaç bir sınıfa yönelir. En yüksek yönelim alan sınıf sonuç olarak değerlendirilir [30].

3.5. Hiper Parametre Optimizasyonu (Hyperparameter Optimization)

Makine öğrenmesinde parametreler ve hiper parametreler ayrı kavramlardır ve birbiri ile karıştırmamak gerekir. Parametreler algoritma işleyişi içinde öğrenilir, tahmin edilir ve bu değerler makine öğrenmesi eğitimi devam ederken güncellenir. Algoritma eğitimi bitiğinde ise parametreler modelin bir parçasını oluşturur. Tahmin modeline ait parametreleri bulmak için ise hiper parametre değerleri kullanılır. Hiper parametreler algoritmaya özgü verilerdir, bu nedenle bu değerleri verilerden yola çıkarak hesaplanamaz. Hiper parametreler algoritma eğitimi içinde tahmin edilemeyen veya öğrenilemeyen parametrelerdir. Farklı veri setlerinde farklı hiper parametreler elde edilebilir. Hiper parametre ayarları makine öğrenimini kontrollü bir şekilde yapmanın gerekli bir bölümüdür. Doğru biçimde ayarlanmayan parametreler modelimizin yetersiz sonuçlar üretmesine neden olacaktır.

3.5.1. Izgara Arama (Grid Search)

Izgara arama belirlenen alt ve üst sınırlarda ve belirli steplerde kapsamlı parametre belirleyicidir. Belirlenen adımlarda tüm kombinasyonlar değerlendirilir ve değerlendirme sonucuna göre en verimli parametreler elde edilmeye çalışılır. Izgara arama uygulanması kolay bir yöntemdir ancak kapsamlı tarama yaptığı için büyük veri setlerinde maliyet artar.

3.5.2. Yarım Izgara Arama (Halving Grid Search)

Yarım ızgara arama çalışma mantığı Izgara Arama tekniğinden esinlenir ancak Izgara Arama'ya göre çok daha verimli çalışan bir parametre belirleme yöntemidir. İyi performans gösteremeyen veriler filtrelenerek daha iyi adaylar üzerinde çalışan ve değerlendirme yapan tekniktir. Yarım ızgara arama özetle minimal seviyede kaynakla aramaya başlayıp en yüksek düzeyde yineleme yaparak en doğru verileri seçimini yapar.

3.5.3. Rasgele Arama (Random Search)

Rasgele arama, oluşturulan kılavuzda rasgele değerlerde arama yaparak performansı en çok artıran değerleri belirlemeye çalışır. Her tekrarda hiper parametrelerin bir kombinasyonunu rasgele belirler ve modelin verimliliğini tutar. Yapılan tekrarlar neticesinde en iyi sonucu veren parametre grubunu belirler.

3.5.4. Bayes Optimizasyonu (Bayesian Optimization)

Bayes optimizasyon tekniği rasgele ve grid arama tekniklerinin aksine belirli bir standartla modeli sonuçlandırmak yerine performansı en çok artıran değerlerin bulunduğu alana yoğunlaşarak doğruluğu yüksek olasılıklı modeller oluşturmayı hedefler. Her veri

performansı bir sonraki veri girdisinin düzeyini güncelleyerek devam eder. Bu sayede daha hızlı ve verimli çalışmaktadır.

3.6. Skor Açıklamaları (Score Descriptions)

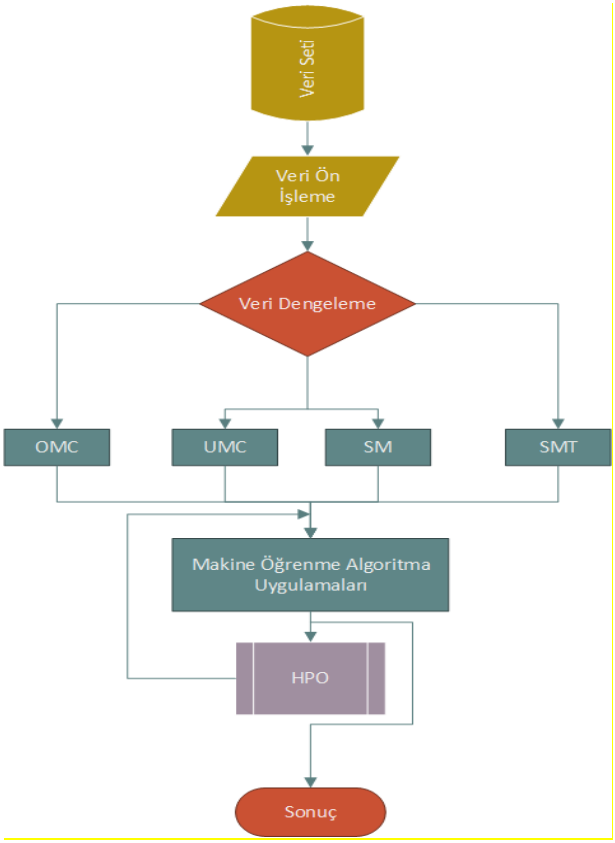
Tablo 2'te algoritma değerlendirmelerinde kullanılan metrikler hakkında bilgi verilmiştir.

Tablo 2. Değerlendirme metrikleri (Evaluation metrics)

Skor	Açıklama
Accuracy	Doğruluk skoru, doğru sınıflandırılmış verinin toplam veriye oranını temsil eder.
Precision	Kesinlik skoru, pozitif tahminlerin ne kadarının gerçek pozitif olduğunu gösterir. Doğru pozitif tahminlerin, doğru pozitif ve yanlış pozitif tahmin toplamına oranıdır.
Recall	Geri çağırma skoru, tüm pozitif veriler içinden ne kadar doğru pozitif tahmin edildiğini gösterir. Doğru pozitif tahminlerin doğru pozitif ve yanlış negatif tahmin toplamına oranıdır.
F1 Score	F1 skoru kesinlik ve geri çağırma skorlarını birleştiren iyi bir ölçüttür. Genel olarak doğruluk skorundan daha güvenilir bir sonuçtur. F1 skoru Precision ve Recall skorlarının harmonik ortalamasıdır.
Cross Validation (CV)	Çapraz doğrulama skoru uygulanan diğer skorları test etmenin, doğruluğunun hesaplamanın bir yoludur. Çapraz doğrulamada test ve eğitim setleri yeniden örneklenerek farklı veri seti oluşturulur ve bu veri setlerinde test edilen skor tekrar uygulanır. Bu çalışmada çapraz doğrulama skoru F1 skorunu test etmek için kullanıldı.

4. BULGULAR ve TARTIŞMA (FINDINGS and DISCUSSION)

Çalışmanın yapıldığı bilgisayar Intel Core i7-9750H 2,60GHz CPU, 16GB Ram, NVIDIA GeForce GTX 1650 (4GB), 512 GB SSD donanım özelliklerine sahiptir. Bilgisayarda Windows 10 Pro 64 Bit işletim sistemi bulunmaktadır. Şekil 7'de verilen şema tüm çalışmanın özeti niteliğindedir.



Şekil 6. Tüm çalışmanın akış şeması (Flow chart of the whole study)

Çalışmada makine öğrenmesi algoritmalarını çalıştırmak için Python dili tercih edilmiştir. Algoritmalar Anaconda3 Navigatör Jupyter derleyici ile derlenmiştir. Çalışmada veri kümeleri ile çalışmak için Pandas kütüphanesi, diziler ile çalışabilmek için Numpy kütüphanesi, etkileşimli görseller oluşturmak için Matplotlib kütüphanesi ve istatistiksel grafikler için Seaborn kütüphanesi, istatiksel modelleme ve makina öğreniminde kullanılan birçok aracı barındıran Sklearn kütüphanesi, veri eşitleme yöntemleri barındıran Imblearn kütüphanesi kullanılmıştır.

Çalışmada bazı kısaltmalar kullanılmıştır. Bu kısaltmalar Tablo 3'te verilmiştir.

Tablo 3. Kısaltmalar (Abbreviations)

Kısaltma	Tanımı	Grubu
OMC	Oversample Minority Class	Dengeleme algoritmaları
SM	SMOTE	
SMT	SMOTETomek	
UMC	Undersample Majority Class	
BAG	Bagging	Yapay zekâ algoritmaları
DT	Decision Trees	
GB	Gradient Boosting	
KNN	K Nearest Neighbors	
LR	Logistic Regression	
NB	Naive Bayes	
SVM	Destek Vektör Makinası	
RF	Random Forest	
XGB	XGBoost	
HPO	Hiper parametre Optimizasyonu	Hiper parametre optimizasyonu algoritmaları
RS	Random Search	
BO	Bayesian Optimization	

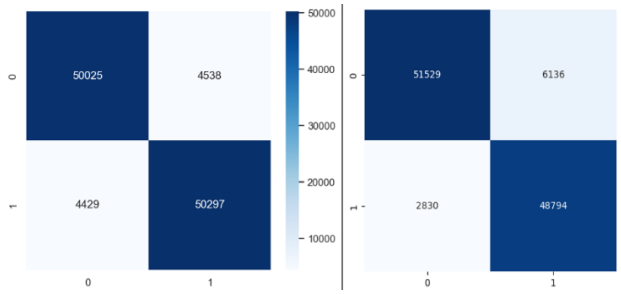
Makine öğrenme algoritmaları veri dengeleme teknikleri ile dengelenerek oluşturulan veri setleri ile eğitilmiş ve sonuçlar bu bölümde verilmiştir. HPO uygulamaları bu bölüm başlığı altında incelenmiştir.

4.1. SM ile Dengelenmiş Veri Seti Algoritma Sonuçları
(SM Balanced Dataset Algorithm Results)

Tablo 4'te SM ile dengelenmiş veri setine uygulanan tahmin algoritmalarına ait sonuçlar gösterilmiştir. F1 skorlarında en iyi sonuçların %92 değeri ile RF ve XGB algoritmalarının olduğu görüldü. RF algoritmasında tüm skorlar yakın değerler alırken XGB Recall skoru diğer skorlardan daha düşük olduğu görülmektedir. Bu durum karışıklık matrisinde daha açık görülmektedir. Şekil 8'de görüleceği üzere XGB algoritmasında tüm pozitif veriler içinde doğru tahmin edilen değerler düşük, yanlış tahmin edilen negatifler daha yüksek çıkmıştır.

Tablo 4. SM uygulanan veri setine ait algoritma skorları
(Algorithm scores of the SM applied data set)

Algoritma	Accuracy	Precision	Recall	F1-Score	CV Mean
LR	%75	%74	%78	%76	%75
DT	%89	%89	%89	%89	%75
RF	***%92***	***%92***	***%92***	***%92***	***%92***
KNN	%86	%81	%94	%87	%85
NB	%71	%74	%66	%70	%71
GB	%87	%88	%87	%87	%87
XGB	***%92***	***%95***	***%89***	***%92***	***%92***
BAG(DT)	%91	%93	%89	%91	%91



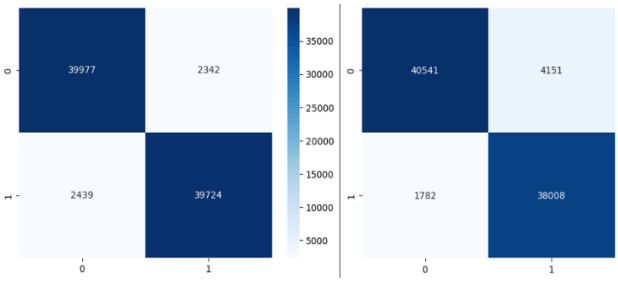
Şekil 7. SM veri seti RF Karışıklık Matrisi (solda) (SM dataset RF Confusion Matrix (on the left)), SM veri seti XGB Karışıklık matrisi (sağda) (SM dataset XGB Confusion Matrix (on the right))

4.2. SMT ile Dengelenmiş Veri Seti Algoritma Sonuçları
(SMT Balanced Dataset Algorithm Results)

Tablo 5'te verilen skor değerlerine göre F1 skorlarında %94 değeri ile RF ve %92 değeri ile XGB'nin en iyi tahmin başarısı elde ettiği söylenebilir. SMT veri dengeleme yöntemi ile oluşturulan veri setinde RF algoritmasında tüm skorlar yakın değerler alırken XGB Recall skoru diğer skorlardan daha düşük olduğu görülmektedir. Şekil 9'da RF algoritmasının yanlış tahmin edilen 0 ve 1 sınıflarının sayılarının yakın olduğu görülmektedir. RF algoritmasının daha dengeli tahmin yaptığı çıkarımında bulunabiliriz. Şekil 9'da XGB'nin 0 sınıfı olarak tahmin edilen ancak 1 sınıfına ait olan 4151 değerinin yanlış tahmin edildiğini görülüyor. Buna karşılık 1 sınıfı olarak tahmin edilen 0 sınıfına ait değerlerin 1782 olduğu görülüyor. Bu durumda SMT dengeleme sisteminin çoğaltma işlemi sonucunda çoğalttığı 1 sınıfını 0 sınıfına çok yaklaştırdığı ve yanlış tahminlere neden olduğu söylenebilir.

Tablo 5. SMT uygulanan veri setine ait algoritma skorları
(Algorithm scores of the SMT applied data set)

Algoritma	Accuracy		Precision	Recall	F1-Score	CV mean
LR	%76		%74	%79	%77	%76
DT	%91		%91	%91	%91	%90
RF	***%94***		***%94***	***%94***	***%94***	***%94***
KNN	%87		%82	%95	%88	%86
NB	%72		%74	%68	%71	%72
GB	%88		%89	%87	%88	%88
XGB	***%93***		***%96***	***%90***	***%93***	***%93***
BAG(DT)	%93		%95	%91	%93	%92



Şekil 8. SMT veri seti RF Karışıklık Matrisi (solda) (SMT dataset RF Confusion Matrix (on the left)), SMT veri seti XGB Karışıklık matrisi (sağda) (SMT dataset XGB Confusion Matrix (on the right))

SMOTE tabanlı dengelenmiş (SM ve SMT) veri setlerinde XGB algoritmasının dengelenmeden önce çoğunluk sınıfın olan negatif (0) sınıfını daha iyi sınıflandırdığı, dengelenmeden önce azınlık sınıfı olup yapay olarak çoğaltılmış pozitif (1) sınıfını (0) sınıfına göre daha kötü sınıflandırdığı görülmüştür. RF algoritmasının ise azınlık veya çoğunluk sınıfı farkı olmaksızın negatif ve pozitif sınıflarını eşit doğrulukta sınıflandırdığı dikkat çekmektedir. XGB algoritmasında kalp hastası olup hasta değil olarak sınıflandırılan kişi sayısı daha fazladır. Kalp hastası olduğu halde kalp hastası olmadığı yönünde yapılan sınıflandırmanın insan hayatı açısından daha riskli bir durum oluşturduğu için SM ve SMT dengelenmiş veri setlerinde RF algoritması XGB algoritmasına daha göre öncelikli tercih sebebidir.

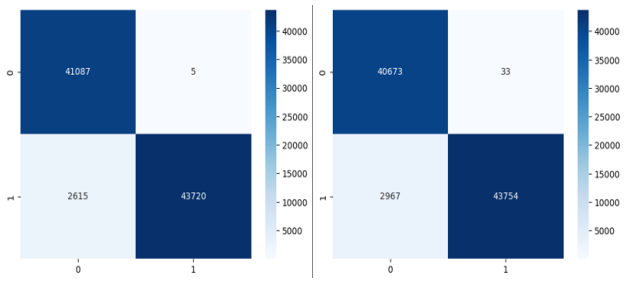
4.3. OMC ile Dengelenmiş Veri Seti Algoritma Sonuçları
(OMC Balanced Dataset Algorithm Results)

Tablo 6'da F1 skorlarında en iyi tahmin olasılıklarını %97 değeri ile RF ve %97 değeri ile BAG algoritmalarının verdiği görülmüştür. BAG algoritması tahmin edici(base_estimator) değeri varsayılan DT algoritması olarak ayarlanmıştır. Recall skorlarının %100'e çok yakın olması algoritmanın aşırı öğrenme yapmış olabileceğini akla getirmektedir. Çünkü OMC dengeleme sisteminde azınlık sınıfı olan '1' sınıfının rasgele kopyalanarak çoğaltılması söz konusudur. Veri setinin dengelenmeden önceki halinde '0' sınıfı 273.222 örnek içerirken '1' sınıfı 27.048 örnek içeriyor. OMC uygulandıktan sonra her iki

sınıf 218.566 örnekte dengelenmiştir. Bu durumda azınlık sınıfı yaklaşık 8 katı kadar çoğaltılmıştır. Dengesizlik oranı yüksek olduğu için veri setinin test kısmı hangi oranda seçilirse seçilsin eğitim seti içerisinde test seti verileri olacaktır. Bu nedenle algoritmaların öğrenme yapmak yerine aşırı öğrenme yapmış olabileceğini söyleyebiliriz. Şekil 10'da karışıklık matrisi verileri bu savı destekler niteliktedir. Hem RF hem de BAG algoritması için SMOTE tabanlı dengeleme sistemlerinin aksine OMC dengeleme sisteminde rasgele çoğaltılan "1" sınıfının hatalı tahminlerinin az, aksine "0" sınıfının yanlış tahminlerinin daha fazla olduğu Şekil 10'da görülmektedir.

Tablo 6. OMC uygulanan veri setine ait algoritma skorları
(Algorithm scores of the OMC applied data set)

Algoritma	Accuracy	Precision	Recall	F1-Score	CV mean
LR	%75	%74	%76	%75	%75
DT	%95	%91	%100	%95	%94
RF	***%97***	***%94***	***%100***	***%97***	***%96***
KNN	%89	%82	%99	%90	%87
NB	%70	%74	%64	%69	%70
GB	%76	%74	%81	%77	%76
XGB	%79	%76	%83	%80	%78
BAG(DT)	***%97***	***%94***	***%100***	***%97***	***%96***



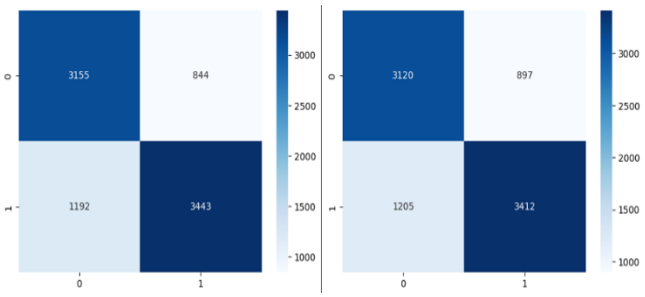
Şekil 9. OMC veri seti RF Karışıklık Matrisi (solda) (OMC dataset RF Confusion Matrix (on the left)), OMC veri seti BAG Karışıklık matrisi (sağda) (OMC dataset BAG Confusion Matrix (on the right))

4.4. UMC ile Dengelenmiş Veri Seti Algoritma Sonuçları
(UMC Balanced Dataset Algorithm Results)

UMC ile dengelenmiş veri setinde dikkat çeken ilk ayrıntı Tablo 7'de görüldüğü gibi diğer veri dengeleme teknikleri ile oluşturulan veri setlerine göre algoritma uygulamalarında düşük skorlar elde edilmiş olmasıdır. UMC veri setinde yaşanan veri kaybının düşük skorlara neden olduğu düşünülmektedir. Zira UMC uygulanmadan önce 300.270 satır örnek barındıran orijinal veri seti UMC uygulandıktan sonra 43.168 örneğe düşmüştür. Bu durum kalp hastalığı tahmininde önemli ölçüde performans kaybına neden olmuştur. Tablo 7'de görüldüğü gibi F1 skorlarında en iyi tahmin başarısı gösteren algoritmalar %77 ile GB ve %76 ile XGB olmuştur. Şekil 11'de UMC setine ait en iyi algoritmanın karışıklık matrisi verilmiştir. Kalp hastası olduğu halde kalp hastası değil olarak etiketlenen Yanlış negatif değerlerinin toplam veri içindeki oranı yüksek olduğu görülmektedir.

Tablo 7. UMC uygulanan veri setine ait algoritma skorları
(Algorithm scores of the UMC applied data set)

Algoritma	Accuracy	Precision	Recall	F1-Score	CV mean
LR	%75	%75	%76	%76	%75
DT	%68	%68	%67	%67	%67
RF	%75	%74	%77	%76	%75
KNN	%72	%72	%73	%72	%72
NB	%70	%73	%63	%68	%71
GB	***%76***	***%74***	***%80***	***%77***	***%76***
XGB	***%76***	***%74***	***%79***	***%76***	***%75***
BAG(DT)	%73	%74	%70	%72	%72



Şekil 10. UMC veri seti GB Karışıklık Matrisi (solda) (UMC dataset GB Confusion Matrix (on the left)), UMC veri seti XGB Karışıklık matrisi (sağda) (UMC dataset XGB Confusion Matrix (on the right))

Tüm veri dengeleme teknikleri (SM, SMT, OMC, UMC) ile oluşturulan veri setleri ile eğitilen makine öğrenme algoritmaları değerlendirildikten sonra her dengelenmiş veri seti için en yüksek skoru veren iki ayrı algoritma seçilmiştir. Bu algoritmalara Hiper parametre optimizasyonu uygulanmıştır. Buna göre SM dengeleme için RF ve XGB, SMT dengeleme için RF ve XGB, OMC dengeleme için RF ve BAG (DT tabanlı), UMC dengeleme için XGB ve GB algoritmaları seçilmiştir.

Hiper parametre optimizasyonu işlemi için iki ayrı optimizasyon tekniği seçilmiştir. Bu tekniklerden diğer arama tekniklerine göre daha hızlı arama yaparak maliyeti azalttığı için Random Search (RS) ve öğrenme tabanlı ve verimli çalıştığı için Bayesian Optimization (BO) tercih edilmiştir. Algoritmaların hiper parametre yapılandırma değerleri Tablo 8'de gösterildiği gibi belirlenerek her iki arama tekniği için uygulama yapılmıştır. Tablo 9'da Hiper parametre optimizasyonu uygulama sonucunda elde edilen optimum hiper parametreler ve bu hiper parametreler ile makine öğrenme algoritmalarının tekrar uygulama sonuçları gösterilmiştir. Buna göre SM ve SMT veri setlerinde RF ve XGB algoritmaları için BO'un daha iyi sonuçları verdiği için bu teknik tercih edilmiştir. Ayrıca, OMC veri setinde RF algoritması için BO, BAG algoritması için RS tercih edilmiş ve UMC veri setinde BO tercih edilmiştir. Bu işlem sonucunda, BO tekniğinin RS tekniğine göre daha iyi sonuçlar verdiği görülmüştür.

Tablo 8. Hiper parametre yapılandırma değerleri
(Recommended hyperparameter values)

	Hiper parametreler	Önerilen hiper parametre değerleri
RF	max_depth	[10, 20, 30, 40, 50]
	min_samples_leaf	[1, 2, 4, 6]
	min_samples_split	[2, 5, 10, 15]
	n_estimators	[200, 400, 600, 800]
XGB	colsample_bytree	[0,3,0,4,0,5,0,7],

	learning_rate	[0,05,0,10,0,15,0,20,0,25,0,30],
	max_depth	[3,4,5,6,8,10,12,15],
	min_child_weight	[1,3,5,7],
	gamma	[0,0,0,1,0,2,0,3,0,4]
BAG	bootstrap	[True, False]
	bootstrap_features	[True, False]
	n_estimators	[5, 10, 15],

	max_samples	[0,6, 0,8, 1,0]
GB	n_estimators	[5,50,150,250,500],
	max_depth	[1,3,5,7,9],
	learning_rate	[0,0001, 0,001, 0,01, 0,1, 1]

Tablo 9. Optimize edilmiş parametrelerle uygulama sonuçları (Application results with optimized parameters)

Dengeleme Sistemi	Algoritma	*HPO Tekniği*	En iyi Parametreler	Skor
SM	RF	*BO*	max_depth=30, min_samples_leaf=1, min_samples_split=2, n_estimators=800	*Train Score: %98* *Test Score: %92*
		RS	max_depth=40, min_samples_leaf=1 min_samples_split=10, n_estimators=600	Train Score: %97 Test Score: %92
	XGB	*BO*	colsample_bytree=0,7, gamma=0, max_depth=15, learning_rate=0,3, min_child_weight= 1	*Train Score: %97* *Test Score: %94*
		RS	colsample_bytree=0,3, gamma=0,1, max_depth=15, learning_rate=0,15, min_child_weight=5	Train Score: %93 Test Score: %92
SMT	RF	*BO*	max_depth=40, min_samples_leaf=1, _samples_split=2, n_estimators= 800	*Train Score: %100* *Test Score: %94*
		RS	max_depth= 40, min_samples_leaf=1 min_samples_split=2, n_estimators=800	Train Score: %100 Test Score: %94
	XGB	*BO*	colsample_bytree=0,7, gamma=0,2, max_depth=15, learning_rate=0,3, min_child_weight=1	*Train Score: %98* *Test Score: %94*
		RS	min_child_weight=1, max_depth=15, gamma= 0,4, learning_rate=0,1, colsample_bytree=0,4	Train Score: %95 Test Score: %93
OMC	RF	*BO*	max_depth=50, min_samples_leaf=1, min_samples_split=2, n_estimators= 600	*Train Score: %100* *Test Score: %97*
		RS	max_depth=50, min_samples_leaf=2 min_samples_split=5, n_estimators=600	Train Score: %98 Test Score: %94
	BAG	*BO*	bootstrap=False, bootstrap_features=True, _samples=1,0, n_estimators= 15	Train Score: %98 Test Score: %96
		RS	bootstrap=False , bootstrap_features=True, _samples=1,0, n_estimators=10	*Train Score: %99* *Test Score: %98*
UMC	XGB	*BO*	colsample_bytree=0,4, gamma=0,1, max_depth=4, learning_rate=0,15, min_child_weight=1	*Train Score: %77* *Test Score: %76*
		RS	min_child_weight=5, max_depth=5, gamma= 0,3, learning_rate=0,1, colsample_bytree= 0,3	Train Score: %77 Test Score: %76
	GB	*BO*	learning_rate=0,1, max_depth=3, n_estimators=150	*Train Score: %77* *Test Score: %76*
		RS	learning_rate=1, max_depth=1, n_estimators=500	Train Score: %76 Test Score: %76

Hiper parametre optimizasyon teknikleri uygulanıp en iyi hiper parametreler bulunduktan sonra tüm veri setlerine makine öğrenme algoritmaları en iyi hiper parametre ile tekrar uygulandı ve sonuçlar Tablo 10'da gösterildi. Çıkan sonuçlar karşılaştırılarak bu çalışma için en iyi algoritma ve tekniklere karar verildi. SM ile dengelenmiş veri setinde HPO'dan sonra modellerin performansının arttığı görülmektedir. SMT veri dengeleme tekniği ile oluşturulan veri setinde ise RF algoritmasının performansı değişmezken XGB algoritmasının F1 skorda performansı artmıştır. OMC veri dengeleme tekniği ile oluşturulan veri setinde model skorlarının HPO ile değişmediği görülmüştür. UMC veri dengeleme tekniği ile oluşturulan veri setinde yine HPO'nun anlamlı bir fark yaratmadığı görülmüştür. Bu değerler skorların tam kısımları olmakla birlikte ondalık kısmında küçük farkların oluştuğu gözlenmiştir. Genel olarak HPO ile model performanslarının arttığı sonucuna ulaşılmıştır.

Tablo 10. En iyi hiper parametreler ile tekrar uygulama sonuçları (Reapplication results with the best hyperparameters)

Dengeleme Sistemi	Skor	RF		XGB		Dengeleme Sistemi	Skor	RF		XGB	
		A	*B*	*A*	*B*			*A*	*B*	*A*	*B*
SM	*Train (%)*	100	98	92	97	SMT	Train (%)	100	100	93	98
	Test (%)	92	93	92	94		Test (%)	94	94	93	94
	CV (%)	92	92	92	93		CV (%)	94	94	93	94
	Accuracy (%)	92	93	92	94		Accuracy (%)	94	94	93	94
	F1 (%)	92	*93*	92	*94*		F1 (%)	94	***94***	93	*94*

Dengeleme Sistemi	Skor	RF		BAG		Dengeleme Sistemi	Skor	XGB		GB	
		A	*B*	*A*	*B*			*A*	*B*	*A*	*B*

OMC	*Train (%)*	100	100	100	99	UMC	*Train (%)*	81	77	77	77
	Test (%)	97	97	97	97		*Test (%)*	76	76	76	76
	CV (%)	96	96	96	97		*CV (%)*	75	76	76	76
	Accuracy (%)	97	97	97	97		*Accuracy (%)*	76	76	76	76
	F1 (%)	97	*97*	97	*97*		*F1 (%)*	77	*77*	76	76

A: HPO'den önceki skor (Score before HPO) *B: HPO'den sonraki skor (Score after HPO)*

SM veri dengeleme yöntemi ile oluşturulan veri setinde XGB %94, RF %93 tahmin, SMT veri dengeleme yöntemi ile oluşturulan veri setinde XGB ve RF %94 tahmin, OMC veri dengeleme yöntemi ile oluşturulan veri setinde RF ve BAG %97 tahmin ve UMC veri dengeleme yöntemi ile oluşturulan veri setinde XGB %77 tahmin başarısı elde edilmiştir. Buradan UMC ile dengelenmiş veri setinde yaşanan kayıplardan dolayı performans düşüklüğü yaşanmış ve bu veri seti elenmiştir. OMC veri setinde yaşanan aşırı öğrenme nedeniyle bu veri seti elenmiştir. SM ve SMT veri setlerinde XGB algoritmasının Recall skor sonuçlarının tatmin edici olmaması nedeniyle XGB algoritması elenmiştir. Bu çalışmada RF makine öğrenme algoritmasının en güvenilir skoru verdiği sonucuna ulaşılmıştır. Sonuca ulaşmak için yapılan işlem Şekil 12'de gösterilmiştir.

Şekil 11. Bu çalışmada en verimli tahmin algoritması için iş akışı (This study workflow for the best prediction algorithm)

4.5. Benzer Çalışma Sonuçları ile Karşılaştırma (Comparison with Other Study Results)

Bu bölümde kalp hastalıklarının tahmini ile ilgili benzer çalışmalara ilişki karşılaştırma yapılmıştır. Diğer araştırmacıların en iyi sonucu veren algoritmaları ile bu çalışmadaki aynı algoritmanın en iyi skoru karşılaştırılarak Tablo 11'de gösterilmiştir. Bu çalışma ile kalp hastalıklarının tahmin edilmesinde verimli sonuçlar veren modeller elde edilmiştir. Bu çalışmanın diğer çalışmalara göre üstünlükleri de bulunmaktadır. Model skorları karşılaştırıldığında bu çalışmanın benzer çalışmalara göre üstün olduğu göze çarpmaktadır. Tablo 11'deki veriler de bu bilgiyi desteklemektedir. Model skorları ile birlikte bu çalışmada model performanslarını etkileyen veri dengeleme teknikleri ve hiper parametre optimizasyonu teknikleri incelenmiştir. Model performanslarını etkileyen yöntemleri karşılaştırmalı olarak inceleyerek benzer çalışmalardan farklı özgün bir kaynak ortaya koymuş olduk.

Tablo 11. Benzer çalışmaları ile karşılaştırma tablosu
(Table of comparison with the work of others)

Çalışmalar	RF-OMC	NB	DT	LR
Bu Çalışmadaki Sonuçlar	*%97*	*%71*	*%95*	*%77*
Anitha ve ark. [7]	-	%86,6	-	-
Coşar ve ark.[31]	%88	-	-	-
Rajdhan ve ark. [10]	%90,2	-	-	-
Mertcan Görgün [11]	%90,2	-	-	-
Repaka ve ark. [12]	%89,8	-	-	-
Ramalingam ve ark. [14]	%83,5	-	-	-
Kamil [16]	%98	-	-	-
Salman [32]	-	%86	-	-
Konda ve ark. [18]	-	-	%89	-
Tarawneh ve ark [19]	-	%86	-	-
Venkatesh ve ark. [2]	-	%97	-	-
Çil [20]	-	-	-	%90,8
Saqlain ve ark. [33]	-	%86,7	-	-
Taşçı ve ark. [34]	-	%88	-	-
Ekrem ve ark. [32]	%87	-	-	-
Gündoğdu [35]	%89.7			

5. SINIRLILIKLAR (LIMITATIONS)

Veri seti Amerika Birleşik Devletinde yapıldığı için hem yaşam biçimi hem de bölgesel farklılık gösterebilecek bazı özniteliklerin çalışmayı kısıtladığı söylenebilir. "Irk" özniteliği çok fazla çeşitlendirilmediği için Amerika kıtası dışında kullanımı bu özniteliğin çalışmayı kısıtladığı söylenebilir. Veri setinin büyük boyutlu olması bazı algoritmaların uygulanmasını zorlaştırmaktadır. Destek Vektör Makinası veri setine uygulanmış ancak günlerce süren işlemlerden sonuç alınamadığı için bu çalışmadan çıkarılmıştır.

6. SONUÇLAR (CONCLUSION)

Bu çalışma, farklı makine öğrenmesi algoritmalarının verimini artırmak üzere uygulanan veri düzenleme ve hiper parametre optimizasyonu tekniklerinin karşılaştırılarak kalp hastalığı tahmininde en verimli modelin ortaya konduğu bir çerçeve sunmaktadır. Çalışmadaki asıl amaç mevcut veri seti üzerinden kalp hastalığı tahmini yapan en verimli algoritmanın belirlenmesidir. Bununla birlikte veri dengeleme ve hiper parametre optimizasyonunun kalp hastalığında kullanılan makine öğrenme algoritmaları üzerindeki etkisi ikinci amaç olarak belirlenmiştir. Bunun için mevcut veri seti temizlenerek, dört farklı veri dengeleme tekniği ile (SM, SMT, OMC, UMC) oluşturulmuş her veri setine ayrı ayrı tahmin algoritmaları uygulanmıştır. Hiper parametre optimizasyonu ile en iyi hiper parametreler bulunarak bu parametreler ile tahmin algoritmaları tekrar eğitilmiştir. XGB algoritması SM ve SMT dengelenmiş veri setlerinde %94 tahmin başarısı elde etmiştir. SM ve SMT veri dengeleme yöntemleri ile oluşturulan veri setlerinde XGB algoritması ile yapılan tahminlerde kalp hastası olduğu halde kalp hastası değil olarak sınıflandırılan örneklerin daha fazla olduğu ve bu durumun insan hayatı açısından riskli bir durum olduğu değerlendirilmiştir. Bu nedenle XGB algoritması bu çalışma için en iyi algoritma olmadığı yönünde

değerlendirme yapılmıştır. OMC algoritması ile dengelenen veri setinde RF ve XGB algoritmaları çoğaltılan azınlık sınıfını orantısız bir şekilde, yüksek doğrulukta tahmin ettiği görülmüştür. Veri setinin orijinal hali çok dengesiz olduğu ve OMC dengeleme tekniğinin rasgele kopyalama yolu dengeleme yaptığı için bu veri seti ile eğitilen tahmin algoritmalarının aşırı öğrenme yaptığı çıkarımında bulunulmuştur. UMC ile dengelenmiş veri setinde veri kaybı çok fazla olduğu için algoritma skorları yüksek çıkmamıştır. UMC dengeleme setinin büyük ve dengesiz veri setlerinde başarılı olamadığı sonucuna ulaşılmıştır. RF algoritması SM dengelenmiş veri setinde %93, SMT dengelenmiş veri setinde %94 ve OMC dengelenmiş veri setinde %97 başarı elde etmiştir. RF algoritmasının her ne kadar OMC ile dengelenmiş veri setinde aşırı öğrenmeye yöneldiği çıkarımı yapılsa da RF algoritmasının tüm veri setlerinde diğer tahmin algoritmalarına oranla yüksek tahmin başarısı gösterdiği sonucuna ulaşılmıştır. SM ve SMT veri dengeleme teknikleri arasında çok fazla fark olmamakla birlikte SMT tekniği ile oluşturulan veri setleri ile eğitilen modellerin daha verimli sonuçlar çıkardığı görülmüştür. RF ve XGB algoritmalarının LR, DT, KNN, NB, GB ve BAG algoritmalarından daha iyi performans gösterdiği görülmüştür. Bu çalışma farklı veri setleri ve HPO sonuçlarına göre için en iyi sonuç veren algoritmanın RF algoritması olduğuna karar verilmiştir. Model performansları göz önüne alınarak en sağlıklı sonucu SMT veri dengeleme tekniği ile oluşturulan veri setine uygulanan RF algoritmasının %94 tahmin başarısı ile verdiği söylenebilir. Bu çalışma kalp hastalıklarının tahmin edilmesinde makine öğrenme algoritmalarının kullanılabileceğini ve çeşitli iyileştirme kombinasyonları ile tahmin performansının artırılabileceğini ortaya koymuştur. Bu çalışmanın bundan sonraki çalışmalara ışık tutarak makine öğrenme tekniklerinin kalp hastalıklarının teşhisinde ön tanı kriteri olarak kullanılabileceğini göstermiştir.

Yapılan bu çalışma veri dengeleme teknikleri ve HPO'nun kalp hastalığının tahmin edilmesinde kullanılan makine öğrenme tekniklerinin performansını etkilediğini ortaya koymuştur. Farklı tekniklerin bir arada kullanılması ile daha verimli modellerin oluşturulabileceğini göstermiştir. Bu çalışma sonraki çalışmalar için kaynak niteliğinde olup geliştirilmeye de açıktır. Farklı veri setleri ile çalışılabileceği gibi bu çalışmada kullanılan veri seti için birtakım iyileştirmeler yapılabilir. Genetik algoritması gibi sezgisel algoritmalar kullanılarak farklı hiper parametre optimizasyonu yöntemleri denenebilir. Öznitelik seçimi gibi yöntemlerle maliyet azaltılabilir. Ayrıca derin öğrenme teknikleri ile kalp hastalıklarının tespitinde farklı modeller oluşturulabilir.

Etik Kurul Onayı ve Çıkar Çatışması Beyanı (Ethics Committee Approval and Conflict of İnterest Statement)

Hazırlanan makalede etik kurul izni alınmasına gerek yoktur. Hazırlanan makalede herhangi bir kişi/kurum ile çıkar çatışması bulunmamaktadır.

Yazarların Katkıları (Authors' Contrıbutıons)

SUNGUR, Araştırma yapmış, deneyleri gerçekleştirip yorumlamış, alıntı olmayan görsel, tablo ve şekilleri oluşturmuş, makalenin yazım işlemini yapmıştır.

BAKIR, Sonuçları yorumlamış, analiz yapmış, gerekli düzenleme, şekillendirme işlemlerini yaparak makalenin yazım işleminde müşavirlik yapmıştır.

KAYNAKLAR (REFERENCES)

[1] "Kardiyovasküler Hastalıklar." Jan. 2021. [Online]. Available: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, "Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique," J Med Syst, vol. 43, no. 8, Jan. 2019, doi: 10.1007/s10916-019-1398-y.

[3] A. L. Yadav, K. Soni, and S. Khare, "Heart Diseases Prediction using Machine Learning," in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023, pp. 1–7. doi: 10.1109/ICCCNT56998.2023.10306469.

[4] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, vol. 16, no. 2, 2023, doi: 10.3390/a16020088.

[5] A. Özdemir, "Makine Öğrenmesi Algoritmaları ile Aritmilerin Sınıflandırılması," Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi, vol. 39, no. 3, pp. 394–402, 2023.

[6] N. Chandrasekhar and S. Peddakrishna, "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," Processes, vol. 11, no. 4, 2023, doi: 10.3390/pr11041210.

[7] S. Anitha and N. Sridevi, "Heart Dısease Predıctıon Usıng Data Mınıng Technıques," Journal of Analysis and Computation, vol. 13, no. 2, 2019, [Online]. Available: www.ijaconline.com,

[8] M. E. Çifci, "Kalp Hastalıklarında Kullanılan Yapay Zekâ Teknikleri Ve Uygulamaları." 2019.

[9] P. Kamat and M. C. Beyene, "Survey on prediction and analysis the occurrence of heart disease using data mining techniques," International Journal of Pure and Applied Mathematics, vol. 18, no. 8, 2018, [Online]. Available: https://www.researchgate.net/publication/323277772

[10] A. Rajdhan, A. Agarwal, and M. Sai, "Heart Disease Prediction using Machine Learning," IJERT Journal International Journal of Engineering Research & Technology. 2020. [Online]. Available: www.ijert.org

[11] M. Görgün, "Makine Öğrenmesi Yöntemleri ile Kalp Hastalığının Teşhis Edilmesi," Yüksek Lisans Tezi, Lisansüstü Eğitim Enstitüsü, İstanbul, 2020.

[12] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives Bayesian," in Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019, Institute of Electrical and Electronics Engineers Inc., Jan. 2019, pp. 292–297. doi: 10.1109/icoei.2019.8862604.

[13] O. Köse, "Sınıflama ve Regresyon Ağaçları Tekniği İle Kalp Hastalıklarına Etki Eden Bazı Faktörlerin Belirlenmesi." 2018.

[14] V. V Ramalingam, A. Dandapath, and M. K. Raja, "Heart disease prediction using machine learning techniques: A survey," International Journal of Engineering and Technology(UAE), vol. 7, no. 2.8 Special Issue 8, pp. 684–687, 2018, doi: 10.14419/ijet.v7i2.8.10557.

[15] H. Sharma and M. A. Rizvi, "Prediction of Heart Diseaseusing Machine Learning Algorithms," national Journal on Recent and Innovation Trends in Computing and Communication , vol. 5, no. 8, 2017.

[16] K. H. Kamil, "Artıfıcıal Neural Network Approach For Heart Dısease Classıfıcatıon." p. 58, 2020.

[17] I. Salman, "Heart attack mortality prediction: An application of machine learning methods," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 27, no. 6, pp. 4378–4389, 2019, doi: 10.3906/ELK-1811-4.

[18] S. Konda, A. Govardhan, and G. R. Rao, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques," K. Srinavas, Ed., 2020, pp. 1953–1957.

[19] M. Tarawneh and O. Embarak, "Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques," Lecture Notes on Data Engineering and Communications Technologies, vol. 29. Springer Science and Business Media Deutschland GmbH, pp. 447–454, 2019. doi: 10.1007/978-3-030-12839-5_41.

[20] E. Çil, "Makine Öğrenmesi Algoritmalarıyla Kalp Hastalıklarının Tespit Edilmesine Yönelik Performans Analizi." 2022.

[21] P. Kamil, "Personal Key Indicators of Heart Disease," https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. Jan. 2022.

[22] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16. pp. 321–357, 2002.

[23] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "SMOTETomek-Based Resampling for Personality Recognition," IEEE Access, vol. 7, pp. 129678–129689, 2019, doi: 10.1109/ACCESS.2019.2940061.

[24] Miss. M. S. Shelke1, Dr. P. R. Deshmukh2, and Prof. V. K. Shandilya, "A Review on Imbalanced Data Handling UsingUndersampling and Oversampling Technique," International Journal of Recent Trends in Engineering and Research, vol. 3, no. 4, pp. 444–449, Jan. 2017, doi: 10.23883/ijrter.2017.3168.0uwxm.

[25] A. Sekulić, M. Kilibarda, G. B. M. Heuvelink, M. Nikolić, and B. Bajat, "Random forest spatial interpolation," Remote Sens (Basel), vol. 12, no. 10, Jan. 2020, doi: 10.3390/rs12101687.

[26] E. Deniz, "Yapay sinir ağları ve K-en yakın komşu algoritması ile toprak çeşitliliğinin belirlenmesi." p. 69, 2021.

[27] E. Akca, "Satış Tahminlemesinde Hibrit Bir Yaklaşım:Pestel, Rfm, Gradıent Boostıng." Jan. 2022.

[28] A. Abraham, · Paramartha, D. Jyotsna, K. Mandal, A. Bhattacharya, and S. Dutta, Advances in Intelligent Systems and Computing 813 Emerging Technologies in Data Mining and Information Security. [Online]. Available: http://www.springer.com/series/11156

[29] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge *Discovery and Data Mining, Association for Computing Machinery*, Jan. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[30] M. Zareapoor and P. Shamsolmoali, "Application of credit card fraud detection: Based on bagging ensemble classifier," in *Procedia Computer Science, Elsevier B.V.*, 2015, pp. 679–685. doi: 10.1016/j.procs.2015.04.201.

[31] M. COŞAR and E. DENİZ, "Makine Öğrenimi Algoritmaları Kullanarak Kalp Hastalıklarının Tespit Edilmesi*," European Journal of Science and Technology,* Jan. 2021, doi: 10.31590/ejosat.1012986.

[32] Ö. Ekrem, O. K. M. Salman, B. Aksoy, And S. A. İnan, "Yapay Zekâ Yöntemleri Kullanılarak Kalp Hastalığının Tespiti," *Mühendislik Bilimleri ve Tasarım Dergisi, vol*. 8, no. 5, pp. 241–254, Jan. 2020, doi: 10.21923/jesd.824703.

[33] M. Saqlain, W. Hussain, N. A. Saqib, and M. A. Khan, "Identification of Heart Failure by Using Unstructured Data of Cardiac Patients," in Proceedings of the International Conference on Parallel Processing Workshops, Institute of Electrical and Electronics Engineers Inc., Jan. 2016, pp. 426–431. doi: 10.1109/ICPPW.2016.66.

[34] M. E. TAŞÇI and R. ŞAMLI, "Veri Madenciliği İle Kalp Hastalığı Teşhisi," *European Journal of Science and Technology*, pp. 88–95, Jan. 2020, doi: 10.31590/ejosat.araconf12.

[35] S. GÜNDOĞDU, "Kalp hastalık risk tahmini için Python aracılığıyla sınıflandırıcı algoritmalarının performans değerlendirmesi," *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi, vol.* 23, no. 69, pp. 1005–1013, 2021, doi: 10.21205/deufmd.2021236926.