T. C.
SİVAS CUMHURİYET ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KALP HASTALARININ SAĞKALIM TAHMİNİ

YÜKSEK LİSANS TEZİ

Fatma AZİZOĞLU

(20209257004)

Bilgisayar Mühendisliği Ana Bilim Dalı

Tez Danışmanı: Doç. Dr. Hidayet TAKCI

SİVAS

OCAK 2023

Fatma AZİZOĞLU'nun hazırladığı ve **"MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KALP HASTALARININ SAĞKALIM TAHMİNİ"** adlı bu çalışma aşağıdaki jüri tarafından **BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI**'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Tez Danışmanı	**Doç. Dr. Hidayet TAKCI** Sivas Cumhuriyet Üniversitesi	……
Jüri Üyesi	**Dr. Öğr. Üyesi Abdulkadir ŞEKER** Sivas Cumhuriyet Üniversitesi	……..……
Jüri Üyesi	**Dr. Öğr. Üyesi Halit BAKIR** Sivas Bilim ve Teknoloji Üniversitesi	……..……..

Bu tez, Sivas Cumhuriyet Üniversitesi Fen Bilimleri Enstitüsü tarafından **YÜKSEK LİSANS TEZİ** olarak onaylanmıştır.

Prof. Dr. Nevcihan GÜRSOY

FEN BİLİMLERİ ENSTİTÜSÜ MÜDÜRÜ

Bu tez, Sivas Cumhuriyet Üniversitesi Senatosu'nun 20.08.2014 tarihli ve 7 sayılı kararı ile kabul edilen Fen Bilimleri Enstitüsü Lisansüstü Tez Yazım Kılavuzu (Yönerge)'nda belirtilen kurallara uygun olarak hazırlanmıştır.

ETİK

Sivas Cumhuriyet Üniversitesi Fen Bilimleri Enstitüsü, Tez Yazım Kılavuzu (Yönerge)'nda belirtilen kurallara uygun olarak hazırladığım bu tez çalışmasında;

- ✓ Bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- ✓ Görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- ✓ Başkalarının eserlerinden yararlanılması durumunda ilgili eserlere, bilimsel normlara uygun olarak atıfta bulunduğumu ve atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi,
- ✓ Bütün bilgilerin doğru ve tam olduğunu, kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- ✓ Tezin herhangi bir bölümünü, Sivas Cumhuriyet Üniversitesi veya bir başka üniversitede, bir başka tez çalışması olarak sunmadığımı; beyan ederim.

12.01.2023
Fatma AZİZOĞLU

TEŞEKKÜR

Bilgi birikimi ve deneyimlerinden her zaman yararlandığım, tezin her aşamasında yardımlarını esirgemeyen, çalışma disiplini ve üretkenliğiyle her zaman örnek aldığım değerli danışman hocam sayın Doç. Dr. Hidayet TAKCI ’ya teşekkürü borç bilirim.

Ayrıca maddi ve manevi destekleriyle bugünlere gelmemi sağlayan babam Davut ÇETİN ve annem Ayşe ÇETİN başta olmak üzere tüm aileme, yıllardır manevi desteğini esirgemeyen sevgili eşim Arş. Gör. Gökhan AZİZOĞLU ’na ve arkadaşlarıma teşekkür ederim.

ÖZET

MAKİNE ÖĞRENMESİ YÖNTEMLERİYLE KALP HASTALARININ SAĞKALIM TAHMİNİ

Fatma AZİZOĞLU

Yüksek Lisans Tezi

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Hidayet TAKCI

2023, 59+xiv sayfa

Kalp hastalıkları, dünya genelinde yaşlanan nüfusla birlikte her geçen gün daha fazla kişiyi etkilemektedir. Bununla birlikte, her yıl yaklaşık 18 milyon insan kalp hastalıklarına bağlı nedenlerle yaşamını yitirmektedir. Bu nedenle kalp hastalığının erken teşhisi ve sağkalımı etkileyen faktörlerin ortaya çıkarılması önemli bir problemdir. Teknolojinin gelişmesiyle birlikte hastaların elektronik sağlık kayıtları toplanıp depolanmakta ve elde edilen veriler makine öğrenmesi yöntemleri ile hastalık teşhisi veya sağkalım tahmini için yaygın olarak kullanılmaktadır. Yapılan çalışma kapsamında kalp hastalarının sağkalımını tahmin etmek için makine öğrenmesi algoritmalarından; Destek Vektör Makinesi, Rastgele Orman, Karar Ağacı, XGBoost, CatBoost ve Lojistik Regresyon kullanılmıştır. Sağkalım tahmin başarısını artırmak için veri dengeleme, özellik seçimi ve normalizasyon yöntemleri tekil ve bütünsel olarak makine öğrenmesi algoritmalarına uygulanmıştır. Ayrıca, deneylerde 5-fold ve 10-fold çapraz doğrulama işlemi kullanılarak doğruluk, kesinlik, duyarlılık ve f1-skor açısından önerilen yöntemlerin ve algoritmaların performans ve çalışma sürelerinin karşılaştırması yapılmıştır. Son olarak veri seti %80 eğitim, %20 test ve %70 eğitim, %30 test olarak bölünmüş ve deneyler tekrar edilmiştir. Elde edilen sonuçlara göre kalp hastalarının sağkalım tahmininde 5-fold çapraz doğrulama ile en yüksek doğruluk başarısına sahip algoritma %86,4649 değeriyle Destek Vektör Makinesi olmuştur. Çapraz doğrulama kullanmadan elde edilen sonuçlara göre %80 eğitim, %20 test işlemi ile en yüksek

doğruluk başarısına ulaşan algoritmalar %93,9 değeri ile Rastgele Orman ve CatBoost olmuştur.

Anahtar kelimeler: Sağkalım tahmini, veri dengeleme, özellik seçimi, normalizasyon, makine öğrenmesi.

ABSTRACT

SURVIVAL PREDICTION OF HEART PATIENTS WITH MACHINE LEARNING METHODS

Fatma AZİZOĞLU

Master of Science Thesis

Department of Computer Engineering

Supervisor: Assoc. Prof. Hidayet TAKCI

2023, 59+xiv pages

Heart disease is affecting more and more people in the ageing population around the world. Nevertheless, about 18 million people die each year from causes related to heart disease. Therefore, early diagnosis of heart disease and revealing the factors affecting survival is an important problem. With the advancement of technology, electronic health records of patients are extracted and stored, and the resulting data is utilized extensively for disease diagnosis or survival prediction using machine learning methods. Within the scope of the study, machine learning algorithms to predict the survival of heart patients; Support Vector Machine, Random Forest, Decision Tree, XGBoost, CatBoost and Logistic Regression are used. Data balancing, feature selection and normalization methods have been applied to machine learning algorithms singly and together to increase the survival prediction success. In addition, the performance and running times of the proposed methods and algorithms were compared in terms of accuracy, precision, sensitivity and f1-score using 5-fold and 10-fold cross validation in the experiments. According to the results obtained, the algorithm with the highest accuracy in predicting the survival of heart patients was the Support Vector Machine with a value of 86.4649%. According to the results obtained without using cross validation, the algorithms that achieved the highest accuracy with 80% training and 20% testing were Random Forest and CatBoost with 93.9%.

Key Words: Survival prediction, data balancing, feature selection, normalization, machine learning.

İÇİNDEKİLER

ÖZET ..... **vii**
ABSTRACT ..... **ix**
ŞEKİLLER DİZİNİ ..... **xii**
ÇİZELGELER DİZİNİ ..... **xiii**
KISALTMALAR DİZİNİ ..... **xiv**
1. GİRİŞ **1**
1.1 Amaç ve Kapsam 3
1.2 Literatür Araştırması 3
1.3 Kalp Hastalıklarının Türleri 13
1.3.1 Koroner kalp hastalığı 13
1.3.2 Serebrovasküler hastalıklar 13
1.3.3 Periferik Arter hastalığı 13
1.3.4 Romatizmal kalp hastalığı 13
1.3.5 Konjenital (Doğumsal) kalp hastalığı 14
1.3.6 Derin Ven Trombozu ve Pulmoner Emboli 14
1.4 Kalp Hastalıklarının Belirtileri 14
1.5 Kalp Hastalıkları için Risk Faktörleri 15
2. MAKİNE ÖĞRENMESİ **16**
2.1 Makine Öğrenmesinin Tanımı 16
2.2 Makine Öğrenmesi Türleri 16
2.2.1 Denetimli öğrenme 17
2.2.2 Denetimsiz öğrenme 18
2.2.3 Yarı denetimli öğrenme 18
2.2.4 Takviyeli öğrenme 19
3. MATERYAL VE YÖNTEM **20**
3.1 Materyal 20
3.1.1 Veri seti 20
3.1.2 Uygulama ortamı 24
3.1.3 Sentetik Azınlık Örneklem Arttırma (SMOTE) yöntemi 24
3.1.4 Veri ön işleme 26
3.1.5 Makine öğrenmesi algoritmaları 28

3.2 Önerilen Yöntem 32
3.2.1 Önerilen yöntemlerde SMOTE ile veri dengeleme 34
3.2.2 Önerilen yöntemlerde özellik seçiminin uygulanması 35
3.2.3 Önerilen yöntemlerde normalizasyonun uygulanması 37
3.2.3 Kalp hastalarının sağkalımının tahmin edilmesi 37
3.3 Geliştirilen Modeller için Performans Değerlendirme Metrikleri 38
3.3.1 Çapraz Doğrulama 38
3.3.2 Karmaşıklık Matrisi 39
3.3.3 Doğruluk 40
3.3.4 Kesinlik 40
3.3.5 Duyarlılık 40
3.3.6 F1-Skor 41
4. BULGULAR 42
4.1 Doğruluğa Dayalı Performans Sonuçları 42
4.2 Kesinliğe Dayalı Performans Sonuçları 43
4.3 Duyarlılığa Dayalı Performans Sonuçları 45
4.4 F1-Skor Dayalı Performans Sonuçları 46
4.5 Algoritmaların Çalışma Sürelerinin Karşılaştırılması 48
4.6 Çapraz Doğrulama Kullanmadan Elde Edilen Doğruluk Değerleri (%80 Eğitim, %20 Test) 49
4.7 Çapraz Doğrulama Kullanmadan Elde Edilen Doğruluk Değerleri (%70 Eğitim, %30 Test) 50
4.8 Literatürdeki Benzer Çalışmalarla Karşılaştırma 51
5. TARTIŞMA VE SONUÇ 53
ÖZGEÇMİŞ 59

ŞEKİLLER DİZİNİ

Şekil 2.1 Makine öğrenmesinin türleri 17
Şekil 2.2 Denetimli makine öğrenmesi çalışma prensibi 17
Şekil 2.3 Denetimsiz makine öğrenmesi çalışma prensibi 18
Şekil 2.4 Takviyeli öğrenme çalışma prensibi 19
Şekil 3.1 Veri seti özelliklerinin sayısal dağılımı 20
Şekil 3.2 Özniteliklerin dağılımlarını gösteren histogram grafikleri 23
Şekil 3.3 Özellikler arası korelasyonları gösteren ısı haritası 24
Şekil 3.4 Veri ön işleme yöntemleri 26
Şekil 3.5 Destek Vektör Makinesi çalışma prensibi 29
Şekil 3.6 Rastgele Orman algoritmasının çalışma prensibi 30
Şekil 3.7 Karar Ağacı algoritmasının çalışma prensibi 31
Şekil 3.8 LR yönteminin çalışma prensibi 32
Şekil 3.9 Önerilen sağkalım tahmin yöntemlerinin genel sistem şeması 34
Şekil 3.10 SMOTE uygulandıktan sonra hedef sınıfın sayısal dağılımı 35
Şekil 3.11 SÖM ve SÖNM yöntemi için her bir algoritmaya ÖÖE tekniği uygulanması ile elde özelliklere ait önem dereceleri 37
Şekil 3.12 5-fold çapraz doğrulama temsili gösterimi 39
Şekil 3.13 Karmaşıklık Matrisi 39

ÇİZELGELER DİZİNİ

Tablo 1.1. Literatürdeki çalışmaların özeti 10
Tablo 3.1. Kalp hastalarının sağkalımına ait veri setinin özellikleri..... 21
Tablo 4.1 Önerilen dört yönteme 5-fold çapraz doğrulama uygulanması sonucu elde edilen doğruluk değerleri 43
Tablo 4.2 Önerilen dört yönteme 10-fold çapraz doğrulama uygulanması sonucu elde edilen doğruluk değerleri 43
Tablo 4.3 Önerilen dört yönteme 5-fold çapraz doğrulama uygulanması sonucu elde edilen kesinlik değerleri..... 44
Tablo 4.4 Önerilen dört yönteme 10-fold çapraz doğrulama uygulanması sonucu elde edilen kesinlik değerleri..... 45
Tablo 4.5 Önerilen dört yönteme 5-fold çapraz doğrulama uygulanması sonucu elde edilen duyarlılık değerleri..... 46
Tablo 4.6 Önerilen dört yönteme 10-fold çapraz doğrulama uygulanması sonucu elde edilen duyarlılık değerleri..... 46
Tablo 4.7 Önerilen dört yönteme 5-fold çapraz doğrulama uygulanması sonucu elde edilen f1-skor değerleri..... 47
Tablo 4.8 Önerilen dört yönteme 10-fold çapraz doğrulama uygulanması sonucu elde edilen f1-skor değerleri..... 47
Tablo 4.9 Dört yöntem için altı farklı makine öğrenmesi algoritmasının..... 49
5-fold çapraz doğrulama ile saniye cinsinden çalışma zamanı..... 49
Tablo 4.10 Dört yöntem için altı farklı makine öğrenmesi algoritmasının..... 49
10-fold çapraz doğrulama ile saniye cinsinden çalışma zamanı..... 49
Tablo 4.11 Dört yöntem için altı farklı makine öğrenmesi algoritmasının..... 50
%80 eğitim %20 test şeklinde bölünmesi ile elde edilen doğruluk değerleri..... 50
Tablo 4.12 Dört yöntem için altı farklı makine öğrenmesi algoritmasının..... 50
%70 eğitim %30 test şeklinde bölünmesi ile elde edilen doğruluk değerleri..... 50
Tablo 4.13 Önerilen SÖNM yönteminin literatürdeki benzer çalışmalarla karşılaştırılması..... 52

KISALTMALAR DİZİNİ

Adaboost	**:**	Adaptive Boosting
ABD	**:**	Amerika Birleşik Devletleri
CatBoost	**:**	Categorical Boosting
Chi2	**:**	Chi-Square
MLP	**:**	Multi Layer Perceptron
DR	**:**	Doğrusal Regresyon
SVM	**:**	Support Vector Machine
EHMRG	**:**	Emergency Heart Failure Mortality Risk Grade
ET	**:**	Extra Trees
GYS	**:**	Gradyan Yükseltme Sınıflandırıcısı
DT	**:**	Decision Tree
KNN	**:**	K Nearest Neighbors
LASSO	**:**	Least Absolute Shrinkage and Selection Operator
LightGBM	**:**	Light Gradiest Boosting Machines
LR	**:**	Lojistik Regresyon
MICE	**:**	Multivariate Imputation by Chained Equations
NB	**:**	Naive Bayes
NIS	**:**	National Inpatient Sample
OVM	**:**	Orijinal Veri Seti Modeli
ÖÖE	**:**	Özyinelemeli Özellik Eleme
RF	**:**	Random Forest
ROÖS	**:**	Rastgele Orman Özellik Seçimi
RSF	**:**	Random Survival Forest
SGS	**:**	Stokastik Gradyan Sınıflandırıcı
SMOTE	**:**	Syntetic Minority Oversampling Technique
SM	**:**	SMOTE Modeli
SÖM	**:**	SMOTE Özellik Seçimi Modeli
SÖNM	**:**	SMOTE Özellik Seçimi Normalizasyon Modeli
UCI	**:**	University of California at Irvine
XGBoost	**:**	Extreme Gradient Boosting
YSA	**:**	Yapay Sinir Ağları

1. GİRİŞ

Her yıl, dünya genelinde milyonlarca insan kalp hastalıkları nedeniyle hayatını kaybetmektedir. Bu da meydana gelen tüm ölüm vakalarının yaklaşık %32'sine karşılık gelmektedir. Bu ölümlerin %75'inden daha fazlası düşük ve orta gelirli ülkelerde meydana gelirken, %85'i kalp krizi ya da felce bağlı olarak gerçekleşmektedir [1].

Kalp hastalıkları, dünyada artan ve yaşlanan nüfusla birlikte her geçen gün daha fazla insanı etkisi altına almaktadır. Tıbbi tedavilerde yaşanan gelişmelere rağmen kalp hastalığı teşhisi konulan ve bu hastalıktan dolayı hayatını kaybeden insan sayısı oldukça fazladır [2]. Bu anlamda, kalp hastalarının sağkalımı veya ölümüne etki eden faktörlerin ortaya çıkarılması oldukça önemli bir konudur.

Son yıllarda makine öğrenmesi, doğal dil işleme, akıllı şehirler, IoT cihazlar, sosyal medya analizleri, siber güvenlik ve medikal alanlar gibi pek çok alanda kullanılmakla birlikte en yaygın kullanıldığı alanlardan birisi sağlık hizmetleridir [3]. Makine öğrenmesi bu alanda, hastalara ait tıbbi kayıtları cinsiyet, yaş, kilo vb. diğer özellikler ile birleştirerek hastaların takip süresince sağkalımını tahmin etmek, hastalığı teşhis etmek veya hastalığa etki eden faktörlerin ortaya çıkarılması amaçlarıyla kullanılabilmektedir [4]. Hastaneye yatış sebepleri bakımından da kalp hastalığı ilk sırada yer almakta ve yüksek oranda ölümle sonuçlanmaktadır [5]. Bu ölümlerin klinik nedenlerinin tespitinde son dönemde özellikle makine öğrenmesi yöntemlerinin sağlık verilerinde kullanılması etkili olmuştur [6].

Literatürde, kalp hastalığının sağkalım tahminine yönelik birçok çalışma olmasına rağmen bu çalışmaların çoğunda yalnızca özellik seçiminin makine öğrenmesi algoritmaları üzerindeki etkisine odaklanılmıştır. Yapılan çalışmalarda, sentetik veri üretme, özellik seçimi ve normalizasyon yöntemlerini

bir arada kullanan yöntemlerin sayısı oldukça azdır. Bu anlamda, kalp hastalarının sağkalım tahmininde doğruluk performansının arttırılmasına yönelik bir yönteme olan ihtiyaç hala devam etmektedir.

Tez kapsamında, Kaliforniya Üniversitesi Irvine (University of California at Irvine, UCI) veri setlerinden "*heart_failure_clinical_records*" isimli veri seti [6] kullanılmış ve 299 kalp hastalığına sahip bireyin tedavi takip süresi içerisindeki sağkalımlarını tahmin etmek için makine öğrenmesi yöntemleri kullanılmıştır. Makine öğrenmesi algoritmalarından; Destek Vektör Makinesi (Support Vector Machine, SVM), Rastgele Orman (Random Forest, RF), Karar Ağacı (Decision Tree, DT), Aşırı Gradiyent Arttırma (Extreme Gradient Boosting, XGBoost), Kategorisel Arttırma (Categorical Boosting, CatBoost) ve Lojistik Regresyon (LR) tercih edilmiştir. Sağkalım tahmin başarısını artırmak için veri setinin hedef sınıfındaki dengesiz veri dağılımı ve aykırı değer aralıklarının olumsuz etkileri giderilmiştir. Sentetik veri artırma için; Sentetik Azınlık Aşırı Örnekleme (Syntetic Minority Oversampling Technique, SMOTE), özellik seçimi için; Özyinelemeli Özellik Eleme (ÖÖE) yöntemi ve normalizasyon için; Yeo-Johnson teknikleri tekil ve birlikte farklı kombinasyonlarla uygulanmıştır. Sonuç olarak, dört farklı model önerilmiş, her model için altı farklı makine öğrenmesi algoritması 5-fold ve 10-fold çapraz doğrulama yöntemi ile çalıştırılarak toplam 48 adet deney gerçekleştirilmiştir. Daha sonra literatürdeki benzer çalışmalarla önerilen yöntemleri kıyaslamak ve çapraz doğrulamanın etkisini incelemek için veri seti %80 eğitim, %20 test ve %70 eğitim, %30 test şeklinde bölerek ve deneyler tekrar edilmiştir.

Tez çalışmasının organizasyonu şu şekilde belirlenmiştir:

- Birinci bölümde, çalışmanın amaç ve kapsamı, literatür araştırması, kalp hastalıklarının türleri, belirtileri ve risk faktörlerine yer verilmiştir.
- İkinci bölümde, makine öğrenmesi ile ilgili genel bilgiler açıklanmıştır.
- Üçüncü bölümde, makine öğrenmesi yöntemleri kullanılarak kalp hastalarının sağkalımına yönelik önerilen yöntemler detaylı bir şekilde verilmiştir.

- Dördüncü bölümde, önerilen yöntemlere ait deneysel sonuçlar paylaşılmıştır.
- Son olarak, beşinci bölümde elde edilen deneyse sonuçlar ve çıkarımlar tartışılmıştır.

1.1 Amaç ve Kapsam

Tez çalışmasının amacı, makine öğrenmesi algoritmaları ile kalp hastalarının sağkalımının tahmin edilmesi ve sağkalımı en fazla etkileyen parametrelerin ortaya çıkarılmasıdır. Bu kapsamda, yapılan çalışmada kalp hastalarının sağkalımına ait veri seti kullanılarak makine öğrenmesi algoritmalarının tahmin başarısının artırılması için veri setindeki ölen ve hayatta kalan veri sınıfları üzerinde veri dengeleme işlemi, özellik seçim yöntemi ve normalizasyon yöntemi tekil ve bütünsel olarak uygulanmış ve deney sonuçları karşılaştırılmıştır.

1.2 Literatür Araştırması

Son yıllarda literatürde kalp hastalarının sağkalım tahmini ve sağkalım üzerinde en fazla etkiye sahip özelliklerinin ortaya çıkarılması üzerine pek çok çalışma yapılmıştır.

Wang vd. Çin Klinik Araştırma Merkezi'ndeki koroner kalp hastalığından kaynaklanan kalp yetmezliği problemi yaşayan 5188 hastanın 3 yıldan fazla bir süre boyunca klinik takibini yapmışlardır. Bu hastalardan 1562'sinin bu süre zarfında hayatını kaybettiği belirlenmiştir. Hastaların sağkalımını tahmin etmek için makine öğrenmesi algoritmalarından LR, K-En Yakın Komşu (K Nearest Neighbors, KNN), SVM, Naive Bayes (NB), Çok Katmanlı Algılayıcı (Multi Layer Perceptron, MLP) ve XGBoost algoritmalarını kullanmışlardır. Yaptıkları çalışmada eğitim veri setindeki negatif ve pozitif veri örneklerinin sayısını dengelemek için SMOTE tekniğini kullanarak sentetik veri çoğaltma işlemini gerçekleştirmişlerdir. Yazarlar, öznitelik seçimi için ÖÖE uygulayarak modele en fazla katkı sağlayan alt özellikleri seçmişlerdir. Sağkalımın sınıflandırılmasında en iyi başarıyı veren algoritma %82 doğrulukla XGBoost olmuştur [4].

Dai vd. yapmış oldukları çalışmada kalp yetmezliğine bağlı olarak gelişen Sarkoidoz hastalarının sağkalımını makine öğrenmesi yöntemleri ile tahmin etmişlerdir. Çalışmalarında kalp hastalarının kayıtlarından oluşan, Ulusal Yatan Hasta Örneği (National Inpatient Sample, NIS) veri tabanından alınan 4659 hastanın tıbbi kayıtlarını kullanmışlardır. Veri setindeki ölüm ve sağkalım oranının dengesizliğini gidermek için veri çoğaltma işlemini uygulayarak hasta sayısını 6360'a çıkarmışlardır. Yazarlar, hastaların sağkalımını tahmin etmek için makine öğrenmesi algoritmalarından; LR, RF ve XGBoost algoritmasını tercih etmişlerdir. Çalışmalarında, makine öğrenmesi algoritmalarının aşırı öğrenmesini (overfitting) önlemek ve modele en fazla katkısı olan öznitelikleri seçmek için En Küçük Mutlak Büzülme ve Seçim Operatörü (Least Absolute Shrinkage and Selection Operator, LASSO) kullanmışlardır. Ayrıca, scikit-learn içerisinde bulunan GridSearchCV fonksiyonu ile her bir algoritmanın optimum parametre değerini bulmuşlardır. Veri setini %70 eğitim ve %30 test olacak şekilde ayırmışlardır. Daha sonra, her algoritma için en iyi parametreyi kullanarak hastaların sağkalımını tahmin etmişlerdir. Elde ettikleri sonuçlara göre kalp hastalarının sağkalımını RF algoritması %71 doğruluk ile diğer algoritmadan daha iyi tahmin ettiği görülmüştür [7].

Austin vd. Akut Dekompanse kalp yetmezliği olan, Kanada'da bulunan 86 hastanenin acil servisine başvuran 18 yaş üzeri 12608 hastanın sağkalımını tahmin etmeye çalışmışlardır. Çalışmalarında hastaların 7 ve 30 günlük süre için sağkalımlarını tahmin eden Lojistik Regresyon tabanlı, Acil Kalp Yetmezliği Mortalite Risk Derecesi (Emergency Heart Failure Mortality Risk Grade, EHMRG) isimli modeli önermişlerdir. Kullandıkları veri setini %80 eğitim ve %20 test olacak şekilde ayırarak modellerini eğitmişlerdir. Çalışmalarında özellik seçimi için, ÖÖE yöntemini uygulayarak XGBoost ve RF makine öğrenmesi algoritmalarını tercih etmişlerdir. Yazarlar, Yapay Sinir Ağlarını (YSA) modellerken daha dengeli bir veri seti ile çalışmak için "*imbalanced-learn"* kütüphanesini kullanarak rastgele örnekleme yoluyla sentetik veri çoğalma işlemini gerçekleştirmişlerdir. Ek olarak önerdikleri modeli YSA, LASSO Regresyonu, RF ve XGBoost algoritması ile karşılaştırmışlardır. Elde ettikleri sonuçlara göre 7 günlük süre için en yüksek sağkalım tahminini

gerçekleştiren model %80 doğruluk ile EHMRG olurken, 30 günlük süre için %77,9 doğruluk ile XGBoost algoritması olmuştur [8].

Chicco ve Jurman, UCI üzerinden halka açık olarak paylaşılan kalp yetmezliğine sahip 299 hastanın sağkalımını tahmin etmek için makine öğrenmesi algoritmalarını kullanmışlardır. Çalışmalarında 13 tıbbi özellik arasından seçim yapmak için Rastgele Orman Özellik Seçimi (ROÖS) yönteminden yararlanmışlardır. Özellik seçiminden sonra 13 olan özellik adedi enjeksiyon fraksiyonu ve serum kreatinin olmak üzere 2 özelliğe düşürülmüştür. Çalışmalarında RF, DT, Gradyan Artırma, Doğrusal Regresyon (DR), YSA, NB, SVM ve KNN algoritmalarını kullanarak hastaların sağkalımını tahmin etmişlerdir. Elde ettikleri sonuçlara göre en yüksek doğruluğa sahip algoritma %75,4 ile RF olmuştur [6].

Angraal vd. Amerika Birleşik Devletleri (ABD), Kanada, Brezilya, Arjantin, Rusya ve Gürcistan'da yaşayan ve kalp yetmezliğine sahip 1767 hastadan toplanan veriler üzerinde çalışarak hastaların hastaneye yatış durumunu ve sağkalımını tahmin etmişlerdir. Yaptıkları çalışmada LASSO yöntemi ile özellik seçimi yaparak makine öğrenmesi algoritmalarından LR, RF, Gradyan İniş Artırma ve SVM kullanarak hastaların sağkalımlarını tahmin etmişlerdir. Elde ettikleri sonuçlara göre hastaların sağkalımını %72 doğruluk değeriyle en iyi RF algoritması tahmin etmiştir [9].

Newaz vd. UCI üzerinden halka açık olarak paylaşılan kalp yetmezliğine sahip 299 hastanın sağkalımını tahmin etmek için SVM, KNN, LR, Adaptif Arttırma (Adaptive Boosting, AdaBoost) ve RF algoritmasını kullanmışlardır. Yaptıkları çalışmada ki-kare (Chi-Square, Chi2) ve ÖÖE kullanarak özellik seçimi yapmışlardır. Daha sonra veri kümesini %80 eğitim ve %20 test olacak şekilde bölerek eğitim veri kümesindeki sınıf dengesizliğini ortadan kaldırmak için SMOTE yöntemini kullanmışlardır. Elde ettikleri sonuçlara göre en iyi sağkalım tahminini Chi2 özellik seçme yöntemi uygulanan RF algoritmasıyla %77,33 doğruluk ile elde etmişlerdir [10].

Park vd. Kore'de yaşayan ve kalp yetmezliğine sahip 4312 hastanın 3 yıllık takip süresi boyunca sağkalımını tahmin etmek için kişilerin 19 klinik ve 8 ekokardiyografi özelliğini birleştirmişlerdir. Veri setinde, kayıp veri miktarı %30'un üzerinde olan özellikleri veri setinden çıkararak kalan özellikler içerisindeki eksik değerleri doldurmak için R paketinden Çoklu Atama Yöntemlerini (Multiple Imputation Methods) kullanmışlardır. Özellik seçimi için R paketinden Rastgele Hayatta Kalma Ormanı (Random Survival Forest, RSF) metodunu kullanarak en önemli özelliklerini belirlemişlerdir. Daha sonra, 5-fold çapraz doğrulama ve CoxBoost yöntemini kullanarak hastaların sağkalımlarını tahmin etmişlerdir. Elde ettikleri sonuçlara göre önerdikleri çalışma ile hastaların 3 yıl boyunca sağkalımlarını ortalama %76 doğruluk değeri ile tahmin etmişlerdir [5].

Ahmad vd. Pakistan'da yaşayan ve hastaneye yatırılan kalp yetmezliğine sahip 299 hastanın tıbbi kayıtları, yaş, cinsiyet, hastanede yatış süreleri ve sağkalımı hakkındaki bilgilerini toplayarak oluşturdukları veri setini UCI üzerinden halka açık bir şekilde paylaşmışlardır. Bu verilerden yola çıkarak Cox Regresyonu ve Kaplan Meier yöntemlerini kullanarak hastaların sağkalımlarını tahmin etmişlerdir [11].

Ishaq vd. UCI üzerinden paylaşılan 299 kalp hastasının sağkalımını tahmin etmek için makine öğrenmesi algoritmalarından; DT, Adaboost, LR, Stokastik Gradyan Sınıflandırıcı (SGS), RF, Gradyan Yükseltme Sınıflandırıcısı (GYS), Ekstra Ağaçlar (Extra Trees, ET), Naive Bayes ve SVM kullanmışlardır. Yaptıkları çalışmada, dengesiz veri kümesi problemini ortadan kaldırmak için SMOTE yöntemini kullanarak sentetik veri çoğaltma işlemini yapmışlardır. Ayrıca hedef sınıf üzerinde en fazla etkiye sahip alt özellikleri ortaya çıkarmak için Rastgele Orman Özellik Seçimi (ROÖS) yöntemini kullanarak toplam 13 özellik arasından en önemli 9 alt özelliği ortaya çıkarmışlardır. Elde ettikleri sonuçlara göre en yüksek doğruluk başarısını tüm özellikleri kullanarak veri setine SMOTE uyguladıktan sonra %92 değeri ile ET Sınıflandırıcı algoritması ile elde etmişlerdir [12].

Mamun vd. UCI üzerinden paylaşılan 299 kalp yetmezliğine sahip hastanın sağkalımını tahmin etmek için makine öğrenmesi algoritmalarından; Hafif Gradyan Arttırma Makineleri (Light Gradiest Boosting Machines, LightGBM), XGBoost, SVM, DT, LR ve Torbalama (Bagging) kullanmışlardır. Çalışmalarında, SMOTE yöntemi yardımıyla sentetik veri çoğaltma yaparak hedef sınıftaki dengesiz veri dağılımını ortadan kaldırmışlardır. Ayrıca çalışmalarında makine öğrenmesi algoritmalarının aşırı öğrenmesini engellemek amacıyla 10-fold çapraz doğrulama kullanmışlardır. Elde ettikleri sonuçlara göre LightGBM %85 doğruluk değeri ile en iyi sınıflandırma performansına sahip model olduğu görülmüştür [13].

Panahiazar vd. Minnesota'da bulunan Mayo Clinic merkezinde tedavi gören kalp hastalarından 1993-2013 yılları arasında 5044 hastadan elektronik sağlık kayıtları ve klinik bakım verilerini toplamışlardır. Kardiyologlara danışarak topladıkları verilerden yaş, cinsiyet, ırk, kolesterol, hemoglobin gibi bazı özellikleri el ile (manuel) çıkarmışlardır. Yaptıkları çalışmada, Cox Regresyonunu kullanarak hastaların 2 ile 5 yıl içerisindeki sağkalımlarını tahmin etmişlerdir. Daha sonra, makine öğrenmesi algoritmalarından; RF, LR, SVM, DT ve Adaboost algoritmalarını kullanarak tedavi gören kalp hastalarından 1 yıl içerisinde ölen ve sağ kalan hastaları sınıflandırmışlardır. Elde ettikleri sonuçlara göre LR %81 tahmin başarısı ile sağ kalımı en iyi tahmin eden model olmuştur [14].

Erdaş ve Ölçer yapmış oldukları çalışmada UCI üzerinde paylaşılmış olan 299 kalp hastalığına sahip hastaların verilerini kullanarak makine öğrenmesi yöntemleri ile hastaların sağkalımlarını tahmin etmişlerdir. Çalışmalarında, ölüm ve sağkalımı sınıflandırmak için makine öğrenmesi algoritmalarından; Tek Kural (One Rule), RF, SVM, MLP ve NB algoritmalarını kullanmışlardır. Daha sonra, Korelasyona Dayalı Öznitelik Seçim (KDÖS) tekniğini kullanarak özellik seçimi yapmış ve 2 farklı alt özellik kümesi oluşturmuşlardır. Birinci alt özellik kümesi için yaş, serum kreatinin ve ejeksiyon fraksiyonu seçilirken ikinci alt özellik kümesi için ise sadece serum kreatinin ve ejeksiyon fraksiyonu

seçilmiştir. Elde ettikleri sonuçlara göre her iki veri seti için de en başarılı model %78 doğruluk değeri ile MLP olmuştur [15].

Samad vd. Pensilvanya bölgesinde 1998-2017 yılları arasında 171510 kalp hastalığına sahip kişinin elektronik tıbbi kayıtları ve ekokardiyografi kayıtlarını toplamışlardır. Oluşturdukları veri setinde bulunan eksik değerleri Zincirli Denklemlerle Çok Değişkenli Değerlendirme (Multivariate Imputation by Chained Equations, MICE) yöntemi ile doldurmuşlardır. Daha sonra makine öğrenmesi algoritmalarından RF ve LR kullanarak hastaların 5 yıllık süre içerisindeki sağkalım ve ölüm durumlarını tahmin etmişlerdir. Elde ettikleri sonuçlara göre RF algoritması sağkalımı %95 doğruluk başarısı ile tahmin eden en iyi algoritma olmuştur [16].

Türkmenoğlu ve Yıldız UCI üzerinden paylaşılan 299 kalp hastasının verilerini kullanarak hedef sınıf üzerinde aşağı örnekleme ve yukarı örnekleme olmak üzere 2 farklı veri dengeleme yöntemi uygulamışlardır. Yukarı örnekleme işlemi için KNN tabanlı SMOTE yöntemini kullanmışlardır. Özellik seçimi için RF ve Korelasyon Matrisini kullanarak hasta takip süresi, serum kreatinin, ejeksiyon fraksiyonu ve yaş bilgilerinden oluşan bir alt özellik kümesi oluşturmuşlardır. Veri ön işleme adımında, serum kreatinin, ejeksiyon fraksiyonu ve yaş özelliklerinin değerini 0 ile 1 arasına yeniden ölçeklemişlerdir. Daha sonra makine öğrenmesi algoritmalarından KNN, RF ve ET yöntemlerini kullanarak hastaların sağkalımını tahmin etmişlerdir. Çalışmalarında k-fold çapraz doğrulama yerine veri setini 8 parçaya bölerek deneyleri 5 defa tekrar etmişlerdir. Yazarlar, aşağı örnekleme yöntemi ile oluşturdukları veri seti üzerinde en iyi sağkalım tahminini %84,58 doğruluk değeri ile ET algoritmasının elde ettiğini bulmuşlardır. Yukarı örnekleme yöntemi için ise kalp hastalarının sağkalımını en başarılı sınıflandıran yöntem %84,51 doğruluk değeriyle RF algoritması olmuştur [17].

Potur ve Erginel UCI veri havuzundan aldıkları 299 kalp hastasına ait veri setini kullanarak makine öğrenmesi yöntemleri ile sağ kalım tahmini yapmışlardır. Çalışmalarında özellik seçimi için WEKA yazılımında bulunan InfoGainAttributeEval özellik seçim yöntemi ile yaş, serum kreatinin, ejeksiyon

fraksiyonu, serum sodyum ve hasta takip süresi olmak üzere 5 alt özellik kümesi belirlemişlerdir. Benzer şekilde, WEKA programında bulunan CfsSubsetEval özellik seçim yöntemini kullanarak; yaş, ejeksiyon fraksiyonu, serum kreatinin ve hasta takip süresi olmak üzere 4 özellikten oluşan bir alt özellik kümesi oluşturmuşlardır. Daha sonra, makine öğrenmesi yöntemlerinden NB, LR, MLP, SVM ve J48 algoritmalarını kullanarak hastaların sağ kalımını tahmin etmişlerdir. Algoritmaların tahmin sonuçlarını doğruluk, f1-skor (f1-score) ve Kapa İstatistiği metriklerini kullanarak karşılaştırmışlardır. Elde ettikleri deneysel sonuçlara göre, MLP %90 doğruluk değeri ile kalp hastalarının sağ kalımını en iyi tahmin eden model olmuştur [18].

Al-Dury vd. İsveç'te 2008 ile 2016 yılları arasında ani kalp durması (cardiac arrest) vakası yaşayan 45067 kişinin kayıtlarını içeren bir veri seti oluşturmuşlardır. Oluşturdukları veri setindeki eksik değerleri tamamlamak için MICE yöntemini kullanmışlardır. Daha sonra, hastaların 30 günlük takip süresi için 16 özellik bilgisine odaklanarak makine öğrenmesi yöntemlerinden; RF ve Gradyan Artırma algoritmasını kullanarak hastaların sağ kalımlarını tahmin etmişlerdir. Çalışmalarında, makine öğrenmesi algoritmalarının aşırı öğrenme probleminin önüne geçmek için 10-fold çapraz doğrulama işlemini uygulamışlardır. Yazarlar, %82,1 tahmin doğruluğuna ulaştıklarını ifade etmiş ancak bu sonucun hangi algoritmaya ait olduğunu belirtmemişlerdir [19].

Literatürde kalp hastalarının sağkalımını makine öğrenmesi yöntemleri ile tahmin eden çalışmalar Tablo 1.1'de özetlenmiştir.

Tablo 1.1. Literatürdeki çalışmaların özeti

Yazarlar	Çalışma Yılı	Veri Toplama Yöntemi	Veri Sayısı	Kullanılan Algoritmalar	Veri Dengeleme	Özellik Seçimi	Çapraz Doğrulama	Normalizasyon Yöntemi	Çalışma Ortamı
Wang vd. [4]	2021	Çin Klinik Araştırma Merkezi	5188	LR, KNN, SVM, NB, MLP, XGBoost	SMOTE	ÖÖE	5-fold	Min-Max Normalizasyonu	Python
Dai vd. [7]	2022	NIS	4659	LR, RF, XGBoost	Yok	LASSO	Yok	Yok	Python
Austin vd. [8]	2022	Kanada'da bulunan acil servisler	12608	RF, XGBoost, YSA, LASSO, EHMRG	Rastgele Örnekleme	ÖÖE	Yok	Yok	Python
Chicco ve Jurman [6]	2020	UCI	299	RF, DT, Gradyan Artırma, Doğrusal Regresyon, YSA, NB, SVM, KNN	Yok	ROÖS	Yok	Yok	Python

Angral vd. [9]	2020	ABD, Kanada, Brezilya, Arjantin, Rusya, Gürcistan'daki kalp hastalarının sağkalımına ait veri seti	1767	LR, RF, Gradyan İniş, SVM	Yok	LASSO	5-fold	Yok	Belirtilmemiş
Newaz vd. [10]	2021	UCI	299	SVM, KNN, LR, RF, Adaboost	SMOTE	Chi2, ÖÖE	5-fold	Yok	Belirtilmemiş
Park vd. [5]	2022	Kore'deki kalp hastalarına ait sağkalım veri seti	4312	CoxBoost	Yok	RSF	5-fold	Yok	Belirtilmemiş
Ahmad vd. [11]	2017	UCI'de paylaşılan veri setini topladılar	299	Cox Regresyonu, Kaplan Meier	Yok	Yok	Yok	Yok	Belirtilmemiş
Ishaq vd. [12]	2021	UCI	299	DT, Adaboost, LR, GYS, SGS, ET, NB, SVM	SMOTE	ROÖS	Yok	Yok	Python
Mamun vd. [13]	2022	UCI	299	LightGBM, XGBoost, SVM, DT, LR,	SMOTE	Yok	10-fold	Yok	Belirtilmemiş

Panahizar vd. [14]	2016	Mayo Clinic Merkezi	5044	RF, LR, SVM, DT, Adaboost	Yok	Manuel bazı özellikleri sildiler	Yok	Yok	Belirtilmemiş
Erdaş ve Ölçer [15]	2020	UCI	299	OneRule	Yok	KDÖS	Yok	Yok	Belirtilmemiş
Samad vd. [16]	2019	Pensilvanya'daki kalp hastalarının sağkalım veri seti	171510	RF, LR	Yok	Yok	10-fold	Yok	Belirtilmemiş
Türkmenoğlu ve Yıldız [17]	2021	UCI	299	KNN, RF, ET	SMOTE	RF, Korelasyon Matrisi	Yok	Min-Max Normalizasyonu	Belirtilmemiş
Potur ve Erginel [18]	2021	UCI	299	NB, LR, MLP, SVM J48	Yok	Info Gain Attribute Eval (WEKA)	Yok	Yok	WEKA
Al-Dury vd. [19]	2020	İsveç'te meydana gelen ani kalp durma vakaları	45067	RF, Gradyan Artırma	Yok	Yok	10-fold	Yok	Belirtilmemiş

1.3 Kalp Hastalıklarının Türleri

Kardiyovasküler hastalıklar kalbi ve kan damarlarından birini veya her ikisini birden etkiyen hastalık olarak tanımlanmaktadır [20]. Kardiyovasküler hastalıklar, dünya genelinde meydana gelen ölümlerin en önemli nedenleri arasında sayılmaktadır [21]. Dünya sağlık örgütü kardiyovasküler hastalıkları, kalp ve kan damarları bozukluğu olarak tanımlamış ve altı farklı kategoriye ayırmıştır [22]. Bu kategoriler aşağıda alt başlıklar halinde verilmiş ve açıklanmıştır.

1.3.1 Koroner kalp hastalığı

Kalp kasını besleyen, büyük veya orta büyüklükteki kan damarlarında oluşan kalınlaşma ve esneklik kaybı gibi patolojik bozukluklardan kaynaklanmaktadır. Koroner kalp hastalığı ilerleyen aşamalarda kalp yetmezliği ve ani kalp ölümlerine neden olabilmektedir [23].

1.3.2 Serebrovasküler hastalıklar

Serebrovasküler hastalıklar, beyni besleyen kan damarlarının bir veya daha fazlasında meydana gelen damar tıkanıklığı veya patolojik hasar sonucunda beynin nörolojik fonksiyonlarının bozulmasıdır. Serebrovasküler hastalık, dünya genelinde meydana gelen ölümler arasında ikinci sırada olan ciddi bir hastalıktır [24].

1.3.3 Periferik Arter hastalığı

Periferik Arter hastalığı, aort ve koroner atardamarlar dışında kolları ve bacakları besleyen arterlerde oluşan damar sertliği veya pıhtı oluşumuna bağlı bir damar tıkanıklığı hastalığıdır [25].

1.3.4 Romatizmal kalp hastalığı

Romatizmal kalp hastalığı, streptokok bakterilerinin neden olduğu enfeksiyonlara bağlı olarak romatizmal ateşin gelişmesi daha sonra bunun sonucunda kalp kası ve kalp kapakçıklarında hasar oluşmasıdır. Romatizmal ateş, genellikle çocuk yaşta streptokok bakterilerinin neden olduğu boğaz ağrısı

ve bademcik iltihabı ile gelişen enfeksiyonlar ile vücudun bağışıklık sisteminin aşırı tepki vermesinden kaynaklanmaktadır [26].

1.3.5 Konjenital (Doğumsal) kalp hastalığı

Doğumsal kalp hastalığı, kalp gelişimi ve kalp fonksiyonlarını etkileyen, yeni doğan bebeklerde görülen bir hastalık türüdür. Doğumsal kalp hastalıklarının; kromozom anomaliler, viral enfeksiyonlar, bazı hormonlar ve anti epileptik ilaçlar gibi birçok kalıtsal ve çevresel etkenden dolayı gelişebildiği düşünülmektedir [27].

1.3.6 Derin Ven Trombozu ve Pulmoner Emboli

Derin Ven Trombozu ve Pulmoner Emboli, bacaklardan çıkıp akciğerlere kanı taşıyan damarlarda (Derin Ven) ve pulmoner arterlerde (akciğer atardamarları) oluşan pıhtıların sebep olduğu bir hastalık türüdür. Derin Ven Trombozu, özellikle ileriki yaş döneminde sıklıkla rastlanılan önlenebilir bir hastalık türüdür [28]. Pulmoner Emboli hastalığında, pulmoner arterlerde oluşan pıhtıların koparak akciğerlere ulaşma ve burada tıkanıklığa sebep olma riski bulunmaktadır. Bu yüzden, acil müdahale gerektiren oldukça önemli bir hastalık türüdür [29].

1.4 Kalp Hastalıklarının Belirtileri

Kan damarlarından kaynaklanan hastalıklar çoğu zaman hiçbir belirti göstermez. Hastanın kalp krizi veya inme geçirmesi kalp hastalığının bir belirtisi olarak gelişebilmektedir. Dünya sağlık örgütü kalp krizi ve inmenin yaygın belirtilerini şu şekilde tanımlamıştır [22]:

• Göğüsün ortasında şiddetli ağrı olması.

• Her iki kolda, sol omuzda, çenede, dirseklerde ve sırtta ağrı veya rahatsızlık hissedilmesi.

• Kollarda, bacaklarda veya yüz bölgesinde uyuşma görülmesi.

• Konuşmakta zorlanma, konuşulanı kavramada güçlük yaşama veya kafa karışıklığı meydana gelmesi.

• Sebebi bilinmeyen şiddetli baş ağrıları.

• Baygınlık geçirme veya bilinç kaybı yaşanması.

• Yürüme güçlüğü, baş dönmesi veya denge kaybı yaşanması.

1.5 Kalp Hastalıkları için Risk Faktörleri

Kalp hastalıkları için değiştirilemez risk faktörlerinin başında ırk, yaş, cinsiyet ve aile öyküsü gibi özellikler bulunmaktadır [30]. Bunun yanı sıra, tütün kullanımı, yüksek tansiyon, kolesterol, obezite, diyabet, dislipidemi, fiziksel hareketsizlik, alkol kullanımı ve hiperlipidemi kalp hastalıklarına sebep olan riskler arasında geleneksel risk faktörleri olarak kabul edilmektedir. Bu risklerin çoğu değiştirilebilir ve kontrol altına alınabilmektedir. Kalp hastalıklarını etkileyen geleneksel risklerin azaltılması kalp hastalıkların artış hızını ve etkilerini önemli ölçüde azaltacağı düşünülmektedir [21, 22, 31]. Bu anlamda, geliştirilecek bir yöntem ile kalp hastalıklarının sağkalım tahminin yapılması ve ölüme etki eden risk faktörlerinin tespit edilerek önlemler alınması oldukça önemli bir konudur.

2. MAKİNE ÖĞRENMESİ

Bu bölümde, makine öğrenmesi tanımlanmış ve makine öğrenmesine ait öğrenme türleri detaylı bir şekilde verilmiştir.

2.1 Makine Öğrenmesinin Tanımı

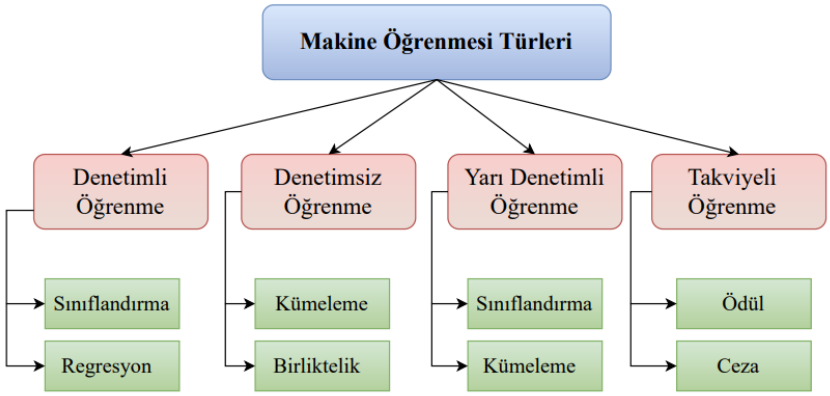
Makine öğrenmesi, insana ait öğrenme yapısından esinlenerek geliştirilmiş, çıkarımlarını matematiksel ve istatistiğe dayalı yöntemler kullanarak gerçekleştiren ayrıca yaptığı bu çıkarımlar ile tahminde bulunan yapay zekânın bir alt dalıdır [32]. Makine öğrenmesinin daha modern tanımını Tom Mitchell şu şekilde yapmıştır; *P* bilgisayar programının ölçülebilen performansını, *G* görev sınıfını ve *D* ise deneyimlerini ifade etmek üzere: Bir bilgisayar programının performansı *P* ile ölçülen *G*'deki bir görevdeki performansı *D* deneyimi ile gelişiyorsa bu programın *D* deneyiminden öğrendiği söylenebilir [33].

Makine öğrenmesi, geleneksel programlama teknikleri ile çözülemeyen karmaşık ve zor problemleri çözmek için mevcut veri kümelerinden öğrenerek modeller oluşturmak için kullanılmaktadır [34].

Makine öğrenmesi günümüzde sağlık hizmetleri, doğal dil işleme, finansal risk değerlendirmesi, robot sistemlerinin kontrolü, görüntü işleme, siber güvenlik, IoT sistemleri, akıllı şehirler, tarım, hava tahmini gibi birçok sınıflandırma, kümeleme ve regresyon probleminde yaygın olarak kullanılmaktadır.

2.2 Makine Öğrenmesi Türleri

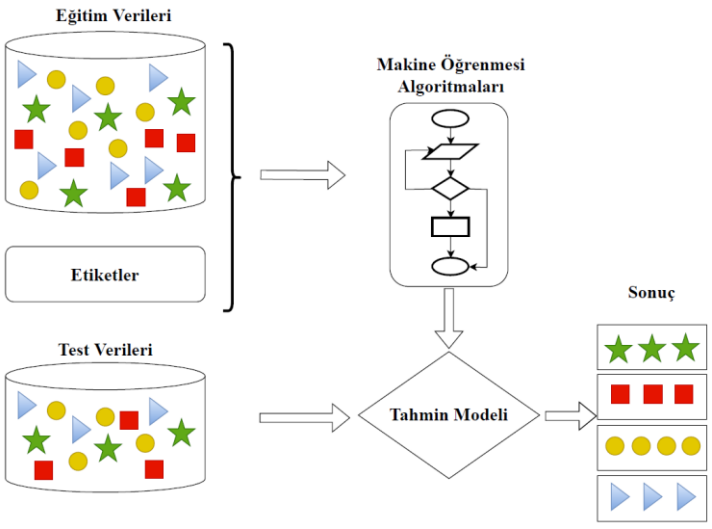
Makine öğrenmesi, öğrenme türlerine göre Denetimli Öğrenme (Supervised Learning), Denetimsiz Öğrenme (Unsupervised Learning), Yarı Denetimli Öğrenme (Semi-Supervised Learning) ve Takviyeli Öğrenme (Reinforcement Learning) olmak üzere dört farklı kategoriye ayrılmaktadır [35, 36]. Yapılan tez çalışması kapsamında bu öğrenme türlerinden denetimli öğrenme algoritmaları kalp hastalarının sağkalım tahmininde kullanılmıştır.



Şekil 2.1 Makine öğrenmesinin türleri

2.2.1 Denetimli öğrenme

Denetimli öğrenme yönteminde, modelin örnek girdi verisine karşılık gelen hedef değişken değerlerini eşleştirerek aralarındaki ilişkiyi öğrenmesi hedeflenir. Eğitimi tamamlanan modele, test verilerinden bir grup girdi değeri verilerek hedefe en yakın çıktıyı elde etmesi beklenir. Denetimli öğrenme problemleri sınıflandırma ve regresyon problemleri olarak ikiye ayrılmaktadır. Regresyon problemlerinde, veriler arasındaki ilişkiye göre matematiksel bir bağıntı oluşturularak tahmin işlemi gerçekleştirilmektedir. Sınıflandırma problemlerinde ise mevcut veri setinden öğrenme gerçekleştirilerek elde edilen model ile gelecek verilerin hedef sınıfının tahmin edilmesi işlemi gerçekleştirilmektedir. Regresyon probleminde, tahmin edilmeye çalışılan hedef değişken sürekli iken sınıflandırma problemlerinde ise ayrık haldedir [35, 37].

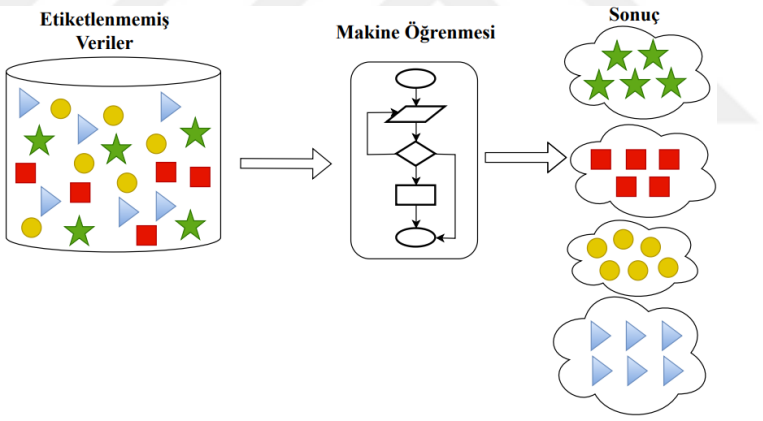


Şekil 2.2 Denetimli makine öğrenmesi çalışma prensibi

2.2.2 Denetimsiz öğrenme

Denetimsiz öğrenme yönteminde, etiketlenmemiş veri kümelerinin arasındaki gizli örüntüler ortaya çıkarılmaya çalışılmaktadır. Dolayısıyla bu yöntemde denetimli öğrenmenin aksine modelin eğitim aşaması bulunmaz. Geliştirilen model, eldeki verilerden yola çıkarak kendi kendine öğrenir ve işlem sonunda birbirine en yakın değerleri gruplandırılarak kümeleme işlemini gerçekleştirmiş olur. Daha önce karşılaşılmayan durumların anlamlandırılmasında kullanılan etkili bir yöntemdir.

Denetimsiz öğrenme yöntemlerinde, Kümeleme ve Birliktelik Kuralı Analizi yaygın olarak kullanılan iki tekniktir [35]. Kümeleme tekniği, hangi grupta olduğu bilinmeyen verinin mevcut veriler ile arasındaki yakınlık ve uzaklık gibi benzerlik ölçütlerine göre analiz edilerek bir gruba atanması işlemidir. Birliktelik Kuralı Analizi, mevcut verilerin analiz edilerek birliktelik davranışlarının tespit eden ve buna göre yapılacak olan analizleri destekleyen bir tekniktir.



Şekil 2.3 Denetimsiz makine öğrenmesi çalışma prensibi

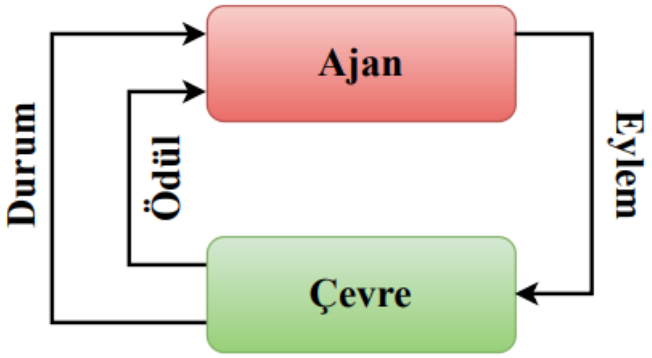
2.2.3 Yarı denetimli öğrenme

Denetimli öğrenme ve denetimsiz öğrenme yöntemleri bazı dezavantajlara sahiptir. Örneğin, denetimli öğrenme algoritmalarında bir uzman tarafından verilerin etiketlenmesine ihtiyaç duyulur. Bu da büyük boyutlu verilerin etiketlenmesi sırasında problem oluşturabilmektedir. Denetimsiz öğrenme algoritmalarında ise uygulama spektrumunun sınırlı olması problemi bulunmaktadır. Yarı denetimli öğrenme, denetimli ve denetimsiz öğrenme yöntemlerinin bu dezavantajlarından baş edebilmek için geliştirilmiştir. Bu

öğrenme yönteminde öncelikle, aynı küme içerisinde bulunan etiketsiz veriler etiketli veriler yardımıyla etiketlenir. Daha sonra, modelin eğitiminde etiketli ve sonradan etiketlenmiş veriler birlikte kullanılır [35, 38].

2.2.4 Takviyeli öğrenme

Takviyeli öğrenmede ajan (agent) adı verilen öğrenme süreci içerisinde olan bir sistem ve ödül-ceza (reward) mekanizması bulunmaktadır. Ajan hedefine ulaşmak için bulunduğu çevre içerisinde aksiyonlar alır ve bu aksiyonlara göre ajana ödül veya ceza verilir. Böylece, öğrenme sürecinde ağırlıklar güncellenerek en iyi sonuca ulaşmaya çalışır. Bu süreç içerisinde ajan yanlış bir seçim yaptığında cezalandırılırken iyi bir seçim yaptığında ise ödüllendirilir. Takviyeli öğrenmenin en iyi örneği markov karar sürecidir. Takviyeli öğrenmede, geçmişte verilen kararlardan elde edilen deneyimler ile hedefe ulaşılmaya çalışır [39].



Şekil 2.4 Takviyeli öğrenme çalışma prensibi

3. MATERYAL VE YÖNTEM

Bu bölümde, çalışmada kullanılan materyaller, makine öğrenmesi algoritmaları ve önerilen yöntemler detaylı bir şekilde açıklanmıştır.

3.1 Materyal

Tez çalışmasında kullanılan materyaller; veri seti, uygulama ortamı, sentetik veri örnekleme yöntemi, kullanılan veri ön işleme adımları ve makine öğrenmesi algoritmalarını içermektedir. Kullanılan materyaller alt başlıklar halinde aşağıda verilmiştir.

3.1.1 Veri seti

Bu çalışmada, UCI Kalp yetmezliği klinik kayıtları veri seti [6] kullanılarak 299 kalp hastasının sağkalımı makine öğrenmesi algoritmaları ile tahmin edilmiştir. Veri setinde hedef değişken olan, sağkalım bilgisi dahil 13 öznitelik bulunmaktadır. Veri setinde yer alan özelliklerin açıklamaları Tablo 3.1'de verilmiştir.

Veri setinin toplam kayıt sayısı, 13 özelliğin her birine ait ortalama ve standart sapma gibi istatiksel değerler Şekil 3.1'de gösterilmiştir.

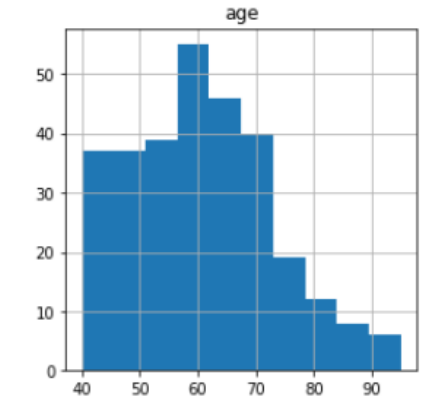
	count	mean	std	min	25%	50%	75%	max
age	299.0	60.833893	11.894809	40.0	51.0	60.0	70.0	95.0
anaemia	299.0	0.431438	0.496107	0.0	0.0	0.0	1.0	1.0
creatinine_phosphokinase	299.0	581.839465	970.287881	23.0	116.5	250.0	582.0	7861.0
diabetes	299.0	0.418060	0.494067	0.0	0.0	0.0	1.0	1.0
ejection_fraction	299.0	38.083612	11.834841	14.0	30.0	38.0	45.0	80.0
high_blood_pressure	299.0	0.351171	0.478136	0.0	0.0	0.0	1.0	1.0
platelets	299.0	263358.029264	97804.236869	25100.0	212500.0	262000.0	303500.0	850000.0
serum_creatinine	299.0	1.393880	1.034510	0.5	0.9	1.1	1.4	9.4
serum_sodium	299.0	136.625418	4.412477	113.0	134.0	137.0	140.0	148.0
sex	299.0	0.648829	0.478136	0.0	0.0	1.0	1.0	1.0
smoking	299.0	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0
time	299.0	130.260870	77.614208	4.0	73.0	115.0	203.0	285.0
DEATH_EVENT	299.0	0.321070	0.467670	0.0	0.0	0.0	1.0	1.0

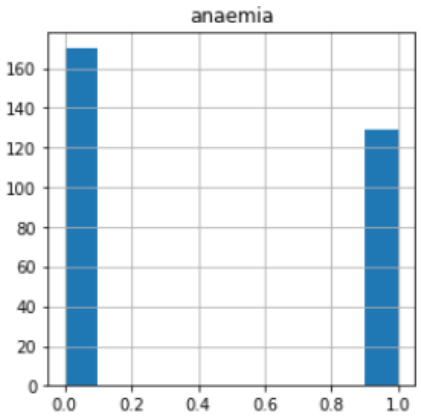
Şekil 3.1 Veri seti özelliklerinin sayısal dağılımı

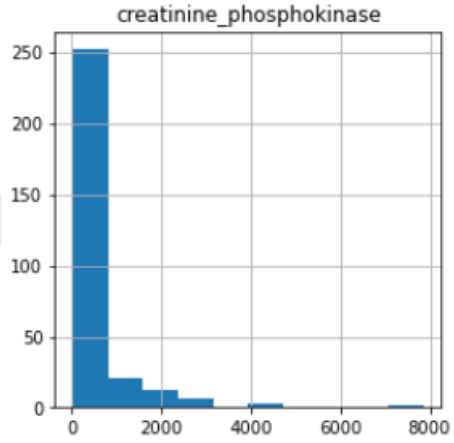
Tablo 3.1. Kalp hastalarının sağkalımına ait veri setinin özellikleri

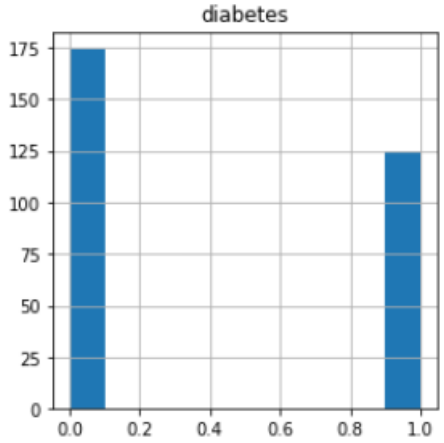
Özellik	Açıklama	Özelliklere Ait İstatistikler
Yaş (age)	Hastaların yaş bilgisi	**Hastaların yaş aralığı:** 40-95
Anemi (anaemia)	Hastalardaki anemi teşhisi	**0:** Hastalık yok (170) **1:** Hastalık var (129)
Kreatin Fosfokinaz (creatinine_phosphokinase)	Kandaki CPK enziminin miktarı (mcg/L)	**En düşük:** 23 **En yüksek:** 7861
Diyabet (diabetes)	Hastalardaki diyabet teşhisi	**0:** Hastalık yok (174) **1:** Hastalık var (125)
Enjeksiyon Fraksiyonu (ejection_fraction)	Kalbin her kasılmada pompaladığı kanın yüzdesi	**En düşük:** 14 **En yüksek:** 80
Hipertansiyon (high_blood_pressure)	Hastalardaki yüksek tansiyon teşhisi	**0:** Hastalık yok (194) **1:** Hastalık var (105)
Trombositler (platelets)	Kanda bulunan trombositler (Kan pulcukları)	**En düşük:** 25,01 **En yüksek:** 850
Serum Kreatinin (serum_creatinine)	Kanda bulunan kreatinin düzeyi	**En düşük:** 0,5 **En yüksek:** 9,40
Serum Sodyum (serum_sodium)	Kanda bulunan sodyum seviyesi (mEq/L cinsinden)	**En düşük:** 114 **En yüksek:** 148
Cinsiyet (sex)	Hastanın cinsiyeti	**0:** Kadın (105) **1:** Erkek (194)
Sigara Kullanımı (smoking)	Hastanın sigara içme alışkanlığının bulunup bulunmaması	**0:** Kullanmıyor (203) **1:** Kullanıyor (96)
Zaman (Time)	Hastanın takip süresi (Gün cinsinden)	**En düşük:** 4 **En yüksek:** 285
Sağkalım Durumu (DEATH_EVENT)	Hastanın takip süresi içerisinde sağkalım veya ölüm durumu	**0:** Hayatta kaldı (203) **1:** Öldü (96)

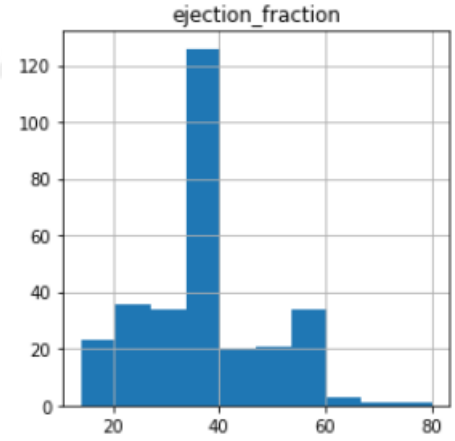
Veri setine ait özelliklerin, kendi içerisindeki sayısal değerlerinin dağılımlarını gösteren histogram grafikleri Şekil 3.2 gösterilmiştir.

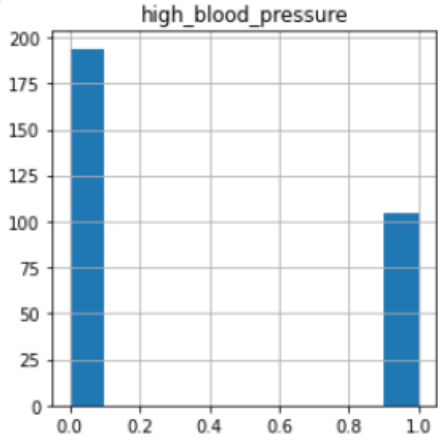

age
50
40
30
20
10
0
40
50
60
70
80
90

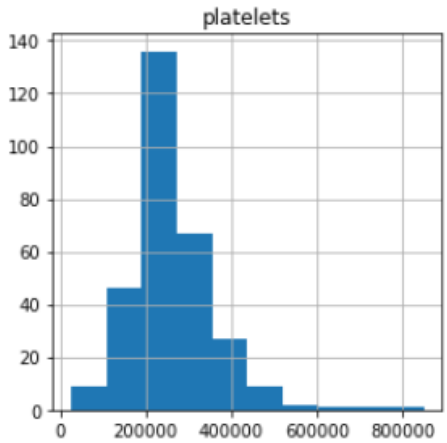

anaemia
160
140
120
100
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0

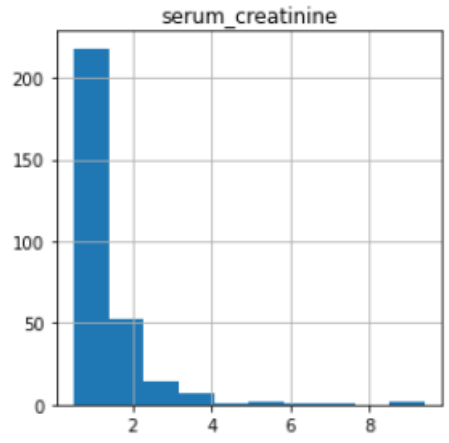

creatinine_phosphokinase
250
200
150
100
50
0
0
2000
4000
6000
8000

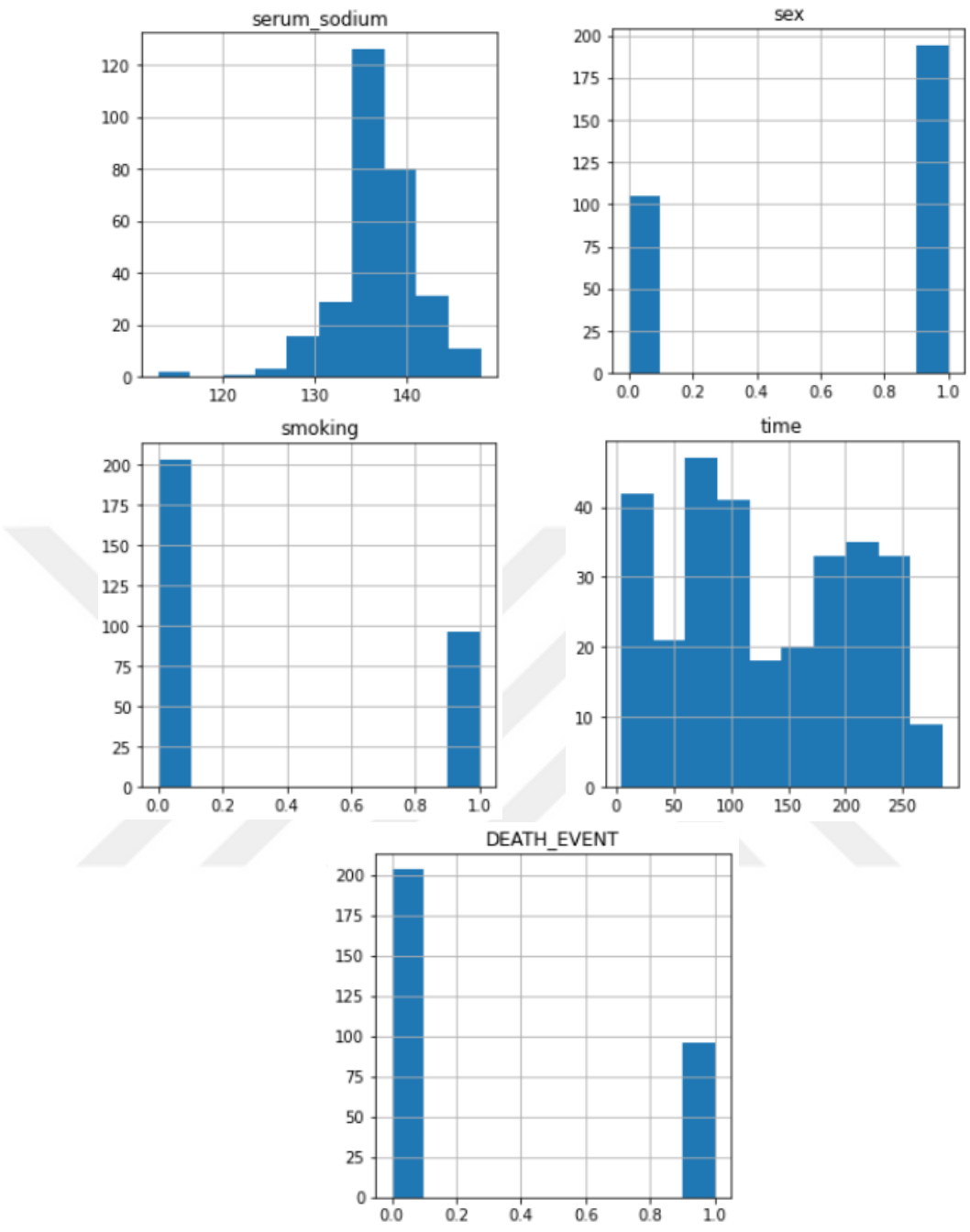

diabetes
175
150
125
100
75
50
25
0
0.0
0.2
0.4
0.6
0.8
1.0


ejection_fraction
120
100
80
60
40
20
0
20
40
60
80


high_blood_pressure
200
175
150
125
100
75
50
25
0
0.0
0.2
0.4
0.6
0.8
1.0


platelets
140
120
100
80
60
40
20
0
0
200000
400000
600000
800000

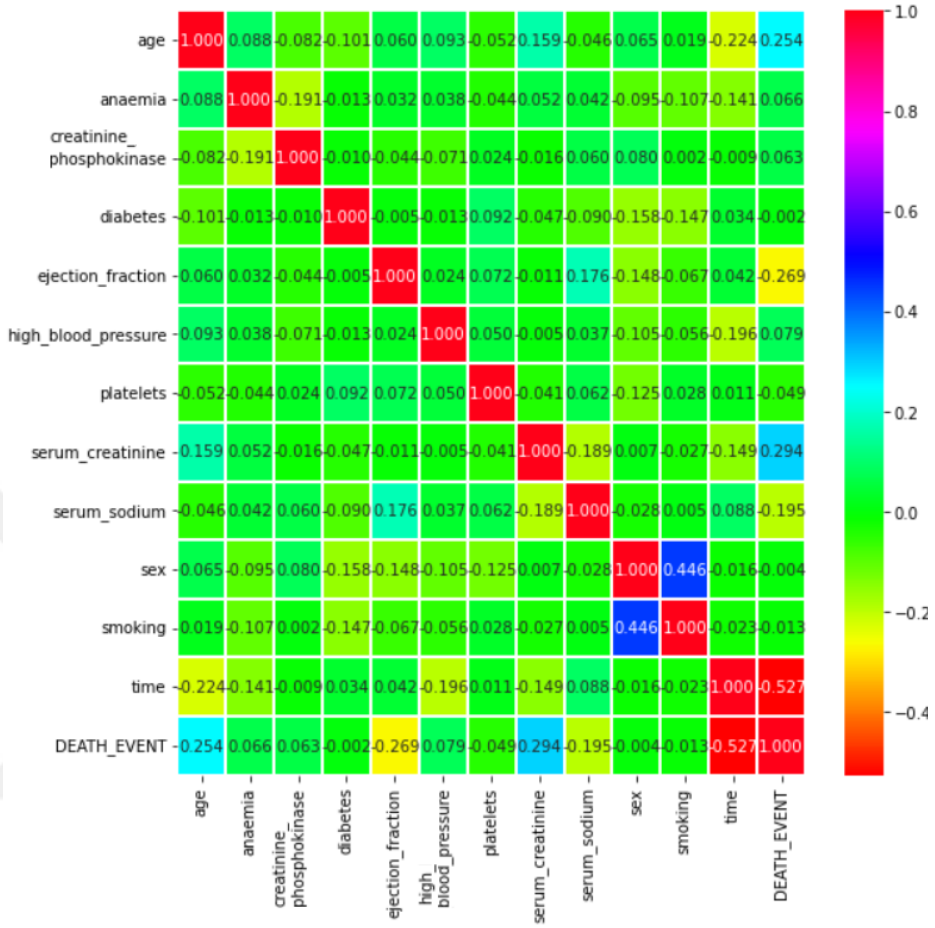

serum_creatinine
200
150
100
50
0
2
4
6
8



Şekil 3.2 Özniteliklerin dağılımlarını gösteren histogram grafikleri

Verilere ait korelasyonları gösteren ısı haritası Şekil 3.3'te verilmiştir. Bu haritaya göre hastaların sağkalım ve ölüm bilgisini tutan hedef sınıfımız "*DEATH_EVENT*" ile diğer sınıflar arasında pozitif ve negatif ilişki olduğu görülmektedir. Isı haritasına göre hedef sınıf ile "*time*", "*serum_sodium*" ve "*ejection_fraction*" özellikleri arasında negatif bir ilişki olduğu dikkat

çekmektedir. Bununla birlikte, hedef sınıf olan "*DEATH_EVENT*" ile "*age*" ve "*serum_creatinine*" arasında pozitif bir ilişki olduğu gözlemlenmektedir.



Şekil 3.3 Özellikler arası korelasyonları gösteren ısı haritası

3.1.2 Uygulama ortamı

Tez çalışması kapsamında yapılan deneyler, Intel i7 işlemci, 16 GB Ram NVIDIA Geforce GTX ekran kartı ve Windows 10 işletim sistemine sahip bilgisayar kullanılarak elde edilmiştir. Çalışma ortamı olarak Jupyter Notebook editörü, programlama dili olarak Python programlama dili tercih edilmiştir. Yöntemlerin geliştirilmesi sırasında Python 3.9.3 sürümü kullanılmıştır.

3.1.3 Sentetik Azınlık Örneklem Arttırma (SMOTE) yöntemi

Veri setinde bulunan hedef sınıftaki kayıtların sayısı, her kategori için yaklaşık olarak eşit dağılımda bulunmuyorsa, bu tür veri setlerine dengesiz veri seti denmektedir. Chawla vd. [40] tarafından 2002 yılında öne sürülen Sentetik

Azınlık Örneklem Artırma (SMOTE) yöntemi, dengesiz veri setlerinde azınlık sayıda bulunan örnek sayısını artırmak için kullanılmaktadır.

SMOTE yönteminde, sentetik veri oluşturulurken seçilen örneklerin en yakın k tane komşusu ele alınarak yapay örnekler üretilmektedir. Bu yöntemin çalışma adımları aşağıda özetlenmiştir:

Azınlık sınıfta bulunan örnek sayısı *T*, *N* değeri SMOTE ile üretilecek sentetik veri yüzdesi ve *k* değeri ise seçilen her bir örneğe en yakın komşu sayısını ifade etmektedir.

- **1.Adım:** Eğer *N* değeri %100'den daha az ise azınlık sınıfının en yakın *k* komşularından rastgele bir örnek seçilerek çoğaltılır. *N* değeri 100'e bölünerek tam sayı kısmı alınır ve tam sayı kısmı 0 olana kadar aşağıdaki adımlar devam eder.
- **2. Adım:** *SÖ* seçilen örneği, *EYÖ* seçilen örneğe en yakın örneği, *ÜÖ* de üretilen sentetik örneği ifade etmek üzere: *N* değeri sıfır değilse 1 ile *k* arasında rastgele örnek seçilir ve bu örneğin en yakın *k* tane komşusu belirlenir. Seçilen örneğe en yakın örnek seçilir. Sentetik örnek üretmek için aşağıdaki 3.1 ve 3.2 ile ifade edilen denklemler kullanılır.
- **3. Adım:** Seçilen örnek ile kendisine en yakın komşusunun farkı (S_{fark}) alınır.

$$S_{fark} = S_{SÖ} - S_{EYÖ} \quad [3.1]$$

- **4. Adım:** Denklem 3.2 kullanılarak 0 ile 1 arasında rastgele bir sayı üretilir ve *a* değerine atanır. Denklemde yer alan *rastgele ()* fonksiyonu verilen girişe göre rastgele sayı üretmektedir.

$$a = rastgele([0,1]) \quad [3.2]$$

- **5. Adım:** Denklem 3.1'de elde edilen S_{fark} değeri, *a* değeri çarpılarak en yakın komşu değeri $S_{EYÖ}$ ile toplanır. Böylelikle sentetik veri üretilmiş olur.

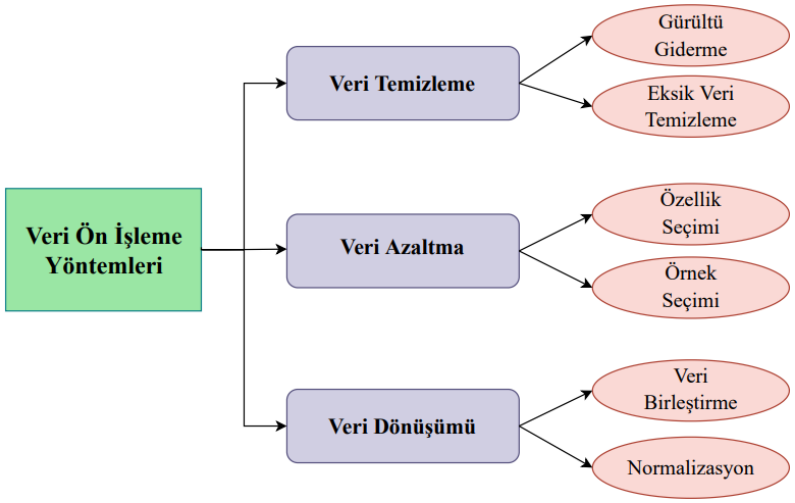
$$S_{\text{ÜÖ}} = S_{EY\text{Ö}} + (S_{fark}) * a \quad [3.3]$$

- **6. Adım:** Yukarıda verilen işlem adımları *N* değeri sıfırdan farklı olduğu sürece devam ederek yeni örnekler üretilir.

Bu çalışmada kullanılan veri setindeki hedef sınıftaki kayıt sayısı, Tablo 3.1'de ifade edildiği gibi hayatta kalan 203 ve ölen 96 olmak üzere toplam 299 hasta kaydından oluşmaktadır. Veri setindeki bu dengesiz dağılımın sağkalımı tahmin etme başarı üzerindeki olumsuz etkisini azaltmak için SMOTE yöntemi kullanılarak sentetik veri üretilmiş ve hedef sınıftaki dağılım 203 sağkalım 203 ölüm olmak üzere toplam 406 olacak şekilde eşitlenmiştir.

3.1.4 Veri ön işleme

Veri ön işleme, makine öğrenmesi algoritmalarının performansını arttırmak için uygulanan en önemli adımlarından biridir. Toplanan ham veriler, eksik veya tutarsız değerler içerme, çok küçük ve çok büyük değer aralığında olma, insan hatası veya hatalı ölçümden kaynaklı gürültülü veriler gibi durumlarla karşı karşıya kalabilmektedir. Verilerin bu kusurlarının giderilmesi kullanılacak olan makine öğrenmesi modelinin performansını önemli ölçüde arttırabilmektedir. Veri ön işleme yöntemleri Şekil 3.4'te gösterildiği gibi veri temizleme, veri azaltma ve veri dönüşümü olmak üzere temelde 3 farklı kategoriye ayrılmaktadır [41].



Şekil 3.4 Veri ön işleme yöntemleri

Tez çalışması kapsamında önerilen yöntemlerde, veri ön işleme adımlarında kullanılan Özyinelemeli Özellik Eleme yöntemi ve Yeo-Johnson veri normalizasyon tekniği kullanılmıştır.

3.1.4.1 Yeo-Johnson normalizasyon yöntemi

Yeo ve Johnson tarafından önerilen normalizasyon yöntemlerinden biri olan güç dönüşüm yöntemi, veri madenciliği ve makine öğrenmesi çalışmalarında verinin ön işleme aşamasında sıklıkla kullanılmaktadır. Veri setinde yer alan özellikler içerisinde bulunan aykırı değerleri normalleştirmek ve Gauss benzeri bir forma dönüştürmek için Yeo-Johnson güç dönüşüm işlemi uygulanmaktadır. Yeo-Johnson güç dönüşüm işleminde aykırı değerler ile yoğun bölgedeki değerler arasındaki mesafenin azaltılması hedeflenmektedir [42-44]. Yeo-Johnson dönüşümünün formülü denklem [3.4]'te ifade edilmiştir. λ değeri dönüşüm işlemi için kullanılan parametreyi ifade eder ve 0 ile 2 aralığında bir değer alır. Denklemde verilen x değeri ise dönüştürülecek olan değeri ifade etmektedir [44].

$$x^{\lambda} = \begin{cases} (x+1)^{\lambda}, & \lambda \neq 0 \; ise, \quad x \geq 0 \\ \ln(x)+1, & \lambda = 0 \text{ ise}, \quad x \geq 0 \\ -\dfrac{\left[(-x+1)^{2-\lambda}-1\right]}{2-\lambda}, & \lambda \neq 2 \text{ ise}, \quad x < 0 \\ -\ln(-x+1), & \lambda \neq 2 \text{ ise}, \quad x < 0 \end{cases} \qquad [3.4]$$

3.1.4.2 Özyinelemeli Özellik Eleme yöntemi

Öznitelik seçme yöntemlerinden biri olan Özyinelemeli Özellik Seçimi, veri setindeki tüm özelliklerden yola çıkarak hedef sınıfın tahmin edilmesine en fazla katkısı olan özellikleri ağırlıklandırmaktadır. Daha sonra, hedef sınıfın tahmin edilmesinde optimum özellik sayısına ulaşıncaya kadar düşük ağırlıklı özellikleri veri kümesinden kaldırmaktadır. Son olarak, veri setindeki özelliklerden en fazla katkısı olanlara en düşük numaradan başlayarak sıra ataması yapmaktadır. ÖÖE yöntemi, veri seti içerisinden daha az özellik sayısı ile tahmin başarısında en iyi performans gösteren alt özellik kümesine ulaşmayı hedeflemektedir [45]. Bu çalışmada kullanılan veri setinde 13 özellik bilgisi bulunmaktadır. Tez çalışmasında, her algoritma ile ÖÖE yöntemi birleştirilmiş

ve özelliklerin önem dereceleri belirlenerek her bir algoritma için alt özellik kümeleri oluşturulmuştur.

ÖÖE yöntemi makine öğrenmesi algoritmasına sarmalanarak veri kümesindeki özelliklerin önem derecelerini tahmin eder. Bu özellik seçim yöntemi algoritma ile kullanılarak veri setindeki tüm özellikleri sırayla dener ve model performansının ne kadar az özellikle daha iyi sonuç verdiğini ölçer. ÖÖE yöntemi özelliklerin tamamını denedikten sonra daha az öneme sahip özellikleri sırayla veri setinden atar ve bu işlemi tekrar eder. Son olarak birlikte kullanıldığı makine öğrenmesi algoritmasının en yüksek performans göstereceği özelliklerin rank değerini belirler. Algoritmanın "*n_features_to_select*" parametresi seçilmesi istenilen alt özellik sayısını ifade eder. Örneğin bir veri kümesinde toplam 13 özellik varsa ve en önemli 5 alt özellik bulunmak isteniyorsa bu parametreye 5 değeri verilir. Bu parametre değerine "none" değeri verilirse varsayılan olarak özellik kümesinin yarısı kadar en önemli alt özellik kümesini belirler. Tez çalışmasında ÖÖE yönteminin "*n_features_to_select*" parametresi varsayılan olarak tercih edilmiştir. Bu nedenle veri kümesinde hedef değişken hariç 12 özellik bulunduğu için ÖÖE yöntemi her algoritma için en önemli 6 özelliği ortaya çıkarmıştır.

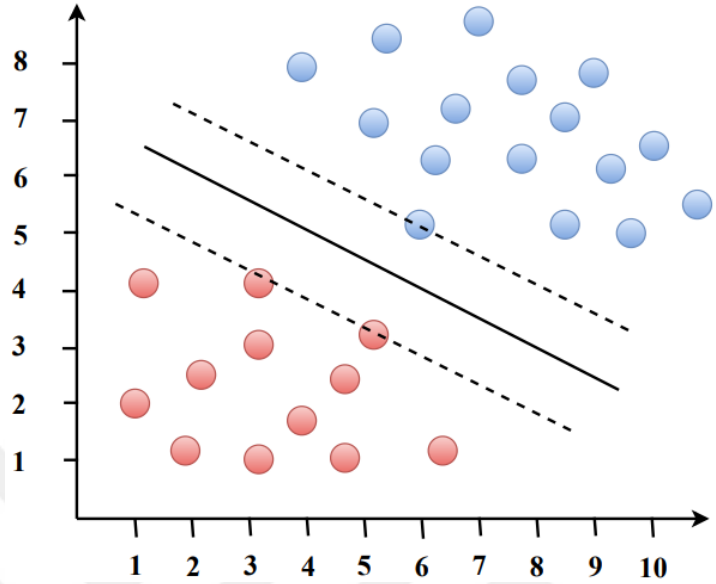
3.1.5 Makine öğrenmesi algoritmaları

Yapılan literatür çalışmasında, sağkalım tahminine yönelik birçok makine öğrenimi algoritması kullanımına rastlanmıştır. Bu algoritmalardan, yaygın olanları seçilerek tez çalışmasında kullanılmıştır. Bu algoritmalara ait bilgiler alt başlıklar halinde açıklanmıştır.

3.1.5.1 Destek Vektör Makinesi

Vapnik vd. tarafından önerilen ve sınıflandırma problemlerinde sıklıkla kullanılan Destek Vektör Makinesi (SVM), istatistiksel öğrenme ve optimizasyon teorisine dayanmaktadır. SVM sınıflandırıcısının temel amacı birbirine yakın verilerden oluşan farklı iki sınıf arasındaki mesafeyi maksimize etmek için paralel hiper düzlemler oluşturmaktır [46, 47]. Bu çalışmada, SVM sınıflandırıcısının çekirdek türü için Lineer Temelli Fonksiyon, çekirdek

katsayısı için scale ve çekirdek ön bellek boyutu için 200 Megabayt tercih edilmiştir. SVM algoritmasına ait çalışma prensibi Şekil 3.5'te gösterilmiştir.

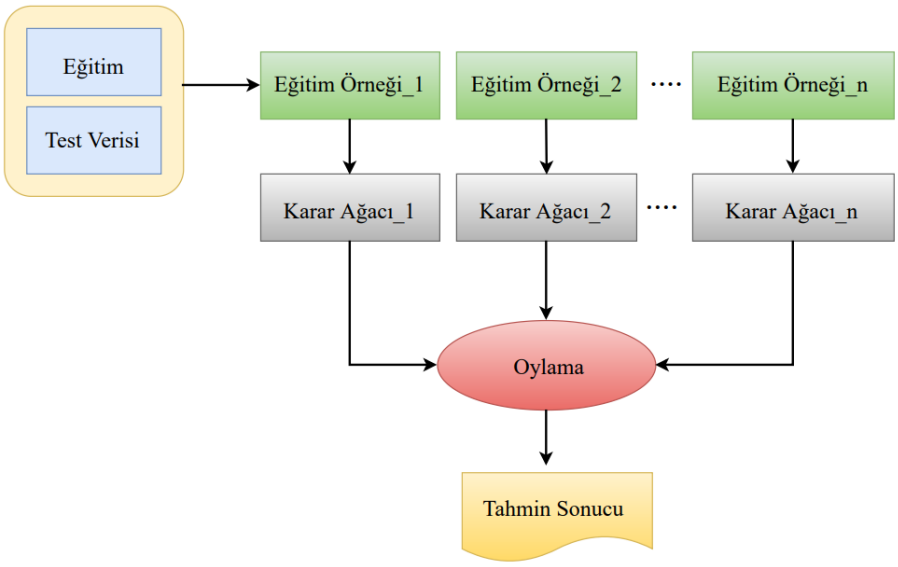


Şekil 3.5 Destek Vektör Makinesi çalışma prensibi

3.1.5.2 Rastgele Orman

Denetimli öğrenme kategorisinde yer alan Rastgele Orman (Random Forest, RF), içerisinde birden fazla karar ağacı bulunmaktadır. RF, bu ağaçlar üzerinde oylama yaparak oylama sonucunda çoğunluk oyu alan sınıf ve problemin sınıf etiketini belirler [48]. RF algoritmasının çalışma prensibi Şekil 3.6'da gösterilmiştir.

Bu çalışmada, RF ile model eğitilirken ormandaki ağaç sayısı (*n_estimators*) 100, maksimum derinlik (*max-depth*) 5 olarak belirlenmiştir.

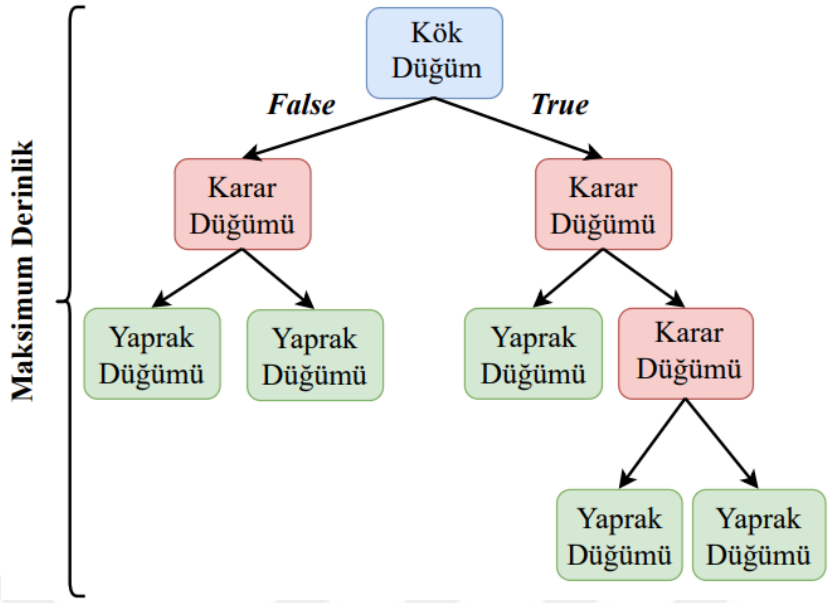


Şekil 3.6 Rastgele Orman algoritmasının çalışma prensibi

3.1.5.3 Karar Ağacı

Ağaç veri yapısına benzer bir yapıda olan denetimli öğrenme kategorisinde yer alan Karar Ağacı (DT) algoritması, kök, düğüm, dallar ve yaprak olmak üzere dört bölümden oluşmaktadır. Her bir yaprak düğüm sonucunu, düğümler özellikleri, dallar ise karar kurallarını temsil etmektedir. Kök düğümler ise veri kümesinden yola çıkarak nihai sınıflandırma sonucunu içermektedir. DT algoritmasının çalışma prensibi Şekil 3.7'de gösterilmiştir [48].

Bu çalışmada DT ile model eğitilirken maksimum derinlik değeri 1 ile 10 arasındaki tüm değerler için denenmiş ve en iyi sonuç 3 ile elde edilmiştir. Bu yüzden, yapılan deneylerde DT için maksimum derinlik değeri 3 olarak belirlenmiştir.



Şekil 3.7 Karar Ağacı algoritmasının çalışma prensibi

3.1.5.4 Aşırı Gradyan Arttırma (XGBoost)

Ölçeklenebilir bir makine öğrenmesi yöntemi olan Aşırı Gradyan Arttırma (XGBoost), ağaç güçlendirme mantığına dayanmaktadır. XGBoost, hızlı işlem yeteneği ve yüksek performansı nedeniyle yüksek başarımda sonuçlar veren bir makine öğrenmesi algoritmasıdır. 2015 yılında Kaggle web sitesi [49] üzerinden düzenlenen makine öğrenmesi yarışmalarında, kazanan 29 çözüm arasından 17'sinde XGBoost algoritması tercih edilmiş ve bunlardan 8'i modeli eğitmek için sadece XGBoost algoritmasını kullanmıştır [50, 51].

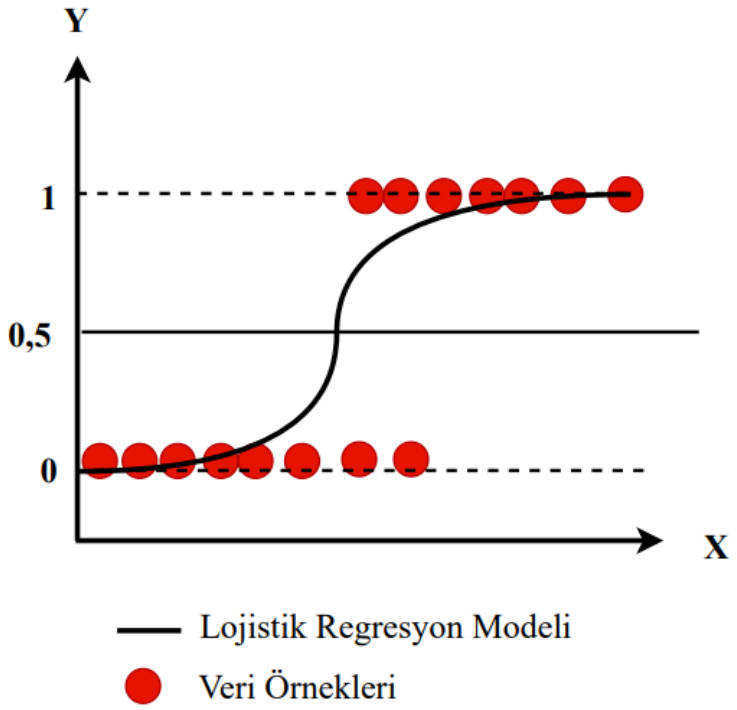
3.1.5.5 Kategorisel Arttırma (CatBoost)

Kategorisel arttırma (CatBoost), Gradyan Arttırma (Gradient Boosting) algoritmasının geliştirilmesiyle 2018 yılında ortaya çıkan bir makine öğrenmesi algoritmasıdır. Bu algoritmanın en büyük avantajı, kategorik verilerin işlenmesine olanak sağlamasıdır. CatBoost, veri sıralamasına olan bağımlılığı engellemek için sınıflandırma işlemini eğitim sürecinde gerçekleştirmektedir [52, 53].

3.1.5.6 Lojistik Regresyon

Lojistik Regresyon (LR), sınıflandırma problemleri için geliştirilmiş denetimli öğrenme kategorisinde yer alan makine öğrenmesi yöntemidir. LR'nin amacı,

bir örneğin hangi hedef sınıfına ait olduğunu tahmin etmektir. Hedef değişkenleri kategorik olduğu durumlarda sıklıkla kullanılmaktadır. LR yönteminin çalışma prensibi Şekil 3.8'de gösterilmiştir [54].



Şekil 3.8 LR yönteminin çalışma prensibi

3.2 Önerilen Yöntem

Bu bölümünde, kalp hastalığı teşhisi konulan 299 hastanın sağkalım tahmini için kullanılan metotlar ve uygulama adımları detaylı bir şekilde açıklanmıştır.

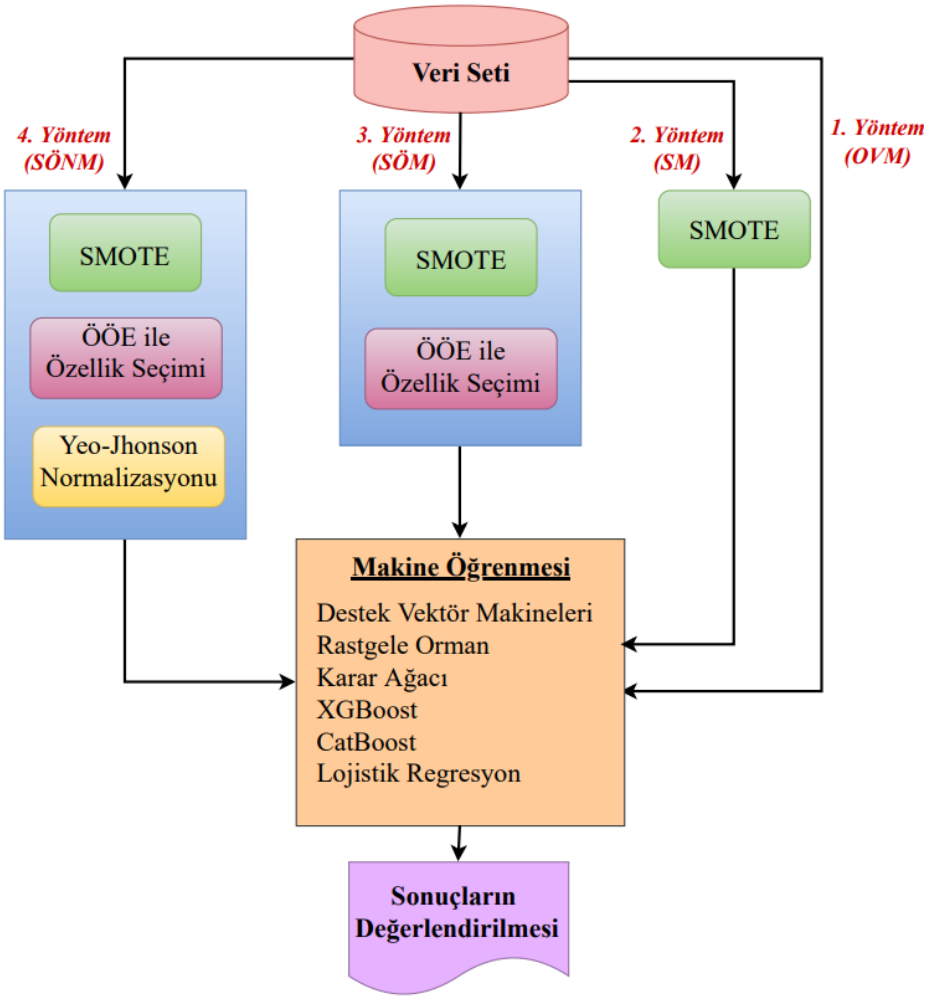
Yapılan deneysel çalışmalarda, makine öğrenmesi algoritmalarından; SVM, RF, DT, XGBoost, CatBoost ve LR kullanılarak dört faklı yöntem önerilmiştir. Sağkalım tahminine yönelik önerilen yöntemlere ait genel sistem şeması Şekil 3.9'da gösterilmiştir.

Orijinal veri ile geliştirilen Orijinal Veri seti Modeli'nde (OVM), herhangi bir özellik seçimi, veri arttırma ve normalizasyon işlemi uygulanmadan ham veri seti ile model oluşturularak sağkalım tahmini gerçekleştirilmektedir.

Veri setine SMOTE uygulanarak geliştirilen SMOTE Modelinde (SM), öncelikle dengesiz bir durumda olan veri seti SMOTE tekniği kullanılarak sentetik veri üretimi ile dengeli bir hale getirilir. Ardından, oluşturulan veri seti ile eğitim gerçekleştirilerek oluşturulan model ile sağkalım tahmini gerçekleştirilir.

SMOTE uygulanmış veri setine ÖÖE özellik seçimi uygulanması ile elde edilen üçüncü yöntem olan SMOTE özellik seçimi Modeli (SÖM), SM yöntemine benzer olarak SMOTE tekniği ile veri dengeli hale getirildikten sonra her makine öğrenimi algoritması için ÖÖE özellik seçimi uygulanarak sağkalım tahminine etki eden en önemli özellikler ortaya çıkarılır. Ardından, daha az öneme sahip olan özellikler veri setinden çıkarılarak modeller oluşturulur ve sağkalım tahmini gerçekleştirilir.

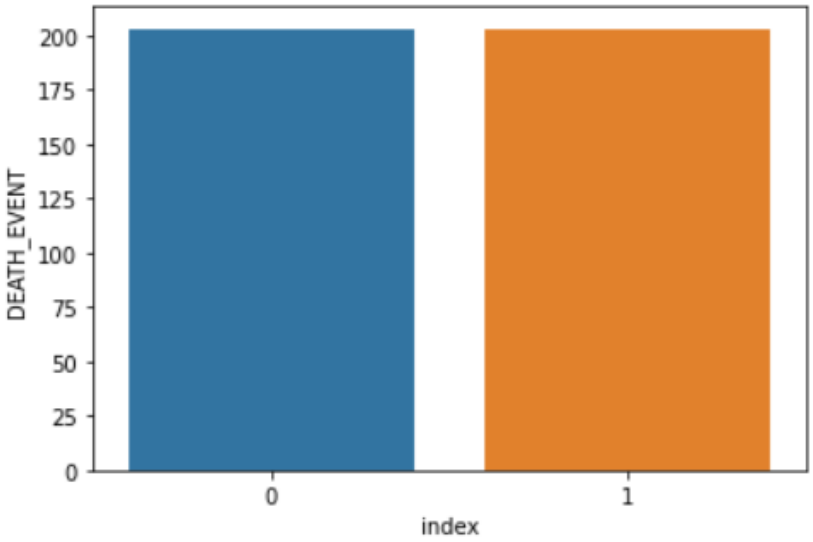
Son olarak, SÖM yöntemine normalizasyon eklenerek elde edilen; SMOTE Özellik seçimi Normalizasyon Modeli'nde (SÖNM) ise SMOTE ile dengelenmiş veri setine ÖÖE ile özellik seçimi uygulanır. Her algoritma için elde edilen alt özellik kümesi elde edilir. Modelin oluşturulmasında veri setindeki alt özellik kümesi korunurken ranking değeri 1'den büyük olan değişkenler veri setinden çıkarılır. Daha sonra elde edilen veri setine Yeo-Johnson dönüşümü uygulanır. Bu teknik uygulanırken özellik değeri 0 ve 1'den oluşan sütunlara normalizasyon işlemi uygulanmamıştır. Son olarak normalizasyon uygulanmış veri seti makine öğrenmesi algoritmasına giriş olarak verilerek sağkalım tahmini gerçekleştirilir.



Şekil 3.9 Önerilen sağkalım tahmin yöntemlerinin genel sistem şeması

3.2.1 Önerilen yöntemlerde SMOTE ile veri dengeleme

Bölüm 3.1.1'de verildiği gibi orijinal veri setinin hedef sınıfında 203 sağkalan ve 96 hayatını kaybeden hasta bulunmaktadır. Python programlama dili ile "*imbalanced-learn"* kütüphanesinin "*SMOTE*" sınıfından faydalanarak veri setine SMOTE tekniği uygulanarak sentetik veri üretimi yapılmıştır. Veri üretiminden sonra hedef değişkendeki örnek sayısı 203 sağkalım ve 203 ölüm olacak şekilde toplam 406 kayıt ile eşitlenmiştir. Veri setine, SMOTE işlemi uygulandıktan sonraki hedef sınıfın sağkalım grafiği Şekil 3.10'da gösterilmiştir.



Şekil 3.10 SMOTE uygulandıktan sonra hedef sınıfın sayısal dağılımı

3.2.2 Önerilen yöntemlerde özellik seçiminin uygulanması

Kalp hastalarının sağkalımını en fazla etkileyen özelliklerin bulunması ve sağkalımın tahmininin başarısını artırmak için SÖM ve SÖNM yöntemlerinde özellik seçimi uygulanmıştır. Bu yöntemlerde, özellik seçimi ÖÖE tekniğini kullanılarak gerçekleştirilmiştir. Altı farklı makine öğrenmesi algoritması ile elde edilen özelliklere ait önem derecesi sonuçları Şekil 3.11'de verilmiştir.

Şekillerde her bir algoritmaya ait özellik "Feature" sütunlarında gösterilirken özelliğin önem derecesi "Ranking" sütunlarında verilmiştir. Özelliğin 1 derecesine sahip olması sağkalım tahminini etkileyen en önemli faktör olduğunu göstermektedir. Şekil 3.11'de görüldüğü gibi her algoritma için Ranking değeri 1 olan özellikler en önemli özellikler olarak seçilmiştir. Bu yöntemler uygulanırken, önem derecesi yüksek olan özelliklerin modelin oluşturulması sırasında korunumu sağlanırken daha az öneme sahip olan özelliklerin veri setinden çıkarılması sağlanacaktır.

ÖÖE ile seçilen alt özellik kümeleri incelendiğinde "*age*", "ejection_*fraction*", "*serum_creatinine*" ve "*time*" özellikleri tüm algoritmalar tarafından ortak olarak seçilen özellikler olmuştur. Bununla birlikte "*sex*" sütunu RF algoritması hariç diğer tüm algoritmaların en önemli özellik listesinde seçilmiştir. Ayrıca

"*creatinin_phosphokinase*" özelliği LR ve RF algoritmaları hariç diğer tüm algoritmalar tarafından seçilen önemli özellik olmuştur.

SVM

	Feature	Ranking
0	age	1
2	creatinine_phosphokinase	1
4	ejection_fraction	1
7	serum_creatinine	1
9	sex	1
11	time	1
8	serum_sodium	2
3	diabetes	3
10	smoking	4
5	high_blood_pressure	5
6	platelets	6
1	anaemia	7

RF

	Feature	Ranking
0	age	1
2	creatinine_phosphokinase	1
4	ejection_fraction	1
6	platelets	1
7	serum_creatinine	1
11	time	1
8	serum_sodium	2
9	sex	3
1	anaemia	4
5	high_blood_pressure	5
3	diabetes	6
10	smoking	7

DT

	Feature	Ranking
0	age	1
4	ejection_fraction	1
7	serum_creatinine	1
9	sex	1
10	smoking	1
11	time	1
8	serum_sodium	2
6	platelets	3
5	high_blood_pressure	4
3	diabetes	5
2	creatinine_phosphokinase	6
1	anaemia	7

XGBoost

	Feature	Ranking
0	age	1
2	creatinine_phosphokinase	1
4	ejection_fraction	1
7	serum_creatinine	1
9	sex	1
11	time	1
6	platelets	2
8	serum_sodium	3
1	anaemia	4
3	diabetes	5
5	high_blood_pressure	6
10	smoking	7

CatBoost	Feature	Ranking
0	age	1
2	creatinine_phosphokinase	1
4	ejection_fraction	1
7	serum_creatinine	1
9	sex	1
11	time	1
6	platelets	2
8	serum_sodium	3
10	smoking	4
1	anaemia	5
5	high_blood_pressure	6
3	diabetes	7

LR	Feature	Ranking
0	age	1
4	ejection_fraction	1
7	serum_creatinine	1
8	serum_sodium	1
9	sex	1
11	time	1
2	creatinine_phosphokinase	2
6	platelets	3
5	high_blood_pressure	4
3	diabetes	5
10	smoking	6
1	anaemia	7

Şekil 3.11 SÖM ve SÖNM yöntemi için her bir algoritmaya ÖÖE tekniği uygulanması ile elde özelliklere ait önem dereceleri

3.2.3 Önerilen yöntemlerde normalizasyonun uygulanması

SÖNM yönteminde, aykırı değerlerin modelin tahmin performansı üzerindeki etkisini araştırmak ve normalizasyon uygulamak için Yeo-Johnson dönüşüm tekniğinden faydalanılmıştır. Normalizasyon işlemi, veri setinde yer alan özelliklerden 0 veya 1 değerine sahip özellikler dışındaki özelliklere uygulanmıştır. Örneğin, SVM algoritması için seçilen özelliklerden biri cinsiyet özelliğidir. Cinsiyet özelliği veri setinde kadın için 0 erkek için 1 değeri atandığı için bu özelliğe normalizasyon uygulanmamıştır.

3.2.3 Kalp hastalarının sağkalımının tahmin edilmesi

Kalp hastalarının sağkalım tahminini gerçekleştirmeye yönelik OVM, SM, SÖM ve SÖNM olmak üzere dört farklı yöntem önerilmiştir. Bu yöntemler, Bölüm 3.2'de detaylı bir şekilde verilmiştir. Önerilen bu yöntemler, SVM, RF, DT, XGBoost, CatBoost ve LR olmak üzere her bir algoritma ile kalp hastalarının sağkalım tahminini gerçekleştirmek için uygulanmıştır. Ayrıca, modelin eğitiminde 5-fold ve 10-fold olmak üzere iki farklı çapraz doğrulama uygulanmıştır. Bu anlamda, dört yöntem, altı farklı makine öğrenmesi algoritması ve iki farklı çapraz doğrulama için toplamda 48 farklı deney gerçekleştirilmiştir. Daha sonra algoritmaların çalışma süreleri hesaplanmış ve önerilen dört yöntem için altı makine öğrenmesi algoritmalarının çalışma

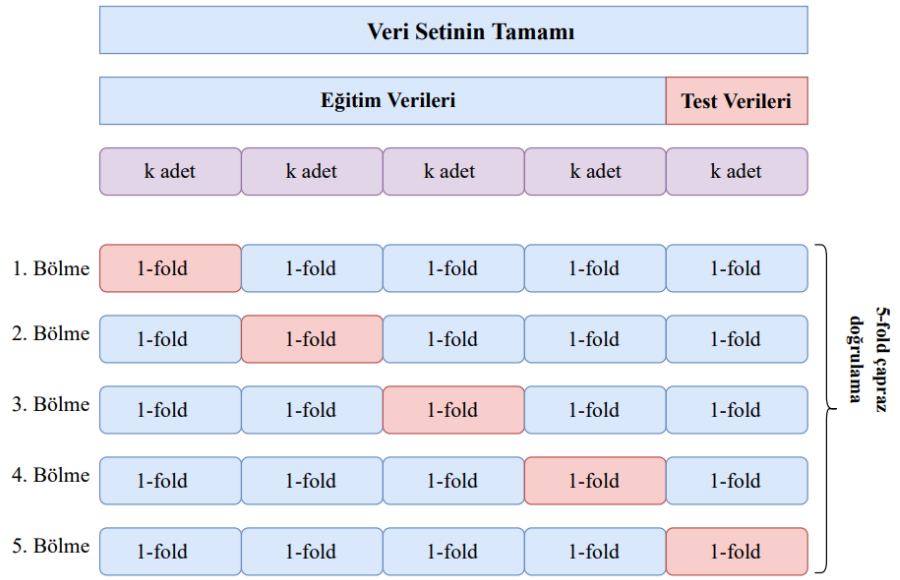
zamanları karşılaştırılmıştır. Son olarak veri setini %80 eğitim, %20 test ve %70 eğitim, %30 test şeklinde bölerek dört yöntem için altı makine öğrenmesi algoritması kullanılarak sağkalım tahmini yapılmıştır. Deneysel çalışmalar sonucunda elde edilen bulgular Bölüm 4'te paylaşılmıştır.

3.3 Geliştirilen Modeller için Performans Değerlendirme Metrikleri

Kalp hastalarının sağkalım tahminini gerçekleştirmeye yönelik geliştirilen modellerin performansının değerlendirilmesinde literatürde yaygın olarak kullanılmakta olan karmaşıklık matrisi (confusion matrix), doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve f1-skor'dan (f1-score) faydalanılmıştır.

3.3.1 Çapraz Doğrulama

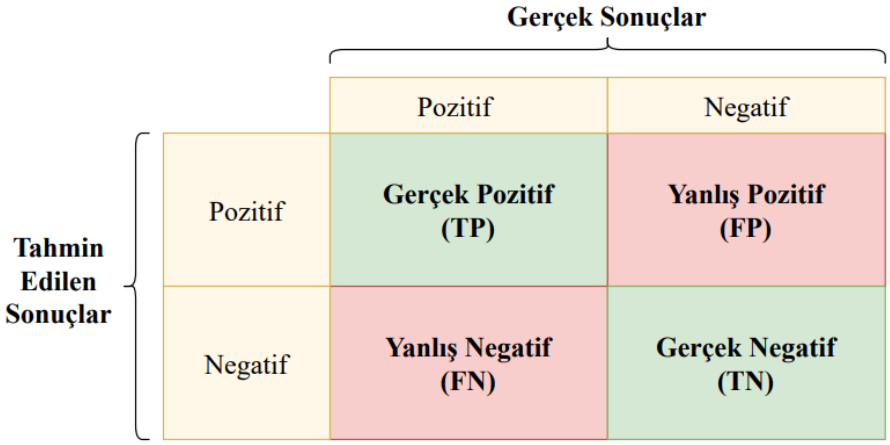
K-fold çapraz doğrulama işleminde veri seti *k* tane alt veri kümesine bölünür. Daha sonra, rastgele bir *k* veri kümesi doğrulama işlemi için belirlenerek kalan *k-1* adet veri kümesi modelin eğitimi için kullanılır ve doğruluk değeri hesaplanır. Ardından, başka bir *k* veri kümesi test işlemi için ayrılarak kalan *k-1* tane veri kümesi ile model eğitilir. Bu işlem bütün *k* veri kümeleri doğrulama işleminde kullanılana kadar yani *k* defa tekrar edilir. Son olarak elde edilen *k* adet doğruluk değerinin ortalaması alınır ve nihai sonuç elde edilir [55]. 5-fold çapraz doğrulamaya ait temsili görsel Şekil 3.13'te ifade gösterilmiştir [56]. Sınıflandırma problemlerinde çapraz doğrulamanın kullanılmasının amacı veri setindeki her bir örneğin doğrulamada kullanılmasını sağlayarak modelin aşırı öğrenmesinin önüne geçilmektir. Tez çalışmasında 5-fold ve 10-fold çapraz doğrulama işlemi uygulanmıştır.



Şekil 3.12 5-fold çapraz doğrulama temsili gösterimi

3.3.2 Karmaşıklık Matrisi

Sınıflandırma problemlerinde modelin başarısını ölçmek için gerçek değerler ve tahmin edilen değerlere ait sayısal bilgileri ifade eden Karmaşıklık Matrisi (Confusion Matrix) adı verilen bir matris oluşturulmuştur [57]. Geliştirilen modelin performansının değerlendirilmesinde kullanılan metriklerden, doğruluk (accuracy), kesinlik (precision), duyarlılık (recall) ve f1-skor (f-score) hesaplanırken karmaşıklık matrisi kullanılır. Karmaşıklık matrisi Şekil 3.12'de gösterilmiştir. Tez çalışmasında kullanılan veri setinde sağkalan hastalar 0 (negatif) ölen hastalar 1 (pozitif) olarak etiketlenmiştir.



Şekil 3.13 Karmaşıklık Matrisi

Şekil 3.12'de görüldüğü gibi karmaşıklık matrisinde yer alan ifadeler:

- **Gerçek Pozitif (True Positives, TP):** Gerçek sonucun ölüm olduğu ve modelin de ölüm olarak tahmin ettiği değerlerdir.
- **Gerçek Negatif (True Negatives, TN):** Gerçek sonucun hayatta kalma olduğu ve modelin de hayatta kaldı olarak tahmin ettiği değerlerdir.
- **Yanlış Pozitif (False Positives, FP):** Gerçek sonucun hayatta olduğu ancak model tarafından sonucun ölüm olarak tahmin edildiği değerlerdir.
- **Yanlış Negatif (False Negatives, FN):** Gerçek sonucun ölüm olduğu ancak model tarafından sonucun hayatta kaldı olarak tahmin edildiği değerlerdir.

3.3.3 Doğruluk

Doğruluk değeri, doğru tahmin edilen ölüm sayısı ve hayatta kalma değerlerin toplam tahmin sayısına bölünmesi ile elde edilmektedir. Doğruluk değeri, Denklem 3.4'te ifade edildiği gibi hesaplanmaktadır. Makine öğrenmesi algoritmaların performans değerlendirmesinde sıklıkla kullanılan bir metriktir [58].

$$Doğruluk = \frac{(|TP| + |TN|)}{(|TP| + |TN| + |FP| + |FN|)} \quad [3.5]$$

3.3.4 Kesinlik

Kesinlik değeri, sınıflandırma sonucunda gerçek pozitif olarak sınıflandırılan değerlerin, tahmin edilen toplam pozitif sayısına bölünmesiyle elde edilir. Kesinlik değeri, Denklem 3.5'te gösterildiği şekilde hesaplanmaktadır [58].

$$Kesinlik = \frac{(|TP|)}{(|TP| + |FP|)} \quad [3.6]$$

3.3.5 Duyarlılık

Duyarlılık değeri, doğru tahmin edilen pozitif değerlerin sayısının gerçekte pozitif olarak etiketlenmiş örnek sayısına bölünmesi ile elde edilir. Gerçekte pozitif olarak etiketlenen örnekler *TP* ve *FN* değerlerin toplamından oluşmaktadır [58].

$$Duyarlılık = \frac{(|TP|)}{(|TP| + |FN|)} \quad [3.7]$$

3.3.6 F1-Skor

Makine öğrenmesi yöntemlerinde sınıflandırma başarısını ölçmek için doğruluk, kesinlik ve duyarlılık değerleri çoğu zaman tek başına yeterli değildir. Bu nedenle model değerlendirmelerinde f1-skor (f-score) sıklıkla kullanılmaktadır. F1-skor değeri kesinlik ve duyarlılık değerlerinin harmonik ortalamasının alınması sonucu hesaplanmaktadır. F1-skor değeri denklem 3.6' da gösterildiği gibi hesaplanmaktadır [58].

$$F1 - Skor = \frac{(2 * Kesinlik * Duyarlılık)}{(Kesinlik + Duyarlılık)} \quad [3.8]$$

4. BULGULAR

Bu bölümde, 3. Bölümde detaylı bir şekilde açıklanan yöntemlere ait deneysel çalışma sonuçları verilmiş ve yorumlanmıştır. Elde edilen sonuçlar, doğruluğa dayalı, kesinliğe dayalı, duyarlılığa dayalı ve f1-skor dayalı sonuçlar olmak üzere alt başlıklar halinde açıklanmıştır.

4.1 Doğruluğa Dayalı Performans Sonuçları

Doğruluğa dayalı deneyler OVM, SM, SÖM ve SÖNM olmak üzere önerilen dört yöntem ve altı farklı makine öğrenmesi algoritmasıyla birlikte çalıştırılmıştır. Ayrıca her bir deney için 5-fold ve 10-fold çapraz doğrulama işlemi ayrı ayrı çalıştırılmıştır. Önerilen yöntemlerin doğruluğa dayalı sonuçları Tablo 4.1 ve Tablo 4.2'de verilmiştir.

Bu sonuçlara göre, OVM yöntemi ile çalıştırılan bütün algoritmaların performansı diğer üç yönteme kıyasla en kötü performansı göstermiştir. Ayrıca OVM yöntemi içerisinde en başarılı algoritma %77,2259 doğruluk değeri ile LR olmuştur.

SM yöntemi ile SVM algoritmasının doğruluk performansı düşerken diğer beş algoritmanın performansı önemli ölçüde yükselmiştir. Bununla birlikte, SM yöntemi için en yüksek doğruluk başarısına sahip algoritma 10-fold çapraz doğrulama ile %83,6341 doğruluk değeri ile RF olmuştur.

SÖM yönteminde SVM algoritmasının başarısı yaklaşık %20 oranında artarak 5-fold için %71,4754'e 10-fold için %72,0914'e yükselmiştir. Buna ek olarak SÖNM yönteminde DT ve LR algoritmasının doğruluk başarısı hem 5-fold hem de 10-fold için yükselirken RF, XGBoost ve CatBoost'un performansı kısmen düşüş göstermiştir. Ayrıca, SÖM yönteminde en başarılı sağkalım tahmini yapan algoritma 10-fold çapraz doğrulama sonucunda %84,3414 doğruluk başarısı ile LR olmuştur.

SÖNM yönteminde hem 5-fold hem de10-fold için SVM, DT, XGBoost ve CatBoost algoritmalarının performansı önemli ölçüde yükselirken LR ve RF algoritmalarının doğruluk başarısı düşüş göstermiştir. Bu durumda Yeo-Johnson normalizasyon yönteminin SVM, DT, XGBoost ve CatBoost algoritmalarını olumlu yönde etkilerken LR ve RF algoritmalarını olumsuz yönde etkilediği söylenebilir. Altı farklı algoritmanın önerilen dört farklı yöntem için çalıştırılması sonucu kalp hastalarının sağkalımını en iyi tahmin eden model SVM algoritmasının SÖNM yöntemi ile kullanılması sonucunda 5-fold çapraz doğrulama ile %86,4649 elde edilmiştir.

Tablo 4.1 Önerilen dört yönteme 5-fold çapraz doğrulama uygulanması sonucu elde edilen doğruluk değerleri

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	67,8926	51,7163	71,4754	86,4649
RF	71,2316	81,1020	80,6022	79,3646
DT	63,5593	74,1915	75,4260	79,8524
XGBoost	60,2259	73,6976	70,4817	80,8491
CatBoost	70,5593	79,8614	79,3736	81,3369
LR	77,2259	78,6118	84,0250	81,5477

Tablo 4.2 Önerilen dört yönteme 10-fold çapraz doğrulama uygulanması sonucu elde edilen doğruluk değerleri

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	67,8965	52,4329	72,0914	85,6090
RF	77,2298	83,6341	82,6402	82,1219
DT	73,2298	74,5060	77,6707	80,8658
XGBoost	71,8965	80,9329	78,4878	80,9329
CatBoost	76,8965	82,6463	80,1768	81,6463
LR	79,5632	78,4695	84,3414	84,3354

4.2 Kesinliğe Dayalı Performans Sonuçları

Kesinlik değeri denklem 3.5'te ifade edildiği gibi gerçekte ölen hasta sayısının (*TP*) tahmin edilen toplam ölüm (*TP+FP*) değerine bölünmesiyle elde edilir. Kesinlik değeri ölüm olarak tahmin edilen değerlerin ne kadarının gerçek ölüm olduğunun oranını ifade etmektedir.

OVM yönteminde, en yüksek kesinlik değerini hem 5-fold hem de 10-fold için RF algoritması göstermiştir.

SM yönteminde, aynı şekilde ölüm tahminleri içerisinde gerçek ölüm sayısını en iyi belirleyen algoritma 5-fold için %84,2213 kesinlik değeriyle LR olurken 10-fold için en iyi değeri %86,0624 kesinlik sonucu ile RF olmuştur.

SÖM yöntemi için kesinlik değerleri incelendiğinde 5-fold için RF, XGBoost, CatBoost algoritmalarının kesinlik değeri düşerken; SVM, DT ve LR için kesinlik değerlerinin arttığı görülmüştür. Bu durum RF, XGBoost, CatBoost algoritmalarının özellik seçimi ile kesinlik değerlerinin düştüğünü, SVM, DT ve LR algoritmalarının değerlerinin olumlu yönde etkilendiğini göstermektedir. SÖM yönteminde hem 5-fold hem de 10-fold için en iyi kesinlik sonucu sırasıyla %87,5351 ile %86,9507 olmak üzere LR ile elde edilmiştir.

Hem SÖNM yöntemi hem de tüm yöntemler arasında kesinlik değeri açısından en iyi sonucu 5-fold ve 10-fold çapraz doğrulama ile sırasıyla %88,9285 ve %87,8082 kesinlik sonucuyla SVM algoritmasının elde ettiği görülmektedir. RF algoritmasının kesinlik performansı normalizasyonla da düşerken diğer tüm algoritmaların kesinlik başarısı normalizasyonla artış göstermiştir.

Tablo 4.3 Önerilen dört yönteme 5-fold çapraz doğrulama uygulanması sonucu elde edilen kesinlik değerleri

Algoritmalar	**OVM**	**SM**	**SÖM**	**SÖNM**
SVM	0,0000	56,2099	77,8752	88,9285
RF	74,3121	83,0656	81,3674	79,6319
DT	36,4478	75,7990	78,9719	81,2968
XGBoost	30,6645	71,0687	69,0630	81,5715
CatBoost	47,6844	82,3613	81,6936	82,5483
LR	62,0329	84,2213	87,5351	85,4801

Tablo 4.4 Önerilen dört yönteme 10-fold çapraz doğrulama uygulanması sonucu elde edilen kesinlik değerleri

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	0,0000	59,0317	75,3054	87,8082
RF	74,5811	86,0624	84,0296	84,1105
DT	52,869	75,5224	81,1791	81,5554
XGBoost	50,3928	80,0528	78,2018	81,7879
CatBoost	59,5970	83,6986	82,1123	83,7423
LR	66,8686	83,5989	86,9507	87,6965

4.3 Duyarlılığa Dayalı Performans Sonuçları

Duyarlılık değeri doğru tahmin edilen ölüm sayısının gerçek ölüm sayısının tamamına bölünmesi ile elde edilir.

OVM yönteminde, doğruluk ve kesinlik değerlerinde olduğu gibi aynı şekilde tüm algoritmalar düşük bir performans sergilemiştir. Ancak OVM yöntemini duyarlılık açısından ele alırsak en iyi performansı 5-fold için LR, 10-fold için ise CatBoost sırasıyla %61,1578 ve %59,3333 değerleriyle elde etmiştir.

SM yönteminde duyarlılık başarısı açısından tüm algoritmalarda yaklaşık %20'lik bir performans artışı görülmektedir. Bu yöntemde hem 5-fold hem de 10-fold için en iyi duyarlılık sonucunu sırasıyla %89,2439 ve %87,1428 değeriyle XGBoost algoritması göstermiştir.

SÖM yönteminin, duyarlılık sonuçlarını incelediğimizde SVM algoritmasının performansında %50 oranı gibi yüksek bir başarı artışı dikkat çekmektedir. Bununla birlikte, LR algoritmasının duyarlılık başarısı %4 puanlık bir oranda artış gösterirken RF, DT, XGBoost ve CatBoost algoritmalarının performansında kısmen düşüş gözlemlenmiştir. Bu durum, modele ÖÖE ile özellik seçiminin eklenmesinin SVM ve LR algoritmalarının duyarlılık başarısını ciddi ölçüde yükseltirken diğer algoritmaların performansında kısmen düşüşe neden olmuştur. SÖM yönteminde, en yüksek performansı 5-fold için %87,7439 değeriyle RF, 10-fold için %86,1190 değeriyle XGBoost göstermiştir.

SÖNM yönteminde, SVM, DT, XGBoost ve CatBoost algoritmalarının performansı yükselirken, RF ve LR algoritmalarının performansı kısmen

düştüğü gözlemlenmiştir. Veri setine önce normalizasyon işlemi uygulayarak verilerin aykırı durumlarının giderilmesi daha sonra da ÖÖE yöntemiyle özellik seçiminin uygulanması SVM, DT, XGBoost ve CatBoost algoritmalarının duyarlılık performansını artırdığını göstermektedir. Bu yöntemde, en yüksek duyarlılık başarısını gösteren algoritmalar 5-fold için %87,7682 değeriyle XGBoost, 10-fold için %88,0714 değeriyle DT olmuştur.

Tablo 4.5 Önerilen dört yönteme 5-fold çapraz doğrulama uygulanması sonucu elde edilen duyarlılık değerleri

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	0,0000	20,1951	79,7804	85,7804
RF	55,7894	87,7804	87,7439	87,1829
DT	42,1052	80,8048	79,8170	84,8292
XGBoost	47,3684	89,2439	84,7804	87,7682
CatBoost	52,6315	85,3048	85,2682	87,2317
LR	611578	80,8292	84,7926	81,7804

Tablo 4.6 Önerilen dört yönteme 10-fold çapraz doğrulama uygulanması sonucu elde edilen duyarlılık değerleri

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	0,0000	19,7380	79,5238	84,5952
RF	57,5555	85,0952	85,0714	84,5714
DT	52,5555	79,0952	77,5714	88,0714
XGBoost	55,3333	87,1428	86,1190	86,2190
CatBoost	59,3333	85,5714	81,6190	85,0714
LR	56,6666	81,5476	84,5952	82,6190

4.4 F1-Skor Dayalı Performans Sonuçları

Elde edilen deney sonuçları f1-skor açısından değerlendirildiğinde diğer tüm performans metriklerinde olduğu gibi OVM modelinin sonuçlarının oldukça düşük olduğu gözlemlenmektedir.

SM yönteminde, veri setinin hedef sınıfındaki dengesiz dağılım giderildiğinde tüm algoritmaların performansının yükseldiği görülmüştür. Hem 5-fold hem de 10-fold için en yüksek f1-skor değerini RF algoritması sırasıyla %83,4201 ve %83,5198 değerleriyle elde ettiği görülmüştür.

SÖM yönteminde, SVM, DT ve LR algoritmalarının f-ölçüsüne bağlı başarısı artış gösterirken RF, XGBoost ve CatBoost algoritmasının performansının kısmen düştüğü gözlemlenmiştir. Bu durum, veri dengeleme işleminden sonra ÖÖE ile özellik seçiminin uygulanmasının SVM, DT ve LR algoritmalarının performansını iyileştirirken diğer algoritmaların performansını olumsuz yönde etkilediğini göstermektedir. SÖM yönteminde, en iyi f1-skor başarısını hem 5-fold hem de 10-fold için LR algoritması sırasıyla %84,5463 ve %83,8068 değerleriyle elde etmiştir.

SÖNM yönteminde, SVM, DT, XGBoost, CatBoost algoritmalarının f1-skor değerleri yükselirken RF ve LR algoritmasının performansı düşüş göstermiştir. Bununla birlikte, SVM algoritması SÖNM yöntemi 5-fold çapraz doğrulama sonucunda %86,3092 değeriyle en yüksek f1-skor performansını göstermiştir.

Tablo 4.7 Önerilen dört yönteme 5-fold çapraz doğrulama uygulanması sonucu elde edilen f1-skor değerleri

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	0,0000	29,2349	75,1431	86,3092
RF	50,5250	83,4201	82,9686	81,8911
DT	35,6489	76,5381	77,4423	81,2181
XGBoost	34,9536	78,0514	74,9529	83,0733
CatBoost	44,8810	81,8276	81,7344	83,4441
LR	58,0237	79,6443	84,5463	81,9179

Tablo 4.8 Önerilen dört yönteme 10-fold çapraz doğrulama uygulanması sonucu elde edilen f1-skor değerleri

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	0,0000	28,9567	73,9175	84,2926
RF	54,9045	83,5198	82,3625	82,5009
DT	49,1059	75,5996	76,8296	83,1776
XGBoost	48,7839	82,4027	80,4774	82,3854
CatBoost	53,2999	82,3907	80,2045	82,1252
LR	56,9935	79,2797	83,8068	83,2533

4.5 Algoritmaların Çalışma Sürelerinin Karşılaştırılması

Deneysel çalışmalarda altı farklı makine öğrenmesi algoritmasının dört farklı yöntemle çalışma süreleri saniye türünden hesaplanmıştır. Elde edilen sonuçlar 5-fold çapraz doğrulama için Tablo 4.9'da 10-fold çapraz doğrulama için Tablo 4.10'da sunulmuştur.

Algoritmaların çalışma zamanı değerlendirildiğinde SVM algoritmasının çalışma süresinin verinin dengelemesinden sonra kısaldığı, özellik seçimi eklenen SÖM yönteminde SM yöntemine göre daha hızlı performans gösterdiği ve en hızlı çalışma zamanını SÖNM yöntemi için gösterdiği görülmektedir. SVM algoritması için veri dengeleme, özellik seçimi ve normalizasyon işlemlerinin uygulanmasının hem doğruluk başarısı hem de çalışma hızı açısından olumlu yönde etki ettiği söylenebilir. RF algoritması için veri dengeleme, özellik seçimi ve normalizasyon yöntemlerinin tekil veya bütünsel uygulanmasının çalışma zamanını olumsuz etkilediği görülmektedir. DT algoritmasının çalışma süresinin veri dengeleme, özellik seçimi ve normalizasyon yöntemlerinin uygulanması ile olumlu etkilenerek kısaldığı görülmektedir. XGBoost algoritmasının SM yöntemi ile OVM yöntemine kıyasla daha hızlı çalıştığı, ancak SÖM yönteminde özellik seçiminden olumsuz etkilendiği ve en hızlı çalışma performansını SÖNM yöntemi ile elde ettiği görülmektedir. CatBoost algoritmasının çalışma hızının veri dengeleme, özellik seçimi ve normalizasyon yöntemlerinden olumsuz etkilendiği ve en yavaş çalıştığı yöntemin SÖNM yöntemi olduğu görülmektedir. LR algoritmasının SM yönteminde OVM yöntemine kıyasla veri dengelemeden olumsuz etkilendiği görülmektedir. Ancak LR algoritmasının çalışma süresinin özellik seçiminin eklenmesiyle olumlu yönde etkilendiği ve normalizasyon yönteminin de eklendiği SÖNM yönteminde en iyi performansı gösterdiği görülmektedir. Yapılan deneysel çalışmalarda algoritmaların çalışma süresi 5-fold ve 10-fold açısından ele alındığında tüm algoritmaların dört farklı yöntem için 5-fold çapraz doğrulama ile daha kısa çalışma süresine sahip olduğu görülmektedir. Bununla birlikte tüm algoritmaların dört yöntem ile çalışması incelendiğinde çalışma sürelerinin 10-fold için 5-fold'a göre yaklaşık olarak iki katı süreye yaklaştığı görülmektedir.

Tablo 4.9 Dört yöntem için altı farklı makine öğrenmesi algoritmasının 5-fold çapraz doğrulama ile saniye cinsinden çalışma zamanı

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	0,4070	0,2659	0,2340	0,1559
RF	0,3599	2,6090	2,6570	2,6880
DT	0,3589	0,1559	0,1250	0,0789
XGBoost	2,4530	2,2970	2,3119	1,4849
CatBoost	3,3440	3,6250	3,7190	4,2960
LR	0,3750	0,3829	0,1559	0,0940

Tablo 4.10 Dört yöntem için altı farklı makine öğrenmesi algoritmasının 10-fold çapraz doğrulama ile saniye cinsinden çalışma zamanı

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	0,7190	0,4529	0,4369	0,2660
RF	0,6400	5,2809	5,2903	5,4559
DT	0,6410	0,2500	0,2030	0,0940
XGBoost	3,6720	3,0000	3,6099	3,2649
CatBoost	7,2969	8,0470	8,1330	8,2039
LR	0,6250	0,71900	0,2809	0,1099

4.6 Çapraz Doğrulama Kullanmadan Elde Edilen Doğruluk Değerleri (%80 Eğitim, %20 Test)

Bu bölümdeki deneysel çalışmalarda çapraz doğrulama kullanılmadan veri seti %80 eğitim, %20 test olacak şekilde bölünmüş ve kalp hastalarının sağkalımı tahmin edilmiştir. Önerilen dört farklı model için altı farklı makine öğrenmesi algoritması ile sağkalım tahmini yapılmış ve sonuçlar doğruluk metriği açısından kıyaslanmıştır.

Deneysel sonuçlar Tablo 4.11'de verilmiştir. Elde edilen sonuçlara göre SÖM yöntemi için RF ve SÖNM yöntemi için de CatBoost algoritması %93,9 doğruluk değeri ile bütün deneyler içerisinde en yüksek tahmin başarısına ulaşmıştır.

Tablo 4.11 Dört yöntem için altı farklı makine öğrenmesi algoritmasının %80 eğitim %20 test şeklinde bölünmesi ile elde edilen doğruluk değerleri

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	61,7	62,2	73,2	89
RF	87,8	90,2	93,9	92,7
DT	83,3	89	86,6	90,2
XGBoost	86,8	92,7	89	91,5
CatBoost	84,4	92,6	90,2	93,9
LR	78,9	83,6	73,2	88,5

4.7 Çapraz Doğrulama Kullanmadan Elde Edilen Doğruluk Değerleri (%70 Eğitim, %30 Test)

Bu bölümdeki deneysel çalışmalarda çapraz doğrulama kullanılmadan veri seti %70 eğitim, %30 test olacak şekilde ayrılarak kalp hastalarının sağkalımı tahmin edilmiştir. Önerilen dört farklı model için altı farklı makine öğrenmesi algoritması ile sağkalım tahmini yapılmış ve sonuçlar doğruluk metriği açısından kıyaslanmıştır.

Elde edilen sonuçlara göre veri seti %70 eğitim %30 test olarak ayrıldığında elde edilen doğruluk başarısı %80 eğitim %20 test ayrımından daha düşük olmuştur. Veri seti %70 eğitim %30 test olarak ayrıldığında Tablo 4.12'de gösterildiği gibi en yüksek başarıya SM yöntemiyle XGBoost algoritması %91 doğruluk değeri ile elde etmiştir.

Tablo 4.12 Dört yöntem için altı farklı makine öğrenmesi algoritmasının %70 eğitim %30 test şeklinde bölünmesi ile elde edilen doğruluk değerleri

Algoritmalar	OVM	SM	SÖM	SÖNM
SVM	68,9	49,2	73	82
RF	85	88,5	83,6	86,9
DT	82,9	83,6	84,4	85,2
XGBoost	80	91	89,3	87,7
CatBoost	83,3	90,2	88,5	87,7
LR	73,3	76,8	70,5	87,8

4.8 Literatürdeki Benzer Çalışmalarla Karşılaştırma

Tez çalışmasında, UCI veri havuzundan alınan "*heart_failure_clinical_records*" isimli veri seti kullanılarak kalp hastalığı teşhisi konulan kişilerin tedavi takip süresi içerisindeki sağkalımları tahmin edilmiştir. Sağkalım tahmini için altı farklı makine öğrenmesi algoritması için dört farklı yöntem önerilmiştir. Literatürde, aynı veri setini kullanarak makine öğrenmesi yöntemleri ile sağkalım tahmini yapan çalışmalar bulunmaktadır. Önerilen yöntemlerin literatürdeki benzer çalışmalarla kıyaslaması Tablo 4.13'te verilmiştir.

Tez çalışması kapsamında önerilen SÖNM yöntemiyle 5-fold çapraz doğrulama kullanılarak Tablo 4.13'daki benzer çalışmalarla kıyaslandığında 1, 2, 4, 5 ve 6 numaralı çalışmalardan daha yüksek bir performans gösterdiği görülmektedir. Çapraz doğrulama ile veri setinin tamamının test işleminde kullanılması sağlanarak 10-fold için 10 adet doğruluk değeri 5-fold için de 5 adet doğruluk değeri elde edilir. Daha sonra bu değerlerin ortalaması alınarak ortalama doğruluk başarısı hesaplanır. Dolayısıyla veri setinin test örnekleri belirlenirken rastsallıktan dolayı elde edilecek yüksek başarı veya düşük sonuçtan kaçınılarak ortalama bir başarı sonucu elde edilmiş olur. Literatürde bulunan bazı çalışmalar Tablo 4.13'te ifade edildiği gibi çapraz doğrulama işlemi kullanmamıştır. Bu nedenle aynı deneyler veri seti %80 eğitim, %20 test olacak şekilde ayrılarak tekrar edilmiş ve elde edilen bulgular literatürdeki çalışmalarla kıyaslanmıştır. Elde edilen sonuçlara göre SÖM yönteminde RF algoritması, SÖNM yönteminde ise CatBoost algoritması %93,9 doğruluk değerine ulaşarak literatürdeki benzer çalışmaların tamamından daha iyi bir tahmin başarısına ulaşmıştır.

Tablo 4.13 Önerilen SÖNM yönteminin literatürdeki benzer çalışmalarla karşılaştırılması

Çalışma Numarası	Yazarlar	Çalışma Yılı	Veri Dengeleme Yöntemi	Özellik Seçim Yöntemi	Çapraz Doğrulama	Normalizasyon Yöntemi	Kullanılan Algoritma	Doğruluk (%)
1	Chicco ve Jurman [6]	2020	Yok	ROÖS	Yok	Yok	RF	75,40
2	Newaz vd. [10]	2021	SMOTE	Chi2	5-fold	Yok	RF	77,33
3	Ishaq vd. [12]	2021	SMOTE	ROÖS	Yok	Yok	ET	92,00
4	Mamun vd. [13]	2022	SMOTE	Yok	10-fold	Yok	LightGBM	85,00
5	Erdaş ve Ölçer [15]	2020	Yok	KDÖS	Yok	Yok	MLP	78,00
6	Türkmenoğlu ve Yıldız [17]	2020	SMOTE	Korelasyon Matrisi	Yok	Min-Max Normalizasyonu	ET	84,58
7	Potur ve Erginel [18]	2021	Yok	Info Gain Attribute Eval (WEKA)	Yok	Yok	MLP	90
8	Önerilen SÖNM Yöntemi	2022	SMOTE	ÖÖE	5-fold	Yeo-Johnson Normalizasyonu	SVM	86,47
9	**Önerilen SÖM Yöntemi**	**2022**	**SMOTE**	**ÖÖE**	**Yok**	**Yok**	**RF**	**93,9**
10	**Önerilen SÖNM Yöntemi**	**2022**	**SMOTE**	**ÖÖE**	**Yok**	**Yeo-Johnson Normalizasyonu**	**CatBoost**	**93,9**

5. TARTIŞMA VE SONUÇ

Kalp hastalıkları mortalite oranı yüksek ve her yıl milyonlarca insanın hayatını kaybetmesine neden olan önemli bir hastalıktır. Bu nedenle kalp hastalığına sahip bireylerin sağkalımını tahmin etmek ve sağkalıma en fazla etki eden parametrelerin ortaya çıkarmak önemli bir araştırma konusudur. Tez çalışmasında, kalp hastalarının sağkalımını tahmin etmek için UCI'den alınan veri seti kullanılmış ve önerilen dört farklı yöntem için Destek Vektör Makineleri, Rastgele Orman, Karar Ağacı, XGBoost, CatBoost ve Lojistik Regresyon algoritmaları 5-fold ve 10-fold çapraz doğrulama ile kullanılmıştır. Daha sonra çapraz doğrulamanın etkisini araştırmak için veri seti %80 eğitim, %20 test ve %70 eğitim, %30 test olacak şekilde bölünerek deneyler tekrar edilmiştir. Elde edilen sonuçlara göre 5-fold çapraz doğrulama ile önerilen SÖNM yöntemi ve SVM algoritması birlikte çalıştırıldığında %86,4649 doğruluk, %88,9285 kesinlik, %85,7804 duyarlılık ve %86,3092 f1-skor değeri elde edilmiştir. Çapraz doğrulama kullanılmadan veri seti %80 eğitim ve %20 test olarak bölündüğünde SÖM yöntemi ile RF algoritmasının birlikte kullanımı ve SÖNM yöntemi ile CatBoost algoritmasının birlikte kullanımı sonucu %93,9 doğruluk değerine ulaşılmıştır. Elde edilen %93,9 doğruluk başarısı değeri ile Tablo 4.13'te verilen literatürdeki benzer çalışmaların tamamının üzerinde performans göstererek en yüksek tahmin başarısına ulaşılmıştır.

Kullanılan UCI veri seti, Pakistan'da bulunan kalp hastalarından toplanarak oluşturulmuştur. Kalp hastalıklarını ve sağkalımı medikal kayıtların yanı sıra ırk, yaş, cinsiyet, genetik faktörler kronik hastalıklar ve beslenme alışkanlıkları gibi pek çok faktörün etkileyebileceği tahmin edilmektedir. Bu nedenle, önerilen yöntemin başka medikal hastalıklar, kalp hastalıkları ve farklı hastalıklara bağlı sağkalım veri setleri üzerinde gerçekleştirilebilme potansiyeline sahiptir. Ayrıca gelecekte, literatürde aykırı değerleri normal dağılıma dönüştüren güç dönüşümü normalizasyonları, farklı özellik seçim yöntemleri birlikte kullanılarak hastalık ve sağkalım tahminine yönelik başarı performansının arttırmasına yönelik çalışmalar gerçekleştirilecektir.

KAYNAKLAR

[1] **World Health Organization**. (2022). "Cardiovascular diseases." https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. Erişim Tarihi: 25.12.2022.

[2] **Savarese G. ve Lund L. H**., (2017). Global Public Health Burden of Heart Failure, *Cardiac Failure Review*, 3(1), 7.

[3] **Kondababu A., Siddhartha V., Kumar B. B. ve Penumutchi B.**, (2021). WITHDRAWN: A comparative study on machine learning based heart disease prediction, *Materials Today: Proceedings.*

[4] **Wang, K., Tian, J., Zheng, C., Yang, H., Ren, J., Liu, Y., ... ve Zhang, Y.** (2021). Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. *Computers in Biology and Medicine*, 137, 104813.

[5] **Park, J., Hwang, I. C., Yoon, Y. E., Park, J. B., Park, J. H., ve Cho, G. Y**. (2022). Predicting Long-Term Mortality in Patients With Acute Heart Failure by Using Machine Learning. *Journal of Cardiac Failure.*

[6] **Chicco, D., ve Jurman, G.** (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, 20(1), 1-16.

[7] **Dai, Q., Sherif, A. A., Jin, C., Chen, Y., Cai, P., ve Li, P.** (2022). Machine learning predicting mortality in sarcoidosis patients admitted for acute heart failure. *Cardiovascular Digital Health Journal.*

[8] **Austin, D. E., Lee, D. S., Wang, C. X., Ma, S., Wang, X., Porter, J., ve Wang, B**. (2022). Comparison of machine learning and the regression-based EHMRG model for predicting early mortality in acute heart failure. *International Journal of Cardiology*, 365, 78-84.

[9] **Angraal, S., Mortazavi, B. J., Gupta, A., Khera, R., Ahmad, T., Desai, N. R., ... ve Krumholz, H. M.** (2020). Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *JACC: Heart Failure*, 8(1), 12-21.

[10] **Newaz, A., Ahmed, N., ve Haq, F. S.** (2021). Survival prediction of heart failure patients using machine learning techniques. *Informatics in Medicine Unlocked*, 26, 100772.

[11] **Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., ve Raza, M. A**. (2017). Survival analysis of heart failure patients: A case study. *PloS one*, 12(7), e0181001.

[12] **Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., ve Nappi, M**. (2021). Improving the prediction of heart failure patients' survival using

SMOTE and effective data mining techniques. *IEEE access*, 9, 39707-39716.

[13] **Mamun, M., Farjana, A., Al Mamun, M., Ahammed, M. S., ve Rahman, M. M.** (2022, June). Heart failure survival prediction using machine learning algorithm: am I safe from heart failure?. *In 2022 IEEE World AI IoT Congress (AIIoT)* 194-200pp.

[14] **Panahiazar, M., Taslimitehrani, V., Pereira, N., ve Pathak, J.** (2015). Using EHRs and machine learning for heart failure survival analysis. *Studies in health technology and informatics*, 216, 40.

[15] **Erdaş, Ç. B., & Ölçer, D.** (2020, November). A machine learning-based approach to detect survival of heart failure patients. I*n 2020 Medical Technologies Congress (TIPTEKNO)* 1-4pp.

[16] **Samad, M. D., Ulloa, A., Wehner, G. J., Jing, L., Hartzel, D., Good, C. W., ... ve Fornwalt, B. K.** (2019). Predicting survival from large echocardiography and electronic health record datasets: optimization with machine learning. *JACC: Cardiovascular Imaging*, 12(4), 681-689.

[17] **Türkmenoğlu, B. K., ve Yildiz, O.** (2021, June). Predicting the survival of heart failure patients in unbalanced data sets. *In 2021 29th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

[18] **Potur, E. A., ve Erginel, N.** (2021). Kalp Yetmezliği Hastalarının Sağ Kalımlarının Sınıflandırma Algoritmaları ile Tahmin Edilmesi. *Avrupa Bilim ve Teknoloji Dergisi*, (24), 112-118.

[19] **Al-Dury, N., Ravn-Fischer, A., Hollenberg, J., Israelsson, J., Nordberg, P., Strömsöe, A., ... ve Rawshani, A.** (2020). Identifying the relative importance of predictors of survival in out of hospital cardiac arrest: a machine learning study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 28(1), 1-8.

[20] **Mózsik, G., ve Díaz-Soto, G. (Eds.).** (2021). Mineral Deficiencies: Electrolyte Disturbances, Genes, Diet and Disease Interface. *BoD–Books on Demand*. 13, 99p.

[21] **Liang, X., Chou, O. H. I., ve Cheung, B. M.** (2022). The Effects of Human Papillomavirus Infection and Vaccination on Cardiovascular Diseases. The US National Health and Nutrition Examination Survey 2003-2016. *The American Journal of Medicine.*

[22] **Cardiovascular diseases (CVDs).,** https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), Erişim Tarihi:25.12.2022.

[23] **Cihan, Ş. , Karabulut, B. , Arslan, G. ve Cihan, G.** (2017). Koroner Arter Hastalığı Riskinin Veri Madenciliği Yöntemleri İle İncelenmesi . *International Journal of Engineering Research and Development* , 10 (1) , 85-93 .

[24] **Akoca, A.** (2022). Hastane Öncesi Serebrovasküler Hastalıklara Yaklaşım Organizasyonu. *Hastane Öncesi Dergisi*, 7(2), 261-273.

[25] **Kaptan Ozen, D., ve Turan, B.** (2019). Alt ekstremite periferik arter hastalığında endovasküler tedavi: Tek merkez deneyimi. *Kocaeli Tıp Dergisi*, 8(3), 142-147.

[26] **Teker, A. B., ve Teker, E.** (2022). Türk kadın hastalarda IL-10 promoter polimorfizmlerinin romatizmal kalp hastalığı ile ilişkisi üzerine bir çalışma. *Turk Kardiyol Dern Ars*, 50(1), 14-21.

[27] **Varal, İ. G., Köksal, N., Özkan, H., Bostan, Ö., SIĞINAK, İ. Ş., BAĞCI, O., ... ve Uysal, F.** (2015). Yenidoğan yoğun bakım ünitemizde izlenen konjenital kalp hastalıkları: Sıklığı, risk faktörleri ve prognoz. *Güncel Pediatri*, 13(3), 159-164.

[28] **BOLAT, A., ve GÜLTEKİN, Y.** (2021). Vena Safena Magnanın Anatomik Varyasyonu Derin Ven Trombozu için Bir Risk Faktörü Müdür?, *Kırıkkale Üniversitesi Tıp Fakültesi Dergisi*, 23(2), 343-350.

[29] **Erdinç, K.,** (2021). Pulmoner Emboli Hastalarının Yoğun Bakım Takibi. *Kırıkkale Üniversitesi Tıp Fakültesi Dergisi*, 23(2), 262-269.

[30] **Kumsar, A. K., ve Yılmaz, F. T.** (2017). Kardiyovasküler hastalıklar risk faktörlerinden korunmada hemşirenin rolü. *Online Türk Sağlık Bilimleri Dergisi*, 2(4), 18-27.

[31] **Kokubo, Y., ve Matsumoto, C.** (2016). Hypertension is a risk factor for several types of heart disease: review of prospective studies. *Hypertension: from basic research to clinical practice*, 419-426.

[32] **Géron, A.** (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Unsupervised learning techniques. *O'Reilly Media, Incorporated*, 25p.

[33] **Mitchell, T. M., & Mitchell, T. M.** (1997). Machine learning (Vol. 1, No. 9). N*ew York: McGraw-hill*, 1-20pp.

[34] **Rebala, G., Ravi, A., & Churiwala, S.** (2019). An introduction to machine learning. Springer, 4pp.

[35] **Takcı, H**. (2020). Teori ve uygulamada veri madenciği (1. baskı). *Nobel Akademik Yayıncılık*, 20s.

[36] **Sarker, I. H.** (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3), 1-21pp.

[37] **Atalay, M., ve Çelik, E.** (2017). Büyük veri analizinde yapay zekâ ve makine öğrenmesi uygulamalari-artificial intelligence and machine learning applications in big data analysis. *Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 9(22), 155-172.

[38] **Ağca, K., ve Takcı, H.** (2022). Hibrit Bir Model Oluşturarak Diyabetik Retinopati Sınıflandırılması. *Avrupa Bilim ve Teknoloji Dergisi*, (36), 227-236pp.

[39] **Dukkancı, S. A.** (2021). Level generation using genetic algorithms and difficulty testing using reinforcement learning in match-3 game., *Middle East Technical University*, (Master of Science Thesis), 12p, Ankara.

[40] **Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.** (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357pp.

[41] **Çetin, V., ve YILDIZ, O**. (2022). A comprehensive review on data preprocessing techniques in data analysis. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 28(2), 299-312pp.

[42] **Sun, L., Hu, N., Ye, Y., Tan, W., Wu, M., Wang, X., ve Huang, Z.** (2022). Ensemble stacking rockburst prediction model based on Yeo–Johnson, K-means SMOTE, and optimal rockburst feature dimension determination. *Scientific Reports*, 12(1), 1-16pp.

[43] **Danacı, Ç.** (2022). Covid-19 Tanısında Biyokimya Parametre Baskınlığının Makine Öğrenimi Yöntemleri Kullanılarak Belirlenmesi, Fırat Üniversitesi Fen Bilimleri Enstitüsü, (Yüksek Lisans Tezi), 38-41s, Elazığ.

[44] **Pedregosa, F., Varoquaux, G., Gramfort, A. … ve Duchesnay, E.** (2011), Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12(1), 2825-2830pp.

[45] **Kannari, P. R., Chowdary, N. S., ve Biradar, R. L.** (2022). An anomaly-based intrusion detection system using recursive feature elimination technique for improved attack detection. *Theoretical Computer Science*, 931, 56-64pp.

[46] **Vapnik, V.** (1999). The nature of statistical learning theory. *Springer science & business media*, 138p.

[47] **Ma, J., Yang, L., ve Sun, Q.** (2021). Adaptive robust learning framework for twin support vector machine classification. *Knowledge-Based Systems*, 211, 106536.

[48] **Barik, S., Mohanty, S., Rout, D., Mohanty, S., Patra, A. K., ve Mishra, A. K.** (2020). Heart disease prediction using machine learning techniques. *In Advances in Electrical Control and Signal Systems, Springer*, Singapore, 879-888pp.

[49] **Kaggle**, (2022). https://www.kaggle.com/, Erişim Tarihi: 25.12.2022.

[50] **Prabha, A., Yadav, J., Rani, A., ve Singh, V.** (2021). Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier. *Computers in Biology and Medicine*, 136, 104664.

[51] **Chen, T., ve Guestrin, C.** (2016). Xgboost: A scalable tree boosting system., *In Proceedings of the 22nd acm sigkdd international conference on knowledge*

discovery and data mining 785-794pp.

[52] **Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., ve Gulin, A.** (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

[53] **Lu, C., Zhang, S., Xue, D., Xiao, F., ve Liu, C.** (2022). Improved estimation of coalbed methane content using the revised estimate of depth and CatBoost algorithm: A case study from southern Sichuan Basin, *Computers & Geosciences*, China, 158, 104973.

[54] **Bisong, E.** (2019). Logistic regression. In Building Machine Learning and Deep Learning Models on Google Cloud Platform, 243-250pp.

[55] **Berrar, D.** (2019). Cross-Validation., *Encyclopedia of Bioinformatics and Computational Biology*., (1), 542-545pp.

[56] (https://scikit-learn.org/stable/modules/cross_validation.html)., Erişim Tarihi: 25.12.2022.

[57] **Ekrem, Ö., Salman, O. K. M., Aksoy, B., ve İnan, S. A.** (2020). Yapay Zekâ Yöntemleri Kullanilarak Kalp Hastaliğinin Tespiti. *Mühendislik Bilimleri ve Tasarım Dergisi*, 8(5), 241-254.

[58] **Nizam, H., ve Akın, S. S.** (2014). Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. *XIX. Türkiye'de İnternet Konferansı*, 1(6).