T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİNDE PERFORMANS METRİKLERİ İLE TEST TEKNİKLERİNİN FARKLI VERİ SETLERİ ÜZERİNDE DEĞERLENDİRİLMESİ

Abdullah ALAN

Yüksek Lisans Tezi

YAZILIM MÜHENDİSLİĞİ ANABİLİM DALI

MAYIS 2020

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Yazılım Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİNDE PERFORMANS METRİKLERİ İLE TEST TEKNİKLERİNİN FARKLI VERİ SETLERİ ÜZERİNDE DEĞERLENDİRİLMESİ

Tez Yazarı
Abdullah ALAN

Danışman
Doç. Dr. Murat KARABATAK

MAYIS 2020
ELAZIĞ

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Yazılım Mühendisliği Anabilim Dalı

Yüksek Lisans Tezi

Başlığı: Makine Öğrenmesi Sınıflandırma Yöntemlerinde Performans Metrikleri ile Test Tekniklerinin Farklı Veri Setleri Üzerinde Değerlendirilmesi

Yazarı: Abdullah ALAN

İlk Teslim Tarihi: 22.1.2020

Savunma Tarihi: 21.5.2020

TEZ ONAYI

Fırat Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına göre hazırlanan bu tez aşağıda imzaları bulunan jüri üyeleri tarafından değerlendirilmiş ve akademik dinleyicilere açık yapılan savunma sonucunda OYBİRLİĞİ ile kabul edilmiştir.

		İmza
Danışman:	Doç. Dr. Murat KARABATAK Fırat Üniversitesi, Teknoloji Fakültesi	Onayladım
İkinci Danışman:		Onayladım
Başkan:	Doç. Dr. Muhammed Fatih TALU İnönü Üniversitesi, Mühendislik Fakültesi	Onayladım
Üye:	Doç. Dr. Erkan TANYILDIZI Fırat Üniversitesi, Teknoloji Fakültesi	Onayladım
Üye:	Unvan Adı SOYADI ... Üniversitesi, ... Fakültesi	Onayladım
Üye:	Unvan Adı SOYADI ... Üniversitesi, ... Fakültesi	Onayladım
Üye:	Unvan Adı SOYADI ... Üniversitesi, ... Fakültesi	Onayladım

Bu tez, Enstitü Yönetim Kurulunun ...../...../20..... tarihli toplantısında tescillenmiştir.

İmza

Prof. Dr. Soner ÖZGEN
Enstitü Müdürü

BEYAN

Fırat Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım " Makine Öğrenmesi Sınıflandırma Yöntemlerinde Performans Metrikleri ile Test Tekniklerinin Farklı Veri Setleri Üzerinde Değerlendirilmesi " Başlıklı Yüksek Lisans Tezimin içindeki bütün bilgilerin doğru olduğunu, bilgilerin üretilmesi ve sunulmasında bilimsel etik kurallarına uygun davrandığımı, kullandığım bütün kaynakları atıf yaparak belirttiğimi, maddi ve manevi desteği olan tüm kurum/kuruluş ve kişileri belirttiğimi, burada sunduğum veri ve bilgileri unvan almak amacıyla daha önce hiçbir şekilde kullanmadığımı beyan ederim.

21/5/2020

Abdullah ALAN

ÖNSÖZ

Yüksek lisans öğrenimim boyunca öğrencisi olmaktan büyük onur duyduğum, bilgi ve deneyimlerinden her zaman yararlandığım, değerli danışman hocam Sayın Doç. Dr. Murat KARABATAK'a teşekkürlerimi borç bilirim.

Öğrenciliğim süresince bana her zaman inanan, destekleyen, bugünlere gelmemi sağlayan, haklarını ödeyemeyeceğim annem, babam ve kardeşlerime sonsuz teşekkür ederim.

Bu süreci en az benim kadar yaşayan, manevi desteğini esirgemeyen, içten desteğiyle beni ayakta tutan ve süreç boyunca bilgilerinden yararlandığım eşim Burcu ALAN' a çok teşekkür ederim.

Tezimin yazım sürecinde hayatımıza dâhil olan, motivasyonumu ve çalışma hırsımı artıran canım kızım Eylül ALAN' a çok teşekkür ediyorum.

Abdullah ALAN
ELAZIĞ, 2020

İÇİNDEKİLER

Sayfa

ÖZET

Makine Öğrenmesi Sınıflandırma Yöntemlerinde Performans Metrikleri ile Test Tekniklerinin Farklı Veri Setleri Üzerinde Değerlendirilmesi

Abdullah ALAN

Yüksek Lisans Tezi

FIRAT ÜNİVERSİTESİ
Fen Bilimleri Enstitüsü

Yazılım Mühendisliği Anabilim Dalı

Mayıs 2020, Sayfa: xiii + 62

Makine öğrenmesi sınıflandırma yöntemlerinde bir model oluşturulurken en önemli sorunlardan birisi en iyi sınıflandırıcının seçimi sürecidir. Doğru sınıflandırıcının seçiminde, veri setinin özellikleri ve modeli oluşturmak için kullanılan eğitim ve test verilerinin seçimi süreci büyük önem arz etmektedir. Bu tezde, test tekniklerinden hold-out ve k-katlı çapraz doğrulama yöntemleri kullanılmıştır. Model oluşturulduktan sonra ise sınıflandırıcının performansını değerlendirmek için bazı metrikler kullanılmaktadır. Bu amaçla, çalışmada veri dağılımı ve karar sınıfı dağılımı birbirinden farklı olan 32 veri setine dokuz farklı sınıflandırıcı uygulanmıştır. Model oluşturmak için açık kaynak kodlu bir dil olan Python programlama dili ve Sklearn, Pandas, Numpy, Seaborn ve Matplotlib kütüphanelerinden faydalanılmıştır. Bu sınıflandırıcılar ile hold-out ve çapraz doğrulama yöntemleri kullanılarak modeller oluşturulmuştur. Sınıflandırıcıların performanslarını değerlendirmek için karmaşıklık matrisinden faydalanılmış ve karmaşıklık matrisi yardımı ile her bir model için doğruluk, kesinlik, anma, F1, Matthews korelasyon katsayısı ve ROC eğrisi altında kalan alan değerleri hesaplanmıştır. Dengesiz veri setlerinde Matthews korelasyon katsayısı ve ROC eğrisi altında kalan alan değerlerinin sınıflandırıcı seçiminde daha doğru sonuçlar verdiği gözlemlenmiştir. Elde edilen sonuçlar incelendiğinde hold-out yöntemi, yirmi veri setinde k-katlı çapraz doğrulama yönteminden daha iyi sonuçlar vermiş olsa da eğitim ve test aşamasında çapraz doğrulama yöntemi seçilmiştir. Bunun nedeni, hold-out yönteminde model eğitilirken ya da test edilirken verilerin dengesiz dağılabilmesidir. Bu durum, elde edilen sonuçların güvenirliliğini azaltmaktadır. Tezin sonunda, veri setinin özellikleri, test tekniğinin seçimi ve sonuçların değerlendirilmesi konusunda elde edilen sonuçlar detaylı bir şekilde incelenmiştir.

Anahtar Kelimeler: Karmaşıklık matrisi, Çapraz doğrulama, Hold-out, Sınıflandırma

ABSTRACT

Evaluation of Performance Metrics and Test Techniques on Various Data Sets in Machine Learning Classification Methods

Abdullah ALAN

Master's Thesis

FIRAT UNIVERSITY
Graduate School of Natural and Applied Sciences

Department of Software Engineering

May 2020, Pages: xiii + 62

One of the most important problems when creating a model in machine learning classification methods is the selection process of the best classifier. In the selection of the correct classifier, properties of the data set and process of selecting training and test data used to create the model have great importance. In the study, hold-out and k-fold cross-validation methods were used for testing stage. After the training stage, some metrics are used to evaluate the performance of the classifier. Within the scope of this thesis, nine classifiers have been applied to 32 data sets whose data distribution and decision class distribution are different from each other. The Python programming language, which is an open source language, and the Sklearn, Pandas, Numpy, Seaborn and Matplotlib libraries were used to create models. With these classifiers, models were created using hold-out and k-fold cross validation methods. The confusion matrix was used to evaluate the performance of the classifiers and the accuracy, precision, recall, F_1, Matthews correlation coefficient and ROC curve values were calculated for each model with the help of the confusion matrix. It was observed that Matthews correlation coefficient and ROC curve values gave more accurate results in the classifier selection on unbalanced data sets. When the obtained results are analyzed, although the hold-out method has yielded better results than the k-fold cross validation method on twenty datasets, the k-fold cross validation method was chosen while training and testing stage. The reason of this is the data can be split unbalanced while the model is being trained and tested in the hold-out method. This case reduces the reliability of the obtained results. In the end of thesis, the results obtained on the properties of the data set, the selection of the test technique and the evaluation of the results are examined in detail.

Keywords: Confusion matrix, Cross validation, Hold-out, Classification

ŞEKİLLER LİSTESİ

Sayfa

TABLOLAR LİSTESİ

Sayfa

EKLER LİSTESİ

Sayfa

KISALTMALAR

Kısaltmalar

KNN : k-En Yakın Komşu
KA : Karar Ağaçları
RO : Rasgele Orman
NB : Naive Bayes
LRS : Lojistik Regresyon Sınıflandırıcı
DVM : Destek Vektör Makineleri
GA : Gradyan Artırma
AB : AdaBoost
YSA : Yapay Sinir Ağları
DP : Doğru Pozitif
DN : Doğru Negatif
YP : Yanlış Pozitif
YN : Yanlış Negatif
DPO : Doğru Pozitif Oran
DNO : Doğru Negatif Oran
YPO : Yanlış Pozitif Oran
YNO : Yanlış Negatif Oran
ROC : Alıcı İşlem Karakteristikleri
AUC : Alıcı İşlem Karakteristikleri Altında Kalan Alan
MCC : Matthews korelasyon katsayısı
ÇD : Çapraz Doğrulama
PKD : Pozitif Kestirim Değeri
NKD : Negatif Kestirim Değeri

1. GİRİŞ

Bilişim teknolojileri alanında son yıllarda farklı kaynaklardan toplanan oldukça fazla veri bulunmaktadır. Bu veriler web sayfalarından, bloglardan, sosyal ağlardan, farklı amaçlar için üretilen sensörlerden, alışveriş platformlarından ve bunlara benzer pek çok farklı kaynaktan elde edilmektedir [1]. Bu veriler günümüzde neredeyse her gün yaklaşık olarak 2.5 exabyte boyutunda üretilmektedir. Elde edilen veri yaklaşık olarak 2.5 milyon tane 1 terabyte'lık sabit diske tekabül etmektedir [2]. Toplanan veriler, pazarlama stratejileri oluşturmada, halkla ilişkilerde, bankacılık uygulamalarında, güvenlik vb. gibi alanların yanı sıra bilim insanlarının yaptıkları araştırmalarda da kullanılmaktadır.

İnsanlar farkında olmasa da yapay öğrenme hayatımızın çoğu alanına girmiş bulunmaktadır. Makine nasıl öğrenmekte, insanlar bu sürecin neresinde bulunmakta gibi sorular yapılan çalışmalarda en sık karşılaşılan sorular olarak karşımıza çıkmaktadır. Aslında günlük olarak yapılan çalışmalar yapay öğrenme sürecine katkı sağlamaktadır. Örneğin günlük çekilen fotoğraflar, sosyal medya üzerinden yapılan herhangi bir yorum, bir internet adresinde bulunan anket çalışması, netflix'de izlenilen bir film, herhangi bir alışveriş sitesi üzerinden yapılan bir alışveriş bunların hepsi yapay öğrenme ya da daha sık karşılaşılan ifadeyle makine öğrenmesi süreçlerine yapılan katkılar olarak karşımıza çıkmaktadır [3].

Üretilen bu verilerin çoğunluğu istatistiksel veri olarak saklanmak dışında başka bir amaçla kullanılmamaktadır. Örneğin hastanelerde üretilen verilerin herhangi bir işlem görmeden sadece istatistiksel veri olarak saklandığını söyleyen Profesör Regina Barzilay, yaptığı bir çalışmada göğüs kanseri ile uğraşmış ama hastalardan toplanan verilerin büyük bir bölümünün kullanılmadığını fark etmiştir. Oysa ki veri ve veriden çıkarımlarda bulunmak o kadar önemli bir noktaya geldi ki NASA uzaya göndereceği astronotların yanı sıra veri bilimi üzerine çalışmalar yapmak üzere R veya Python programlama dillerine hakim veri bilimciler yetiştirmektedir [4].

Bilgisayarlar milyonlarca kişinin ürettiği bu verilerden çıkarımlarda bulunmaktadır. Makine elinde bulunan veri ne kadar çoksa o kadar iyi öğrenmektedir. Makinelerin insanların yaptıklarından çıkarımda bulunmasına en güzel örneklerden bir tanesi Google'ın geliştirdiği DeepMind'ın Go Şampiyonunu yenmesidir. Aslında DeepMind'ın yaptığı insanların geçmişte yaptığı hamlelerden çıkarımlarda bulunmaktır [2, 5].

Makine ile öğrenme gerçekleştirilirken bu öğrenmenin ne oranda doğru olduğu, makinenin ne kadar doğru tahminde bulunduğu büyük öneme sahiptir. Örneğin bir hastanın hasta ya da hasta olmadığını tahmin eden bir makine öğrenmesi modeli hayati önem taşıyan bir karara imza atmaktadır. Modeli oluşturmak için kullanılan makine öğrenmesi yöntemi burada verinin fazlalığı, dağılımı gibi etkenler ile sonuçlar üretmektedir. Burada ki en önemli sorun modeli oluştururken

belli bir sınıfa ait veriden modeli oluşturan algoritmaya fazla verilmesi modelin başarımını ölçerken yanılgıya düşülmesine sebep olabilmektedir.

Bu tez çalışmasında yukarıda bahsedilen soruna çözüm olması için modelin değerlendirilmesinde sadece doğruluk metriği değil bunun yanı sıra bazı ek metriklerin de kullanılması gerektiği savunulmaktadır. Ayrıca model oluşturulurken veri setinin eğitim ve test şeklinde ayrılma sürecinde de bazı ek çalışmalar yapılmasının yine modelin başarım ölçümünde doğru sonuçlar alınmasına katkıda bulunup bulunmayacağı tespit edilecektir. Bu işlemler yapılırken kullanılacak veri setleri kaggle açık kaynak veri seti sağlayan internet adresinden farklı dağılım ve karar sınıflarına sahip olacak şekilde seçilmiştir [6]. Bu veri setlerine açık kaynak kodlu Python programlama dili ile Scikit learn, Pandas, Numpy, Matplotlib ve Seaborn kütüphaneleri kullanılarak makine öğrenme süreci adımları ve algoritmaları uygulanacaktır.

1.1. Problemin Tanımı

Makine öğrenmesi alanında literatür taraması yapıldığında sınıflandırma çalışmaları ile ilgili olarak ya sadece tek veri seti kullanılmış ya da sadece doğruluk metriği kullanılarak performans ölçümü yapılıp karşılaştırma niteliğinde çalışmaları yapıldığı tespit edilmiştir. Sınıflandırma modellerinin performanslarını ölçerken, farklı metriklerin kullanımı ve farklı veri dağılımı olduğu durumlarda başarımın nasıl değiştiği hakkında yeterince çalışmaya rastlanmamıştır. Yapılan sınıflandırma çalışmalarında ise genelde sonuçlar sadece sayısal olarak verilmekte, sonuçlar hakkında değerlendirme yeterince yapılmamaktadır. Bu çalışmada kullanılacak ek metrikler sınıflandırma sonuçlarının değerlendirilmesinde önemli bir kaynak oluşturacaktır.

1.2. Çalışmanın Önemi

Makine Öğrenmesi ile ilgili literatür taraması yapıldığında bu konuyla ilgili yapılan çalışmaların sayısında artma olduğu gözlemlenmiştir. Bu tez için, sınıflandırma algoritmaları kullanılan araştırmalar incelenmiştir [7-13]. Bu çalışmalar da, genel olarak sadece tek veri seti üzerinde incelemeler yapılmış ve başarım kriteri olarak da doğruluk metriği kullanılmıştır. Bu tez çalışmasında model seçiminin doğru yapılabilmesi için doğruluk metriğine ek metrikler üzerinde durulmuştur. Bunun için farklı dağılım ve büyüklükteki veri setlerine sınıflandırma algoritmaları uygulanmış, sonucunda hangi modelin nasıl seçileceği hakkında değerlendirmeler yapılmıştır. Doğruluk metriğine ek olarak duyarlılık, anma, kesinlik, özgüllük, doğru pozitif oranı (DPO), doğru negatif oranı (DNO), yanlış negatif oranı (YNO), yanlış pozitif oranı (YPO), F ölçütü, Matthews korelasyon katsayısı (MCC), Alıcı İşlem Karakteristikleri Altında Kalan Alan (AUC) ve Alıcı İşlem Karakteristikleri (ROC) eğrisi metrikleri kullanılmıştır. Böylelikle model seçimi yapılırken en iyi modelin seçimi için bir yol haritası oluşturulmuştur.

1.3. Çalışmanın Amacı

Bu tez çalışması kapsamında, sınıflandırma algoritmalarının seçimini sağlarken hangi performans metriklerinin daha faydalı olacağının belirlenmesi amaçlanmaktadır. Bu kapsamda farklı karar sınıfı dağılımı olan veri setleri incelenerek bu veri setlerine sınıflandırma algoritmaları uygulanacak ve her bir modelin performans değerlendirmeleri yapılacaktır. Bu değerlendirmeleri yaparken karmaşıklık matrisinin ve ondan elde edilecek metriklerin ne kadar önemli olduğuna yönelik çalışma yapılması amaçlanmaktadır.

1.4. Hipotezler

$H_0$1: Doğruluk metriği sınıflandırıcının performansını ölçmek için tek başına yeterli değildir. Diğer metriklerin bazı durumlarda önemi çok daha fazla olabilmektedir.

$H_0$2: Test-Eğitim verisinin belirlenmesine uygun yöntemin seçilmesi çok önemlidir ve performansı büyük ölçüde etkilemektedir.

$H_0$3: Sınıf sayısının ikiden fazla olduğu durumlarda karmaşıklık matrisinin kullanımı modelin başarımının değerlendirilmesinde daha anlamlı sonuçlar çıkmasını sağlamaktadır.

1.5. Literatür Taraması

Makine öğrenmesi ile ilgili olarak literatürde çok fazla çalışma olup Türkiye'de yapılmış olan bazı çalışmalar aşağıda verilmiştir.

Aydın [7], çalışmasında yapay zeka ve makine öğrenmesi teknikleri kullanarak bir tıp bilişimi uygulaması geliştirmeyi amaçlamıştır. Geliştirilmiş olan sistemde, K-en yakın komşu (KNN) makine öğrenmesi modeli olarak kullanılmıştır. KNN kullanılmasının sebebi yapılan tahminlerde yüksek oranda doğruluk sağlamasıdır.

Haciefendioğlu [8], glokom hastalığının göz sinirleri zarar görmeden önce teşhis edilebilmesi ve dünya çapında körlük nedenleri arasında ilk sıralarda bulunan bu hastalığın tahmin edilmesini amaçlamıştır. Çalışmada veri olarak Pamukkale Üniversitesi Göz Hastalıkları Anabilim Dalından alınan hastalara ait bilgiler kullanılmıştır. Destek Vektör Makineleri (DVM), Karar Ağaçları (KA) ve Yapay Sinir Ağları (YSA) çalışmada üç farklı makine öğrenmesi sınıflandırma yöntemi olarak kullanılmıştır. Bu yöntemlerle hastalığın başlangıç safhasında teşhisi için sınıflandırma yapılmış ve daha sonra birbirleriyle karşılaştırılmıştır. Bahsedilen makine öğrenmesi yöntemlerinin performansları X-Validation ile belirlenmiş ve DVM'nın en yüksek sınıflandırma başarısı elde edebileceği görülmüştür.

Kartal [9], tezinde kalp ameliyatı sırasında ya da kalp ameliyatı geçirdikten kısa bir süre sonra hastaya ait hayati riskin, sınıflandırmaya dayalı makine öğrenmesi teknikleri kullanılarak belirlenmesini amaçlamıştır. Çalışmasında Acıbadem Maslak Hastanesinden veri setini temin etmiş

ve KNN, Lojistik Regrasyon Sınıflandırıcı (LRS), Naive Bayes (NB), KA algoritmalarını kullanarak farklı modeller oluşturmuştur. Oluşturulan modellere ait başarım sonucu farklı metrikler kullanılarak ölçülmüştür. Çalışma sonucunda en iyi performansı gösteren modelin karar ağacı modeli olduğunu tespit etmiştir.

Ünal [10], doktor, radyolog ve uzmanların bel bölgesi hastalıklarının teşhisine yardımcı olması amaçlanan bir bilgisayar destekli teşhis sistemi geliştirmeyi amaçlamıştır. Çalışmada veri kümelerinin sınıflandırılması işlemi yapılmıştır. Bu aşamada YSA, DVM, LVQ sinir ağları, KNN, RTFA sınıflandırma algoritmaları uygulanmış ve her biri için sonuçlar ayrı ayrı kaydedilmiştir.

Şeker [11], çalışmasında iyi-kötü koku verilerine yönelik Elektro Ensefalo Grafi (EEG) işaretlerinin analizi ve sınıflandırılmasını amaçlamıştır. Katılımcılara değerlendirme anketleri ile kokulara ait güç spektrum grafikleri yardımıyla en baskın olan ikişer iyi-kötü koku belirlenmiştir. Sınıflandırıcı olarak WEKA veri madenciliği programına ait olan YSA, NB, KNN ve Rasgele Orman (RO) algoritmaları kullanılmıştır. Diferansiyel gelişim algoritmaları kullanılarak kanal seçimi yapılmış ve sınıflandırma işlemleri tekrarlanarak sonuçlar karşılaştırılmıştır. Çalışmanın sonucunda tüm kanalların kullanıldığı aşamada ki sınıflandırmaya bakıldığında; NB %70.93, KNN %92.76, YSA %92.73 ve RO algoritması %99.19 başarım oranları tespit edilmiştir. Seçilmiş olan 5 kanalın kullanıldığı aşamadaki sınıflandırma sonuçlarına bakıldığında ise; NB %68.45, KNN %88.95, ÇKA %88.83 ve RO algoritması %97.58 başarım oranları tespit edilmiştir. Mevcut çalışma ile beynin hangi bölgelerinin ve frekans bantlarının koku ile ilişkili olduğu kestirilmeye çalışılmıştır. Çalışmada kullanılan yöntemin klinik tedavilerde bazı nörolojik hastalıkların erken teşhis edilmesinde kullanılabileceği düşünülmektedir.

Turgut [12], mikrodizi verileri kullanarak makine öğrenmesi yöntemleri ile sınıflandırma yapmıştır. Uygulamaları Python programlama dili kullanarak gerçekleştirmiştir. Uygulanan makine öğrenmesi algoritmaları YSA, DVM, KNN, KA, LRS, RO, Adaboost (AB) ve Gradyan Artırma (GA)'dır. Katman sayısı yükseldikçe, başlangıçta değişmeyen doğruluk oranı, belirli bir aşamadan sonra azalmaya başlamıştır. Bulunan en yüksek doğruluk oranı, ilk veride %97.69, ikinci veride ise %68.72 olmuştur. Sonuç olarak ise derin öğrenmede sınıflandırmanın doğruluk oranının katman sayısının yükselmesiyle doğru orantılı olduğu tespit edilmiştir.

Pekel [13], çalışmasında sınıflandırma problemi üzerinde durmuş ve 4 temel sınıflandırma algoritması sunarak hazır veri setindeki performansları karşılaştırılmıştır. Bunlar, NB, KA, YSA ve DVM'dir. Çalışmanın sonucunda veri setine uygulanan diğer 3 sınıflandırma yöntemlerinden daha iyi performans gösterenin NB algoritması (%70.29) olduğu tespit edilmiştir. Buna ek olarak temel sınıflandırma algoritmalarının performansını yükseltmek amacıyla, Genetik Algoritma ile melez modelleri önerilerek performansları karşılaştırılmıştır. En yüksek performans değerine sahip olan algoritmanın Genetik Algoritmalı KA (%92.57) olduğu tespit edilmiştir.

1.6. Yaygın Etki

Tez çalışmamın sonucunda, veri setinin dağılımına göre hem test-eğitim veri setinin seçiminde hem de başarım değerlendirmesinde hangi modellerin seçileceği noktasında makine öğrenmesi alanında çalışmak isteyen araştırmacılar için tavsiyeler yer alacaktır. Özellikle bu alanda çalışmalarına yeni başlayan araştırmacılara bu tez çalışması iyi bir kaynak niteliği taşıyacaktır.

2. MATERYAL VE YÖNTEM

Bu tez çalışmasında, makine öğrenmesi yöntemlerinden olan sınıflandırma süreci üzerinde durulmuştur. Bu bölümde öncelikle makine öğrenmesinin tanımı, süreci ve çalışmada kullanılacak makine öğrenmesi algoritmaları hakkında bilgi verilmiştir. Ardından çalışmada yer alan veri setlerinin ne amaçla oluşturulduğu ve veri setlerine ait özellikler hakkında bilgi verilmiştir. Daha sonra ise oluşturulan modellere ait performans sonuçları verilmiştir.

2.1. Makine Öğrenmesi

Makine öğrenmesi, algoritma, tecrübe ve yapacağı iş olarak ele alındığında, etkili ve yetenekli algoritmaların tasarlanmasıdır. Bu algoritmalar, makinelerin deneyimlerini artırarak öğrenme işleminin gerçekleştirilmesi, veri seti üzerinde çalışmalar yaparak kurallar üreten, veri setinde yapılan değişimlere uyarlanabilen, tecrübesi arttıkça başarım yüzdesi de artan yazılımlar planlanması ile ilgilenmektedir. Yapılan araştırmalarda 2011 yılından itibaren arama motorlarında "Büyük veri" ifadesi "Makine öğrenmesi" ifadesinden daha fazla aranırken, 2017 yılından itibaren bu durum eşitlenmektedir. Arama motorlarında yapılan incelemelerde sadece ABD'de yapılan aramalara bakıldığında ise durum Makine öğrenmesi kavramının daha fazla arandığın göstermektedir [14].

Yapay zeka kavramı henüz tam olarak anlamını ifade edecek duruma gelmediği için bunun yerine makine öğrenmesi ve derin öğrenme kavramlarının kullanılması daha doğru olmaktadır [15]. Makine öğrenmesinin yapay zekanın, derin öğrenmenin de makine öğrenmesinin bir alt kümesi olduğu söylenmektedir [16]. Mail hesaplarımızdaki e-postalarımız makine öğrenmesi kullanılarak filtreleniyor, film izlediğimiz netflix benzeri siteler izleyeceğimiz filmler konusunda tavsiyeler vermek için makine öğrenmesini kullanmakta, müzik dinleme uygulamaları da aynı şekilde makine öğrenmesi kullanarak tavsiyelerde bulunmaktadır. Apple 2017 yılında piyasaya sürdüğü IPhone X adlı modelinde yine makine öğrenmesi algoritmaları kullandığı teknolojisi FaceID ile yüzdeki herhangi bir değişikliğe rağmen yüzü algılayabilmektedir [17].

Makine öğrenmesini sağlayan aslında kullanılan algoritmalardır. Bu algoritmaların seçiminde elimizde bulunan veri setinin önemi çok fazladır. Elimizde bulunan veri setinde çıktı değerleri biliniyorsa denetimli öğrenme, çıktı değerleri olmadan bir öğrenme modeli geliştireceksek denetimsiz öğrenme algoritmaları kullanılmaktadır. Yapılan çalışmalarda, denetimli öğrenme sınıflandırma, denetimsiz öğrenme ise kümeleme ile temsil edilmektedir [18]. Eğitim için etiketlenmiş verilerin kullanımını da sağlayan diğer bir yöntem ise yarı denetimli öğrenme modelidir. Bu hibrit yöntemde etiketlenmiş ve etiketlenmemiş verilerin bir kombinasyonu kullanılarak model oluşturulmaktadır [19].

Sınıflandırma, bir model oluştururken girdi ve çıktı değerlerinin bilindiği durumlarda kullanılan modelleri temsil etmektedir. Model oluşturulurken öncelikle veri iki parçaya bölünür ve bunlar eğitim ve test verisi olarak isimlendirilir. Eğitim verisi ile model oluşturulur, test verisi ile de bu oluşturulan modelin başarımı test edilir.

Kümeleme ise model oluştururken karar sınıfının belli olmadığı modelin veri setindeki verileri belirli mesafe ölçüm araçları kullanılarak verinin kümelenmesi işlemini içermektedir. Burada birbirine benzerlik gösteren veriler ölçüm araçları ile birbirinden ayrılmaktadır.

2.1.1. Kullanım Alanları

Makine öğrenmesi, öneri sistemlerinde sürekli kullanılan bir modeldir. Öneri sistemleri, kullanıcıya sistemde bulunan ve beğenmesi muhtemel olan ürünleri (film, müzik, kitap vb.) önerme özelliği barındıran sistemler olarak karşımıza çıkmaktadır. Yapılan öneriler ile kişinin ne tarz ürünler alacağı tahmin edilmektedir. Öneri sistemleri şu anda en aktif olarak online alışveriş platformları, müzik ve film izleme platformlarında kullanılmaktadır.

Makine öğrenmesi uygulamaları gelecek ile ilgili tahminler yapmak için de sıklıkla kullanılmaktadır. Özellikle hava durumunu belirlerken meteoroloji sistemleri şu anda var olan hava durumu ve geçmişteki durumları değerlendirerek geleceğe yönelik hava tahminleri yapmaktadır.

Görüntü işleme alanında ise makine öğrenmesi; yüz tanıma, parmak izi tanıma, hareketli nesne tanıma, tıbbi uygulamalar gibi alanlarda kullanılmaktadır. Bu yapılan işlemler günümüzde hastane, güvenlik, bankacılık vb. alanlarda sıklıkla kullanılan işlemler olmaktadır.

Makine öğrenmesinin günümüzde en güncel olarak kullanıldığı alan ise otonom araçlar olarak karşımıza çıkmaktadır. Otonom araçlar; sürücüsüz araç, kendi kendine çalışan araç, robotik araç olarak karşımıza çıkmaktadır. Bu araçlar dışardan bir müdahale olmaksızın kendi kendine karar verip uygun bir sürüş modeli ortaya çıkarabilmektedir. Bu işlemi sağlayabilmek için dışarıdan sensörler ve kameralar yardımıyla aldığı verilerle kararlar vermektedir. Hatta Tesla markalı araç diğer Tesla markalı araçlarla bağlantı kurarak sürüş şeklinde değişiklikler yapabilmektedir.

2.1.2. Makine Öğrenmesi Süreci

Makine öğrenmesi ve veri madenciliği yakın ilişki içerisinde olan disiplinlerdir. Veri madenciliğinde makine öğrenmesi algoritmaları, kredi başvurusu, finansal yatırım, tıbbi kayıtlar vb. , büyük veri tabanlarında değerli bilgiyi keşfetme işlemlerinde sıklıkla kullanılmaktadır. Makine öğrenmesi süreci, Veri madenciliği süreçlerine benzer işlemleri içermektedir. Makine öğrenmesi süreci olarak literatürde benzer adımlar verilmektedir. Bunlar;

- problemin tanımlanması,
- veriyi anlama,

- veri hazırlama,
- model kurma,
- model değerlendirme ve seçimi,
- modelin uygulanması olmak üzere 6 aşamadan oluşmaktadır [20-22].

Bir problemi çözmek için problemi iyi okumak ve anlamak çözüme kişiyi yaklaştırmaktadır. Bu işlemin gerçekleştirilebilmesi için problemin tanımının iyi bir şekilde yapılmış olması gerekmektedir. Öncelikle makinenin neyi öğreneceğinin iyi açıklanması gerekmektedir. Daha sonra problemimizi çözmek için bize yardımcı olacak veri seti ile ilgili çalışmaların yapılması gerekmektedir. Bu işleme veri ön işlem süreci ismi verilmekte ve bu süreçte ham verinin analiz edilebilmesi için hazır hale getirilme çalışmaları yapılmaktadır. Burada ki en önemli sorun, eğer veri seti problemimizi çözmeye uygun değilse yapılan çalışmanın hiçbir değerinin olmamasıdır. Bu nedenle çalışma yapılacak veri setinin kaliteli olmasının önemi büyüktür. Verinin kalitesi; doğruluk, tamlık, tutarlılık, güncellik, inanılırlık, katma değer, yorumlanabilir olması ve ulaşılabilirlik gibi farklı durumlar incelenerek belirlenmektedir. Veri ön işlem süreci 5 temel adımı içermektedir [18]. Bunlar;

- veri özetleme,
- veri temizleme,
- veri bütünleştirme ve dönüştürme,
- veri indirgeme,
- veri ayrıklaştırma ve kavram hiyerarşisi oluşturmaktadır [23].

Bu işlemlerin yapılabilmesi için de verinin iyi anlaşılması gerekmektedir. Burada hangi özelliğin hangi tipten veriler içerdiği, hangi özelliğin ne ifade ettiği gibi sorulara cevap aranmaktadır. Bir sonraki aşama olan veri setini hazırlama sürecinde ise, verinin modele uygulanma aşamasından önce birtakım işlemlerden geçmesi gerekebilir. Bunlara örnek olarak; aykırı değerler, tekrar eden veriler ve eksik veri sorunu verilebilir. Bu aşama da bu tarz sorunlara çözümler bulunmaktadır.

2.1.3. Makine Öğrenmesi Yöntemleri

Literatürde öğrenme yöntemleri 4 temel başlıkta incelenmektedir. Bunlar; denetimli öğrenme, denetimsiz öğrenme, yarı denetimli öğrenme ve pekiştirmeli öğrenmedir.

Denetimli öğrenme, verilerin karar sınıflarının yani etiketlerinin, hangi gruba ait olduklarının belirli olduğu verilere ait oluşturulan modelleri içermektedir. Burada modele öğrenmesi için karar sınıfı belli olan eğitim verisi verilir. Daha sonra yine karar sınıfı belirli olan fakat modelin bu sonuçları görmeden test etmesini sağladığı test verisi ile oluşturulan modelin test edilmesi hedeflenmektedir. Burada amaçlanan daha önce bilinen ve modelin tahmin ettiği sonuçlardan elde edilen hatanın en aza indirgenmesinin sağlanmasıdır [24]. Verinin etiketini içeren her bir veriye

sınıf, bu sınıflandırma problemini çözmek için kullanılan öğrenme algoritmasına ise sınıflandırıcı ismi verilmektedir. Örneğin; bir hastanedeki hastaların hasta olup olmadıklarına dair bilgiler toplanırken, kişinin demografik bilgilerinin yanında biyolojik veriler ve sonuç olarak hasta olup olmadığını gösteren verileri de toplanmaktadır. Burada kişinin hasta olup olmadığı verisi bizim sınıfımızı ifade etmektedir. Toplanan bu veriler ile gelecekte hastaneye başvuracak hastaların hasta olup olmadıklarına dair sonuçlar elde etmek hedeflenmektedir. Yapılması gereken bu problemi çözecek olan modelin oluşturulmasıdır. Bunun için elimizdeki veriyi eğitim ve test verisi şeklinde iki parçaya bölmemiz gerekmektedir. Eğitim verisi ile model oluşturulacak ve test verisi ile de modelin ne kadar doğru tahminler de bulunduğu test edilecektir. Denetimli öğrenme modelleri, geçmişteki verilerle diğer bir ifade ile geçmiş deneyimlerle geleceğe yönelik tahminlerde bulunmaktadır. Önemli olan bu işlemi en iyi yapacak modelin seçiminin doğru yapılmasıdır.

Denetimsiz öğrenme, yalnızca giriş bilgilerinin olduğu çıkış bilgilerinin olmadığı durumlarda kullanılan bir öğrenme yöntemidir. Bir başka ifade şekliyle, etiketli veriler olmadan veri kümelerindeki girdiler üzerinden sonuç üretmek için kullanılan bir makine öğrenme yöntemidir [25]. Denetimsiz Öğrenme, kümeleme, aykırılık tespiti, sinir ağları, gizli değişken modellerin öğrenilmesi gibi çalışmaları içermektedir. En yaygın kullanım alanı ise, gizli model bulma veya etiketli olmayan verileri gruplandırma aşaması için kullanılmakta olan kümeleme analizidir [26].

Yarı denetimli öğrenme modeli, elde az miktarda etiketlenmiş ve çok miktarda etiketlenmemiş veri bulunduğunda kullanılabilecek bir yöntemdir. Bu hibrit yöntem denetimli ve denetimsiz öğrenme yöntemlerinin arasında bir yöntemdir. Burada hedeflenen az miktarda etiketlenmiş olan veriden yola çıkarak etiketlenmemiş veriler hakkında bilgi sahibi olmaya çalışmak yani onları sınıflandırmaktır. Denetimli öğrenme yöntemi ile en önemli farkı etiketlenmiş veri sayısıdır [27].

Bunların dışında son zamanlarda büyük bir önem kazanan pekiştirmeli öğrenme türü de bulunmaktadır. Takviyeli öğrenme, çevre ile etkileşime dayalı amaca yönelik bir öğrenme yöntemidir. Bu yöntemde bilgisayarda bir ajan ödüllerini maksimize edecek şekilde, deneme yanılma ile hareket etmektedir. Bu deneme yanılma sürecinde ajanın yaptığı her doğru ve yanlışta bilgisayarda bir değer hesaplanır. Bu değer bir tabloda saklanır. Ajan öğrenme işlemi gerçekleştirdikçe, bu değerler sürekli güncellenir ve süreç sürekli olarak olasılıkların son değerlerden hesaplandığı bir süreç olarak karşımıza çıkmaktadır. Son yıllarda Derin öğrenme olarak karşımıza çıkan model pekiştirmeli öğrenme sürecine en güzel örnek olarak verilebilir [28].

2.2. Sınıflandırma

Denetimli öğrenme yöntemi olan sınıflandırma, makine öğrenmesinde en sık kullanılan yöntemlerden birisidir. En sık kullanıldığı alanlar; örüntü tanıma, dolandırıcılık tespiti, hastalık

tanıları ve pazarlama konularıdır. Verinin sınıflandırılması işlemi, daha önce sınıf etiketleri belli olan veri setini eğitim ve test olmak üzere iki parçaya bölme işlemi ile başlamaktadır. Burada eğitim verisi ile oluşturulmak istenen model tasarlanmaktadır. Test verisi, oluşturulan modelin test edilerek uygun modelin oluşturulup oluşturulmadığı denetlenmesini sağlamaktadır. Oluşturulan model yardımıyla yeni bir örnek ile karşılaşıldığında örneğin hangi sınıfa ait etiketle etiketleneceği kolaylıkla belirlenebilmektedir [29].

Sınıflandırmanın amacı, benzer özellikteki verilerin önceden etiketlenmiş veri gruplarından hangisine ait olduğunu tahmin edilmesi işlemidir. Literatürde sınıflandırma işlemi için çok fazla algoritma bulunmaktadır. Bazıları şunlardır; Karar Ağaçları, Yapay Sinir Ağları, Bayes, Destek Vektör Makineleri, K-en Yakın Komşu algoritmalarıdır.

2.2.1. Sınıflandırma Yöntemleri

Makine öğrenmesinde verilerin etiketlendiği durumda kullanılan sınıflandırma yöntemi için farklı algoritmalar geliştirilmiştir. Bu algoritmaların temel amacı, verinin içerisinde gizlenmiş gizli örüntüyü bulmaktır [30]. Sınıflandırma algoritmaları içerisinden hangisinin kullanılacağı, veri setini test ve eğitim verisi olarak ayırma işlemi, daha sonra eğitim verisi ile modeli oluşturup test verisi ile de modeli test etme aşamasından sonra belirlenmektedir. Burada hangi model daha doğru tahminlerde bulunuyorsa o model sınıflandırma modelimiz olmaktadır [31].

K-en yakın komşu algoritması

Mesafeye dayalı algoritmalardan birisidir. Temel mantığı, verilerin birbirleriyle olan uzaklıkları ve benzerliklerini kullanarak sınıflama işlemi gerçekleştirmekdir. Örneklem yoluyla öğrenmeye dayalı bir algoritmadır. Bu teknikte tüm veri bir örüntü uzayında saklanır. Bilinmeyen bir veri geldiğinde verinin hangi sınıfa ait olduğunu belirlemek için örüntü uzayı araştırılarak sınıfı bilinmeyen veriye en yakın k adet veri belirlenir. Yakınlık ölçüsü olarak farklı ölçüm teknikleri bulunmaktadır. En sık kullanılan mesafe ölçütleri Öklid ve Manhattan ölçütleridir. Daha sonra ise, veri kendisine yakın olan bu k adet veriden hangisine daha çok benziyorsa onun sınıfıyla etiketlenir [32]. Bu yöntemdeki en önemli sorunlardan bir tanesi k değerinin seçimidir. Literatürde k değerinin seçiminde bir ön çalışma yapılması ve çapraz geçerlilik ile incelenilmesi önerilmektedir. Bunun yanında örnek sayısının çok olduğu veri setlerinde k değeri için büyük değer kullanılması önerilmektedir [33].

Naive Bayes

Naive Bayes sınıflandırma algoritması, Bayes teoreminden faydalanılarak oluşturulmuş bir algoritmadır. Literatürde en sık kullanılan algoritmalardan biri olmasının sebebi kolay anlaşılabilir

olması ve uygulanabilirliğidir. Bu yöntem kullanılarak bir verinin hedeflenen niteliğin sınıf değerlerine ait olma olasılığı bulunabilmektedir [34]. Diğer bir deyişle, elde var olan, sınıflandırılmış veriler kullanılarak yeni gelen verinin mevcut bulunan sınıf etiketlerinden birisi olma ihtimalini hesaplayan bir yöntemdir.

Karar Ağaçları

Sınıflandırma problemlerinde, oluşturulması, yorumlanması ve veri seti ile bütünleştirilmesinin kolaylığından dolayı tercih edilen bir yöntemdir. KA, veri yapılarındaki ağaç yapısı ile benzerlik göstermektedir. KA düğüm ve dallardan oluşan anlaşılması kolay olan bir algoritmadır. KA'da her dal bir olasılık durumunu temsil etmektedir. Bir KA, prensip olarak veriyi rekürsif olarak alt gruplara dallanma yaparak bölmektedir. Bu ayrım aşamasında oluşan dalların her biri bir kuralı temsil etmektedir [34]. İki aşamadan oluşmaktadır. İlk olarak modelin oluşturulması, ikinci aşamada ise oluşturulan modele veri tabanından gerekli bilgileri alarak verinin hangi sınıfta olduğunu belirleme işlemi yapılmaktadır. KA oluşturma algoritmalarına örnek olarak; CHAID, CART, ID3, C4.5, SLIQ vb. verilebilir [35].

Rasgele Orman

RO sınıflandırma algoritması, ağaç tipindeki sınıflandırma algoritmalarının topluluğu olarak tanımlanabilmektedir. RO, tüm öznitelikleri ile en iyi dalı kullanarak her bir düğümü dallara ayırmak yerine, her bir düğümden rastgele seçilen öznitelikler arasından en iyisini kullanarak her bir düğümü dallara ayırır. Model oluşturulurken kullanılan her bir veri seti orijinal veri setinden yer değiştirilerek elde edilir. Sonra rastgele öznitelik seçimi yapılarak ağaçlar geliştirilir. Geliştirilen ağaçlar budanmaz ve yapılan işlem RO'ın doğruluğunu eşsiz yapar. RO algoritması çok hızlı ve aşırı öğrenmeye karşıda dayanıklı bir algoritmadır. Algoritma uygulanırken istenilen sayıda ağaç kullanılabilmektedir [36-38].

Lojistik Regresyon Sınıflandırıcı

Regresyon analizi, iki ya da daha fazla değişken arasındaki ilişkiyi ölçme amacıyla kullanılmakta ve tanımlayıcı çıkarımsal istatistik sağlamaktadır [18]. LRS'de hedeflenen, bağımlı ve bağımsız değişkenler arasındaki ilişkiyi en az sayıda değişken ile en iyi uyuma sahip olacak şekilde oluşturabilecek en iyi modeli kurmaktır.

Yapay Sinir Ağları

YSA, insan beynini model olarak alarak geliştirilen bir teknolojidir. YSA pek çok teknolojide kullanıldığı gibi veri madenciliğinde de kullanılmaktadır. YSA öğrenme yöntemiyle, yeni bilgileri üretme, bu bilgileri keşfetme, düşünebilen sistemler tasarlamak için geliştirilmiştir. YSA için özel bir yöntem geliştirmeye gerek yoktur. YSA kendi iç kurallarını üretir ve bu kurallarla

elde edilen sonuçları karşılaştırma yaparak kendisini geliştirir. YSA deneme yanılma yoluyla kendisini eğitir ve bir işin nasıl yapılması gerektiğini öğrenir [39].

Destek Vektör Makineleri

DVM, 1992 yılında Boser ve arkadaşları tarafından bulunmuştur [40]. DVM, son yıllarda regresyon ve sınıflandırma ile alakalı problemlerin çözümünde kullanılmaktadır [41]. DVM sınıflandırma işlemi yaparken yüksek düzeyde başarım sağlamak için yüksek boyut özelliklerine sahip çekirdek fonksiyonları kullanmaktadır. Diğer sınıflandırma algoritmalarına nazaran DVM'nin daha başarılı sınıflandırma yaptığı gözlemlenmiştir.

DVM'nin temel amacı, öznitelikler içerisinden sınıfları en uygun biçimde ayırabilecek bir hiperdüzlem bulmaktır. DVM'nin diğer doğrusal yöntemlerden en önemli farkı, test verisi için sınıflandırma yaparken yanlış yapma olasılığının en aza indirilecek çözümler sunmasıdır.

AdaBoost

AB algoritması zayıf sınıflandırıcıların bir araya gelmesiyle ortaya çıkan güçlü bir sınıflandırıcıyı temsil eden topluluk sınıflandırıcısı olarak isimlendirilmektedir. Modelin genel çalışma mantığı her aşamada, bir önceki aşamanın sonucunda yapılan yanlış tahminlerin ağırlığını artırarak, sınıflandırıcının tekrar çalıştırılması ile başlamaktadır. Yapılan bu işlemlerle yanlış yapılan tahminlere odaklanıp, oluşturulan modelin sınıflandırmadaki doğruluk oranını yükseltmek amaçlanmaktadır [42].

Gradyan Artırma

GA sınıflandırma algoritması, 2001 yılında Friedman tarafından ortaya atılmıştır. Yöntemin çalışma şekli; iteratif olarak hatayı tahmin edip, hatanın iyileştirilmesi hedeflenmektedir. GA yöntemi, AB sınıflandırma algoritmasında olduğu gibi örnekleri tekrar ağırlıklandırmaya çalışmamaktadır. Bir önceki aşamada bulunan kayıp fonksiyonunu negatif gradyan vektörüne uydurmaya çalışır. Hatayı azaltmak için, dereceli azaltma sonuçlarını fonksiyona ekler [43].

2.2.2. Model Değerlendirme ve Seçimi

Sınıflandırma algoritmaları ile oluşturulan modellerin değerlendirilmesi yani hangi sınıflandırma modelinin daha doğru sonuçlar ürettiğinin belirlenmesinde bazı yöntemler ve değerlendirme metrikleri kullanılmaktadır. Bu bölümde bu konular ile alakalı literatür çalışmasına kısaca yer verilecektir.

Model Performans Değerlendirme Yöntemleri

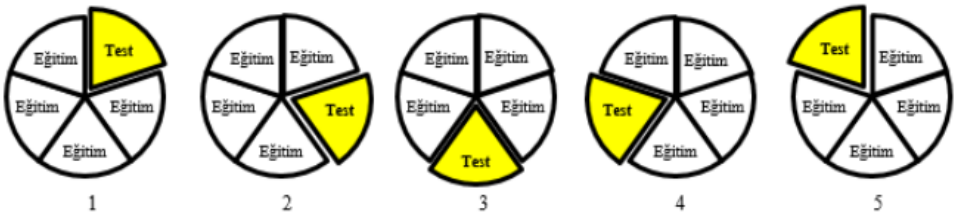
Bir sınıflandırıcı modelin performansında kullanılan algoritma dışında, verinin sınıf etiketlerinin dağılımı, yanlış sınıflandırma veya eğitim ve test verilerinin seçim şekli etkili olmaktadır [8]. Böyle durumlarda çözüm üretebilmek için çeşitli yöntemler geliştirilmiştir. Geliştirilen yöntemler hakkında bilgiler aşağıdaki bölümde verilmektedir [44-48]:

- Hold-out yönteminde veri, eğitim ve test verisi olacak şekilde ikiye ayrılır. Eğitim verisi model olarak seçilen sınıflandırıcının eğitilmesi aşamasında kullanılmaktadır. Model parametreleri için en iyi değerler ve performans ölçütleri bu adımda belirlenmektedir. Test verisi ise, oluşturulmuş olan modelin genel performansını ölçme amaçlı kullanılır. Bu yöntemin dezavantajları, elde bulunan veri az ise test için yeterli düzeyde veri ayırma şansı kalmamaktadır. Diğer bir dezavantajı ise, eğitim ve test verisini ilk başta ayırdığımız için ortaya çıkan performans kriterlerin de bu ayırma işleminden kaynaklanan sorunlar ile karşılaşılabilmektedir.
- Tekrarlı Hold-out yönteminin Hold-out yönteminden farkı, farklı alt veri setleri oluşturularak her seferinde belirli bir oran eğitim verisi ve test verisi olarak ayrılarak hold-out yöntemini birden fazla tekrar ettirmesidir. Bu yöntemde seçme işlemi her ne kadar rasgele yapılsa da, seçilmiş farklı test veri setleri üst üste binebileceğinden, tekrarlı hold-out yöntemi elverişli bir yöntem değildir.
- Tabakalı Örnekleme yöntemi; veri setlerinde çıktı değeri birer sınıf olduğu durumlarda, bazı sınıfların örnekleri az sayıda olabilir. Bu nedenle bu yöntem tercih edilmektedir. Fakat çıktı değeri bir sınıf değil de bir nümerik değer ise o zaman bu yöntem kullanılamaz.
- Üçlü Ayırma yöntemi sınıflandırıcı seçimi ve performans ölçümü aynı anda yapmaktadır. Veri seti, eğitim, doğrulama ve test verisi olacak şekilde üçe ayrılmaktadır. Hold-out yönteminden farklı olarak, doğrulama, veri setindeki örnekler ile eğitim veri setinden seçilen parametrelerin son ayarları yapılmaktadır. Test veri seti ise son ayarlamaları yapılmış modelin son performansını ölçmek için kullanılmaktadır.
- Çapraz doğrulama(ÇD) yönteminde;
    - Bir algoritma kullanılarak mevcut veriden öğrenilen modelin performansını ölçmek,
    - İki ya da daha fazla algoritmanın performansını ölçmek ve mevcut veri için en iyi algoritmayı seçmek ya da parametreli bir modelin iki ya da daha fazla değişkenin performansını kıyaslamak, şeklinde iki temel hedef mevcuttur [49].

ÇD, *k*-kat ÇD (*k*-fold cross validation) ve birini dışarıda bırak ÇD (leave one out cross validation – LOOCV) olmak üzere iki farklı şekilde kullanılmaktadır. ***k***-kat ÇD, veri öncelikle k adet parçaya bölünmektedir. Bölünen parçalardan bir tanesi test geri kalanlar yani k-1 parça eğitim amaçlı kullanılmaktadır. Bu aşamaların her birinde bir değerlendirme sonucu diğer bir ifade ile

modelin başarım oranı ortaya çıkmaktadır. Modelin genel performansını ölçmek için bu k adet parçalanan verinin her biri için çıkan sonuç toplanıp *k*'ya bölünmesi gerekmektedir. Burada seçilen *k* kat sayısı ne kadar büyük seçilirse veri o kadar fazla parçaya ayrılıp, sınıflandırıcının tahmininin daha doğru, gerçek hata tahmincisinin sapması küçük; varyans ve hesaplama zamanı ise büyük olacaktır [46,50]. Bir başka deyişle, *k* kat sayısı küçük seçilirse, tahmincinin sapması küçük ve sapması gerçek hata tahmincisinden büyük olacaktır [46,51]. Sapmadan kaynaklanan hata gerçek değer ile tahmin edilen değer arasındaki farktır, varyanstan kaynaklanan hata ise verilen bir nokta için model tahminindeki değişikliktir [52].

k değerinin en sık kullanılan değeri 10'dur [51, 52-56]. Kohavi [57], model seçimi için tabakalı 10-kat çapraz geçerlemeyi önermiştir. 5-kat çapraz geçerleme Rogers ve Girolami [58]'den faydalanılarak Şekil 2.1'de gösterilmiştir.



Şekil 2.1. 5 - kat çapraz geçerleme

Birini dışarıda bırak yönteminde ise, veri setinde her seferinde sadece 1 örnek test veri seti, geri kalan örnekler ise eğitim veri seti olarak kullanılmaktadır. Bu yöntem özellikle veri setindeki örnek sayısının çok az olduğu durumlarda tercih edilmektedir.

Monte-Carlo ÇD olarak da bilinen rasgele örnekleme yönteminde ise veri seti *k* defa bölünmektedir. Veri setindeki örnekler kullanıcının belirlediği oranda eğitim ve test veri setine dâhil edilmektedir. ÇD'den farkı, analiz süresince kullanılacak k adet test verisinde aynı noktaların tekrar edebilmesidir.

m adet örnekten oluşan bir veri seti ile modelleme çalışması yapıldığı durumda, Bootstrap yöntemi, veri setinden eğitim veri seti için m defa rasgele örnek seçmektedir; ancak her defasında seçilen örnek veri setinden çıkarılmaz. Bu nedenle aynı örnek, eğitim veri setinde birden fazla tekrar edebilmektedir. Eğitim veri seti için örnek seçim işlemi bittikten sonra, eğitim veri setinde olmayan asıl veri setindeki tüm örnekler test veri setine atanmaktadır. Test veri setinde her örnek sadece bir defa tekrar etmektedir.

Model Performansını Değerlendirme Metrikleri

Hangi sınıflandırma algoritmasının daha iyi performans verdiğini bulabilmek için çeşitli yöntemler geliştirilmiştir. Bunlardan bazıları, literatürde kontenjans tablosu, olabilirlik çizelgesi,

hata matrisi, karışıklık matrisi, karmaşıklık matrisi olarak isimlendirilen bir tablonun oluşturulmasına dayanmaktadır [23, 59]. Makine öğrenimi alanında ve özellikle istatistiksel sınıflandırma probleminde, bir hata matrisi olarak da bilinen karmaşıklık matrisi, bir algoritmanın performansını görselleştirmeyi sağlayan tipik bir tablo düzenidir.

Matrisin her satırı, tahmin edilen sınıftaki örnekleri temsil ederken, her sütun gerçek bir sınıftaki örnekleri temsil eder (veya tam tersi). İki boyutlu ("gerçek" ve "öngörülen") ve her iki boyutta da aynı "sınıflar" kümeleriyle (her bir boyut ve sınıf birleşimi olasılık tablosunda bir değişken olan) özel durum tablosudur. Örnek karmaşıklık matrisleri Tablo 2.1 ve Tablo 2.2'de verilmiştir.

Birçok değerlendirme ölçütünde kullanılan bazı terimler aşağıda listelenmektedir;

Doğru pozitif (DP): Sınıflandırıcı tarafından doğru şekilde etiketlenmiş pozitif veri gruplarını ifade eder.

Doğru Negatif (DN): Sınıflandırıcı tarafından doğru şekilde etiketlenmiş negatif veri gruplarını ifade eder.

Yanlış Pozitif (YP): Gerçekte negatif veri grubu olarak etiketlenmiş olan ama sınıflandırıcı tarafından pozitif olarak etiketlenmiş veri gruplarını ifade eder. (Örnek olarak; kanser olan bir hastanın kanser değil olarak etiketlenmesi verilebilmektedir.)

Yanlış Negatif (YN): Gerçekte pozitif olarak etiketlenmiş olan ama sınıflandırıcı tarafından negatif olarak tahmin edilmiş veri gruplarını ifade eder.(Örnek olarak; kanser olmayan bir hastanın kanser olarak etiketlenmesi verilebilmektedir.)

Tablo 2.1. İki sınıflı karmaşıklık matrisi

		Tahmin Edilen Sınıf		
		Sınıf=Evet	Sınıf=Hayır	Toplam
Gerçek Sınıf	Sınıf=Evet	DP	YN	P
	Sınıf=Hayır	YP	DN	N
	Toplam	P^I	N^I	P+N

Tablo 2.2. Çok sınıflı karmaşıklık matrisi

		Tahmin edilen sınıf					Toplam
		C_1	C_2	C_3	…	C_N	
Gerçek Sınıf	C_1	D_1	Y_{12}	Y_{13}	…	Y_{1n}	N_1
	C_2	Y_{21}	D_2	Y_{23}	…	Y_{2n}	N_2
	C_3	Y_{31}	Y_{32}	D_3	…	Y_{3n}	N_3
	…	…	…	…	…	…	…
	C_N	Y_{n1}	Y_{n2}	Y_{n3}	…	Dn	n_n
Toplam		N_1	N_2	N_3	…	N_n	

Karmaşıklık matrisi, sınıflandırıcının farklı sınıfların verilerini ne kadar iyi tanıyabileceğini analiz etmek için yararlı bir araçtır. *YP* ve *YN,* sınıflandırıcının ne zaman yanlış gittiğini bize bildirirken (yani, yanlış etiketleme) *DP* ve *DN,* sınıflandırıcının doğru tahminde bulunduğunu ifade etmektedir. *m* sınıfı verildiğinde (burada $m \geq 2$), bir karmaşıklık matrisi en azından m x m boyutlu bir tablodur. İlk m satır ve m sütunlarındaki bir giriş, $CM_{i,j}$, sınıflandırıcı tarafından sınıf *j* olarak etiketlenen sınıf *i*'nin veri grupları sayısını belirtir. Bir sınıflandırıcının doğru tahminlerde bulunduğunu göstermesi için doğru tahminlerin genel olarak karmaşıklık matrisinin diyagonali boyunca olması yani $CM_{1,1}$ girişinden $CM_{m,m}$ girişine kadar, diğer tahminlerin ise sıfır veya sıfıra yakın olacak şekilde temsil etmesi gerekir. Yani, ideal olarak *YP* ve *YN* sıfır civarında olması beklenmektedir.

Karmaşıklık matrisinde bazı durumlarda ek satır ve sütunlar kullanılabilir. Bunlar Tablo 2.1'de de gösterildiği gibi toplam pozitif sınıf değerleri için *P*, Toplam negatif sınıf değerleri için *N,* tahmin edilen toplam pozitif sınıf değerleri için P^I, tahmin edilen toplam negatif sınıf değerleri için N^I ve toplam veri sayısını ifade etmek için $P+N$. Tablo 2.1'de gösterilen karmaşıklık matrisi 2 sınıflı bir model için temsili olarak gösterilmektedir. Tablo 2.2'deki karmaşıklık matrisi ise çok sınıflı bir model için örnek olarak verilmiştir. Burada C_N, karar sınıflarımızı, D_N, doğru tahmin edilen veri sayılarını, Y_N ise yanlış tahmin edilen verileri temsil etmektedir.

Sınıflandırıcıların kullanım amaçlarına, problemin büyüklüğüne ve beklenen doğruluk düzeyine göre farklı skor değerleri kullanılarak sınıflandırıcılar değerlendirilebilir. Karmaşıklık matrisi kullanılarak sınıflandırıcının performansını ölçmek için bazı ölçütler oluşturulmuştur. Bunlar; doğruluk, duyarlılık, anma, kesinlik, özgüllük, DPO, DNO, YNO, YPO, F ölçütü, MCC, AUC ve ROC eğrisi metrikleridir.

Doğruluk ve Hata Oranı

Eğitim kümesi kullanılarak oluşturulan modelin test kümesindeki verileri doğru sınıflandırma oranına doğruluk adı verilmektedir. Örüntü Tanımada, bu sınıflandırıcının genel tanıma oranı olarak da adlandırılmaktadır. Bütün hata tipleri dikkate alınarak, pozitif ve negatif

örnekleri aynı derecede önemsemeyi sağlar. Sınıflandırıcının toplam performansını değerlendirmeye yardımcı olur. Fakat doğruluk ölçütü, veri kümesinde dengesiz dağılım varsa yeterli olmamaktadır. Çünkü doğruluk ölçütü veri prevalansından etkilenmektedir. Örneğin bir tanı testi ile hastalık prevalansı %3 olan bir toplumda tüm deneklere hastalık yok tanısı konulması %97'lik bir genel doğruluk oranı elde edilmesini sağlar. Fakat bu testin hastaları sağlıklılardan ayıramadığı açıktır. Doğruluk, doğru olarak tahmin edilip sınıflandırılmış veri sayısının toplam veri sayısına bölümü ile bulunur. Doğruluk ölçütü iki sınıflı bir durum için Denklem 2.1'deki denklem ile hesaplanmaktadır.

$$Doğruluk = \frac{DP + DN}{DP + DN + YP + YN} \tag{2.1}$$

Bir sistemin hata oranı veya yanlış sınıflandırma oranı ise, sınıflandırıcımızın tespit ettiği doğruluk oranının birden çıkartılması ile elde edilir. Hata Oranı Denklem 2.2 veya Denklem 2.3'deki denklemlerle hesaplanmaktadır.

$$Hata\ oranı = \frac{YP + YN}{P + N} \tag{2.2}$$

$$Hata\ oranı = 1 - Doğruluk \tag{2.3}$$

Bir modelin hata oranını tahmin etmek için bir test seti yerine eğitim seti kullanılırsa, bu değer yeniden aktarma hatası olarak adlandırılır [59].

Duyarlılık (Doğru Pozitif Oran, Anma) ve Özgüllük (Doğru Negatif Oran)

Eğer bir sınıflandırıcı tüm pozitif sınıfları pozitif olarak ve tüm negatif sınıfları negatif olarak etiketlemişse o sınıflandırıcıya altın standart denir. Bir sınıflandırıcı oluştururken hedeflenen altın standart bir sınıflandırıcı oluşturmaktır.

Sınıflandırıcının, duyarlılık ve özgüllük değerlerini saptamak dengesiz veri dağılımı olduğu durumlarda önem taşımaktadır. Sınıflandırıcının duyarlılığı, incelenen verilerde olayın gerçekte var olma durumunu, yani gerçek olumluları saptama yeteneğini belirler. Başka bir şekilde ifade edilecek olursa, pozitif sınıfa ait olduğu bilinen bir verinin, test sonucunda da pozitif sınıfa etiketlenme olasılığıdır. Örneğin, gerçekte kanser hastası olan bir hastanın test sonucunda da hasta olarak etiketlenmesidir. Özgüllük ise, incelenen verilerde olayın bulunmaması durumunu, yani gerçek olumsuzluğu saptama yeteneği olarak ifade edilmektedir. Başka bir şekilde ifade edilecek olursa, kendi sınıfından olmayan verilerin, kendi sınıfı dışındaki sınıflarda sınıflandırılmasıyla, doğru olarak sınıflandırılmasıdır. Özgüllük, doğru olarak sınıflandırılan negatif sınıfa ait verilerin oranıdır. Ölçümün bu iki özelliği, bir sınıflandırıcının doğru sınıfı yanlış sınıftan ne derece

ayırabildiğini ve ne küçüklükte bir ayrımın belirlenip ölçülebildiğini anlatır. Duyarlılık ve özgüllük ikili sınıflandırmada istatiksel olarak performans ölçme göstergeleridir.

Dengesiz veri dağılımı olan veri gruplarında mümkün olduğu kadar duyarlılığın ve seçiciliğin yüksek olması arzu edilir.

Bir testin duyarlılığı ne kadar yüksekse, yanlış negatiflik olasılığı o kadar azalır. Bir testin seçiciliği ne kadar yüksekse, yanlış pozitiflik olasılığı o kadar azalır. Duyarlılık Denklem 2.4'deki denklem ile hesaplanmaktadır.

$$Duyarlılık = Anma = DPO = \frac{DP}{DP + YN} \tag{2.4}$$

Sınıflandırıcının pozitif sınıf etiketlerini tahmin etmedeki etkililiğine duyarlılık denmektedir.

DP+YN adet pozitif veri içerisinde test sonucu pozitif olan *DP* adet veri vardır. Bu yüzden duyarlılık, DPO olarak da adlandırılır.

Mükemmel bir sınıflandırıcı %100 duyarlılık ve özgüllük oranına sahiptir. Duyarlılık için ideal olan *"1"* değerini elde etmektir, ancak pratikte bu mümkün olmayabilir. Yüksek duyarlılığa sahip sınıflandırıcılar, pozitif sınıfa ait verileri yüksek doğrulukta sınıflandırırlar. Duyarlılığa benzer olarak *"1"* özgüllük değerini yakalamak mümkün olmayabilir. Yüksek özgüllüğe sahip sınıflandırıcılar, negatif sınıfa ait verileri yüksek doğrulukta sınıflandırırlar.

Özgüllük Denklem 2.5'daki denklem ile hesaplanmaktadır. Sınıflandırıcının negatif sınıf etiketlerini tahmin etmedeki etkililiğidir.

$$Özgüllük = DNO = \frac{DN}{DN + YP} \tag{2.5}$$

Duyarlılık ve özgüllük değerleri, sınıflandırma yapılacak verinin prevalansından etkilenmez. Bu değerler, verideki gerçek pozitifler ve gerçek negatifler üzerinden hesaplandığından, veri dağılımında ki değişimden etkilenmezler [59].

Yanlış Pozitif Oranı ve Yanlış Negatif Oranı

Yukarıda tanımlanan duyarlılık ve özgüllük değerlerinin tamamlayıcıları olarak YPO ve YNO bulunmaktadır.

YNO Tablo 2.1'deki değerlere göre Denklem 2.6'deki gibi hesaplanmaktadır.

$$YNO = \frac{YN}{DP + YN} \tag{2.6}$$

Sınıflandırma sonucu negatif sınıfta etiketlenen fakat gerçekte pozitif sınıfa ait olan *FN* adet verinin tüm pozitif sınıftaki elemanlara bölümü ile bulunur. Diğer bir ifade ile Duyarlılık ve YNO'nin toplamı bire eşit olmaktadır ve bu Denklem 2.7'de gösterilmektedir.

$$Duyarlılık + YNO = 1 \tag{2.7}$$

YPO, Tablo 2.1'deki karmaşıklık matrisindeki değerlere göre Denklem 2.8'daki gibi hesaplanmaktadır.

$$YPO = \frac{FP}{TN + FP} \tag{2.8}$$

Sınıflandırma sonucu pozitif sınıfta etiketlenen fakat gerçekte negatif sınıfa ait olan *YP* adet verinin tüm negatif sınıfa ait elemanlara bölümü ile bulunur. Diğer bir ifade ile özgüllük ile YPO'nın toplamı bire eşit olmaktadır ve bu Denklem 2.9'da gösterilmektedir [60].

$$Özgüllük + YPO = 1 \tag{2.9}$$

Kesinlik (Pozitif Kestirim Değeri) ve Negatif Kestirim Değeri

Doğruluk ölçütü, veri kümesinde dengesiz dağılım var ise yeterli olmamaktadır. Bu durumda kullanılan anma ve kesinlik ölçütleri, sırasıyla, pozitif örneklerin negatif olarak sınıflandırılmasından oluşan hatalar ile negatif örneklerin pozitif olarak sınıflandırılmasından oluşan hataları belirtmektedirler.

Bir örüntü tanımada, bir sınıfın kesinlik değeri ve Pozitif Kestirim Değeri (PKD), gerçek sınıfa ait olarak etiketlenmiş elemanların, toplam pozitif sınıf olarak etiketlenmiş eleman sayısına oranı olarak hesaplanmaktadır. Bahsi geçen gerçek pozitif sınıf eleman sayısı; doğru olarak etiketlenmiş öğelerin sayısı iken toplam pozitif sınıf eleman sayısı; gerçek pozitiflerin ve yanlış pozitiflerin yani yanlış bir sınıfa ait olarak etiketlenmiş öğelerin toplamına eşittir. "*Sınıflandırıcı, bir sınıf için yaptığı sınıflandırmalarda ne kadar hassastır?"* sorusunun cevabı kesinlik değerlendirme ölçütü ile hesaplanmaktadır [60]. Sınıflandırıcı tarafından doğru negatif olarak tahmin edilenlerin, sınıflandırıcı tarafından negatif olarak tahmin edilen tüm veri grubuna oranına Negatif Kestirim Değeri (NKD) denmektedir.

İki sınıflı bir sınıflandırıcıda kesinlik ve NKD değerleri Tablo 2.1'e göre Denklem 2.10 ve Denklem 2.11'deki gibi hesaplanmaktadır.

$$Kesinlik = PKD = \frac{DP}{DP + YP} \tag{2.10}$$

$$NKD = \frac{DN}{DN + FN} \tag{2.11}$$

Kesinlik bir olduğunda sınıflandırıcının ilgili sınıfta tahmin ettikleri içerisinde olumsuz örnek olmadığı anlamına gelmektedir. Ama ilgili sınıfa ait bazı veriler bu sınıfta tahmin edilmemiş olabilir. Örneğin bir arama motorunda yapılan aramada gelen sonuçların hepsinin aradığımız konu ile alakalı olması ama bunun yanında ilgili olan bazı sayfaların ise listelenmemesi gösterilebilir. NKD'nin yüksek olması sınıflandırma sonucu negatif çıkan bir sonucun gerçekte pozitif olma olasılığının düşük olduğunu gösterir. NKD sınıf dağılımından etkilenir.

F-Ölçütü

Esas olarak istatistik biliminin bir ölçme kavramı olan ve literatürde f_1 ölçütü veya f-ölçütü olarak geçen kavram, bilgisayar bilimlerinde özellikle veri çıkarımı ve veri getirimi konularında kullanılmaktadır.

Kesinlik ve anma performans ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli olmamaktadır. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar vermektedir. Bunun için F-ölçütü tanımlanmıştır. F-ölçütü kesinlik ve anma değerlerinin harmonik ortalamasıdır. F-ölçütü genel olarak, F_1, $F_{0.5}$ ve F_2 olarak kullanılmaktadır. F-ölçütü genel formülü ise F_β ile gösterilir ve Denklem 2.12'deki gibi hesaplanmaktadır.

$$F_\beta - \text{ölçütü} = \frac{(1 + \beta^2)x(kesinlik * anma)}{\beta^2 + kesinlik * anma} \tag{2.12}$$

Burada β değeri genel olarak $\beta = 1$ olarak alınarak, kesinlik ve anma değerlerinin eşit ağırlıklı olarak hesaplanması sağlanmaktadır. F-ölçütü, 0-1 aralığında değerler almaktadır ve yüksek başarımlı bir sınıflandırmada F-ölçütünün bire yakın bir değer alması beklenmektedir. β değeri birin altında seçilirse anma, birin üzerinde seçilirse kesinlik değeri F-ölçütü'nün değerini belirlemektedir [59].

Matthews korelasyon katsayısı

İki ve çok sınıflı sınıflandırma çalışmalarında sınıflandırıcının performansını ölçmek için Matthews korelasyon katsayısı (MCC) olarak adlandırılan bir yöntem önerilmektedir. Bu ölçüm metriği karmaşıklık matrisine ait tüm bilgileri kullanmaktadır. Dengesiz sınıflar söz konusu olduğunda, tahminlerin tüm sonuçlarını kullanarak dengesiz sınıfları işleyen MCC gibi özel ölçüm teknikleri gerekmektedir. MCC değeri -1 ile +1 arasında sonuç üretmektedir. Elde edilen değer +1'e ne kadar yakınsa sınıflandırıcının o kadar doğru tahmin yaptığını göstermektedir. MCC değeri 0'a yakınsa, sınıflandırıcının yaptığı tahminlerin sonuçlarının rasgele olduğunu göstermektedir. Eğer elde edilen değer -1'e yakın ise gerçek ve tahmin sonuçları arasındaki uyumsuzluğu göstermektedir. Çok dengesiz sınıflar söz konusu olduğunda, F_1 değeri yanıltıcı olabilir, çünkü F_1 değeri doğru ölçümün hesaplanmasındaki tüm tahmin sonuçlarını dikkate almaz. Denklem 2.13, MCC değerinin nasıl hesaplanacağını gösterir [61].

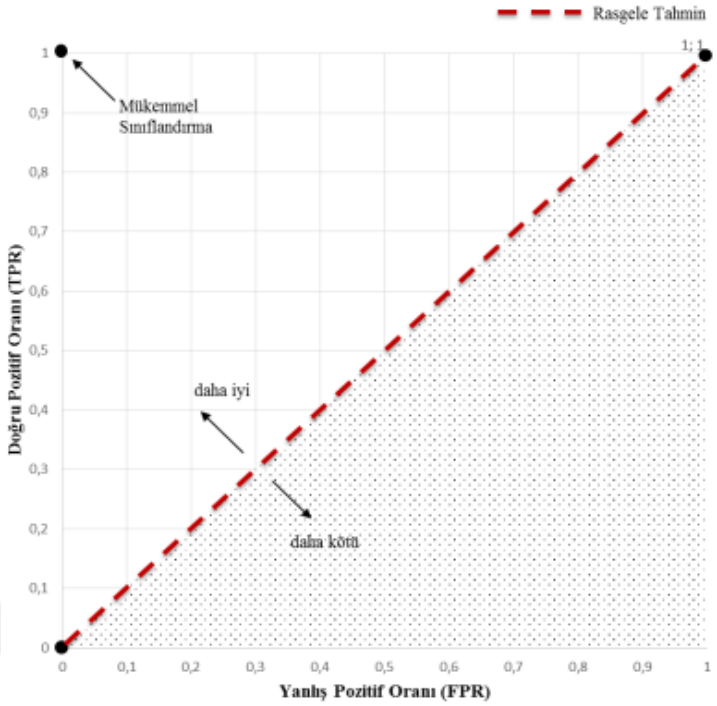
$$MCC = \frac{DP * DN - YP * YN}{\sqrt{(DP + YP) * (DP + YN) * (DN + YP) * (DN + YN)}} \tag{2.13}$$

Çok sınıflı bir veri setin için elde edilen karmaşıklık matrisinden MCC hesaplaması yapmak için Denklem 2.14 kullanılmaktadır.

$$MCC = \frac{\sum_{k,l,m=1}^{N} C_{kk} * C_{ml} - C_{lk} * C_{km}}{\sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{lk}\right) * \left(\sum_{f,g=1\, f\neq k}^{N} C_{gf}\right)\right]} * \sqrt{\sum_{k=1}^{N}\left[\left(\sum_{l=1}^{N} C_{kl}\right) * \left(\sum_{f,g=1\, f\neq k}^{N} C_{fg}\right)\right]}} \tag{2.14}$$

Alıcı İşlem Karakteristikleri Eğrileri

ROC analizi, ikinci dünya savaşı yıllarında sinyallerin doğru tanımlanabilmesi için geliştirilmiştir [62]. *y* ekseninde *DP* ve *x* ekseninde ise *YP* oranlarının gösterilmesiyle çizilmektedir. 1967 yılında Lusted, tıpta karar verme için ROC'u önermiş; 1969 yılında medikal görüntüleme cihazlarında bu analiz kullanılmıştır ve bu kullanım sonraki yıllarda da tıpta tanı testlerinin performans değerlendirilmesinde giderek artmıştır [62]. ROC uzayı Şekil 2.2'deki grafikte gösterilmiştir [29]:



Şekil 2.2. ROC Uzayı [29]

ROC uzayı ile ilgili bilinmesi gereken önemli noktalar şu şekilde sıralanmaktadır [29]

- (0,0) noktası, hiçbir zaman pozitif bir sınıflandırmanın verilmeyeceğini,
- (1,1) noktası, mutlak pozitif sınıflandırmaların elde edileceğini,
- (0,1) noktası, mükemmel sınıflandırmayı göstermektedir.

ROC uzayında *DP* oranı yüksek, *YP* oranının düşük ya da her ikisinin olduğu durumlarda bir sınıflandırıcının diğerinden iyi olduğunu söyleyebilmek mümkündür. Bu nedenledir ki; genellikle bir sınıflandırıcının Şekil 2.2'de verilen grafikte hep sol üst köşeye yakın olması tercih edilmektedir [63].

ROC grafiğinin sol tarafında kalan sınıflandırıcılar, yalnızca güçlü kanıtlar ışığında pozitif sınıflandırmalar yaptıklarından conservative, üst sağda kalan sınıflandırıcılar ise zayıf kanıtlar ışığında pozitif sınıflandırmalar yaptıklarından liberal olarak nitelendirilmektedir.

Grafikteki köşegen ise sınıf tahmininin rasgele yapıldığını göstermektedir. Köşegenin altında kalan ise rasgele tahminden bile kötü olarak nitelendirilmektedir. Bir ayrık sınıflandırıcı, çıktı değeri olarak bir sınıf değeri verdiğinden sınıflandırıcıya ait yalnızca tek bir *YP* ve *DP* oranı elde edilmektedir. Bu nedenle, bir ayrık sınıflandırıcıya ait elde edilen bir tek *YP* ve *DP* ikilisi ROC uzayında tek bir noktaya karşılık gelmektedir [50].

Genellikle belirli bir eşik değere (θ) göre çıktı değerinin hangi sınıfa dâhil olduğunu belirten olasılı sınıflandırıcıların ROC eğrisinin çizilebilmesi için gereken algoritma aşağıda verilmiştir [29]:

Gerçek sınıf değerleri *f(x)*, sınıflandırıcının tahmin ettiği değerler $\hat{f}(x)$, R ROC eğrisini oluşturacak noktaların kümesi olsun.

1.Adım: *YP* ve *DP* değerleri sıfır kabul edilir. $\hat{f}(x)$ değerleri küçükten büyüğe sıralanır.

2.Adım:

- Eğer $\hat{f}(x) > \theta$ ve *f(x)* pozitifse *DP*'nin değeri 1 artırılır.
- Eğer $\hat{f}(x) > \theta$ ve *f(x)* negatifse *YP*'nin değeri 1 artırılır.

3.Adım: $(\frac{YP}{NEG}, \frac{DP}{POZ})$ noktası hesaplanır ve ROC'u oluşturacak noktalardan biri olarak *R* kümesine eklenir.

4.Adım: test veri setindeki bütün örneklere 2. Adım ve 3. Adım uygulanır.

5.Adım: R kümesindeki noktalar yardımı ile ROC çizilir.

ROC eğrisi altında kalan alanı temsil eden AUC metriğine sınıflandırıcıların performansını test etmede sıklıkla başvurulmaktadır. Yukarıda bahsedilen ROC eğrisi (0,0) ile (1,1) noktaları arasında çizilen bir eğridir. Bu eğrinin altında kalan alanı temsil eden AUC ise 0 ile 1 arasında değerler almaktadır. AUC değerinin 1'e yaklaşması sınıflandırıcının doğru tahminlerde bulunduğunu, 0.5 değerini alması sınıflandırıcının rasgele tahminlerde bulunduğunu ve 0 değerine yakın değerler alması da sınıflandırıcının yanlış tahminlerde bulunduğunu göstermektedir [64].

Şimdiye kadar değerlendirilen formüller genel olarak 2 sınıflı sınıflandırıcıları temsil etmektedir. Çok sınıflı sınıflandırıcılarda, modelin performansını değerlendirmek için geliştirilen formüller ise ikili sınıflandırma için verilen formüllerin genelleştirilmiş hali olmaktadır [65-68]. Öncelikle çoklu sınıflar birden fazla ikili sınıflara dönüştürülmekte ve bunların sırayla performans ölçümleri yapılmaktadır. Son olarak ise bu yapılan ölçümlerin ölçüm sayısına bölümüyle ortalama performans ölçüm değerleri bulunmaktadır [69]. $C = \{C_1, C_2, C_3\}$ sınıf değerleri kümesi olsun. Şekil 2.3'deki karmaşıklık matrisi ikili sınıflandırıcı için verilmiştir. C_1 pozitif, C_2 ve C_3 de negatif sınıfına aittir.

		Gerçek Sınıf		
		C_1	C_2	C_3
Tahmin Edilen Sınıf	C_1	C_{11}	C_{12}	C_{13}
	C_2	C_{21}	C_{22}	C_{23}
	C_3	C_{31}	C_{32}	C_{33}

Şekil 2.3. İkili sınıflandırıcı biçiminde verilen karmaşıklık matrisi [71]

Yukarıda anlatılan sınıflandırıcı performans ölçüm formüllerinin çok sınıflı bir sınıflandırıcıdaki karşılıkları aşağıda listelenmektedir.

μ mikro, M makro ortalama, l sınıf sayısı olmak üzere;

$$Ortalama\ Doğruluk(aveACC) = \frac{\sum_{i=1}^{l} \frac{(TP_i+TN_i)}{TP_i+TN_i+FP_i+FN_i}}{l} \quad (2.15)$$

$$Ortalama\ Hata\ Oranı(aveERR) = \frac{\sum_{i=1}^{l} \frac{(FP_i+FN_i)}{TP_i+TN_i+FP_i+FN_i}}{l} \quad (2.16)$$

$$Pozitif\ Öngörü\ Değeri_{\mu}(mPPV) = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l}(TP_i+FP_i)} \quad (2.17)$$

$$Duyarlılık_{\mu} = \frac{\sum_{i=1}^{l} TP_i}{\sum_{i=1}^{l}(TP_i+FN_i)} \quad (2.18)$$

$$F-Ölçütü_{\mu}(mF) = \frac{2xmPPVxmTPR}{mPPV+mTPR} \quad (2.19)$$

$$Pozitif\ Öngörü\ Değeri_{M}(MPPV) = \frac{\sum_{i=1}^{l}(\frac{TP_i}{TP_i+FP_i})}{l} \quad (2.20)$$

$$Duyarlılık_{M}(MTPR) = \frac{\sum_{i=1}^{l}(\frac{TP_i}{TP_i+FN_i})}{l} \quad (2.21)$$

$$F-Ölçütü_{M}(MF)\frac{2xMPPVxMTPR}{MPPV+MTPR} \quad (2.22)$$

Yukarıda çoklu-sınıf sınıflandırma performans ölçüleri arasında listelenmeyen ikili sınıflandırma için verilmiş diğer performans değerlendirme ölçüleri de benzer biçimde mikro ve makro düzeyde formüle edilmiştir:

$$Seçicilik_{\mu}(mSPC) = \frac{\sum_{i=1}^{l} TN_i}{\sum_{i=1}^{l}(TN_i+FP_i)} \quad (2.23)$$

$$Seçicilik_{M}(MSPC) = \frac{\sum_{i=1}^{l}(\frac{TN_i}{TN_i+FP_i})}{l} \quad (2.24)$$

$$Yanlış\ Pozitif\ Oranı_{\mu}(mFPR) = \frac{\sum_{i=1}^{l} FP_i}{\sum_{i=1}^{l}(FP_i+TN_i)} \quad (2.25)$$

$$Yanlış\ Pozitif\ Oranı_{M}(MFPR) = \frac{\sum_{i=1}^{l}(\frac{FP_i}{FP_i+TN_i})}{l} \quad (2.26)$$

$$Yanlış\ Negatif\ Oranı_{\mu}(mFNR) = \frac{\sum_{i=1}^{l} FN_i}{\sum_{i=1}^{l}(FN_i+TP_i)} \quad (2.27)$$

$$Yanlış\ Negatif\ Oranı_{M}(MFNR) = \frac{\sum_{i=1}^{l}(\frac{FN_i}{FN+TP_i})}{l} \quad (2.28)$$

$$Negatif\ Öngörü\ Değeri_{\mu}(mNPV) = \frac{\sum_{i=1}^{l} TN_i}{\sum_{i=1}^{l}(TN_i+FN_i)} \quad (2.29)$$

$$Negatif\ Öngörü\ Değeri_{M}(MNPV) = \frac{\sum_{i=1}^{l}(\frac{TN_i}{TN_i+FN_i})}{l} \quad (2.30)$$

$$Pozitif\ Olabilirlik\ Oranı_{\mu}(mLR+) = \frac{mTPR}{1-mSPC} \quad (2.31)$$

$$Pozitif\ Olabilirlik\ Oranı_{M}(MLR+) = \frac{MTPR}{1-MSPC} \quad (2.32)$$

$$Negatif\ Olabilirlik\ Oranı_{\mu}(mLR-) = \frac{1-mTPR}{mSPC} \quad (2.33)$$

$$Negatif\ Olabilirlik\ Oranı_{M}(MLR-) = \frac{1-MTPR}{MSPC} \quad (2.34)$$

Çok sınıflı bir sınıflandırıcının ikili sınıflara ayrılmasına örnek ise Tablo 2.3 ve Tablo 2.4'te gösterilmektedir [69].

Burada Tablo 2.3'te 5 sınıflı bir karmaşıklık matrisi örnek olarak verilmiştir. Bu karmaşıklık matrisini 2 sınıflı hale dönüştürme işlemi yapmak için yani Tablo 2.4'teki haline dönüştürmek için yapılması gereken öncelikle sınıf ayrımının gerçekleştirilmesini sağlamaktır. Verilen örnekte *A* sınıfı pozitif sınıf olarak kabul edilmekte ve diğer kalan sınıflarda negatif sınıf olarak kabul edilmektedir. Burada sınıflandırıcı tarafından ve gerçekte pozitif olan verilere bakıldığı zaman on adet veri doğru olarak sınıflandırılmıştır. Dikey olarak *A* sınıfına bakıldığında ise aslında farklı sınıflarda olması gerekirken sınıflandırıcı tarafından yanlışlıkla *A* sınıfı olarak etiketlenen on iki adet veri bulunmaktadır. Bunlar yanlış pozitif olarak isimlendirilmektedir. Yani aslında negatif sınıfta etiketlenmesi gerekirken pozitif olarak etiketlenen veri grubunu temsil etmektedir. Yatay olarak tabloya bakıldığında ise gerçekte *A* sınıfında olması gerekirken yanlışlıkla farklı sınıflarda etiketlenen veri grupları olduğu görülmektedir. Bunlar yanlış negatif olarak isimlendirilmektedir. Yani aslında pozitif veri grubunda olması gerekirken sınıflandırıcı tarafından farklı sınıfta etiketlenmiş olan verilerdir. Geri kalan bölüm ise *A* sınıfında olmayan ve tahmin sonucunda da *A* sınıfında değil olarak etiketlenen yani doğru negatifler gurubudur. Bunların iki sınıflı olarak temsili Tablo 2.4'te listelenmektedir [70].

Tablo 2.3. Çok sınıflı Karmaşıklık Matrisinin iki sınıfa ayrılması

		TAHMİNİ SINIFLANDIRMA				
		A	B	C	D	E
GERÇEK SINIFLANDIRMA	A	10	3	2	8	2
	B	2	11	1	6	4
	C	0	3	18	1	1
	D	8	0	0	14	0
	E	2	0	0	0	19

Tablo 2.4. Çok sınıflı karmaşıklık matrisinin iki sınıflı temsili

		Tahmin Edilen Sınıf		
		Sınıf=Evet	Sınıf=Hayır	Toplam
Gerçek Sınıf	Sınıf=Evet	10	15	25
	Sınıf=Hayır	12	78	90
	Toplam	22	93	115

2.3. Veri setleri

Bu çalışmada kullanılacak veri setleri, veri setlerinde yer alan örneklere ait özellikler ve denge oranları Tablo 2.5'de görülmektedir. Denge düzeyi olarak kullanılan değer, veri setinde yer alan örneklerden en fazla sayıda örnek içeren sınıfa ait örnek sayısının, en az örnek içeren sınıfa ait örnek sayısına bölerek bulunmuştur [71].

Tablo 2.5. Çalışmada kullanılan veri setleri ve özellikleri

Veri seti	**Özellik Sayısı**	**Örnek Sayısı**	**Sınıf Sayısı**	**Sınıf Dağılımı**	**Denge Düzeyi**
Avustralya'da yağmur	24	142193	2	%77.58, %22.42	3.46
Wisconsin Meme Kanseri	31	569	2	%63, %37	1.702
Tic-Tac-Toe Oyunu	10	958	2	%34.65, %65.35	1.886
Sloan gökyüzü araştırması	18	10000	3	%8.5, %41.52, %49.98	5.88
Pulsar yıldızı tahmini	9	17898	2	%90.84, %9.16	9.91
Ortopedik Hastaların Biyomekanik Özellikleri	7	310	2	%68, %32	2.125
Ortopedik Hastaların Biyomekanik Özellikleri	7	310	3	%19.35, %32.25, %48.4	2.501
Kalp Hastalığı	14	303	2	%45.54, %54.46	1.195
Yapısal protein dizileri	6	111338	3	%38.8, %31.4, %29.8	1.302
Farelerin protein ekspresyonu	82	1080	8	%16.3, %13.59, %13.59, %13.59, %13.04, %10.87, %10.87, %8.15	2
Seyahat sigortası	11	63326	2	%1.54, %98.54	63.98

Tablo 2.5. Çalışmada kullanılan veri setleri ve özellikleri (Devamı)

Veri seti	Özellik Sayısı	Örnek Sayısı	Sınıf Sayısı	Sınıf Dağılımı	Denge Düzeyi
Kas aktivitesini okuyarak jestlerin sınıflandırılması	64	11678	4	%24.92, %24.86, %25.2, %25.02	1.0
Parkinson hastalığı sınıflandırma	755	756	2	%25.4, %74.6	2.9
Portekiz bankası pazarlama	18	40841	2	%11.36, %88.64	7.8
Genetik çeşitlilik sınıflandırma	46	65188	2	%25.21, %74.79	2.9
Mobil cihaz fiyat sınıflandırması	21	2000	4	%25, %25, %25, %25	1.0
Türkiye siyasi görüşleri	16	885	6	%2.26, %3.05, %14.24, %21.81, %28.81, %29.83	13.1
Banka pazarlama	17	11162	2	%47.38, %52.62	1.1
İris çiçeği	6	150	3	%33.33,%33.33,%33.33	1.0
Şarap kalitesi	13	6497	7	%0.46, %3.32, %32.91, %43.65, %16.61, %2.97, %0.08	545.6
Hepatoselüler Karsinom(HCC)	50	165	2	%61.81, %38.19	1.6
Bireysel kredi sınıflandırma problemi	14	5000	2	%8.4, %91.6	10.9
Sahte şirketleri denetlemek için denetim riski	25	776	2	%60.69, %39.31	1.5
İnternetten alışverişte kullanıcı tercihleri	18	12330	2	%15, %85	5.6
Şarap için müşteri segmentasyonu	14	178	3	%33.14, %39.9, %26.96	1.4
Ph tanıma	4	653	15	%5.819, %6.738, %6.738, %6.738, %6.738, %6.738, %6.738, %6.738, %6.738, %6.738, %6.738, %6.738, %6.738, %6.584, %6.738	1.1
Gelir sınıflandırması	15	32561	2	%24, %76	3.1
Kriyoterapi analiz	7	90	2	%46.67, %53.33	1.1
Kredi kartı sahtekarlığı tespiti	31	284807	2	%99.82, %0.18	554.5
Deniz kulağı	9	4177	3	%33.68, %31.67 %34.65	1.0
Sesle cinsiyet tanıma	21	3168	2	%50, %50	1.0
Pima Kızılderilileri diyabet	9	768	2	%34.9, 65.1	1.8

Tablo 2.5'de yer alan veri setleri ve özellikleri hakkında bilgilere aşağıda yer verilmiştir.

- Avustralya'da yağmur veri seti, Avustralya'da bulunan çok sayıdaki hava istasyonlarından alınan günlük hava gözlem raporlarını içermektedir. Burada hedeflenen ertesi gün yağmur yağıp yağmayacağını tahmin etmektir. Burada karar sınıfımıza ait etiketleri içeren özellik *"RainTomorrow"* dır [6].
- Tic-tac-toe oyunu veri seti, oyunu oynayan iki kişiden ilk oynayan kişinin x seçtiği varsayılarak elde edilebilecek bütün ihtimallerin olduğu bir tabloyu içermektedir. Tablonun son sütunu haricindeki veriler oyun stratejilerini, son sütun ise bu stratejilere göre elde edilen sonucu göstermektedir [6].
- Wisconsi meme kanseri veri seti, Wisconsin Üniversitesi araştırmacıları tarafından oluşturulmuş ve göğüsten alınan kitlenin ince iğne aspirasyonunun sayısallaştırılmış görüntülerinden yapılan ölçümleri içermektedir. Veri setindeki özelliklerden ilki biyopsiye ait kimlik numarasını, bir diğeri biyopsi sonucunu ve geriye kalan 30 özellik ise laboratuvar sonuçlarını gösteren sayısal değerleri içermektedir [6].
- Sloan dijital gökyüzü araştırması veri seti, ABD'nin New Mexcico eyaletinde yer alan Apache Point gözlemevinde yer alan 2.5 metrelik devasa teleskop kullanılarak yürütülmekte olan bir projeden elde edilen değerleri içermektedir. Buradan elde edilen verilerle evrenin oluşumu, galaksi ve kuazarların kökeni, samanyolu galaksisinin oluşum süreci ve evrim konularında çalışmalar yapılması amaçlanmaktadır [6].
- Pulsar yıldız tahmini veri seti, Yüksek Çözünürlüklü Zaman Evren Anketi sırasında toplanan bir pulsar adayı örneğini tanımlayan bir veri setidir. Pulsarlar, dünyada tespit edilebilen radyo emisyonu üreten nadir görülen bir Nötron yıldızı tipidir. Makine öğrenmesi modelleri artık hızlı analizi kolaylaştırmak için pulsar adaylarını otomatik olarak etiketlemede kullanılmaktadır [6].
- Ortopedik hastaların biyomekanik özellikleri veri seti, Dr. Henrique da Mota tarafından Fransa'nın Lyon şehrinde yer alan Group of Applied Research in Orthopaedies(GARO) kliniğinde tedavi gören hastalardan elde edilen verilerle oluşturulmuştur. Veriler 2 farklı şekilde etiketlenmiştir. Birincisinde hastalar 3 farklı grupta etiketlenmiştir. Bunlar; Normal, Disk fıtığı ve Spandilolistezdir. İkinci etiketleme türünde ise disk fıtığı ve Spandilolistez hastalarını 'Abnormal' olarak etiketleyerek karar sınıfına ait etiketleri ikiye düşürülmesi sağlanmıştır [6].
- Yapısal protein dizileri veri seti, Research Collaboratory for Structural Bioinformatics (RCSB)'in Protein veri bankasından(Protein Data Bank-PDB) alınan verileri içermektedir. PDB arşivi, atomik koordinatlar ve diğer bilgileri açıklayan proteinler ve diğer önemli biyolojik makromoleküllerin bir havuzdur [6].

- Kalp hastalığı veri seti, 76 öznitelikten oluşmaktadır. Fakat yapılan bütün deneysel çalışmalarda bu 76 özniteliğe ait 14 alt öznitelik ile çalışılmıştır. Kalp hastalığı veri seti 4 ayrı veritabanından oluşmakta ve kalp hastalığı ile ilgili bilgileri içermektedir. Bu 4 veritabanı, statlog, Cleveland, hungary ve switzerland'dır. Hepsi aynı öznitelikleri içermektedir. Bu çalışmada bu 4 veritabanından Cleveland veritabanı kullanılmıştır [6].
- Farelerin protein ekspresyonu veri seti, korteksin nükleer fraksiyonunda tespit edilebilir sinyaller üreten 77 protein/protein modifikasyonunun ekspresyon seviyelerinden oluşmaktadır. 38 kontrol faresi ve 34 down sendromlu toplam 72 fare bulunmaktadır. Deneylerde numune/fare başına her proteinin 15 ölçümü yapılmıştır [6].
- Seyahat sigortası veri seti, seyahat sigortası yapan farklı firmaların yaptıkları işlere göre müşterilerinin sigorta isteyip istemediklerinin ve yaptıkların işlemlerin örneklerini içermektedir [6].
- Kas aktivitesini okuyarak jestlerin sınıflandırılması veri seti, bir grup araştırmacı tarafından yürütülen proje sonucunda elde edilen verilerle oluşturulmuştur. Projede, protez cihazların çoklu serbestlik derecelerine sahip olmalarını sağlayacak açık kaynak kodlu bir protez kontrol sistemi oluşturulmaktadır. Sistem birkaç bileşenden oluşmaktadır. İlk bileşen bir kullanıcı aracılığı ile oluşturulan android uygulamasına sensörden gelecek kas aktivitesi bilgilerini aktarmaktadır. Uygulama veri toplayıp bunlarla sunucuda tensorflow yardımıyla bir model oluşturmaktadır. Bundan sonra model, motorları veya diğer eklentileri kontrol etmek için verileri cihazdan indirip kullanabilmektedir [6].
- Parkinson hastalığı sınıflandırma veri seti, İstanbul Üniversitesi Cerrahpaşa Tıp Fakültesi Nöroloji Anabilim Dalı'nda yaşları 33 ile 87 arasında değişen 188 parkinson hastasından toplanan verilerle oluşturulmuştur [6].
- Pertekiz bankası pazarlama veri seti, Mayıs 2008'den Kasım 2010'a kadar Portekiz bankası tarafından mevcut müşteriler arasında vadeli mevduatı teşvik etmeyi amaçlayan doğrudan telefon görüşmesi pazarlama kampanyalarıyla ilgili örnekleri içermektedir [6].
- Genetik çeşitlilik sınıflandırması veri seti, clinVar, insan genetik çeşitliliği ile ilgili bilgiler içeren bir kamu kaynağıdır. Bu çeşitlilik, klinik laboratuvarları tarafından iyi huylu, muhtemel iyi huylu, belirsiz, potansiyel patojenik ve patojenik arasında değişen kategorik spektrumda sınıflandırılmaktadır [6].
- Mobil cihaz fiyat sınıflandırması veri seti, bir telefon üreticisinin fiyat aralığını belirlerken hangi kriterlerden nasıl faydalandığını göstermektedir. Burada diğer büyük telefon üreticilerinin fiyatları karşısında daha uygun fiyatlı telefon nasıl satılacağını belirlemek için farklı firmaların farklı özelliklerdeki telefonlarına ait bilgiler toplanmaktadır [6].
- Türkiye siyasi görüşleri veri seti, Türkiye'de politik yönelim üzerine 11 Mayıs 2018 ile 13 Mayıs 2018 tarihleri arasında yapılan bir anketin sonuçlarını içermektedir. Burada kişilere

10 soru sorulup bunun karşılığında evet/hayır şeklinde cevaplar alınmaktadır. Bunların dışında kişilerin demografik bilgileri de veri setinin içerisinde yer almaktadır [6].

- Banka pazarlama veri seti, vadeli mevduat işlemleri için gerçekleştirilen telefon görüşmelerine ait bilgileri içermektedir. Veri setinde yer alan örnekler, kampanya döneminde ürün satmak için bir müşteri listesine telefon görüşmesi yaparak veya herhangi bir müşterinin bir işlem yapmak için iletişim merkezini aradığı durumlarda mevduat hesabı açıp açmayacağına yönelik bilgilerdir [6].
- İris çiçeği veri seti, literatürde iris çiçeği olarak isimlendirilen bitkinin 3 farklı türü ile ilgili örnekler içermektedir. Bunlar, setosa, cersicolor ve virginica'dır. Bu türlerin her birinden veri setinde 50'şer örnek bulunmaktadır. Bu örnekler toplanırken iris çiçeğine ait taçyaprak uzunluğu, taçyaprak genişliği, çanakyaprak genişliği ve çanakyaprak uzunluğuna ait cm olarak veriler alınmıştır [6].
- Şarap kalitesi veri seti, Portekiz şarabı olan 'vinho verde' şarabı ile ilgili örnekleri içermektedir. Burada veriler hazırlanırken belirli gizlilik kuralları nedeniyle şaraplara ait sadece fizyokimyasal ve duyusal bilgiler veri setinde yer almaktadır. Yani üretici firma, üzüm türleri, şarap satış fiyatı gibi bilgiler bulunmamaktadır [6].
- Hepatoselüler Karsinom(HCC) veri seti, Portekiz de yer alan bir üniversite hastanesinde HCC tanısı konmuş 165 hastanın gerçek klinik verilerini içermektedir. Bu hastaların, çeşitli demografik bilgileri, risk faktörleri, laboratuvar ve genel sağlık özelliklerini içermektedir [6].
- Bireysel kredi sınıflandırma problemi veri seti, büyüyen müşteri veri tabanına sahip olan bir bankaya aittir. Bu müşterilerin büyük çoğunluğu farklı boyutlarda mevduatlara sahip müşterilerden oluşmaktadır [6].
- Sahte şirketleri sınıflandırmak için denetim riski veri seti, bir firmanın sahte firma olup olmadığını mevcut ve tarihsel risk faktörleri temelinde tahmin edebilecek bir sınıflandırma modeli oluşturarak denetleyicilere yardımcı olmak için oluşturulmuştur [6].
- İnternetten alışverişte kullanıcı tercihleri veri seti, belirli bir kampanyaya, özel bir güne, kullanıcı profiline veya döneme yönelik herhangi bir eğilimi önlemek için her oturumun 1 yıllık süre içerisinde farklı bir kullanıcıya ait olması ile oluşturulmuştur [6].
- Şarap için müşteri segmentasyonu veri seti, italya'da aynı bölgede üretilen ancak 3 farklı çeşidi olan şarapların kimyasal analizlerinin sonuçlarından elde edilmiştir. Analiz 3 şarap türünün her birinde bulunan 13 bileşenden alınan değerlerle yapılmaktadır [6].
- Ph tanıma veri seti, ph ölçümü yapılmış ve bunun sonucunda oluşan renge ait kırmızı, yeşil ve mavi renklerin dağılımlarına göre sonuçların etiketlendiği örnekleri içermektedir [6].
- Gelir sınıflandırması veri seti, 1994 yılında ABD'de gerçekleştirilen nüfus sayımında ki veri tabanından elde edilmiştir. Nüfus sayımı veri setinin oluşturulduğu dönem için 10 yılda bir

tekrarlanmaktadır. Amaç ülkedeki nüfusun konut koşullarını ve demografik, sosyal ve ekonomik özelliklerini sunmak için genel nüfus hakkında bilgi toplamaktır [6].

- Kriyoterapi analiz veri seti, siğil tedavisi için meşhed'deki Ghaem hastanesinin dermotoloji kliniğine başvuran hastalardan elde edilen verilerle oluşturulmuştur. Gelen hastalara sıvı nitrojen kullanılarak kriyoterapi yöntemiyle tedavi uygulanmıştır [6].
- Kredi kartı sahtekarlığı tespiti veri seti, bankalar için kredi kartı işlemlerinde sahte işlemleri tanımlayabilmek önemlidir. Böylece, müşteriler yapmadıkları işlemler için ücret ödemek zorunda kalmazlar. Bu veri seti, Eylül 2013 tarihinde Avrupa'daki kredi kartı sahiplerinin yaptıkları işlemlere ait verileri içermektedir. Veri setinde 2 gün içerisinde yapılan 284807 işlem ve bu işlemlerden dolandırıcılık olarak etiketlenen 492 işlemi içermektedir [6].
- Deniz kulağı veri seti, abolone(Deniz Kulağı) veri setinde, deniz kulaklarına ait bazı özellikler ve yaşları bulunmaktadır. Deniz kulaklarının yaşları hesaplanırken, kabuk dikey olarak kesilmesinden sonra boyanması ve halka sayısını mikroskop yardımıyla sayılması işlemleri yapılmaktadır. Yapılan işlem finansal olarak sorunlu ve zaman alan bir işlemdir. Bundan dolayı deniz kulalarının yaşlarının sayımı işleminde bilgisayardan yardım almak etkili bir yöntem olarak karşımıza çıkmaktadır [6].
- Sesle cinsiyet tanımlama veri seti, sesin ve konuşmanın akustik özelliklerini kullanarak bir sesi erkek veya kadın olarak tanımlamak için oluşturulmuştur. Veri seti oluşturulurken her biri erkek veya kadın cinsiyet ile etiketlenmiş binlerce erkek ve kadın sesi örneği kullanılmıştır [6].
- Pima Kızılderilileri diyabet veri seti, ulusal diyabet, sindirim ve böbrek hastalıkları enstitüsünden alınan verilerle oluşturulmuştur. Veri setinin oluşturulma amacı, veri setinde yer alan belirli teşhis ölçümlerine dayanarak bir hastanın diyabet hastası olup olmadığını teşhis etmektir [6].

2.4. Kullanılan Program ve Kütüphaneler

Bu tez çalışmasında veri setlerini analiz etmek, ilgili makine öğrenmesi algoritmalarını uygulamak ve görselleştirmek için açık kaynak kodlu Python programlama dili kullanılmıştır. Bu işlemleri yaparken veri setine ön işlem uygulama, makine öğrenmesi algoritmalarının uygulanması ve değerlendirilmesi sürecinde Pandas, Numpy ve Sci-kit learn kütüphaneleri kullanılmıştır. Sonuçların sunulması ve değerlendirilmesi sürecinde görselleştirme işlemleri için de matplotlib ve seaborn görselleştirme kütüphanelerinden faydalanılmıştır.

3. DENEYSEL SONUÇLAR

Bu tez çalışması kapsamında materyal ve yöntem bölümünde yer alan veri setleri ile ilgili kaggle ortamında makine öğrenmesi uygulamaları yapılmıştır. Makine öğrenmesi algoritmaları uygulanmadan önce bütün veri setleri makine öğrenmesi ön işlem sürecinden geçirilmiştir.

Bu bölümde ise materyal ve yöntem alanında yer alan veri setlerine sırasıyla KNN, KA, RO, NB, LRS, DVM, GA, AB ve YSA algoritmaları uygulanmıştır. Model oluşturulurken iki farklı model performans değerlendirme yöntemi kullanılmıştır. Bunlar, %70 eğitim ve %30 test verisi olacak şekilde hold-out yöntemi ve k=10 değeri seçilen ÇD yöntemidir. Bu sınıflandırıcılar ile oluşturulan modellerin başarımları bazı metrikler ile ölçülerek hangi sınıflandırıcının daha iyi tahminlerde bulunduğunun testi yapılmıştır. Sınıflandırıcıların performanslarını değerlendirirken sınıflandırıcının tahminlerinin yer aldığı karmaşıklık matrisinden faydalanılmıştır. Karmaşıklık matrisi yardımıyla dokuz farklı sınıflandırıcının performansı doğruluk, duyarlılık, anma, kesinlik, özgüllük, DPO, DNO, YNO, YPO, F ölçütü, MCC, AUC ve ROC eğrisi metrikleri ile değerlendirilmiştir.

İlk olarak ÇD ile elde edilen sınıflandırma sonuçları, daha sonra ise ÇD yöntemine göre daha iyi sonuçlar veren hold-out yöntemine ait sonuçlar verilecektir. Model seçiminde ÇD yönteminin tercih edilmesinin nedeni, Bölüm 2.2.2'de bahsedilmiş olan hold-out yönteminin dezavantajlarıdır. Bunların başında, eğer veri setinde yer alan veri sayısı kısıtlı ise test için yeterli verinin olmayacağıdır. Diğer bir dezavantaj ise eğitim ve test verisinin ilk başta model oluşturulmadan ayrılması durumunda modelin performansının ölçümünde doğru sonuçlar elde edilemeyeceğidir. Çünkü modelin test aşamasında modelin eğitim aşamasında yer alan bir değerin olmama durumu ile karşılaşılabilmektedir. ÇD yönteminde ise veri setinin bütün parçaları hem eğitim hem de test aşamasında kullanılarak model oluşturulmaktadır.

3.1. Sınıflandırma Sonuçları

Bu bölümde ÇD sonucunda elde edilen modellere ait sonuçlar verilecektir. Kaggle'dan elde edilen 32 farklı veri seti için k=10 olacak şekilde ÇD yöntemi kullanılarak 9 farklı sınıflandırıcı ile modeller oluşturulmuştur. Oluşturulan modeller ait en yüksek doğruluk değerini veren modele ait karmaşıklık matrisi ve MCC, F_1, ve AUC metrikleri ile en yüksek sonuçlar elde edilen karmaşıklık matrisi verilmiştir. Ek-1'de 32 veri seti için 9 sınıflandırıcıya ait doğruluk, kesinlik, anma, özgüllük, duyarlılık, MCC, F_1 ve AUC değerleri verilmiştir.

3.1.1. Avustralya'da Yağmur Veri Seti

Bu veri seti Avustralya da bir gün sonra yağmur yağıp yağmamasına ait örnekleri içeren iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %83.54 doğruluk değeri ile KA ve GA sınıflandırıcılarının bu veri seti için en iyi sınıflandırıcılar olduğu tespit edilmiştir. Bu sınıflandırıcılardan elde edilen karmaşıklık matrisi Tablo 3.1'de verilmektedir .

Tablo 3.1. KA ve GA Sınıflandırıcıları ile elde edilen Karmaşıklık Matrisleri

		Tahmin	
		Hayır	**Evet**
Gerçek	**Hayır**	82.906	2.569
	Evet	15.181	7.212

Tablo 3.1'deki karmaşıklık matrisi incelendiğinde, Hayır'a ait 85.475 verinin 82.906'sı, Evet'e ait 22.393 verinin ise 15.181'inin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.4124, 0.646 ve 0.7113 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcıların rasgele sınıflandırmaya göre daha iyi sınıflandırma yaptığını göstermektedir.

3.1.2. Tic-Tac-Toe Oyunu Veri Seti

Bu veri seti Tic-tac-toe oyununa ait hamleler ve sonuçlarının yer aldığı iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %100 doğruluk değeri ile GA sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.2'de verilmektedir.

Tablo 3.2. GA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Negatif	**Pozitif**
Gerçek	**Negatif**	332	0
	Pozitif	0	626

Tablo 3.2'deki karmaşıklık matrisi incelendiğinde, pozitif ve negatif sınıfa ait verilerin tamamı doğru olarak sınıflandırıldığı görülmüştür .

3.1.3. Wisconsin Meme Kanseri Veri Seti

Bu veri seti Wisconsin üniversitesinde oluşturulan meme kanserine ait örneklerin yer aldığı iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %97.73 doğruluk değeri ile YSA sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.3'de verilmektedir.

Tablo 3.3. YSA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		İyi huylu	**Kötü Huylu**
Gerçek	**İyi Huylu**	365	2
	Kötü Huylu	10	202

Tablo 3.3'deki karmaşıklık matrisi incelendiğinde, iyi huylu'ya ait 367 verinin 365'i, Kötü huylu'ye ait 212 verinin ise 202'sinin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.9549, 0.9713 ve 0.9775 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.4. Sloan Dijital Gökyüzü Araştırması Veri Seti

Bu veri seti devasa bir teleskop kullanılarak yürütülmekte olan bir projeden elde edilen değerlerden oluşan üç sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %98.89 doğruluk değeri ile GA sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.4'te verilmektedir.

Tablo 3.4. GA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin		
		Galaxy	**Qso**	**Star**
Gerçek	**Galaxy**	4946	44	8
	Qso	52	797	1
	Star	5	1	4146

Tablo 3.4'teki karmaşıklık matrisi incelendiğinde, Galaxy'ye ait 4998 verinin 4946'sı, Qso'ya ait 850 verinin ise 797'sini, Star'a ait 4152 verinin 4146'sını doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.9805, 1 ve 0.9765 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir .

3.1.5. Pulsar Yıldız Tahmini Veri Seti

Bu veri seti Pulsar adayı örneğini tanımlamak için oluşturulan verilerle oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %97.94 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.5'de verilmektedir.

Tablo 3.5. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Negatif	**Pozitif**
Gerçek	**Negatif**	16151	108
	Pozitif	261	1378

Tablo 3.5'deki karmaşıklık matrisi incelendiğinde, Negatif'e ait 16259 verinin 16151'i, Pozitif'e ait 1639 verinin ise 1378'ini doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.8719, 0.9171 ve 0.936 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.6. Ortopedik Hastaların Biyomekanik Özellikleri Veri Seti(2 Sınıflı)

Bu veri seti GARO kliniğinde tedavi gören hastalardan elde edilen verilerle oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %81.29 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.6'da verilmektedir.

Tablo 3.6. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Anormal	**Normal**
Gerçek	**Anormal**	181	29
	Normal	29	71

Tablo 3.6'daki karmaşıklık matrisi incelendiğinde, Anormal'e ait 210 verinin 181'i, Normal'e ait 100 verinin ise 71'ini doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.5719, 0.786 ve 0.786 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre daha iyi sınıflandırma yaptığını göstermektedir.

3.1.7. Ortopedik Hastaların Biyomekanik Özellikleri Veri Seti(3 sınıflı)

Bu veri seti GARO kliniğinde tedavi gören hastalardan elde edilen verilerle oluşan üç sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %84.52 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.7'de verilmektedir.

Tablo 3.7. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin		
		Disk Fıtığı	**Normal**	**Bel kayması**
Gerçek	**Disk Fıtığı**	38	22	0
	Normal	16	80	4
	Bel Kayması	0	6	144

Tablo 3.7'deki karmaşıklık matrisi incelendiğinde, Disk fıtığı'na ait 60 verinin 38'sı, Normal'e ait 100 verinin ise 80'ini, Bal kayması'na ait 150 verinin 144'ünü doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.752, 0.97 ve 0.8018 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre daha iyi sınıflandırma yaptığını göstermektedir.

3.1.8. Yapısal Protein Dizileri Veri Seti

Bu veri seti RCSB'nin Protein veri bankasından alınan verilerle oluşan üç sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %51.93 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.8'de verilmektedir.

Tablo 3.8. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin		
		Hydrolase	**Transferase**	**Oxidoreductase**
Gerçek	**Hydrolase**	25007	10356	7837
	Transferase	11838	16743	6407
	Oxidoreductase	9782	7295	16073

Tablo 3.8'deki karmaşıklık matrisi incelendiğinde, Hydrolase'e ait 43200 verinin 25007'si, Transferase'ya ait 34988 verinin 16743'ünü ve Oxidoreductase'a ait 33150 verinin 16073'ü doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.271, 0.62 ve 0.5159 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya yakın bir sınıflandırma yaptığını göstermektedir.

3.1.9. Kalp Hastalığı Veri Seti

Bu veri seti Kalp hastalığı ile ilgili verilerle oluşturulan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %84.16 doğruluk değeri ile LRS sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.9'da verilmektedir.

Tablo 3.9. LRS Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Hastalık Yok	**Hastalık Var**
Gerçek	**Hastalık Yok**	110	28
	Hastalık Var	20	145

Tablo 3.9'daki karmaşıklık matrisi incelendiğinde, hastalık yok'a ait 128 verinin 110'u, Hastalık var'a ait 165 verinin ise 145'inin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.68, 0.8379 ve 0.84 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.10. Farelerin Protein Ekspresyonu Veri Seti

Bu veri seti Korteksin nükleer fraksiyonunda tespit edilebilir sinyaller üreten 77 protein/protein modifikasyonunun ekspresyon seviyelerinden oluşan sekiz sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %98.91 doğruluk değeri ile YSA sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.10'da verilmektedir.

Tablo 3.10. YSA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin							
		t-CS-m	**c-CS-s**	**t-CS-s**	**c-SC-s**	**t-SC-s**	**c-SC-m**	**t-SC-m**	**c-CS-m**
Gerçek	**t-CS-m**	88	1	0	0	0	0	0	1
	c-CS-s	0	74	1	0	0	0	0	0
	t-CS-s	0	0	75	0	0	0	0	0
	c-SC-s	0	0	0	75	0	0	0	0
	t-SC-s	0	0	0	0	72	0	0	0
	c-SC-m	0	0	0	0	0	60	0	0
	t-SC-m	0	0	0	0	0	0	60	0
	c-CS-m	1	2	0	0	0	0	0	42

Tablo 3.10'daki karmaşıklık matrisi incelendiğinde, "t-CS-m" sınıfına ait 89 verinin 88'i, "c-CS-s" sınıfına ait 75 verinin 74'ü, "t-CS-s" sınıfına ait 75 verinin hepsi, " c-SC-s" sınıfına ait 75 verinin hepsi, " t-SC-s" sınıfına ait 72 verinin hepsi, " c-SC-m" sınıfına ait 60 verinin hepsi, "t-SC-m" sınıfına ait 60 verinin hepsi ve "c-CS-m" sınıfına ait 45 verinin 42'sinin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.9875, 0.9656 ve 0.9882 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.11. Seyahat Sigortası Veri Seti

Bu veri seti seyahat sigortası yapan farklı firmaların yaptıkları işlere göre müşterilerinin sigorta isteyip istemediklerinin ve yaptıkların işlemlerin örneklerinden oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %98.54 doğruluk değeri ile DVM ve YSA sınıflandırıcılarının bu veri seti için en iyi sınıflandırıcılar olduğu tespit edilmiştir. Bu sınıflandırıcılardan elde edilen karmaşıklık matrisi Tablo 3.11'de verilmektedir.

Tablo 3.11. DVM ve YSA Sınıflandırıcıları ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Hayır	**Evet**
Gerçek	**Hayır**	62399	0
	Evet	927	0

Tablo 3.11'deki karmaşıklık matrisi incelendiğinde, evet sınıfına ait 927 verinin hiçbiri, hayır sınıfına ait 62399 verinin ise tamamının doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0, 0.5 ve 0.4949 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler doğruluk değeri %98.54 olmasına rağmen sınıflandırıcıların rasgele sınıflandırma yaptığını göstermektedir. Karmaşıklık matrisi incelendiğinde evet sınıfına ait verilerden hiçbirinin doğru olarak tahmin edilmediği görülmektedir.

3.1.12. Kas Aktivitesini Okuyarak Jestlerin Sınıflandırılması Veri Seti

Bu veri seti kas aktivitesi bilgilerini içeren dört sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %100 doğruluk değeri ile KA, NB, DVM, GA ve YSA sınıflandırıcılarının bu veri seti için en iyi sınıflandırıcılar olduğu tespit edilmiştir. Bu sınıflandırıcılardan elde edilen karmaşıklık matrisi Tablo 3.12'de verilmektedir.

Tablo 3.12.KA, NB, DVM, GA, YSA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin			
		Taş	**Kağıt**	**Makas**	**Tamam**
Gerçek	**Taş**	2910	0	0	0
	Kağıt	0	2903	0	0
	Makas	0	0	2943	0
	Tamam	0	0	0	2922

Tablo 3.12'deki karmaşıklık matrisi incelendiğinde, taş sınıfına ait 2910 verinin, kağıt sınıfına ait 2903 verinin, makas sınıfına ait 2943 verinin ve tamam sınıfına ait 2922 verinin tamamının doğru olarak sınıflandırıldığı görülmüştür.

3.1.13. Parkinson Hastalığı Sınıflandırma Veri Seti

Bu veri seti İstanbul Üniversitesi Cerrahpaşa Tıp Fakültesi Nöroloji Anabilim Dalı'nda yaşları 33 ile 87 arasında değişen 188 parkinson hastasından toplanan verilerle oluşan üç sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %81.48 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.13'de verilmektedir.

Tablo 3.13.RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Parkinson	**Sağlıklı**
Gerçek	**Parkinson**	89	103
	Sağlıklı	37	527

Tablo 3.13'deki karmaşıklık matrisi incelendiğinde, parkinson'a ait 192 verinin 89'u, sağlıklı'ya ait 564 verinin ise 527'sinin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.4647, 0.6989 ve 0.7334 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre daha iyi sınıflandırma yaptığını göstermektedir.

3.1.14. Portekiz Bankası Pazarlama Veri Seti

Bu veri seti Mayıs 2008'den Kasım 2010'a kadar bir Portekiz bankası tarafından mevcut müşteriler arasında vadeli mevduatı teşvik etmeyi amaçlayan doğrudan telefon görüşmesi pazarlama kampanyalarıyla ilgili örneklerden oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %88.75 doğruluk değeri ile DVM sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.14-a'da verilmektedir.

Tablo 3.14. DVM ve RO Sınıflandırıcıları ile elde edilen Karmaşıklık Matrisleri

		Tahmin	
		Hayır	**Evet**
Gerçek	**Hayır**	35901	301
	Evet	4292	347

a)DVM

		Tahmin	
		Hayır	**Evet**
Gerçek	**Hayır**	282397	1918
	Evet	176	316

b)RO

Tablo 3.14-a'daki karmaşıklık matrisi incelendiğinde, hayır sınıfına ait 36202 verinin 35901'i, evet sınıfına ait 4639 verinin ise 347'sinin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.1688, 0.5332 ve 0.6107 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya benzer sınıflandırma yaptığını göstermektedir. Aynı veri seti için oluşturulan RO sınıflandırıcısı ile elde edilen karmaşıklık matrisi Tablo 3.14-b'de yer almaktadır. Bu sınıflandırıcı ile elde edilen doğruluk değeri %86.35, MCC değeri 0.2352, AUC değeri 0.6023 ve F_1 değeri 0.6183 olarak hesaplanmıştır. Oluşturulan bu model de rasgele tahmine yakın tahminde bulunmasına rağmen diğer modelden daha iyi başarım sonucu verdiği görülmektedir.

3.1.15. Genetik Çeşitlilik Sınıflandırması Veri Seti

Bu veri seti insan genetik çeşitliliği ile ilgili verilerle oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %74.93 doğruluk değeri ile DVM ve YSA sınıflandırıcılarının bu veri seti için en iyi sınıflandırıcılar olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.15'de verilmektedir.

Tablo 3.15. DVM ve YSA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Tutarlı	**Çakışan**
Gerçek	**Tutarlı**	37847	0
	Çakışan	12661	0

Tablo 3.15'deki karmaşıklık matrisi incelendiğinde, çakışan'a ait 12661 verinin hiçbiri, tutarlı'ya ait 37847 verinin ise hepsinin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0, 0.5 ve 0.4253 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler doğruluk değeri %74.93 olmasına rağmen sınıflandırıcıların rasgele sınıflandırma yaptığını göstermektedir. Karmaşıklık matrisi incelendiğinde çakışan sınıfına ait verilerden hiçbirinin doğru olarak tahmin edilmediği görülmektedir.

3.1.16. Mobil cihaz fiyat sınıflandırması veri seti

Bu veri seti bir telefon üreticisinin fiyat aralığını belirlerken hangi kriterlerden nasıl faydalandığını gösteren verilerle oluşan dört sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %95.85 doğruluk değeri ile YSA sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.16'da verilmektedir.

Tablo 3.16. YSA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin			
		Düşük Maliyet	**Orta Maliyet**	**Yüksek Maliyet**	**Çok Yüksek Maliyet**
Gerçek	**Düşük Maliyet**	492	8	0	0
	Orta Maliyet	25	461	14	0
	Yüksek Maliyet	0	13	471	16
	Çok Yüksek Maliyet	0	0	10	490

Tablo 3.16'daki karmaşıklık matrisi incelendiğinde, düşük maliyete ait 500 verinin 492'si, orta maliyete ait 500 verinin 461'ini, yüksek maliyete ait 500 verinin 471'ini ve çok yüksek maliyete ait 500 verinin 490'ını doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.9427, 0.9846 ve 0.9585 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.17. Türkiye siyasi görüşleri veri seti

Bu veri seti Türkiye'de politik yönelim üzerine yapılan bir anketin sonuçları ile oluşan altı sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %43.62 doğruluk değeri ile LRS sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.17'de verilmektedir.

Tablo 3.17. LRS Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin					
		AKP	**CHP**	**IYI PARTI**	**MHP**	**HDP**	**DİĞER**
Gerçek	**AKP**	4	2	15	1	1	13
	CHP	4	50	81	0	0	20
	IYİ PARTI	14	9	129	1	0	31
	MHP	7	5	8	0	0	7
	HDP	5	7	7	0	0	1
	DİĞER	28	75	67	0	0	23

Tablo 3.17'deki karmaşıklık matrisi incelendiğinde, AKP'ye ait 126 verinin 84'ü, CHP'ye ait 255 verinin 150'si, IYI PARTI'ye ait 264 verinin 129'u, MHP'ye ait 27 verinin hiçbiri, HDP'ye ait 20 verinin hiçbiri ve DİĞER'e ait 193 verinin 23'ünün doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.242, 0.5075 ve 0.2961 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırma yaptığını göstermektedir. Karmaşıklık matrisi

incelendiğinde MHP ve HDP sınıfına ait verilerden hiçbirinin doğru olarak tahmin edilmediği görülmektedir.

3.1.18. Banka pazarlama veri seti

Bu veri seti vadeli mevduat işlemleri için telefon görüşmeleri yoluyla gerçekleştirilen işlemlerin sonuçlarıyla elde edilen değerlerden oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %80.52 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.18'de verilmektedir.

Tablo 3.18. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Hayır	**Evet**
Gerçek	**Hayır**	4875	998
	Evet	1176	4113

Tablo 3.18'deki karmaşıklık matrisi incelendiğinde, evet sınıfına ait 5873 verinin 4875'i, hayır sınıfına ait 5289 verinin ise 4113'ünün doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.609, 0.8038 ve 0.8045 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre daha iyi sınıflandırma yaptığını göstermektedir.

3.1.19. İris Çiçeği Veri Seti

Bu veri seti literatürde iris çiçeği olarak bilinen bitkiye ait verileri içeren üç sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %96.673 doğruluk değeri ile DVM ve YSA sınıflandırıcılarının bu veri seti için en iyi sınıflandırıcılar olduğu tespit edilmiştir. Bu sınıflandırıcılardan elde edilen karmaşıklık matrisleri Tablo 3.19 ve Tablo 3.20'de verilmektedir.

Tablo 3.19. DVM Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin		
		Setosa	**Versicolor**	**virginica**
Gerçek	**Setosa**	50	0	0
	Versicolor	0	49	1
	virginica	0	4	46

Tablo 3.20. YSA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin		
		Setosa	**Versicolor**	**Virginica**
Gerçek	**Setosa**	50	0	0
	Versicolor	0	47	3
	virginica	0	2	48

Tablo 3.19 ve Tablo 3.20'deki karmaşıklık matrisleri incelendiğinde, setosa sınıfına ait 50 verinin hepsini DVM ve YSA sınıflandırıcılarının doğru sınıflandırdığı, versicolor sınıfınaa ait 50 veriyi DVM 49 ve YSA 47'sini doğru sınıflandırdığı ve virginica sınıfına ait 50 veriyi DVM 46 ve YSA 48'ini doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar DVM için sırasıyla 0.9505, 0.955 ve 0.9672, YSA için sırasıyla 0.95, 0.965 ve 0.9667 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcılarının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.20. Şarap Kalitesi Veri Seti

Bu veri seti Portekiz şarabı olan 'vinho verde' şarabı ile ilgili örneklerden oluşan yedi sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %53.53 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.21'de verilmektedir.

Tablo 3.21. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin						
		3	**4**	**5**	**6**	**7**	**8**	**9**
Gerçek	**3**	0	1	17	12	0	0	0
	4	1	9	132	73	1	0	0
	5	0	14	1289	816	19	0	0
	6	0	3	747	1874	209	3	0
	7	0	0	70	712	291	6	0
	8	0	0	2	99	77	15	0
	9	0	0	0	3	2	0	0

Tablo 3.21'deki karmaşıklık matrisi incelendiğinde, "3" sınıfına ait 30 verinin hiçbiri, "4" sınıfına ait 216 verinin 9'u, "5" sınıfına ait 2138 verinin 1289'u, "6" sınıfına ait 2836 verinin 1874'ü, "7" sınıfına ait 1079 verinin 291'i, "8" sınıfına ait 193 verinin 15'i ve "9" sınıfına ait 5 verinin hiçbiri doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.266, 0.5 ve 0.2852 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırma yaptığını göstermektedir. Karmaşıklık matrisi incelendiğinde 3 ve 9 sınıfına ait verilerden hiçbirinin doğru olarak tahmin edilmediği görülmektedir.

3.1.21. Hepatoselüler Karsinom(HCC) Veri Seti

Bu veri seti Portekiz de yer alan bir üniversite hastanesinde HCC tanısı konmuş 165 hastanın gerçek klinik verileri ile oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %100 doğruluk değeri ile KA, RO, NB, LRS, DVM, GA, AB ve YSA sınıflandırıcılarının bu veri seti için en iyi sınıflandırıcılar olduğu tespit edilmiştir. Bu sınıflandırıcılardan elde edilen karmaşıklık matrisi Tablo 3.22'de verilmektedir.

Tablo 3.22.KA, RO, NB, LRS, DVM, GA, AB ve YSA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Hayatta değil	**Hayatta**
Gerçek	**Hayatta değil**	59	0
	Hayatta	0	97

Tablo 3.22'deki karmaşıklık matrisi incelendiğinde, hayatta değil'e ait 59 verinin hepsi, hayattaya ait 97 verinin de hepsinin doğru olarak sınıflandırıldığı görülmüştür.

3.1.22. Bireysel Kredi Sınıflandırma Problemi Veri Seti

Bu veri seti büyüyen müşteri veri tabanına sahip olan bir bankaya ait değerlerden oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %98.76 doğruluk değeri ile GA sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.23'de verilmektedir.

Tablo 3.23.GA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Kredi almamış	**Kredi almış**
Gerçek	**Kredi almamış**	4506	14
	Kredi almış	48	432

Tablo 3.23'deki karmaşıklık matrisi incelendiğinde, kredi almamışa ait 4520 verinin 4506'sı, kredi almışa ait 480 verinin ise 432'sinin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.9269, 0.9484 ve 0.9635 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.23. Sahte Şirketleri Sınıflandırmak İçin Denetim Riski Veri Seti

Bu veri seti bir firmanın sahte firma olup olmadığını mevcut ve tarihsel risk faktörlerinden elde edilen değerlerden oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %100 doğruluk değeri ile KA, GA ve AB sınıflandırıcılarının bu veri seti için en iyi sınıflandırıcılar olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.24'de verilmektedir.

Tablo 3.24. KA, GA, AB Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Dolandırıcılık yok	**Dolandırıcılık var**
Gerçek	**Dolandırıcılık yok**	471	0
	Dolandırıcılık var	0	305

Tablo 3.24'deki karmaşıklık matrisi incelendiğinde, dolandırıcılık yok'a ait 471 verinin hepsi, dolandırıcılık var'a ait 305 verinin de hepsi doğru olarak sınıflandırıldığı görülmüştür.

3.1.24. İnternetten Alışverişte Kullanıcı Tercihleri Veri Seti

Bu veri seti bir online satış yapan internet sitesinden alınan verilerle oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %88.95 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.25'de verilmektedir.

Tablo 3.25. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Yanlış	**Doğru**
Gerçek	**Yanlış**	10000	422
	Doğru	940	968

Tablo 3.25'deki karmaşıklık matrisi incelendiğinde, yanlış sınıfına ait 10422 verinin 10000'i, doğru sınıfına ait 1908 verinin ise 968'inin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.5338, 0.7334 ve 0.7677 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre daha iyi sınıflandırma yaptığını göstermektedir.

3.1.25. Şarap İçin Müşteri Segmentasyonu Veri Seti

Bu veri seti İtalya'da aynı bölgede üretilen ancak 3 farklı çeşidi olan şarapların kimyasal analizlerinin sonuçlarından oluşan üç sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %98.88 doğruluk değeri ile YSA sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.26'da verilmektedir.

Tablo 3.26. YSA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin		
		1	**2**	**3**
Gerçek	**1**	59	0	0
	2	0	69	2
	3	0	0	48

Tablo 3.26'daki karmaşıklık matrisi incelendiğinde, "1" sınıfına ait 59 verinin hepsi, "2" sınıfına ait 71 verinin 69'unun ve "3" sınıfına ait 48 verinin hepsi doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.9831, 0.9923 ve 0.9886 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.26. pH Tanıma Veri Seti

Bu veri seti pH ölçümü yapılmış ve bunun sonucunda oluşan renge ait kırmızı, yeşil ve mavi renklerin dağılımlarına göre sonuçların etiketlendiği on beş sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %75.19 doğruluk değeri ile RO

sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.27'de verilmektedir.

Tablo 3.27. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin														
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Gerçek	**0**	32	5	1	0	0	0	0	0	0	0	0	0	0	0	0
	1	5	36	2	1	0	0	0	0	0	0	0	0	0	0	0
	2	1	5	34	2	1	1	0	0	0	0	0	0	0	0	0
	3	1	3	6	29	4	0	1	0	0	0	0	0	0	0	0
	4	0	2	2	5	33	2	0	0	0	0	0	0	0	0	0
	5	0	0	2	1	5	28	4	2	0	2	0	0	0	0	0
	6	0	0	0	3	1	6	31	3	0	0	0	0	0	0	0
	7	0	0	0	0	1	1	6	33	2	1	0	0	0	0	0
	8	0	0	0	0	2	0	1	1	36	2	1	1	0	0	0
	9	0	0	0	0	0	2	0	2	1	33	3	3	0	0	0
	10	0	0	0	0	0	0	0	3	2	0	32	3	3	0	1
	11	1	0	0	0	0	0	0	0	2	1	3	30	5	1	1
	12	0	0	0	0	0	0	0	0	0	1	3	0	34	3	3
	13	0	0	0	0	0	0	0	0	0	0	0	1	3	33	6
	14	0	0	0	0	0	0	0	0	0	0	0	0	2	5	37

Tablo 3.27'deki karmaşıklık matrisi incelendiğinde, "0" sınıfına ait 38 veriden 32'si, "1" sınıfına ait veriden 44 veriden 36'sı, "2" sınıfına ait 44 veriden 34'ü, "3 sınıfına ait 44 veriden 29'u, "4" sınıfına ait 44 veriden 33'ü, "5" sınıfına ait 44 veriden 28'i, "6" sınıfına ait 44 veriden 31'i, "7" sınıfına ait 44 veriden 33'ü, "8" sınıfına ait 44 veriden 36'sı, "9" sınıfına ait 44 veriden 33'ü, "10" sınıfına ait 44 veriden 32'si, "11" sınıfına ait 44 veriden 30'u, "12" sınıfına ait 44 veriden 34'ü, "13" sınıfına ait 43 veriden 33'ü ve "14" sınıfına ait 44 veriden 27'sinin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.734, 0.9114 ve 0.5293 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcılarının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.27. Gelir Sınıflandırması Veri Seti

Bu veri seti 1994 yılında ABD'de gerçekleştirilen nüfus sayımında ki veri tabanından alınan bilgilerle oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %87.27 doğruluk değeri ile GA sınıflandırıcısının bu veri seti

için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.28'de verilmektedir.

Tablo 3.28. GA Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		<=50K	**>50K**
Gerçek	**<=50K**	23326	1394
	>50K	2751	5090

Tablo 3.28'deki karmaşıklık matrisi incelendiğinde, "<=50K" ya ait 24720 verinin 23326'sı, ">50K" ya ait 7841 verinin ise 5090'ını doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.6346, 0.7963 ve 0.8175 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre daha iyi sınıflandırma yaptığını göstermektedir.

3.1.28. Kriyoterapi Analiz Veri Seti

Bu veri seti siğil tedavisi için meşhed'deki Ghaem hastanesinin dermotoloji kliniğine başvuran hastalardan elde edilen verilerle oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %94.44 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.29'da verilmektedir.

Tablo 3.29. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Kötü Huylu	**İyi huylu**
Gerçek	**Kötü Huylu**	41	1
	İyi Huylu	4	44

Tablo 3.29'daki karmaşıklık matrisi incelendiğinde, iyi huylu sınıfına ait 48 verinin 44'ü, Kötü huylu sınfına ait 42 verinin ise 41'inin doğru olarak sınıflandırıldığı görülmüştür. Veri seti

için MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.8908, 0.9464 ve 0.9454 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.29. Kredi Kartı Sahtekârlığı Tespiti Veri Seti

Bu veri seti Eylül 2013 tarihinde Avrupa'daki kredi kartı sahiplerinin yaptıkları işlemlere ait verilerden oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %99.83 doğruluk değeri ile LRS, DVM ve YSA sınıflandırıcılarının bu veri seti için en iyi sınıflandırıcılar olduğu tespit edilmiştir. Bu sınıflandırıcılardan elde edilen karmaşıklık matrisi Tablo 3.30'de verilmektedir.

Tablo 3.30. LRS, DVM, YSA ve NB Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Hileli işlem var	Hileli işlem yok
Gerçek	Hileli işlem var	284315	0
	Hileli işlem yok	492	0

a)LRS, DVM, YSA

	Sınıf	**Tahmin**	
		Hileli işlem var	Hileli işlem yok
Gerçek	Hileli işlem var	282397	1918
	Hileli işlem yok	176	316

b)NB

Tablo 3.30'deki karmaşıklık matrisi incelendiğinde, Hileli işlem var'a ait 492 verinin hiçbiri, Hileli işlem yok'a ait 284315 verinin ise hepsinin doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0, 0.5 ve 0.4996 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırma yaptığını göstermektedir. Karmaşıklık matrisi incelendiğinde hileli işlem var sınıfına ait verilerden hiçbirinin doğru olarak tahmin edilmediği görülmektedir. Aynı veri seti için oluşturulan NB sınıflandırıcısı ile elde edilen karmaşıklık matrisi Tablo 3.14-b'de yer almaktadır. Bu sınıflandırıcı ile elde edilen doğruluk değeri %99.26, MCC değeri 0.2991, AUC değeri 0.8268 ve F_1 değeri 0.6721 olarak hesaplanmıştır. Oluşturulan bu model de rasgele tahmine yakın tahminde bulunmasına rağmen diğer modelden daha iyi başarım sonucu verdiği görülmektedir.

3.1.30. Deniz Kulağı Veri Seti

Bu veri seti deniz kulaklarına ait bazı özellikler ve yaşlarından oluşan üç sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %63.83

doğruluk değeri ile KNN(k=16) sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.31'de verilmektedir.

Tablo 3.31. KNN Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin		
		Genç	**Orta**	**Yaşlı**
Gerçek	**Genç**	1071	267	69
	Orta	275	718	330
	Yaşlı	100	470	877

Tablo 3.31'deki karmaşıklık matrisi incelendiğinde, Genç sınıfına ait 443 verinin 1071'i, Orta sınıfına ait 1323 verinin 718'ini ve Yaşlı sınıfına ait 1447 verinin 877'si doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.458, 0.6422 ve 0.6386 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre daha iyi sınıflandırma yaptığını göstermektedir.

3.1.31. Sesle Cinsiyet Tanımlama Veri Seti

Bu veri seti sesin ve konuşmanın akustik özelliklerini kullanarak bir sesi erkek veya kadın olarak tanımlamak için oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %96.78 doğruluk değeri ile RO sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.32'de verilmektedir.

Tablo 3.32. RO Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Kadın	**Erkek**
Gerçek	**Kadın**	1534	50
	Erkek	52	1532

Tablo 3.32'deki karmaşıklık matrisi incelendiğinde, Kadın'a ait 1584 verinin 1534'ü, Erkek'e ait 1584 verinin ise 1532'sini doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.9356, 0.9678 ve 0.9678 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre çok daha iyi sınıflandırma yaptığını göstermektedir.

3.1.32. Pima Kızılderilileri Diyabet Veri Seti

Bu veri seti, ulusal diyabet, sindirim ve böbrek hastalıkları enstitüsünden alınan verilerle oluşan iki sınıflı bir veri setidir. Bu veri setine, tez çalışmasında kullanılan dokuz ayrı sınıflandırıcı uygulanarak 10 katlı ÇD testi ile test edilmiş ve sınıflandırma sonuçları elde edilmiştir. Sınıflandırma sonucunda %76.82 doğruluk değeri ile LRS sınıflandırıcısının bu veri seti için en iyi sınıflandırıcı olduğu tespit edilmiştir. Bu sınıflandırıcıdan elde edilen karmaşıklık matrisi Tablo 3.33'de verilmektedir.

Tablo 3.33. LRS Sınıflandırıcısı ile elde edilen Karmaşıklık Matrisi

		Tahmin	
		Negatif	**Pozitif**
Gerçek	**Negatif**	441	59
	Pozitif	119	149

Tablo 3.33'deki karmaşıklık matrisi incelendiğinde, Negatif sınıfına ait 500 verinin 441'i, Pozitif sınıfına ait 268 verinin ise 149'ünün doğru olarak sınıflandırıldığı görülmüştür. Veri setine ait sınıflarının dengesiz olması dolayısı ile MCC, AUC ve F_1 değerleri hesaplanmış ve bunlar sırasıyla 0.4697, 0.7189 ve 0.7351 olarak elde edilmiş ve Ek-1'de verilmiştir. Bu değerler sınıflandırıcının rasgele sınıflandırmaya göre daha iyi sınıflandırma yaptığını göstermektedir.

3.2. Hold-out Yöntemi ile Sınıflandırma

Bu tez çalışmasında kullanılan 32 veri setinin her birine dokuz sınıflandırıcının hold-out yöntemi %70 eğitim ve %30 test veri seti olacak şekilde uygulaması yapılmıştır. Bu uygulamalar sonucunda bazı veri setleri için oluşturulan modellerde hold-out yöntemi ÇD yönteminden daha iyi sonuçlar vermiştir. Bölüm 2.3.2'de bahsedilen hold-out yönteminin en önemli dezavantajlarından birisi eğer veri setinde yer alan veri sayımız kısıtlı ise test için yeterli verinin olmayacağıdır. Diğer bir dezavantaj ise eğitim ve test verisinin ilk başta model oluşturulmadan ayrılması durumunda

modelin performansının ölçümünde doğru sonuçlar elde edilemeyeceğidir. Çünkü modelin test aşamasında modelin eğitim aşamasında yer alan bir değerin olmama durumu ile karşılaşılabilmektedir. ÇD yönteminde ise veri setinin bütün parçaları hem eğitim hem de test aşamasında kullanılarak model oluşturulmaktadır. Bu sebeple Tablo 3.34'de yer alan değerler aynı veri setleri için hold-out yöntemi kullanılara ÇD'ye göre daha iyi sonuç veren sınıflandırıcıları göstermektedir. ÇD'ye göre daha iyi sonuçlar vermelerine rağmen sınıflandırıcıların gerçek performanslarını temsil etmemektedir.

Tablo 3.34. Hold-out yöntemi ile sınıflandırma sonuçları

Veri seti	Sınıflandırıcı	Yöntem	Doğruluk(%)
Wisconsin Meme Kanseri	KNN	HOLD-OUT	98.25
	YSA	ÇD	97.89
Pulsar yıldızı tahmini	RO, AB	HOLD-OUT	98.01
	RO	ÇD	97.94
Ortopedik Hastaların Biyomekanik Özellikleri(2 sınıf)	RO	HOLD-OUT	86.02
	RO	ÇD	81.29
Ortopedik Hastaların Biyomekanik Özellikleri(3 sınıf)	RO	HOLD-OUT	84.95
	RO	ÇD	84.52
Yapısal protein dizileri	RO	HOLD-OUT	92.69
	RO	ÇD	51.93
Farelerin protein ekspresyonu	LRS	HOLD-OUT	100
	YSA	HOLD-OUT	100
	YSA	ÇD	98.91
Parkinson hastalığı sınıflandırma	RO	HOLD-OUT	86.22
	RO	ÇD	81.48
Genetik çeşitlilik sınıflandırma	LRS	HOLD-OUT	75.32
	DVM	HOLD-OUT	75.32
	YSA	HOLD-OUT	75.32
	DVM	ÇD	74.93
	YSA	ÇD	74.93
Türkiye siyasi görüşleri	DVM	HOLD-OUT	47.58
	LRS	ÇD	43.62
Banka pazarlama	GA	HOLD-OUT	85.85
	RO	ÇD	80.52
İris çiçeği	KNN, GA, AB, YSA	HOLD-OUT	97.78
	DVM, YSA	ÇD	96.67
Şarap kalitesi	RO	HOLD-OUT	69.03
	RO	ÇD	53.53
İnternetten alışverişte kullanıcı tercihleri	RO	HOLD-OUT	90.51
	RO	ÇD	88.95
Şarap için müşteri segmentasyonu	LRS, YSA	HOLD-OUT	100
	RO, LRS, YSA	ÇD	98.31
Gelir sınıflandırması	GA	HOLD-OUT	87.63
	GA	ÇD	87.27
Kredi kartı sahtekarlığı tespiti	RO	HOLD-OUT	99.95
	LRS, DVM, YSA	ÇD	99.83
Pima	YSA	HOLD-OUT	80.95
	LRS	ÇD	76.82
Sesle cinsiyet tanıma	KNN	HOLD-OUT	98
	RO	ÇD	96.78
Deniz kulağı	RO	HOLD-OUT	64.59
	KNN	ÇD	63.83
Ph tanıma	GA	HOLD-OUT	78.57
	RO	ÇD	75.19

4. SONUÇLAR

Makine öğrenmesi model seçim sürecinde en önemli sorunlardan birisi en iyi sınıflandırıcının seçimi sürecidir. Bu tez çalışması kapsamında en iyi modelin seçimi sürecinde, model performans değerlendirme yöntemleri ve model performans değerlendirme metrikleri kullanılmıştır. Model performans değerlendirme yöntemlerinden hold-out ve 10 katlı ÇD yöntemi kullanılmıştır. Model performans değerlendirme metrikleri olarak da karmaşıklık matrisi kullanılarak doğruluk, kesinlik, anma, duyarlılık, MCC, AUC ve F_1 kullanılmıştır.

Bu tez kapsamında 32 adet farklı dağılım ve özellikteki veri setine KNN, KA, RO, NB, LRS, DVM, GA, AB ve YSA sınıflandırıcıları ile yukarıda bahsedilen 2 farklı model değerlendirme yöntemi kullanılarak model oluşturulmuştur. Oluşturulan bu modeller daha sonra karmaşıklık matrisi kullanılarak değerlendirilmiştir. Yapılan değerlendirme sonucunda;

- 32 veri seti için oluşturulan modellerde 20 veri seti için hold-out yöntemi ÇD'ye göre daha iyi sonuç vermiştir. Bunlardan sadece 3 sınıflı veri seti olan yapısal protein dizileri isimli veri setinde belirgin farklılık olduğu görülmüştür. Diğer 19 veri setinde ise hold-out yöntemi ile ÇD yakın değerler elde etmiştir. ÇD kullanılabileceği durumlarda Hold-out yöntemine gerek duyulmadığı gözlemlenmiştir. Fakat ÇD kullanıldığında sınıflandırma çok fazla maliyet gerektiriyorsa ve veri sayısının yetersiz olduğu durumlarda Hold-out yönteminin kullanılabileceği tespit edilmiştir.
- Veri dağılımının dengesiz olduğu durumlarda doğruluk metriğinin yanında MCC, AUC ve F_1 değerlerinin de kullanılması modelin başarımının ölçülmesinde daha önemli olmaktadır. Bölüm 3.1.29'da bahsedilen Kredi kartı sahtekarlığı tespiti veri seti dengesiz bir veri setidir. Burada elde edilen doğruluk oranı %99.83 olmasına rağmen bu veri setinde hileli işlemlerin hiçbiri doğru tahmin edilememiştir. Doğruluk metriği ile birlikte MCC, AUC ve F_1 değerlerine bakıldığında sınıflandırıcımızın rasgele bir sınıflandırma yaptığı görülmektedir. MCC değerinin 0 olması, AUC değerinin 0.5 olması sınıflandırıcının herhangi bir tahminde bulunmadığını göstermektedir. Aynı veri seti için NB sınıflandırıcısının doğruluk değeri %99.26 olduğu görülmektedir. Bu sınıflandırıcı ile ilgili diğer metriklere bakıldığında MCC değerinin 0.2991, AUC değerinin 0.8177 ve F_1 değerinin de 0.6721 olduğu görülmektedir. Bu da %99.83 başarım elde edilen LRS, DVM ve YSA sınıflandırıcılarına göre NB sınıflandırıcısının daha iyi sonuç verdiğini göstermektedir.
- Bazı veri setleri ile oluşturulan modellerde %90 ve üstü başarımlar kötü sonuçlar verirken bazı veri setlerinde ise bu değerden daha kötü sonuç elde edilen modeller daha doğru sonuçlar vermektedir. İki sınıflı bir veri seti olan seyahat sigortası veri setiyle oluşturulan DVM ve YSA modelleriyle elde edilen başarım değeri %98.54'dür. PH tanıma veri seti ise 15 sınıflı bir veri seti olmakla birlikte bu veri setiyle oluşturulan RO modelinin başarım

değeri %75.19 olduğu görülmektedir. Seyahat sigortası veri seti için oluşturulan modellerde MCC, AUC ve F_1 değerlerine bakıldığında sırasıyla 0, 0.5 ve 0.49 olduğu görülmekte iken, ph tanıma veri setinde aynı değerler 0.734, 0.9114 ve 0.5293 olduğu görülmektedir. Bu da daha kötü bir başarım değeri elde edilen ph tanıma veri seti ile oluşturulan modelin seyahat sigortası veri seti ile oluşturulan modellerden daha iyi tahminlerde bulunduğunu göstermektedir.

Sonuç olarak sınıflandırma sürecinde, sınıflandırma başarımının doğru değerlendirilmesi için veri setinin dengeli olup olmama durumu, veri setindeki sınıf sayısı ve kullanılan test yönteminin hangisi olduğu büyük önem taşıdığı tespit edilmiş ve sınıflandırmanın bu parametrelerinin de göz önüne alınarak değerlendirilmesi gerektiği sonucuna varılmıştır.

KAYNAKLAR

[1] Doğan, K. ve Arslantekin, S. (2016). Büyük veri: önemi, yapısı ve günümüzdeki durum. DTCF dergisi, 56(1), 15-36.

[2] Gürsakal, N. (2017). Makine Öğrenmesi ve Derin Öğrenme. Bursa: Dora.

[3] Göbekçin, T. (Ed.)(2017). Master algoritma. İstanbul:Paloma.

[4] https://open.nasa.gov/blog/datanaut-fall-2017-class-announcement/, Erişim: 21.01.2020.

[5] https://www.theguardian.com/media-network/2015/mar/05/digital-oligarchy-algorithms-personal-data, Erişim: 21.01.2020

[6] https://www.kaggle.com/datasets, Erişim: 21.01.2020

[7] Aydın,F. (2011). Kalp ritim bozukluğu olan hastaların tedavi süreçlerini desteklemek amaçlı makine öğrenmesine dayalı bir sistemin geliştirilmesi, Yüksek Lisans Tezi, Trakya Üniversitesi, Fen Bilimleri Üniversitesi, Edirne.

[8] Hacıefendioğlu, Ş. (2012). Makine öğrenmesi yöntemleri ile glokom hastalığının teşhisi, Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.

[9] Kartal, E. (2015). Sınıflandırmaya dayalı makine öğrenmesi teknikleri ve kardiyolojik risk değerlendirmesine ilişkin bir uygulama, Doktora Tezi, İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

[10] Ünal, Y. (2015). Makine öğrenmesi yöntemleriyle bel bölgesi rahatsızlıklarının tanısı, Doktora Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.

[11] Şeker, M. (2017). İyi–kötü kokular ile ilişkili EMOTIV-EPOC tabanlı EEG kayıtlarının makine öğrenmesi yöntemleri ile sınıflandırılması, Yüksek Lisans Tezi, Dicle Üniversitesi, Fen Bilimleri Enstitüsü, Diyarbakır.

[12] Turgut, S. (2017). Makine öğrenmesi yöntemleri kullanarak kanser teşhisi, Yüksek Lisans Tezi, İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, İstanbul.

[13] Pekel, E. (2018). Farklı makine öğrenmesi algoritmalarının karşılaştırılması, Yüksek Lisans Tezi, Ondokuz Mayıs Üniversitesi, Fen Bilimleri Enstitüsü, Samsun.

[14] https://www.kdnuggets.com/2017/05/machine-learning-overtaking-big-data.html,Erişim:21.01.2020.

[15] https://www.pcmag.com/article/353293/7-tips-for-machine-learning-success, Erişim: 21.01.2020.

[16] https://dzone.com/articles/demystifying-ai-machine-learning-and-deep-learning, Erişim: 21.01.2020.

[17] https://www.wired.com/story/apples-faceid-could-be-a-powerful-tool-for-mass-spying/, Erişim: 21.01.2020.

[18] Balaban, M.E. ve Kartal, E. (2015). Veri madenciliği ve makine öğrenmesi Temel Algoritmalar ve R Dili ile Uygulamaları. İstanbul: Çağlayan Kitapevi

[19] Onan, A., & Korukoğlu, S. (2016). Makine öğrenmesi yöntemlerinin görüş madenciliğinde kullanılması üzerine bir literatür araştırması. Pamukkale University Journal of Engineering Sciences, 22(2).

[20] Solve Machine Learning Problems Step-by-step, https://machinelearningmastery.com/working-machine-learning-problem/ , Erişim: 31.08.2018,

[21] About Data Mining: Basic steps of applying machine learning methods, http://www.aboutdm.com/2013/03/basic-steps-of-applying-machine.html, Erişim: 31.08.2018.

[22] Rossi, J.J. (2007). MicroRNA Methods. Academic Press.

[23] Han, J., Kamber, M. ve Pei, J., (2012). Data mining: Concepts and techniques. Morgan Kaufmann, 740.

[24] Bishop, C. M., 2007, Pattern Recognition and Machine Learning, Springer, ISBN: 0-387-31073-8.

[25] Wong, M. A., ve Hartigan, J. A. (1979). A k-means clustering algorithm. Journal of the Royal Statistical Society, 28(1), 100-108.

[26] Polat, S. (2017), "Yazılım Hata Kayıtlarının Makine Öğrenmesi Yöntemleriyle Kümelenerek, Hataya Sebep Olan Bileşenlerin Tespit Edilmesi", 11. Ulusal Yazılım Mühendisliği Sempozyumu, 18-20 Ekim 2017, Alanya, 444-453.

[27] Kızılkaya, Y. M., ve Oğuzlar, A. (2018), Bazı Denetimli Öğrenme Algoritmalarının R Programlama Dili İle Kıyaslanması. Karadeniz Uluslararası Bilimsel Dergi, 37(37), 90-98.

[28] Deepak,P., Pulkit, A., Alexei, E.A., Trevor, D.(2017), "Curiosity-driven Exploration by Self-supervised Prediction", arXiv:1705.05363v1 [cs.LG] 15 Mayıs 2017, https://arxiv.org/pdf/1705.05363.pdf

[29] Fawcett, T., (2006). An introduction to ROC analysis, Pattern recognition letters, 27 (8), 861–874.

[30] Vellido, A., Martin-Guerrero, J. D. ve Lisboa, P. J. G., 2012, Making machine learning models interpretable,2012 Louvain-La-Neuve, I6doc.com, ISBN: 978-2-87419-049-0.

[31] Model Selection and Feature Selection (CS5350/6350: Machine Learning), http://www.cs.utah.edu/~piyush/teaching/22-9-print.pdf, Erişim: 02.06.2015.

[32] Ayık, Y. Z., Özdemir, A. ve Yavuz, U. (2007). Lise türü ve lise mezuniyet başarısının, kazanılan fakülte ile ilişkisinin veri madenciliği tekniği ile analizi. Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10(2), 441-454.

[33] Steinbach, M., & Tan, P. N. (2009). kNN: k-nearest neighbors. In The top ten algorithms in data mining (pp. 165-176). Chapman and Hall/CRC.

[34] Bozkır, A. S., Sezer, E., ve GÖK, Bilge.,(2009). Öğrenci seçme sınavında (öss) öğrenci başarımını etkileyen faktörlerin veri madenciliği yöntemleriyle tespiti. 5. Uluslararası İleri Teknolojiler Sempozyumu (IATS'09).

[35] Albayrak, A. S., ve YILMAZ, Ö. G. Ş. K. (2009). Veri madenciliği: Karar ağacı algoritmaları ve İMKB verileri üzerine bir uygulama. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14(1), 31-52.

[36] Breiman L., (2001), Random forests,machine learning, 2001 Kluwer Academic Publishers, 45(1), 5-32.

[37] Archer K.J., (2008).Emprical characterization of random forest variable ımportance measure, computational statistical data analysis, Computational Statistics & Data Analysis, 52(4), 2249-2260.

[38] Random forest, http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm, Erişim: 24 Nisan 2019.

[39] Özdemir, M. E., Yıldırım, E. ve Yıldırım, S., (2015). Classification of emotional valence dimension using artificial neural networks. In Signal Processing and Communications Applications Conference (SIU), 23, 2549-2552.

[40] Boser, B., Guyon, I. ve Vapnik, V., (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory, 144–152.

[41] Melgani, F. ve Bruzzone, L., (2004). Classification of hyperspectral remote sensing ımages with support vector machines. Geoscience and Remote Sensing, IEEE Transactions, 42(8), 1778-1790.

[42] Kégl, B., 2009, Introduction to AdaBoost, 11-14.

[43] Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

[44] Evaluating Machine-Learning Methods, http://pages.cs.wisc.edu/~dpage/cs760/evaluating.pdf, Erişim: 02.11.2017.

[45] Dua, S. ve Chowriappa, P., (2013). Data mining for bioinformatics, CRC Press, ISBN: 978-1-4200-0430-4.

[46] Lecture 13: Validation (Intelligent Sensor Systems), http://research.cs.tamu.edu/prism/lectures/iss/iss_l13.pdf, Erişim: 24 Kasım 2017.

[47] Nordman, A., 2011, Data Mining lecture 6: Evaluating the performance of a model, http://staffwww.itn.liu.se/~aidvi/courses/06/dm/lectures/lec6.pdf, Erişim: 02 Aralık 2017.

[48] http://web.itu.edu.tr/~sgunduz/courses/verimaden/slides/d4.pdf, Erişim: 02 Aralık 2017.

[49] Refaeilzadeh, P., Tang, L. and Liu, H., (2009). Cross-validation, encyclopedia of database systems, Springer, 532–538.

[50] Saharidis, G. K. D., Androulakis, I. P. and Lerapetritou, M. G., (2011). Model building using bi-level optimization, Journal of Global Optimization, 49 (1), 49–67.

[51] Remesan, R. and Mathew, J., (2014). Hydrological data driven modelling: A Case Study Approach, Springer, ISBN: 978-3-319-09235-5.

[52] Martinasek, Z., Hajny, J. and Malina, L., (2014). Optimization of power analysis using neural network, Smart Card Research and Advanced Applications, Springer, ISBN: 3-319-08301-5, 94–107.

[53] Rajeev, S., (2013). Research developments in computer vision and image processing: Methodologies and Applications: Methodologies and Applications, IGI Global, ISBN: 978-1-4666-4559-2.

[54] Guil-Reyes, F., and Daza-Gonzalez, M. T. (2011, October). Summarizing frequent itemsets via pignistic transformation. In Portuguese Conference on Artificial Intelligence (pp. 297-310). Springer, Berlin, Heidelberg.

[55] Olson, D. L., and Delen, D. (2008). Advanced data mining techniques. Springer Science & Business Media.

[56] Baker, D. R. (2007). A hybrid approach to expert and model based effort estimation. West Virginia University. ISBN: 0-549-46778-5.

[57] Kohavi, R., (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, 1137–1145.

[58] Rogers, S. and Girolami, M., (2011). A first course in machine learning, CRC Press, ISBN: 978-1-4398-2414-6.

[59] Japkowicz, N., (2011). Performance evaluation for learning algorithms, Cambridge University Press, Cambridge

[60] Sevindi, B.İ. (2013). Türkçe metinlerde denetimli ve sözlük tabanlı duygu analizi yaklaşımlarının karşılaştırılması, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen bilimleri Enstitüsü, Konya

[61] Akosa, J. (2017). Predictive accuracy: a misleading performance measure for highly imbalanced data. In Proceedings of the SAS Global Forum (pp. 2-5).

[62] Ertorsun, A. D., Bağ, B., Uzar, G. ve Turanoğlu, M. A., (2009). ROC (Receiver Operating Characteristic) eğrisi yöntemi ile tanı testlerinin performanslarının değerlendirilmesi, XIII. Öğrenci Sempozyumu, 2009, Ankara.

[63] Alpaydin, E., (2004). Introduction to machine learning. The MIT Press Cambridge, Massachusetts London, England, 433.

[64] Kılıç, S. (2013). Klinik karar vermede ROC analizi. Journal Of Mood Disorders, 3 (3), 135-40.

[65] Mukkamala, R., (2013). Evaluating a classification model–What does precision and recall tell me?, http://www.cs.odu.edu/~mukka/cs495s13/Lecturenotes /Chapter5/recallprecision.pdf, 02.11.2017.

[66] Justan, M. P., (2002). Integrating real medical studies into teaching biostatistics, Conference Proceedings, Greece.

[67] Sokolova, M. ve Lapalme, G., (2009). A systematic analysis of performance measures for classification tasks, Information Processing & Management, 45 (4), 427–437.

[68] Felkin, M., (2007). Comparing classification results between n-ary and binary problems, Quality Measures in Data Mining, In: Guillet, F. ve Hamilton, H. J. (ed.), Springer, 277–301.

[69] Kazemian, M., Moshiri, B., Palade, V. and Nikbakht, H., (2010). Using classifier fusion techniques for protein secondary structure prediction, International Journal of Computational Intelligence in Bioinformatics and Systems Biology, 1 (4), 418– 434.

[70] Tuncer, E. (2015). Uyku evrelemesinde çeşitli dalgacık ve sınıflandırıcıların performans analizi, Yüksek Lisans, Kocaeli Üniversitesi, Fen Bilimleri Enstitüsü, Kocaeli.

[71] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 27.

EKLER

EK-1: OLUŞTURULAN MODELLERE AİT ELDE EDİLEN PERFORMANS METRİK SONUÇLARI

Veri seti	Sınıflandırıcı	Doğ.(%)	Kes.	Anma	F_1	Özg.	MCC	AUC
Avustralya'da yağmur	KNN	82.74	0.7534	0.6511	0.6985	0.6511	0.3913	0.6511
	KA	83.54	0.7913	0.6460	0.7113	0.6460	0.4124	0.6460
	RO	83.53	0.7894	0.6469	0.7111	0.6469	0.4123	0.6469
	NB	83.51	0.7808	0.6540	0.7118	0.6540	0.4158	0.6540
	LRS	83.49	0.7798	0.6543	0.7116	0.6543	0.4155	0.6543
	DVM	83.41	0.8124	0.6279	0.7083	0.6279	0.3997	0.6279
	GA	83.54	0.7913	0.6460	0.7113	0.6460	0.4124	0.6460
	AB	83.49	0.7874	0.6472	0.7104	0.6472	0.4112	0.6472
	YSA	83.5	0.7984	0.6392	0.7100	0.6392	0.4076	0.6327
Tic-Tac-Toe Oyunu	KNN	90.19	0.9194	0.8662	0.8920	0.8662	0.7838	0.8662
	KA	94.57	0.9368	0.9450	0.9409	0.9450	0.8817	0.9480
	RO	93.64	0.9226	0.9478	0.9350	0.9478	0.8699	0.9485
	NB	64.51	0.5947	0.5827	0.5887	0.5827	0.1770	0.5827
	LRS	62.53	0.5792	0.5753	0.5772	0.5753	0.1544	0.5753
	DVM	79.85	0.8031	0.7390	0.7697	0.7390	0.5383	0.7390
	GA	100	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	AB	62.84	0.5902	0.5903	0.5903	0.5903	0.1806	0.5905
	YSA	83.72	0.8449	0.7891	0.8160	0.7891	0.6315	0.8019
Wisconsin Meme Kanseri	KNN	96.66	0.9708	0.9581	0.9644	0.9581	0.9288	0.9581
	KA	93.32	0.9293	0.9276	0.9285	0.9276	0.8569	0.9163
	RO	96.13	0.9604	0.9567	0.9585	0.9567	0.9170	0.9567
	NB	93.15	0.9286	0.9243	0.9264	0.9243	0.8528	0.9243
	LRS	96.49	0.9708	0.9547	0.9627	0.9547	0.9253	0.9547
	DVM	95.25	0.9633	0.9373	0.9501	0.9373	0.9001	0.9373
	GA	95.61	0.9604	0.9458	0.9531	0.9458	0.9061	0.9458
	AB	97.01	0.9735	0.9628	0.9681	0.9628	0.9362	0.9628
	YSA	97.89	0.9814	0.9736	0.9775	0.9736	0.9549	0.9713
Sloan gökyüzü araştırması	KNN	77.77	0.5880	0.5709	0.5793	0.8614	0.5999	0.84
	KA	98.27	0.9628	0.9667	0.9648	0.9910	0.9700	1
	RO	98.88	0.9830	0.9743	0.9786	0.9936	0.9800	1
	NB	85.56	0.8339	0.8674	0.8503	0.9145	0.7540	0.85
	LRS	88.17	0.8901	0.8369	0.8627	0.9268	0.7910	0.88
	DVM	70.60	0.6892	0.5357	0.6028	0.8175	0.4620	0.75
	GA	98.89	0.9777	0.9753	0.9765	0.9940	0.9805	1
	AB	83.04	0.7262	0.7674	0.7462	0.9181	0.7200	0.96
	YSA	71.18	0.5825	0.5144	0.5463	0.8164	0.3100	0.7
Pulsar yıldızı tahmini	KNN	97.80	0.9565	0.9072	0.9312	0.9072	0.8622	0.9072
	KA	96.79	0.9011	0.9074	0.9042	0.9074	0.8084	0.9090
	RO	97.94	0.9557	0.9171	0.9360	0.9171	0.8719	0.9171
	NB	94.31	0.8169	0.9002	0.8565	0.9002	0.7122	0.9007
	LRS	97.38	0.9588	0.8783	0.9168	0.8783	0.8332	0.8783
	DVM	97.40	0.9640	0.8752	0.9174	0.8752	0.8344	0.8752
	GA	97.58	0.9361	0.9156	0.9258	0.9156	0.8514	0.9156
	AB	97.83	0.9552	0.9107	0.9324	0.9107	0.8647	0.9107
	YSA	96.74	0.9621	0.8361	0.8947	0.8361	0.7882	0.8361
Ortopedik Hastaların Biyomekanik Özellikleri(2 sınıf)	KNN	78.39	0.7532	0.7619	0.7575	0.7619	0.5150	0.7619
	KA	78.39	0.7530	0.7593	0.7561	0.7593	0.5122	0.7493
	RO	81.29	0.7860	0.7860	0.7860	0.7860	0.5719	0.7860
	NB	78.06	0.7694	0.8067	0.7876	0.8067	0.5748	0.8067
	LRS	72.26	0.6834	0.6276	0.6543	0.6276	0.3059	0.6276
	DVM	67.42	0.5628	0.5081	0.5340	0.5081	0.0450	0.5081
	GA	77.74	0.7458	0.7519	0.7488	0.7519	0.4976	0.7519
	AB	80.65	0.7783	0.7812	0.7798	0.7812	0.5595	0.7812
	YSA	75.81	0.7240	0.7010	0.7123	0.7010	0.4243	0.6936
Ortopedik Hastaların	KNN	76.45	0.7297	0.7333	0.7315	0.8815	0.6247	0.88
	KA	77.42	0.7184	0.7189	0.7186	0.8903	0.6745	0.95

Biyomekanik Özellikleri(3 sınıf)	RO	84.52	0.8058	0.7978	0.8018	0.9259	0.7520	0.97
	NB	83.55	0.7987	0.7967	0.7977	0.9177	0.7360	0.95
	LRS	71.94	0.7528	0.6189	0.6793	0.8421	0.5416	0.83
	DVM	68.71	0.4551	0.5622	0.5030	0.8319	0.5012	0.85
	GA	82.26	0.7744	0.7667	0.7705	0.9141	0.7145	0.96
	AB	74.19	0.7065	0.7144	0.7104	0.8777	0.5982	0.90
	YSA	79.35	0.7735	0.7422	0.7576	0.8923	0.6771	0.91
Yapısal protein dizileri	KNN	43.60	0.4319	0.4277	0.4298	0.7139	0.1430	0.56
	KA	46.81	0.4655	0.4655	0.4655	0.7321	0.1960	0.60
	RO	51.93	0.5178	0.5141	0.5159	0.7564	0.2710	0.62
	NB	42.14	0.4174	0.4181	0.4177	0.7082	0.1260	0.55
	LRS	42.23	0.4255	0.4029	0.4139	0.7002	0.1110	0.55
	DVM	45.05	0.4486	0.4439	0.4462	0.6820	0.1660	0.58
	GA	50.04	0.5010	0.4928	0.4969	0.7460	0.2410	0.62
	AB	45.71	0.4530	0.4534	0.4532	0.7264	0.1790	0.57
	YSA	43.83	0.4336	0.4329	0.4333	0.7157	0.1470	0.55
Kalp Hastalığı	KNN	83.50	0.8348	0.8319	0.8333	0.8319	0.6666	0.8319
	KA	72.94	0.7272	0.7266	0.7269	0.7266	0.4538	0.7333
	RO	82.51	0.8256	0.8210	0.8233	0.8210	0.6466	0.8210
	NB	79.21	0.7904	0.7913	0.7908	0.7913	0.5816	0.7913
	LRS	84.16	0.8422	0.8379	0.8400	0.8379	0.6800	0.8379
	DVM	83.17	0.8312	0.8289	0.8300	0.8289	0.6600	0.8289
	GA	78.55	0.7840	0.7829	0.7834	0.7829	0.5668	0.7829
	AB	78.55	0.7882	0.7787	0.7834	0.7787	0.5668	0.7787
	YSA	83.83	0.8400	0.8337	0.8368	0.8337	0.6736	0.8337
Farelerin protein ekspresyonu	KNN	73.55	0.7476	0.7384	0.7430	0.9621	0.6981	0.8007
	KA	76.27	0.7603	0.7516	0.7559	0.9660	0.6951	0.7704
	RO	91.67	0.9196	0.9149	0.9172	0.9880	0.9044	0.9071
	NB	79.35	0.8041	0.7926	0.7983	0.9705	0.7654	0.8296
	LRS	98.19	0.9775	0.9854	0.9814	0.9975	0.9794	0.9930
	DVM	96.74	0.9682	0.9660	0.9671	0.9953	0.9626	0.9646
	GA	88.41	0.8811	0.8822	0.8816	0.9834	0.8671	0.8901
	AB	34.18	0.3100	0.3112	0.3106	0.9042	0.2513	0.5274
	YSA	98.91	0.9892	0.9872	0.9882	0.9984	0.9875	0.9656
Seyahat sigortası	KNN	98.38	0.4977	0.4998	0.4987	0.4998	-0.0014	0.4997
	KA	96.73	0.5111	0.5148	0.5129	0.5148	0.0256	0.5147
	RO	98.16	0.5176	0.5050	0.5112	0.5050	0.0187	0.5050
	NB	5.53	0.5035	0.5095	0.5065	0.5095	0.0114	0.5094
	LRS	98.53	0.4927	0.5000	0.4963	0.5000	-0.0010	0.4999
	DVM	98.54	0.4900	0.5000	0.4949	0.5000	0	0.5000
	GA	98.31	0.4961	0.4994	0.4977	0.4994	-0.0030	0.4993
	AB	98.53	0.4927	0.5000	0.4963	0.5000	-0.0006	0.4999
	YSA	98.54	0.4900	0.5000	0.4949	0.5000	0	0.5000
Kas aktivitesini okuyarak jestlerin sınıflandırılması	KNN	99.86	0.9986	0.9986	0.9986	0.9995	0.9981	0.9994
	KA	100	1.0000	1.0000	1.0000	1.0000	1	1
	RO	99.98	0.9998	0.9998	0.9998	0.9999	0.9997	1
	NB	100	1.0000	1.0000	1.0000	1.0000	1	1
	LRS	99.70	0.9970	0.9970	0.9970	0.9990	0.9956	1
	DVM	100	1.0000	1.0000	1.0000	1.0000	1	1
	GA	100	1.0000	1.0000	1.0000	1.0000	1	1
	AB	60.02	0.6300	0.7500	0.6848	0.8749	0.7312	1
	YSA	100	1.0000	1.0000	1.0000	1.0000	1	1
Parkinson hastalığı sınıflandırma	KNN	76.59	0.6833	0.6353	0.6584	0.6353	0.3149	0.6352
	KA	73.15	0.6468	0.6483	0.6475	0.6483	0.2950	0.6691
	RO	81.48	0.7714	0.6990	0.7334	0.6990	0.4647	0.6989
	NB	80.95	0.7533	0.7126	0.7324	0.7126	0.4640	0.7126
	LRS	80.69	0.7871	0.6559	0.7155	0.6559	0.4230	0.6558
	DVM	74.60	0.3700	0.5000	0.4253	0.5000	0	0.5000
	GA	79.09	0.7149	0.6756	0.6947	0.6756	0.3885	0.6899
	AB	79.63	0.7320	0.6986	0.7149	0.6986	0.4292	0.6985
	YSA	68.56	0.6103	0.6254	0.6177	0.6254	0.2351	0.6267
Portekiz bankası pazarlama	KNN	88.27	0.5997	0.5096	0.5510	0.5096	0.0620	0.5096
	KA	82.52	0.5907	0.6042	0.5974	0.6042	0.1944	0.6042
	RO	86.35	0.6353	0.6023	0.6183	0.6023	0.2352	0.6022
	NB	81.93	0.5553	0.5564	0.5558	0.5564	0.1116	0.5563

	LRS	88.61	0.5632	0.5004	0.5300	0.5004	0.0098	0.5003
	DVM	88.75	0.7144	0.5332	0.6107	0.5332	0.1688	0.5332
	GA	87.68	0.5844	0.5178	0.5491	0.5178	0.0775	0.5177
	AB	87.77	0.5624	0.5101	0.5350	0.5101	0.0502	0.5101
	YSA	84.35	0.5274	0.5163	0.5218	0.5163	0.0423	0.5163
Genetik çeşitlilik sınıflandırma	KNN	64.81	0.4893	0.4913	0.4903	0.4913	-0.0193	0.4912
	KA	68.21	0.4874	0.4927	0.4901	0.4927	-0.0191	0.4925
	RO	68.70	0.4858	0.4924	0.4891	0.4924	-0.0207	0.4923
	NB	67.87	0.4989	0.4993	0.4991	0.4993	-0.0018	0.4992
	LRS	74.52	0.4873	0.4995	0.4933	0.4995	-0.0050	0.4994
	DVM	74.93	0.3700	0.5000	0.4253	0.5000	0	0.5000
	GA	70.88	0.4916	0.4968	0.4942	0.4968	-0.0102	0.4968
	AB	74.52	0.4855	0.4994	0.4924	0.4994	-0.0057	0.4994
	YSA	74.93	0.3700	0.5000	0.4253	0.5000	0	0.5000
Mobil cihaz fiyat sınıflandırması	KNN	44.05	0.4580	0.4405	0.4491	0.8135	0.2551	0.6956
	KA	82.95	0.8302	0.8295	0.8299	0.9432	0.7700	0.9233
	RO	88.40	0.8835	0.8840	0.8837	0.9613	0.8453	0.9530
	NB	81.20	0.8132	0.8120	0.8126	0.9373	0.7493	0.9306
	LRS	79.20	0.7858	0.7920	0.7889	0.9307	0.7283	0.9513
	DVM	90.45	0.9073	0.9045	0.9059	0.9682	0.8730	0.9423
	GA	91.40	0.9145	0.9140	0.9143	0.9713	0.8853	0.9540
	AB	59.74	0.6902	0.7100	0.7000	0.8608	0.6214	0.8343
	YSA	95.85	0.9585	0.9585	0.9585	0.9862	0.9427	0.9846
Türkiye siyasi görüşleri	KNN	40.56	0.2800	0.2788	0.2794	0.8637	0.1944	0.5069
	KA	36.50	0.2685	0.2693	0.2689	0.8595	0.1541	0.5416
	RO	39.66	0.2822	0.2845	0.2833	0.8641	0.1915	0.5530
	NB	41.24	0.3009	0.3120	0.3063	0.8698	0.2345	0.5259
	LRS	43.62	0.2829	0.3105	0.2961	0.8718	0.242	0.5075
	DVM	42.37	0.2900	0.3005	0.2952	0.8683	0.2231	0.5141
	GA	38.64	0.2826	0.2807	0.2817	0.8624	0.1807	0.5437
	AB	32.77	0.2638	0.3055	0.2832	0.8599	0.1611	0.5255
	YSA	40.45	0.2723	0.2898	0.2808	0.8648	0.2176	0.5246
Banka pazarlama	KNN	71.36	0.7238	0.7068	0.7152	0.7068	0.4303	0.7068
	KA	75.20	0.7531	0.7492	0.7511	0.7492	0.5022	0.7491
	RO	80.52	0.8052	0.8039	0.8045	0.8039	0.6090	0.8038
	NB	69.29	0.7198	0.6825	0.7007	0.6825	0.4005	0.6825
	LRS	79.08	0.7943	0.7874	0.7908	0.7874	0.5816	0.7873
	DVM	79.94	0.8001	0.7974	0.7987	0.7974	0.5974	0.7973
	GA	80.06	0.8021	0.7981	0.8001	0.7981	0.6001	0.7980
	AB	78.55	0.7898	0.7817	0.7857	0.7817	0.5714	0.7817
	YSA	80.21	0.8031	0.7999	0.8015	0.7999	0.6030	0.7998
İris çiçeği	KNN	96.00	0.9605	0.9600	0.9602	0.9800	0.9402	0.9500
	KA	95.33	0.9534	0.9533	0.9534	0.9767	0.93	0.9450
	RO	96.00	0.9600	0.9600	0.9600	0.9800	0.94	0.9550
	NB	95.33	0.9534	0.9533	0.9534	0.9767	0.93	0.9450
	LRS	84.00	0.8658	0.8400	0.8527	0.9200	0.7769	0.8650
	DVM	96.67	0.9678	0.9667	0.9672	0.9700	0.9505	0.9550
	GA	96.00	0.9600	0.9600	0.9600	0.9800	0.94	0.9550
	AB	94.67	0.9471	0.9467	0.9469	0.9733	0.9202	0.9350
	YSA	96.67	0.9668	0.9667	0.9667	0.9833	0.95	0.9650
Şarap kalitesi	KNN	45.53	0.2400	0.2332	0.2365	0.8826	0.1848	0.5000
	KA	43.13	0.2187	0.2218	0.2203	0.8795	0.1497	0.4996
	RO	53.53	0.3600	0.2361	0.2852	0.8926	0.2660	0.5000
	NB	33.65	0.2221	0.2520	0.2361	0.8723	0.1017	0.5984
	LRS	52.70	0.2300	0.1995	0.2137	0.8871	0.2353	0.5000
	DVM	52.12	0.1500	0.1924	0.1686	0.8851	0.2219	0.5000
	GA	50.84	0.2872	0.2528	0.2689	0.8902	0.2417	0.4997
	AB	37.60	0.2100	0.2015	0.2057	0.8745	0.1173	0.5000
	YSA	52.62	0.2300	0.2011	0.2146	0.8873	0.2455	0.5000
Hepatoselüler Karsinom (HCC)	KNN	91.67	0.9340	0.8932	0.9131	0.8932	0.8261	0.8931
	KA	100	1.0000	1.0000	1.0000	1.0000	1	1
	RO	100	1.0000	1.0000	1.0000	1.0000	1	1
	NB	100	1.0000	1.0000	1.0000	1.0000	1	1
	LRS	100	1.0000	1.0000	1.0000	1.0000	1	1
	DVM	100	1.0000	1.0000	1.0000	1.0000	1	1

	GA	100	1.0000	1.0000	1.0000	1.0000	1	1
	AB	100	1.0000	1.0000	1.0000	1.0000	1	1
	YSA	100	1.0000	1.0000	1.0000	1.0000	1	1
Bireysel kredi sınıflandırma problemi	KNN	96.56	0.9561	0.8385	0.8935	0.8385	0.7858	0.8385
	KA	98.10	0.9464	0.9439	0.9451	0.9439	0.8902	0.9438
	RO	98.62	0.9800	0.9393	0.9592	0.9393	0.9183	0.9392
	NB	88.28	0.6867	0.7471	0.7156	0.7471	0.4295	0.7471
	LRS	95.00	0.9099	0.7796	0.8397	0.7796	0.6770	0.7796
	DVM	95.76	0.9675	0.7848	0.8666	0.7848	0.7296	0.7847
	GA	98.76	0.9790	0.9485	0.9635	0.9485	0.9269	0.9484
	AB	96.82	0.9256	0.8847	0.9047	0.8847	0.8091	0.8846
	YSA	95.68	0.9590	0.7852	0.8635	0.7852	0.7236	0.7852
Sahte şirketleri denetlemek için denetim riski	KNN	94.72	0.9579	0.9339	0.9458	0.9339	0.8915	0.9339
	KA	100	1.0000	1.0000	1.0000	1.0000	1	1.
	RO	99.87	0.9989	0.9984	0.9987	0.9984	0.9973	0.9983
	NB	95.23	0.9524	0.9474	0.9499	0.9474	0.8998	0.9474
	LRS	96.01	0.9682	0.9498	0.9589	0.9498	0.9177	0.9497
	DVM	96.78	0.9701	0.9625	0.9663	0.9625	0.9325	0.9624
	GA	100	1.0000	1.0000	1.0000	1.0000	1	1.
	AB	100	1.0000	1.0000	1.0000	1.0000	1	1.
	YSA	95.75	0.9645	0.9476	0.9560	0.9476	0.9119	0.9476
İnternetten alışverişte kullanıcı tercihleri	KNN	86.59	0.7963	0.6017	0.6855	0.6017	0.3472	0.6017
	KA	84.42	0.7039	0.7059	0.7049	0.7059	0.4097	0.7046
	RO	88.95	0.8052	0.7334	0.7677	0.7334	0.5338	0.7334
	NB	71.55	0.5836	0.6309	0.6064	0.6309	0.2092	0.6308
	LRS	87.44	0.7915	0.6547	0.7166	0.6547	0.4246	0.6546
	DVM	84.52	0.6450	0.5008	0.5638	0.5008	0.0216	0.5008
	GA	87.91	0.7742	0.7304	0.7517	0.7304	0.5027	0.7308
	AB	86.85	0.7509	0.7062	0.7279	0.7062	0.4548	0.7062
	YSA	86.91	0.7587	0.6753	0.7146	0.6753	0.4259	0.6753
Şarap için müşteri segmentasyonu	KNN	93.82	0.9393	0.9461	0.9427	0.9696	0.9096	0.9780
	KA	89.39	0.8957	0.8894	0.8925	0.9452	0.8376	0.8974
	RO	98.31	0.9828	0.9850	0.9839	0.9915	0.9744	0.9961
	NB	96.07	0.9593	0.9652	0.9622	0.9805	0.9408	0.9884
	LRS	98.31	0.9811	0.9859	0.9835	0.9921	0.9747	0.9923
	DVM	97.75	0.9792	0.9758	0.9775	0.9770	0.9659	0.9753
	GA	94.38	0.9500	0.9399	0.9449	0.9700	0.915	0.9440
	AB	86.59	0.8885	0.8546	0.8713	0.9282	0.7976	0.8607
	YSA	98.88	0.9867	0.9906	0.9886	0.9949	0.9831	0.9923
Ph tanıma	KNN	71.98	0.7215	0.7213	0.7214	0.9800	0.7000	0.8667
	KA	71.67	0.5185	0.5312	0.5248	0.5214	0.6990	0.8806
	RO	75.19	0.5225	0.5363	0.5293	0.5125	0.7340	0.9114
	NB	61.26	0.5389	0.5565	0.5476	0.5433	0.5860	0.9397
	LRS	61.26	0.4764	0.4678	0.4721	0.4135	0.5897	0.9453
	DVM	54.21	0.4371	0.6451	0.5211	0.3821	0.5117	0.8951
	GA	73.66	0.5183	0.5205	0.5194	0.5212	0.7178	0.9000
	AB	25.27	0.3263	0.2600	0.2894	0.3690	0.2131	0.7878
	YSA	61.41	0.4739	0.5710	0.5179	0.4115	0.5902	0.9381
Gelir sınıflandırması	KNN	85.94	0.8128	0.7914	0.8020	0.7914	0.6038	0.7913
	KA	81.90	0.7525	0.7533	0.7529	0.7533	0.5058	0.7533
	RO	85.40	0.8080	0.7759	0.7916	0.7759	0.5830	0.7759
	NB	79.87	0.7439	0.8050	0.7732	0.8050	0.5454	0.8050
	LRS	85.09	0.8071	0.7636	0.7847	0.7636	0.5689	0.7636
	DVM	86.58	0.8428	0.7685	0.8039	0.7685	0.6067	0.7684
	GA	87.27	0.8398	0.7964	0.8175	0.7964	0.6346	0.7963
	AB	86.50	0.8295	0.7830	0.8056	0.7830	0.6107	0.7830
	YSA	84.20	0.7934	0.7503	0.7713	0.7503	0.5420	0.7503
Kriyoterapi analiz	KNN	90.00	0.9062	0.9048	0.9055	0.9048	0.8109	0.9047
	KA	87.78	0.8778	0.8765	0.8771	0.8765	0.7542	0.8764
	RO	94.44	0.9444	0.9464	0.9454	0.9464	0.8908	0.9464
	NB	84.44	0.8519	0.8393	0.8455	0.8393	0.6910	0.8392
	LRS	82.22	0.8225	0.8199	0.8212	0.8199	0.6424	0.8199
	DVM	84.44	0.8477	0.8408	0.8442	0.8408	0.6884	0.8407
	GA	93.33	0.9328	0.9345	0.9337	0.9345	0.8673	0.9345
	AB	93.33	0.9328	0.9345	0.9337	0.9345	0.8673	0.9345

	YSA	88.89	0.8883	0.8899	0.8891	0.8899	0.7782	0.8898
Kredi kartı sahtekarlığı tespiti	KNN	23.74	0.4972	0.2041	0.2894	0.2041	-0.0576	0.2041
	KA	55.14	0.5008	0.6079	0.5491	0.6079	0.0180	0.6078
	RO	82.96	0.5035	0.7919	0.6156	0.7919	0.0643	0.7919
	NB	99.26	0.5704	0.8178	0.6721	0.8178	0.2991	0.8177
	LRS	99.83	0.4991	0.5000	0.4996	0.5000	0	0.5000
	DVM	99.83	0.4991	0.5000	0.4996	0.5000	0	0.5000
	GA	90.29	0.5043	0.7201	0.5932	0.7201	0.0616	0.7200
	AB	90.13	0.5050	0.7578	0.6061	0.7578	0.0716	0.7577
	YSA	99.83	0.4991	0.5000	0.4996	0.5000	0	0.5000
Deniz kulağı	KNN	63.83	0.6405	0.6367	0.6386	0.8201	0.458	0.6422
	KA	56.79	0.5641	0.5653	0.5647	0.7844	0.338	0.5670
	RO	63.56	0.6308	0.6321	0.6314	0.8182	0.452	0.6230
	NB	56.64	0.5512	0.5622	0.5567	0.7835	0.349	0.5494
	LRS	62.15	0.6025	0.6152	0.6088	0.8105	0.434	0.5889
	DVM	59.37	0.5814	0.5881	0.5847	0.7553	0.39	0.5661
	GA	63.41	0.6299	0.6305	0.6302	0.8176	0.449	0.6164
	AB	63.08	0.6245	0.6271	0.6258	0.8159	0.445	0.6139
	YSA	61.65	0.6024	0.6109	0.6066	0.8082	0.436	0.5972
Sesle cinsiyet tanıma	KNN	95.96	0.9597	0.9596	0.9596	0.9596	0.9192	0.9599
	KA	94.89	0.9489	0.9489	0.9489	0.9489	0.8977	0.9488
	RO	96.78	0.9678	0.9678	0.9678	0.9678	0.9356	0.9678
	NB	86.21	0.8621	0.8621	0.8621	0.8621	0.7241	0.8620
	LRS	96.31	0.9634	0.9631	0.9633	0.9631	0.9265	0.9630
	DVM	96.37	0.9641	0.9637	0.9639	0.9637	0.9278	0.9636
	GA	96.72	0.9672	0.9672	0.9672	0.9672	0.9343	0.9671
	AB	95.96	0.9596	0.9596	0.9596	0.9596	0.9192	0.9595
	YSA	96.37	0.9639	0.9637	0.9638	0.9637	0.9276	0.9636
Pima Kızılderilileri diyabet	KNN	76.17	0.7389	0.7244	0.7316	0.7244	0.4630	0.7243
	KA	68.75	0.6589	0.6630	0.6610	0.6630	0.3219	0.6630
	RO	75.13	0.7268	0.7120	0.7194	0.7120	0.4386	0.7120
	NB	74.90	0.7234	0.7145	0.7190	0.7145	0.4378	0.7153
	LRS	76.82	0.7519	0.7190	0.7351	0.7190	0.4697	0.7189
	DVM	75.52	0.7363	0.7029	0.7192	0.7029	0.4379	0.7029
	GA	74.22	0.7158	0.7094	0.7126	0.7094	0.4250	0.7093
	AB	73.31	0.7053	0.6929	0.6990	0.6929	0.3979	0.6928
	YSA	75.91	0.7367	0.7180	0.7273	0.7180	0.4543	0.7180

ÖZGEÇMİŞ

Abdullah ALAN

KİŞİSEL BİLGİLER

Doğum Yeri : Kelkit
Doğum Yılı : 1983
Uyruğu : Türkiye Cumhuriyeti
Adres : Ataşehir mah. İmam efendi bul. No:88/7 ELAZIĞ/Merkez
E-posta : alanabdullah2001@hotmail.com
Yabancı Diller : İngilizce

EĞİTİM BİLGİLERİ

Lisans : Fırat Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 2016
Lisans : Fırat Üniversitesi, Teknik Eğitim Fakültesi, Bilgisayar Öğretmenliği Bölümü, 2006
Lise : Elazığ Fatih Lisesi, Elazığ, 2001

ARAŞTIRMA DENEYİMİ

- ✓ Arduino, Raspberry pi
- ✓ C#, Java, Python,
- ✓ AutoCad, Adobe CC, Microsoft Office, Windows, Linux

İŞ DENEYİMİ

2012 – 2020 :Karşıyaka Mesleki ve Teknik Anadolu Lisesi
2006 – 2012 : Tevfik Yaramanoğlu İlköğretim Okulu
2006 – 2006 : İstiklal İlköğretim Okulu