T.C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
SAYISAL YÖNTEMLER ANABİLİM DALI

YÜKSEK LİSANS TEZİ

DENGESİZ VERİLERE SINIFLANDIRMA ALGORİTMALARI UYGULANIRKEN KULLANILAN YÖNTEMLERİN KARŞILAŞTIRILMASINA AİT BİR ANALİZ

Kutluk ATALAY

2501201212

TEZ DANIŞMANI

Prof. Dr. Emrah ÖNDER

İSTANBUL – 2023

ÖZ

DENGESİZ VERİLERE SINIFLANDIRMA ALGOTİRMALARI UYGULANIRKEN KULLANILAN YÖNTEMLERİN KARŞILAŞTIRILMASINA AİT BİR ANALİZ

KUTLUK ATALAY

Makine öğreniminin öğrenme türlerinden biri olan denetimli öğrenmede yer alan sınıflandırma algoritmalarıyla kurulan modellerin tahminler gerçekleştirdiği uygulamalarda, veri setlerinin hedef özniteliğinde yer alan sınıfların dağılımının, birbirlerine eşit veya yakın olduğu varsayılmaktadır. Gerçek hayat uygulamalarında ise, bu varsayımın çoğu zaman sağlanmadığı görülmektedir. Varsayımın sağlanmadığı durumlarda veri seti, dengesiz veri seti olarak nitelendirilmekte olsa da dengesiz veri seti problemi çoğunlukla, hedef özniteliğinin sınıf dağılımları arasında anlamlı farklılık bulunan veri setlerinde yaşanan bir problem olarak tanımlanmaktadır.

Dengesiz veri seti problemi, makine öğreniminde, sınıflandırma algoritmaları ile kurulan modellerin eğitim aşamalarında olumsuz bir faktör olarak modelin eğitimini etkilemekte ve modelin test aşamasında gerçekleştirdiği tahmin etme işleminin performansını düşürmektedir. Olumsuz faktör olarak modelin performansını etkilemeleri nedeniyle, dengesiz veri setleri problemi ile sınıflandırma uygulamalarında çözülmesi gereken önemli bir sorun olarak karşılaşılmaktadır. Dengesiz veri seti probleminin temel olarak iki farklı çözüm yaklaşımı vardır. Bu çözüm yaklaşımları; veri seti seviyesinde çözüm yaklaşımları ve maliyet duyarlı çözüm yaklaşımları olarak sıralanmaktadır.

Bu çalışmada araç sigorta poliçelerine ait bilgileri içeren veri seti, beyin felci hastalarının bilgilerini içeren veri seti ve felçli askeri gaziler için bağış yapan donörlere ait bilgileri içeren veri seti olmak üzere, 3 farklı veri seti kullanılmıştır. Her bir veri setinin detaylı analizi sonucunda, dengesiz veri seti probleminin varlığı tespit edilmiş ve dengesiz veri seti problemini çözmek için veri seti seviyesinde çözüm yaklaşımları içinde yer alan farklı yaklaşımlar uygulanmıştır. Uygulanan farklı yaklaşımlardan

sonra sınıflandırma modellerine ait performans metrikleri ve karmaşıklık matrisleri incelenerek yorumlanmıştır. Dengesiz veri seti problemini çözme yaklaşımlarından sonra oluşturulup uygulanan sınıflandırma modellerinin başarılı performans metriklerine sahip oldukları görülmüştür.

Anahtar Kelimeler: (Makine Öğrenimi, Denetimli Öğrenme, Sınıflandırma Algoritmaları, Dengesiz Veri Setleri, Performans Metrikleri)

ABSTRACT

AN ANALYSIS OF COMPARISON OF METHODS USED WHEN APPLYING CLASSIFICATION ALGORITHMS TO IMBALANCED DATASETS

KUTLUK ATALAY

In applications where models established with classification algorithms in supervised learning, one of the learning types of machine learning that make predictions, it is assumed that the distribution of the classes in the target features of the datasets is equal or close to each other. In real life applications, however, this assumption is often not met. Although the dataset is described as an imbalanced dataset in cases where the assumption is not met, the imbalanced dataset problem is mostly defined as a problem experienced in datasets where there are significant differences between the class distributions of the target feature.

The imbalanced dataset problem affects the training of the model as a negative factor in the training stages of the models established with classification algorithms in machine learning and reduces the performance of the estimation process performed by the model in the testing phase. The problem of imbalanced datasets is encountered as an important problem to be solved in classification applications because they affect the performance of the model as a negative factor. There are basically two different solution approaches to the imbalanced dataset problem. These solution approaches are; are listed as dataset level solution approaches and cost sensitive solution approaches.

In this study, 3 different datasets were used: the dataset containing the information about the vehicle insurance policies, the dataset containing the information of the cerebral stroke patients, and the dataset containing the information of the donors who donated for the paralyzed military veterans. As a result of the detailed analysis of each dataset, the existence of the imbalanced dataset problem was determined and different approaches included in the dataset level solution approaches were applied to solve the imbalanced dataset problem. After different approaches applied, performance metrics and confusion matrices of classification models were

examined and interpreted. It has been seen that the classification models that are created and applied after the approaches to solving the imbalanced dataset problem have successful performance metrics.

Keywords: (Machine Learning, Supervised Learning, Classification Algorithms, Imbalanced Datasets, Performance Metrics)

ÖNSÖZ

Lisans eğitimim ve yüksek lisans eğitimimle birlikte yüksek lisans tez sürecimde de destek olup bilgileriyle yol gösteren, kıymetli fikirleriyle yeni bakış açıları kazandıran, desteğini her zaman sürdüren, çok değerli danışman hocam Prof. Dr. Emrah Önder'e,

Hem lisans hem de yüksek lisans akademik eğitim süreçlerimde benden desteğini ve güvenini hiçbir zaman esirgemeyen, her konuda bilgilerinden ve tecrübelerinden yararlandığım, çok kıymetli ve çok değerli hocam Prof. Dr. Umman Tuğba Gürsoy'a,

Tez sürecimde değerli katkılarını ve desteklerini sunan hocam Prof. Dr. Özgür Çakır'a,

Yüksek lisans eğitimim boyunca, her konuda yardımını ve desteğini hep sürdüren, tanıdığıma çok memnun olduğum çok değerli arkadaşım Arş. Gör. Berke Akkaya'ya,

Anaokulundan yüksek lisans mezuniyetime kadar her zaman eğitimimi çok önemseyen, bugüne kadar çok emek veren, fedakarlıklar yapan ve aynı zamanda ilk öğretmenim olan babam Yaşar Atalay'a, bugünlere gelmemde çok büyük fedakarlıklarda bulunan annem Afsaneh Atalay'a, sürekli durum değişikliği gösteren bir ilişkiye sahip olmamıza rağmen birbirimize sürekli destek olduğumuz kardeşim Esra Işılay Atalay'a ve tez aşamamda bana destekte bulunan çok değerli arkadaşlarım Yiğit Gökçehan Koçoğlu, Emel Köse ve Ömer Faruk Yıldırım'a teşekkürlerimi sunarım.

İSTANBUL, 2023

KUTLUK ATALAY

İÇİNDEKİLER

ÖZ……ii

ABSTRACT……iv

ÖNSÖZ……vi

TABLOLAR LİSTESİ……x

GRAFİKLER LİSTESİ……...xi

KISALTMALAR LİSTESİ……xiii

GİRİŞ…….....1

BİRİNCİ BÖLÜM

MAKİNE ÖĞRENİMİ

1.1. Makine Öğrenimi Kavramı....2
1.2. Makine Öğreniminin Türleri....3
1.2.1. Denetimli Öğrenme....3
1.2.1.1. Doğrusal Regresyon....4
1.2.1.2. K-En Yakın Komşuluk....5
1.2.1.3. Destek Vektör Makineleri....7
1.2.1.4. Logistik Regresyon....10
1.2.1.5. Karar Ağaçları....12
1.2.1.6. Rassal Orman....15
1.2.1.7. Gradyan Artırılmış Ağaçlar....17
1.2.1.8. Ekstrem Gradyan Artırılmış Ağaçlar....18
1.2.1.9. Naive Bayes....18
1.2.1.10. Yapay Sinir Ağları....19
1.2.1.11. Denetimli Öğrenme Süreçleri....20
1.2.1.12. Denetimli Öğrenmede Performans Değerlendirme Ölçütleri....21
1.2.1.13. Hiperparametre Optimizasyonu....23
1.2.2. Denetimsiz Öğrenme....24

1.2.2.1. K-Ortalamalar 24
1.2.2.2. Hiyerarşik Kümeleme 27
1.2.2.3. Gürültülü Uygulamaların Yoğunluk Bazlı Mekânsal Kümelenmesi 28
1.2.2.4. Temel Bileşenler Analizi 29
1.2.2.5. Doğrusal Ayırma Analizi 30
1.2.2.6. Apriori 31
1.2.3. Pekiştirmeli Öğrenme 32
1.3. Makine Öğreniminin Kullanım Alanları 33

İKİNCİ BÖLÜM
DENGESİZ VERİ SETLERİ PROBLEMİ

2.1. Dengesiz Veri Setleri 42
2.2. Dengesiz Veri Seti Problemini Çözüm Yöntemleri 43
2.2.1. Veri Seti Seviyesinde Çözüm Yöntemleri 43
2.2.1.1. Alt Örnekleme Yöntemleri 44
2.2.1.1.1. Rassal Alt Örnekleme 45
2.2.1.1.2. Tomek Bağlantıları Yöntemi 45
2.2.1.1.3. Düzenlenmiş En Yakın Komşular Kuralı 45
2.2.1.1.4. Yoğunlaştırılmış En Yakın Komşu Kuralı 45
2.2.1.1.5. Tek Taraflı Seçim 46
2.2.1.1.6. Bölge Temizleme Kuralı 46
2.2.1.2. Aşırı Örnekleme Yöntemleri 47
2.2.1.2.1. Rassal Aşırı Örnekleme 47
2.2.1.2.2. Sentetik Azınlık Aşırı Örnekleme 47
2.2.1.2.3. Uyarlanabilir Sentetik Örnekleme 48
2.2.2. Maliyet Duyarlı Çözüm Yöntemleri 48
2.3. Dengesiz Veri Setlerinde Performans Değerlendirme Ölçütleri 49

2.4. Literatür Taraması 51

ÜÇÜNCÜ BÖLÜM
UYGULAMA

3.1. Uygulamaların Kapsamları ve Amaçları 54
3.1.1. Porto Seguro Sigorta Şirketi Verisine ait Dengesiz Veri Seti Uygulaması 54
3.1.2. Beyin Felci Hastaları Verisine ait Dengesiz Veri Seti Uygulaması 63
3.1.3. Bağış Yapan Donörlerin Verisine ait Dengesiz Veri Seti Uygulaması 70
SONUÇ 77
KAYNAKÇA 79

TABLOLAR LİSTESİ

Tablo 3.1.: Porto Seguro Sigorta Şirketi Veri Setinde Dengesiz Veri Seti Problemini Çözme Yaklaşımları Uygulandıktan Sonra Oluşturulan Modellerin F-Skoru Performans Metriklerinin Değerleri 58

Tablo 3.2.: Bölge Temizleme Kuralı Uygulanmasından Sonra Karar Ağaçları Modeline ait Performans Metrikleri 60

Tablo 3.3.: Bölge Temizleme Kuralı Uygulanmasından Sonra Rassal Orman Modeline ait Performans Metrikleri 62

Tablo 3.4.: Beyin Felci Hastaları Veri Setinde, Dengesiz Veri Seti Problemini Çözme Yaklaşımları Uygulandıktan Sonra Oluşturulan Modellerin F-Skoru Performans Metrikleri.....65

Tablo 3.5.: Bölge Temizleme Kuralı Uygulanmasından Sonra Karar Ağaçları Modeline ait Performans Metrikleri 67

Tablo 3.6.: Bölge Temizleme Kuralı Uygulanmasından Sonra Rassal Orman Modeline ait Performans Metrikleri 69

Tablo 3.7.: Bağış Yapan Donörler Veri Setinde, Dengesiz Veri Seti Problemini Çözme Yaklaşımları Uygulandıktan Sonra Oluşturulan Modellerin F-Skoru Performans Metrikleri.....73

Tablo 3.8.: Rassal Aşırı Örnekleme Yaklaşımının Uygulamasından Sonra Gradyan Artırılmış Ağaçlar Modeline ait Performans Metrikleri 75

GRAFİKLER LİSTESİ

Grafik 1.1.: Makine Öğreniminin Türleri 3

Grafik 1.2.: Doğrusal Regresyon Modeli 4

Grafik 1.3.: K-En Yakın Komşuluk Modeli ile Sınıflandırma 6

Grafik 1.4.: Doğrusal Özelliklere Sahip Veri Setlerinde Destek Vektör Makineleri ile Sınıflandırma 8

Grafik 1.5.: Doğrusal Özelliklere Sahip Olmayan Veri Setlerinde Destek Vektör Makineleri ile Sınıflandırma 9

Grafik 1.6.: Logistik Regresyon Sınıflandırıcısının Sigmoid Fonksiyonu 12

Grafik 1.7.: Rassal Orman Sınıflandırma Modeline ait Öznitelik Önemliliği Grafiği 16

Grafik 1.8.: Karmaşıklık Matrisi 21

Grafik 1.9.: Eğrinin Altında Kalan Alan Grafiği 22

Grafik 1.10.: K-Ortalamalar Modeli ile Kümeleme 27

Grafik 1.11.: Hiyerarşik Kümeleme Modeli ile Kümeleme 28

Grafik 1.12.: Gürültülü Uygulamaların Yoğunluk Bazlı Mekânsal Kümelenmesi Modeli ile Kümeleme 29

Grafik 1.13.: Doğrusal Ayırma Analizi 31

Grafik 2.1.: Alt Örnekleme ve Aşırı Örnekleme Yöntemlerinin Çalışma Şekli 43

Grafik 3.1.: Rassal Örnekleme İşlemi Uygulanmasından Önceki Durumda Veri Setinin Hedef Özniteliğinin Sınıf Dağılımı 56

Grafik 3.2.: Rassal Örnekleme İşlemi Uygulanmasından Sonraki Durumda Veri Setinin Hedef Özniteliğinin Sınıf Dağılımı 56

Grafik 3.3.: Bölge Temizleme Kuralı Uygulanmasından Sonra Karar Ağaçları Modeline ait Karmaşıklık Matrisi 61

Grafik 3.4.: Bölge Temizleme Kuralı Uygulanmasından Sonra Rassal Orman Modeline ait Karmaşıklık Matrisi 62

Grafik 3.5.: Beyin Felci Veri Setinin Hedef Özniteliğinin Sınıf Dağılımı 64

Grafik 3.6.: Bölge Temizleme Kuralı Uygulanmasından Sonra Karar Ağaçları Modeline ait Karmaşıklık Matrisi 67

Grafik 3.7.: Bölge Temizleme Kuralı Uygulanmasından Sonra Rassal Orman Modeline ait Karmaşıklık Matrisi 69

Grafik 3.8.: Bağış Yapan Donörler Veri Setinin Hedef Özniteliğinin Sınıf Dağılımı 71

Grafik 3.9.: Rassal Aşırı Örnekleme Yaklaşımının Uygulanmasından Sonra Gradyan Artırılmış Ağaçlar Modeline ait Karmaşıklık Matrisi 76

KISALTMALAR LİSTESİ

IG	: Information Gain
CART	: Classification and Regression Tree
ID3	: Iterative Dichotomiser 3
CHAID	: Chi-square Automatic Interaction Detection
QUEST	: Quaternion Estimator Algorithm
CRUISE	: Classification Rule with Unbiased Interaction Selection and Estimation
GUIDE	: Generalized, Unbiased, Interaction Detection and Estimation.
CTREE	: Conditional Inference Trees
SARS-CoV-2	: Severe Acute Respiratory Syndrome Coronavirus 2
COVID-19	: Coronavirus Disease 2019
CatBoost	: Category Boosting
ACO	: Ant Colony Optimization
CBEUS	: Cluster-Based Evolutionary Undersampling
IRUS	: Inverse Random Under Sampling
DSUS	: Diversified Sensitivity-based Undersampling
SMOTE	: Synthetic Minority Oversampling Technique
ADOMS	: Adaptive Synthetic Over-sampling Using Minority Search
ADASYN	: Adaptive Synthetic Sampling Approach
ROSE	: Random Oversampling Examples
DBSMOTE	: Density-Based Synthetic Minority Over-sampling
MWMOTE	: Majority Weighted Minority Oversampling Technique
BKS	: Borderline Kernel Based Over-sampling
LLE-SMOTE	: Locally Linear Embedding Synthetic Minority Oversampling Technique
CBSO	: Cluster-based Synthetic Oversampling

DSRBF : Dynamic Smote Radial Fasis Function

LVQ-SMOTE : Learning Vector Quantization based Synthetic Minority Over–sampling Technique

PDFOS : Probability Density Function Estimation Based Over-sampling

OUPS : Over-sampling Using Propensity Scores

NEATER : Non-cooperative Game Theory

ERS-1 : European Remote-Sensing Satellite-1

GİRİŞ

Günümüzde veri kavramının önemi gün geçtikçe artmaktadır. Veri bilimi alanında alt dallardan biri olan makine öğrenimi çeşitli sektörlerde kullanım alanını genişletmektedir. Veri setlerine çeşitli makine öğrenimi modelleri uygulanarak, farklı öngörüler ve çıkarımlar elde edilmekte ve bu modellerin performansları farklı ölçütlerle değerlendirilmektedir. Her ne kadar makine öğrenimi modelleri günlük hayatta meydana gelen sorunlara çözümler bulmak için kullanılsa da performanslarını etkileyecek farklı sorunlar da uygulamalarda ortaya çıkmaktadır.

Makine öğrenimi modellerinin uygulandığı veri setlerinin, modelin uygun şekilde çalışması için her zaman elverişli olmadığı durumlarla da güncel hayatta karşılaşılmaktadır. Bu durumlardan biri de dengesiz veri seti problemidir. Dengesiz veri seti probleminin, sınıflandırma modelleri uygulanırken ortaya çıktığı görülmektedir. Sınıflandırma modellerinin uygulanacağı veri setlerinde, hedef özniteliğindeki gözlemlerin çok yüksek oranda sadece bir sınıfa ait olması, dengesiz veri seti problemine sebep oluşturmaktadır. Bu tip veri setlerinde sınıflandırma algoritmaları ile oluşturulan makine öğrenimi modelleri, gerçekleştirdikleri tahminlerde, çoğunluğa sahip sınıfa daha çok yakınlık göstermektedir. Sınıflara ait etiket bilgisinin yeterli seviyede olmadığı veri setlerinde model yeterince eğitilemediği için modelin azınlık sınıfa ait gözlemleri tahmin performansı, oldukça düşük olarak ortaya çıkmaktadır.

Bu çalışmada farklı sınıflandırma algoritmaları ile oluşturulmuş makine öğrenimi modellerinin performans değerleri ile dengesiz veri setlerinde çözüm yaklaşımlarından olan alt örnekleme ve aşırı örnekleme yöntemleri uygulanıp sonrasında oluşturulan modellerin performans değerleri karşılaştırılmış ve yorumlanmıştır. Çalışmada Kaggle isimli veri bilimi platformundan elde edilen, Porto Seguro's Safe Driver Prediction veri seti, beyin felci hastalarının bilgilerini içeren veri seti ve felçli askeri gaziler için bir bağış toplama talebinde, bağış yapan donörlerin bilgilerini içeren veri seti kullanılmıştır. Her bir veri setinin, sınıf dağılımlarının incelenmesi sonucu, dengesiz veri seti olarak nitelendirilebileceği görülmüştür.

BİRİNCİ BÖLÜM

MAKİNE ÖĞRENİMİ

1.1 Makine Öğrenimi Kavramı

Makine öğrenimi, bilgisayarları verilerden içgörüler kazanarak öğrenim gerçekleştirebilecekleri şekilde programlama bilimidir (Géron, 2019). Makine öğrenimi ayrıca edindiği deneyimlerle kendini geliştiren ve bir öğrenme işlemi gerçekleştiren bilgisayar sistemleri oluşturma süreci olarak tanımlanabilir (Ayodele, 2010). Makine öğrenimi geçmişte kaydedilmiş gözlemlerden öğrenerek geleceğe yönelik doğru tahminler gerçekleştirmeyi sağlayan teknikler bütünüdür (Schapire, 2003).

Makine öğrenimi modelleri ile klasik algoritmalar arasındaki en önemli fark, makine öğrenimi modellerinin verilerden insan yardımı olmadan öğrenebilmesidir. Klasik algoritmalar, bilgisayarla programlama işlemlerinde yazılımcıların doğrudan yönlendirmesi ile karmaşık sistemlerde en iyi çözümü bulmayı hedeflerken, makine öğrenimi modellerinde, modele en iyi çözüm yolu gösterilmemektedir. Bunun yerine çok sayıda benzer problem modele tanıtılıp, modelin kendiliğinden en iyi çözümü bulması beklenmektedir (Özdemir, 2016).

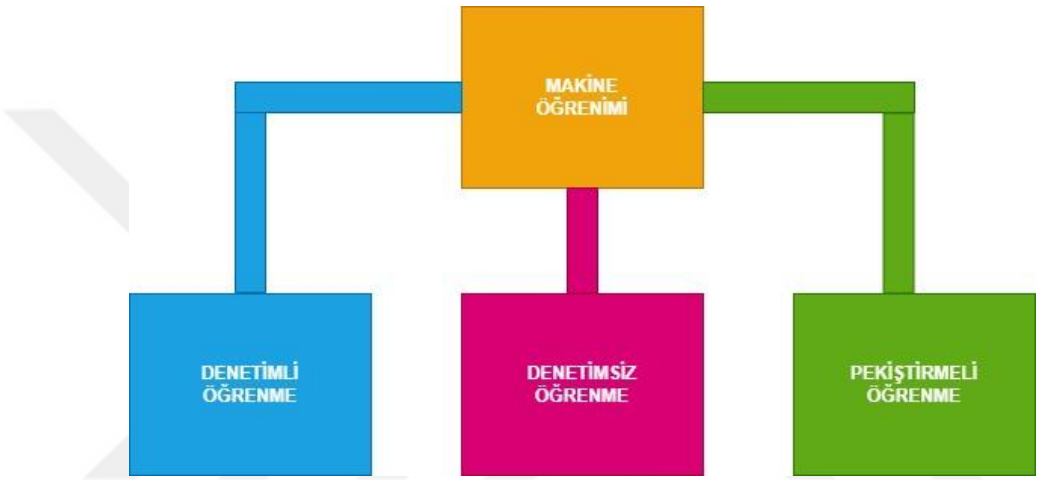
Makine öğrenimi alanında yapılan çalışmaların sayısının artmasıyla birlikte, hesaplanabilir öğrenme teorisi, sinir ağları, istatistik ve örüntü tanıma gibi dallar arasında ilişkiler kurularak bu dalların birbirleriyle ortak olarak çalışması sağlanmıştır. Makine öğrenimi modelleri günümüzde yüz tanıma, el yazısı tanıma, sağlıkta veri bilimi gibi yeni alanların yanı sıra veri tabanlarında bilgi keşfi, dil işleme, robotik gibi birçok farklı alanda da yeni problemleri çözmek için uygulanmaktadır (Dietterich, 1997).

Veri biliminin alt dallarından biri olan makine öğreniminin temelinde denetimli öğrenme, denetimsiz öğrenme ve pekiştirmeli öğrenme olarak 3 farklı öğrenme türü yer almaktadır.

1.2 Makine Öğreniminin Türleri

Makine öğrenimi modelleri, çalışma şekilleri ve elde edilmek istenen sonuca göre temelde 3 farklı türde öğrenme modeli olarak sıralanmaktadır. Bu modeller sırasıyla denetimli öğrenme, denetimsiz öğrenme ve pekiştirmeli öğrenme olarak adlandırılmaktadır.

Grafik 1.1.: Makine Öğreniminin Türleri



1.2.1 Denetimli Öğrenme

Denetimli öğrenmede, model eğitim setinden öğrenmektedir. Eğitim setinin öznitelikleri X iken, hedef öznitelik Y olarak adlandırılmaktadır. Y özniteliği etiket bilgisini taşıyan özniteliktir. Denetimli öğrenmede modelin çalışma amacı, yeni gözlemlerin hedef öznitelik etiketlerine yönelik tahminler geliştirmektir. Hedef öznitelik etiket yapısı niteliğiyle sürekli veya kategorik olabilir. Denetimli öğrenme; regresyon ve sınıflandırma olarak iki alt başlığa ayrılmıştır. Hedef özniteliğin sürekli olduğu problemlerde, modelin amacı regresyon ile tahminler gerçekleştirmek iken, kategorik olduğu problemlerde model, gözlemin dahil olacağı sınıfa yönelik sınıflandırma algoritmaları ile tahminler gerçekleştirir (Önder, 2020).

Regresyon: Hedef özniteliğin sürekli olması durumunda uygulanır.

Sınıflandırma: Hedef özniteliğin kategorik olması durumunda uygulanır.

Denetimli öğrenme modelleri oluşturulurken en çok kullanılan algoritmalar; Doğrusal Regresyon, K-En Yakın Komşuluk (K-NN), Destek Vektör Makineleri, Logistik Regresyon, Karar Ağaçları, Rassal Orman, Gradyan Artırılmış Ağaçlar, Naive Bayes ve Yapay Sinir Ağları olarak sıralanmaktadır.

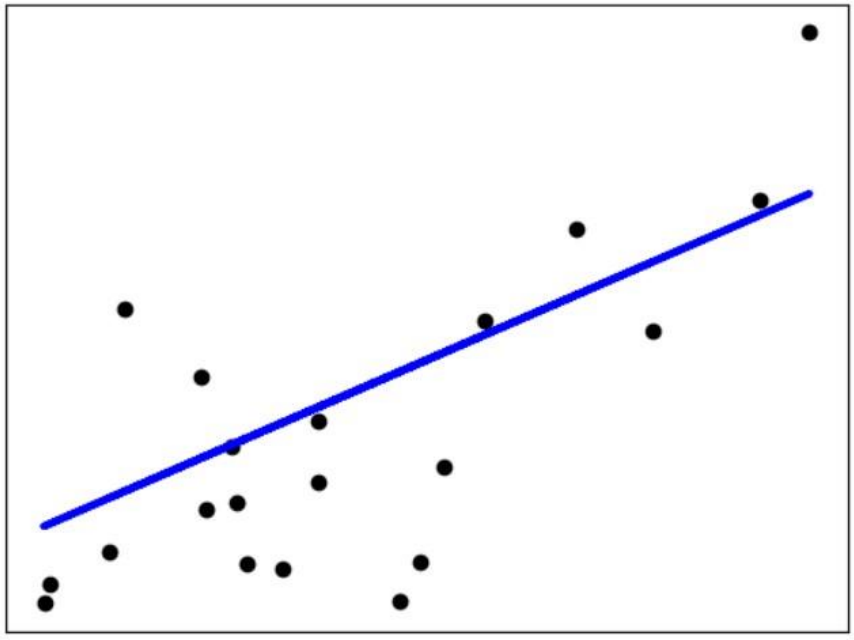
1.2.1.1 Doğrusal Regresyon

Doğrusal regresyon analizinin tarihçesi, 18. yüzyıla kadar uzanmaktadır. İstatistikte, bilgisayar destekli model kurulumuna başlanılmasından çok daha uzun zaman öncelerinde bile doğrusal modellerin yaygın şekilde kullanıldığına denk gelinmektedir. Bu kadar yaygın kullanılmalarına genel olarak, basit şekilde çalışmaları ve modelin girdilerinin, çıktılarını nasıl etkilediklerine dair bilimsel açıdan yeterli ve yorumlanabilir açıklamalar neden olmaktadır (Hastie, Tibshirani, & Friedman, 2017).

Doğrusal regresyon analizi, iki veya daha fazla bağımsız öznitelik ile hedef öznitelik arasında doğrusal ilişki olduğunda, bağımsız öznitelikler ile model oluşturarak, hedef özniteliğinin tahmin edilmesini sağlamaktadır (Siegel, 2016).

Grafik 1.2.'de doğrusal regresyon modeli görülmektedir (Scikit-Learn, 2023).

Grafik 1.2.: Doğrusal Regresyon Modeli



Doğrusal regresyon analizi, hedef öznitelik ile bağımsız öznitelikler arasında ilişki kurarak, veri setine eklenecek olan yeni gözlemlerin hedef özniteliğinin değerlerini tahmin etmek için kullanılır. Basit doğrusal regresyon modelinde, veri setine yeni eklenecek olan gözlemlerin hedef özniteliklerinin değeri, tek bir bağımsız öznitelik kullanılarak oluşturulan model ile tahmin edilmektedir.

$$y_i = a + \beta x_i + \varepsilon_i$$

Yukarıda basit doğrusal regresyon modelinin denkleminde yer alan, a regresyon doğrusunun sabit değerini ifade ederken, ß regresyon doğrusunun eğimini ifade etmektedir. ß değeri 0'dan büyükse doğru pozitif eğime sahipken, 0'dan küçükse doğru negatif eğime sahip olmaktadır. Basit doğrusal regresyon uygulanırken istenen durum, varyasyonu en düşük seviyeye indirmektir. Günümüzde basit doğrusal regresyon modellerinin yanında Lasso regresyon ve ridge regresyon da çeşitli uygulamalarda kullanılmaktadır. Lasso regresyonda amaç hata kareleri toplamını en düşük değere indirmek olurken, ridge regresyon, kendi aralarında yüksek korelasyon değerlerine sahip bağımsız öznitelikler ile bağımlı öznitelik arasındaki ilişkiyi analiz etmektedir (Önder, 2020).

Çoklu doğrusal regresyon modeli, basit doğrusal regresyon modelinden farklı olarak, birden fazla bağımsız öznitelik kullanılarak hedef özniteliğinin değerini tahmin etmek için kullanılmaktadır.

$$y_i = a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k_i} x_{k_i} + u_i$$

Yukarıda çoklu doğrusal regresyon modelinin denkleminde yer alan, a regresyon doğrusunun sabit değerini ifade ederken, her ß değeri, karşılık gelen özniteliklerin katsayısını ifade etmektedir. ß değerleri k sayıda, bağımsız özniteliklerin katsayısı olarak formülde karşılık bulmaktadır. Son olarak u değeri modelin hata değerini ifade etmektedir (Favero & Belfiore, 2019).

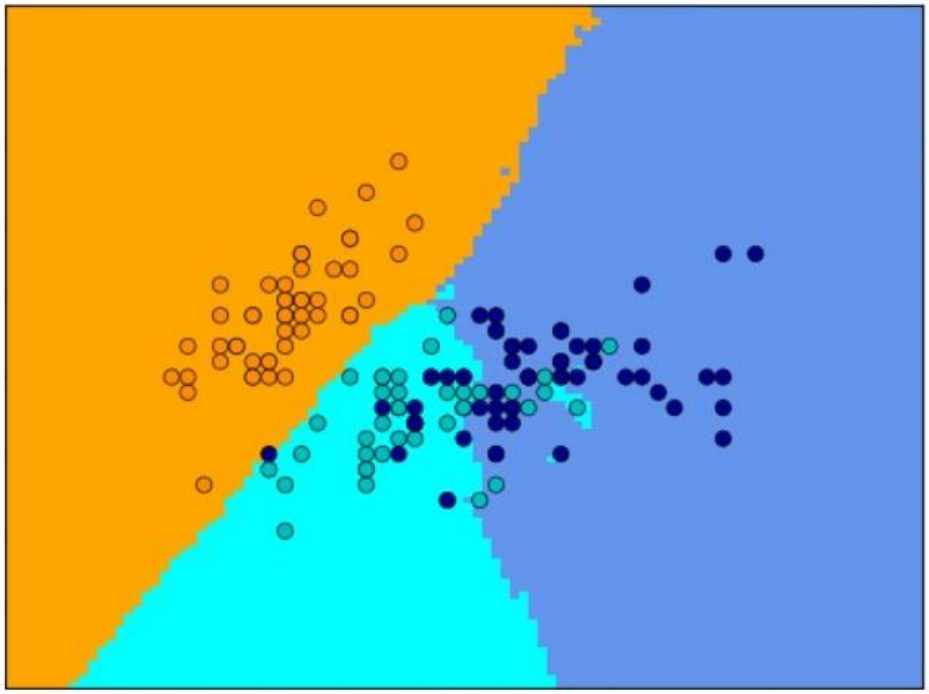
1.2.1.2 K-En Yakın Komşuluk

Denetimli öğrenmede, sınıflandırma algoritmalarından biri olan K-En Yakın Komşuluk (K-NN) algoritması, parametrik olmayan sınıflandırma problemlerinde kullanılan temel algoritmalardan biridir (Fix & Hodges, 1951).

K-En Yakın Komşuluk algoritması her bir gözlemin sınıfını, o gözlemin diğer gözlemlerle arasındaki benzerliklere dayalı olarak tahmin etmektedir. K-En Yakın Komşuluk algoritması oylama sistemi bazlı çalışmaktadır. Bu çalışma şeklinin temelinde, veri setinde sınıfı tahmin edilecek yeni gözlem değeri için *k* sayıda benzer veya en yakın gözlemler seçilip, seçilen *k* sayıda gözlemlerin ait oldukları sınıflar içinde nicelik olarak çoğunluk sınıfın, yeni gözlemin sınıfını oluşturması yer almaktadır (Boehmke & Greenwell, 2020).

K-En Yakın Komşuluk algoritması ile kurulmuş olan sınıflandırma modelinin uygulaması Grafik 1.3.'de görülmektedir (Scikit-Learn, 2023).

Grafik 1.3.: K-En Yakın Komşuluk Modeli ile Sınıflandırma



K-En Yakın Komşuluk algoritmasında gözlemlerin benzerliği, gözlemlerin ikili olarak birbirlerine olan uzaklıklarından yola çıkarak hesaplanmaktadır. Uygulamalarda uzaklık hesaplamalarına yönelik en çok kullanılan mesafe ölçüleri; Öklid, Manhattan, Minkowski ve Canberra olarak öne çıkmaktadır.

Canberra uzaklık ölçüsü Manhattan uzaklık ölçüsünün ağırlıklı şekilde hesaplanmasıyla ölçülürken, Öklid ve Manhattan uzaklıkları, Minkowski uzaklık ölçüsünün özel varyasyonu şeklinde formüle edilmektedir (Önder, 2020).

K-En Yakın Komşuluk sınıflandırma algoritmasının uygulanmasında kullanılan Öklid, Manhattan ve Minkowski uzaklık ölçülerine ait formüller aşağıdaki denklemlerde görülmektedir.

$$\text{Öklid Uzaklığı}: d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$\text{Manhattan Uzaklığı}: \mathrm{d(x,y)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} |\mathrm{x_i} - \mathrm{y_i}|$$

$$\text{Minkowski Uzaklığı}: \mathrm{d(x,y)} = \left(\sum_{\mathrm{i=1}}^{\mathrm{n}} |\mathrm{x_i} - \mathrm{y_i}|^{\mathrm{c}}\right)^{\frac{1}{\mathrm{c}}}$$

K-En Yakın Komşuluk sınıflandırma algoritmasının kullanımının sağladığı avantajlardan bazıları şu şekilde sıralanmaktadır; modelin eğitiminin hızlı, öğrenmesinin kolay olması, veride gürültü olması durumunda, gürültü azaltma teknikleri ile modelin doğruluk ve diğer performans metriklerini olumlu yönde etkilemesi ve parametrik olmaması. Bunların yanında, K-En Yakın Komşuluk sınıflandırma algoritmasının kullanımının sağladığı dezavantajlara, eğitim setinin boyutu büyükse modelin yavaş çalışması ve ilgisiz özniteliklere sahip veri setlerinde duyarlı olması örnek verilmektedir (Cunningham & Delany, 2022).

1.2.1.3 Destek Vektör Makineleri

Destek vektör makineleri, Rus matematikçiler Vladimir Vapnik ve Alexey Chervonenkis tarafından 1960'larda geliştirilen istatistiksel öğrenme teorisine dayalı olarak geliştirilmiştir. Vapnik-Chervonenkis teorisi, öğrenme teorisini istatistiksel olarak açıklamaya çalışmıştır. İstatistiksel öğrenme teorisi olarak adlandırılan bu model, 1960'lardan 1990'lara kadar kullanılmıştır. 1990'larda istatistiksel öğrenme

teorisine dayalı olarak yeni bir öğrenme algoritması olan destek vektör makineleri açıklanmıştır. Destek vektör makineleri, doğrusal olmayan sınıflandırma problemleri, fonksiyon tahmini gibi alanlarda yaygın kullanılan, güçlü bir algoritmadır (Tolun, 2008).

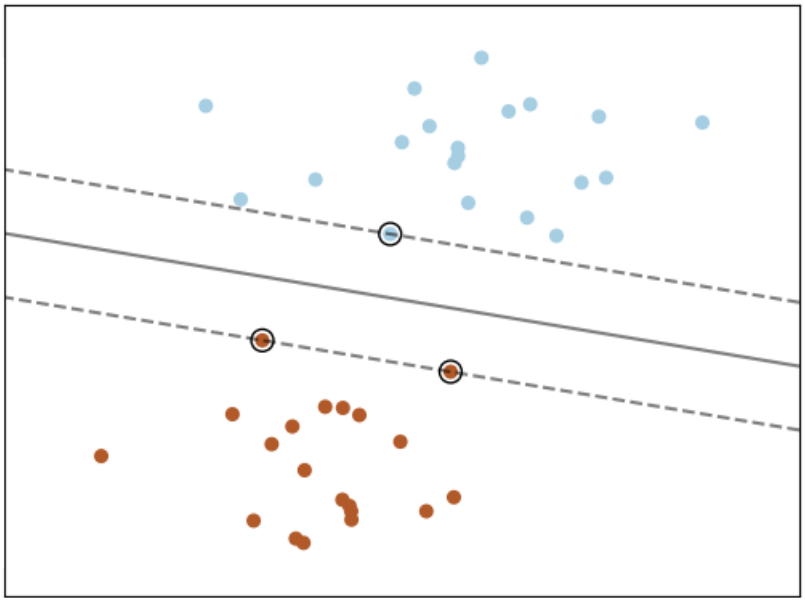
Destek vektör makineleri, veri setinde sınıflandırılacak öznitelikteki gözlemlerin içinde karar sınırları veya hiperdüzlem olarak isimlendirilen sınırlar belirleyip, bu hiperdüzlem ve karar sınırlarının yardımıyla sınıflandırma işlemi gerçekleştirmektedir.

Destek vektör makineleri temel çalışma prensibi olarak, hedef özniteliğindeki iki sınıfı öznitelik uzayında en iyi ayıracak olan hiperdüzlemi bulmak şeklinde çalışmaktadır. Gerçek hayat uygulamalarında öznitelikleri en iyi şekilde ayıracak hiperdüzlemi oluşturma işlemi çok zor olmasına rağmen, destek vektör makineleri bu sorunu; "en iyi ayırma" kavramının tanımını yumuşatarak ve doğrusal olmayan sınıflandırma problemlerinde de "çekirdek hilesi" ismi verilen yöntemi kullanıp doğrusal olmayan sınıfları, doğrusal hâle getirerek çözmektedir (Boehmke & Greenwell, 2020).

Veri setlerinde sınıflandırma problemi için destek vektör makineleri kullanılırken, doğrusal olmayan özelliklere sahip veri setlerinin olduğu durumlarla da karşılaşılabilmektedir. Bu veri setlerinde destek vektör makineleri ile sınıflandırma işlemi gerçekleştirilirken, çekirdek fonksiyonu kullanılarak, çok boyutlu alanda, doğrusal ayırmaya yardımcı olacak bir hiperdüzlem oluşturulmaktadır. Bu hiperdüzlem oluşturuluncaya kadar çekirdek hilesi, tekrarlayan bir şekilde daha yüksek boyutta eşleştirme işlemi gerçekleştirmektedir (Önder, 2020).

Doğrusal özelliklere sahip veri setlerinde, destek vektör makineleri ile kurulmuş olan sınıflandırma modelinin uygulaması Grafik 1.4.'de görülmektedir (Scikit-Learn, 2023).

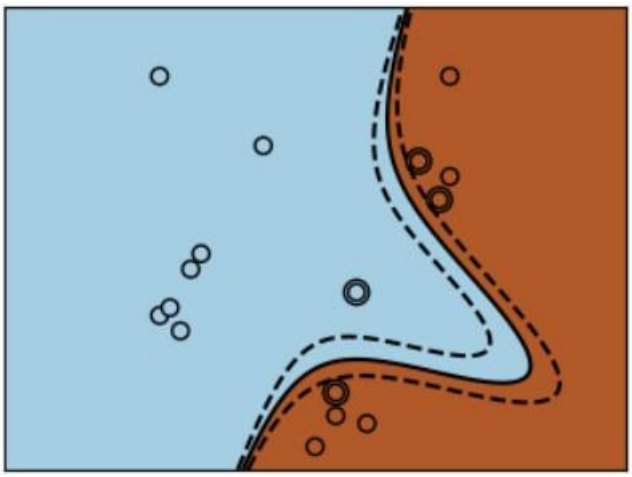
Grafik 1.4.: Doğrusal Özelliklere Sahip Veri Setlerinde Destek Vektör Makineleri ile Sınıflandırma



Grafik 1.4.'te destek vektör makinelerinin, doğrusal özelliklere sahip veri setlerinde oluşturulan sınıflandırma modellerindeki çalışma şekli görülmektedir. Mavi ve kahverengi noktalar, veri setinin hedef özniteliğindeki sınıflara dahil olan gözlemleri temsil etmektedir. Veri setinde yer alan gözlemlerin, iki veya daha fazla sınıfa göre sınıflandırılması için karar sınırı, sınıflara ait gözlemlerin aralarındaki en uygun boşluktan geçerek hiperdüzlemi oluşturmaktadır. Karar sınırına en yakın olan gözlemlere, destek ismi verilmektedir.

Doğrusal özelliklere sahip olmayan veri setlerinde, destek vektör makineleri ile kurulmuş olan sınıflandırma modelinin uygulaması Grafik 1.5.'te görülmektedir (Scikit-Learn, 2023).

Grafik 1.5.: Doğrusal Özelliklere Sahip Olmayan Veri Setlerinde Destek Vektör Makineleri ile Sınıflandırma



Destek vektör makineleri, makine öğreniminde hem regresyon hem de sınıflandırma algoritmaları için kullanılmaktadır. Destek vektör makinelerinin en çok kullanıldığı ve en iyi performans gösterdiği uygulama alanlarının başında örüntü tanıma gelmektedir. Örüntü tanımaya ek olarak, biyoenformatik alanında gen tanım verilerini sınıflandırarak hastalıkların teşhisinde, metin madenciliği ve belgelerin isimlerine göre sınıflandırılmalarında, güvenlik protokolleri ihlali ve doğal afetler gibi nadir gerçekleşen olayların tespitinde destek vektör makinelerinden yararlanılmaktadır (Lantz, 2019).

Destek vektör makinelerinin sağladığı avantajlara, modelin doğruluk ölçütünün yüksek olması, ezberleme sorunun çok düşük sıklıkta yaşanması örnek gösterilmekteyken, dezavantajları olarak çekirdek fonksiyonunun doğru seçimine duyarlı olması, modelin eğitiminin uzun süre tutması, farklı tiplerde verilere sahip veri setlerinde sorun yaşaması örnek gösterilmektedir (Önder, 2020).

1.2.1.4 Logistik Regresyon

Makine öğreniminde sınıflandırma algoritmalarından biri olan logistik regresyon 1944 yılında Joseph Berkson tarafından geliştirilmiştir.

Doğrusal regresyonun aksine hedef özniteliğinin kategorik olduğu sınıflandırma problemlerinde kullanılmaktadır. Logistik regresyonun kullanıldığı sınıflandırma problemlerinde hedef öznitelik, doğrusal regresyonun kullanıldığı sınıflandırma problemlerinde hedef özniteliğin sahip olduğu şekilde sayısal değerler yerine, “evet” ve “hayır” gibi kategorik değerlere sahip olmaktadır (Hoffman, 2019).

Logistik regresyonun varsayımları, hedef öznitelik ile bağımsız öznitelikler arasında doğrusal ilişkinin olmaması, hedef özniteliğin iki kategorik sınıfa sahip olması, bağımsız özniteliklerin normal dağılıma ve eşit varyansa sahip olmasına gerek olmaması, hedef özniteliğin sahip olduğu kategorik sınıfların birbirlerinden tamamen ayrık olması yani bir gözlemin sadece bir sınıfa ait olması ve en yüksek olasılık katsayılarının tahmin edilmesi için çok fazla gözlem gerektiğinden, model için doğrusal regresyon modellerinde gerek görülenden daha fazla sayıda gözleme gerek görülmesi şeklinde sıralanmaktadır (Balaban & Kartal, 2018).

Logistik regresyonun oluşturmuş olduğu sigmoid fonksiyonuna ait denklem aşağıdaki şekilde formülde edilmektedir.

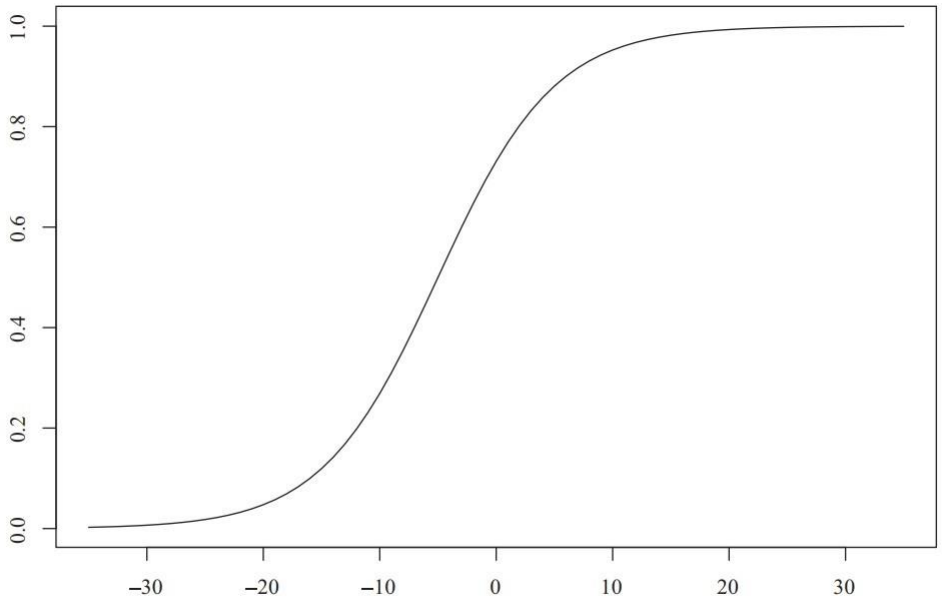
$$P(x) = \frac{1}{1 + e^{-x}}$$

Logistik regresyon sigmoid fonksiyonu oluşturup veri setindeki gözlemler için sınıflandırma işlemini, olasılıklara dayalı olarak gerçekleştirmektedir. Sigmoid fonksiyonunun çıktısı, olasılık değerleri olarak 0 ile 1 arasında değer almaktadır. Olasılık değerleri 0 ile 1 arasında hesaplanırken, orta nokta olan 0.5 değeri, sınıflandırma için eşik değeri olarak yer almaktadır. Veri setinde yer alan her bir gözlemin sınıflandırılması, sigmoid fonksiyonunun çıktı değeri ile gerçekleşmektedir. Bir gözlemin değeri girdi olarak sigmoid fonksiyonunda kullanıldığında, çıktı değeri 0.5'ten büyük ise, gözlemin sınıfı 1 olarak belirlenmektedir. Gözlemin değeri girdi olarak sigmoid fonksiyonunda kullanıldığında, çıktı değeri 0.5'ten küçük ise, gözlemin sınıfı 0 olarak belirlenmektedir (Ranganathan, Gribskov, Nakai, & Schönbach, 2019).

Logistik regresyonun sağladığı avantajlara; çıktılarının kolay yorumlanabilir olması, sınıf tahmini yaparken, gözlemlerin sınıflara dahil olma olasılıklarının hesaplanması örnek verilmektedir. Bunun yanında logistik regresyonun sahip olduğu bazı dezavantajlara ise; veri setinde eksik gözlem olması durumunda modelin performansının düşük kalması ve çoklu doğrusallık problemini önemsemesi örnek verilmektedir (Önder, 2020).

Logistik regresyon sınıflandırıcısının sigmoid fonksiyonu, Grafik 1.6.'da görülmektedir (Riffenburgh & Gillen, 2020).

Grafik 1.6.: Logistik Regresyon Sınıflandırıcısının Sigmoid Fonksiyonu



1.2.1.5 Karar Ağaçları

Makine öğreniminde ağaç temelli algoritmaların temeli olan karar ağaçları, 1963 yılında regresyon hesaplamaları için ağaç modeli olarak James N. Morgan ve John A. Sonquist tarafından geliştirilmiştir (Morgan & Sonquist, 1963).

Günümüzde hem regresyon hem de sınıflandırma problemlerinde en çok kullanılan algoritmalardan biri olan karar ağaçları, veri setlerindeki özniteliklerden, ağaç benzeri yapılar şeklinde karar verme kriterleri oluşturarak, veriyi görsel olarak kolay anlaşılabilir ve açıklanabilir kurallar setine dönüştürmektedir (Lewis, 2017).

Karar ağaçları algoritması, veri setini bir kök düğümünden, dallar sayesinde alt yapraklara kadar ayırarak, karar sürecini ifade edecek şekilde çalışmaktadır. Günlük hayatta gerçek bir ağaçta olduğu gibi karar ağaçları da bir kök düğümü temelinde başlayıp, yapraklara kadar yukarıdan aşağıya inen bir yapıya sahiptir. Bu yapı içinde kök düğümü, tüm veri setini temsil eden bir yapı olarak en tepede yer

almaktadır. Kök düğümün en az 2 farklı düğüme bölünmesiyle karar düğümleri (iç düğümler) oluşmaktadır. Karar düğümlerinin (iç düğümler) bölünme işleminin daha fazla bölünmenin mümkün olmayarak bitmesi durumunda, en son bölünmede elde edilmiş olan düğümler yaprak düğümü (terminal düğümü) olarak adlandırılmaktadır (Belyadi & Haghighat, 2021).

Karar ağaçlarında verilerin bölünmeye başlayacak kök düğümünün seçimi, en düşük sınıflandırma hatasına sahip olan ya da en yüksek bilgi kazanımı sağlayan öznitelikten başlamakta ve sınıflandırma hatası çok düşük seviyelere gelinceye kadar bu işlem devam etmektedir (Önder, 2020).

Çok fazla özniteliğe sahip veri setlerinde kök düğümü ve karar düğümlerinin seçimi için üç farklı yöntem öne çıkmaktadır. Bu yöntemler; entropi, bilgi kazanımı ve gini endeksi olarak adlandırılmaktadır.

Entropi, belirsizliği ve karışıklığı tahminlememize yarayan bir fonksiyondur. Entropi değerinin yükselmesi, karar ağaçlarındaki bölünmelerden sonra ulaşılan son nokta olan yapraklardaki gözlemlerin saflık derecesini düşürmektedir. Sınıf sayısının *c,* veri setindeki bir gözlemin *i* sınıfına ait olma olasılığının ise P_i ile ifade edilmesi durumunda, entropi fonksiyonu aşağıdaki gibi ifade edilmektedir (Belyadi & Haghighat, 2021).

$$E(S) = \sum_{i=1}^{c} -P_i \, log_2 \, P_i$$

Karar ağacı oluşturulurken, en yüksek bilgi kazanımına ve en düşük entropi değerine sahip özniteliği seçmek büyük önem taşımaktadır. Bilgi kazanımı; iki entropi arasındaki fark olarak tanımlanırken, veri setindeki bir özniteliğin, model için değerli olup olmama durumunu belirtmektedir. Modeli eğitim için kullanılan veri setindeki bir özniteliğin bilgi kazanımı değerinin yüksek olması, söz konusu özniteliğin model için benzer derecede önem taşımakta olduğunu göstermektedir. (Aytekin, Sütcü, & Özfidan, 2018).

Kök düğüm seçiminde, bilgi kazanımı değeri dikkate alınmaktadır. Bilgi kazanımı değeri en yüksek olan öznitelik, kök düğüm olarak seçilmektedir (Önder, 2020).

IG ile ifade edilen bilgi kazanımı fonksiyonu aşağıdaki şekilde ifade edilmektedir (Belyadi & Haghighat, 2021).

$$IG(Y,X) = E(Y) - E(Y,X)$$

Karar ağacı algoritmalarından biri olan CART algoritmasının temel çalışma prensibi, karar ağacı yukarıdan aşağıya doğru dallarına ayrılırken, her bir karar düğümünden sonra ağacın iki dala bölünmesi şeklindedir. CART algoritmasının uygulanmasında veri setinde yer alan kategorik öznitelikler için Gini endeksi kullanılmaktadır (Akpınar, 2000).

Gini endeksinin formülü aşağıdaki şekilde ifade edilmektedir.

$$Gini = 1 - \sum_{i=1}^{c} (P_i)^2$$

Gini endeksinin 0 olması, karar ağaçları sınıflandırma modelinin, veri seti üzerindeki performansının çok yüksek olduğunu göstermektedir.

Karar ağaçlarında modelin aşırı uyum probleminin önüne geçmek için budama işlemi gerçekleştirilmektedir. Bölünme sürecine etki edecek çeşitli durdurma kriterleri kullanılmaktadır (Önder, 2020).

Karar ağaçlarında budama; modelin doğruluk ölçütüne ve başarısına yeterince katkıda bulunmayan dallarda yer alan özniteliklerin modelden çıkarılmasını ifade etmektedir. Ayrıca karar ağaçları algoritmaları uygulanan sınıflandırma problemlerinde, çapraz doğrulama uygulanması da modelin aşırı uyum probleminin önüne geçilmesine yardımcı olmaktadır (Belyadi & Haghighat, 2021).

Karar ağaçlarında ID3, C4.5, CART, CHAID, QUEST, CRUISE, GUIDE ve CTREE adlarında farklı sınıflandırma modelleri de kullanılmaktadır (Loh, 2014).

Karar ağaçlarının sınıflandırma problemlerinde uygulanmalarının avantajları arasında, karar ağaçları modellerinin, veri seti ile ilgili "alan" bilgisine ihtiyaç

duymamaları, karmaşık durumları en düşük seviyeye indirmeleri, yüksek boyutlu veri setlerinde uygulanma sorunu yaşamamaları, kolayca anlaşılabilir kurallar oluşturmaları, sürekli ve kategorik özniteliklerle çalışabilmeleri ve eksik verilere duyarlı olmamaları yer almaktadır (Önder, 2020).

Bunların yanında; güçlü performans göstermeleri, çok fazla işlem gerçekleştirmeksizin hızlı çalışmaları, günümüzde popülerliklerinin artması, sınıflandırma için özniteliklerin önem sıralamasını oluşturmaları da karar ağaçlarının sınıflandırma problemlerinde kullanılmalarının avantajları arasında yer almaktadır (Balaban & Kartal, 2018).

Karar ağaçlarının sınıflandırma problemlerinde uygulanmalarının dezavantajları arasında, modellerin tek çıktı özniteliği ile sınırlı çalışmaları, sınıflandırma uygulamalarında modelin performansının veri setine çok bağlı olması, sürekli özniteliklerde karmaşık ağaçlar oluşturması ve aşırı uyum sorunu yaşama ihtimalinin olması yer almaktadır (Önder, 2020).

1.2.1.6 Rassal Orman

Makine öğreniminde topluluk yöntemlerinden biri olan rassal orman sınıflandırma algoritması 2001 yılında Leo Breiman tarafından geliştirilmiştir.

Rassal orman modelinin teorik olarak açıklanmasından önce, topluluk öğrenimi kavramının tanımlanması gerekmektedir.

Topluluk öğrenimi modelleri, zayıf performansla öğrenen çok sayıda modelin birleştirilmesiyle oluşturulan, güçlü performansla öğrenebilen tek bir model şeklinde tanımlanmaktadır. Topluluk öğrenimi modelleri oluşturulduktan sonra, yeni gözlemlere yönelik tahminler gerçekleştirirken, oluşumuna dahil edilen modellerin her biri probleme uygulanarak, farklı tahminler elde edilmektedir. Topluluk öğrenimi modellerinde sınıf tahmini yapılırken, çeşitli yöntemler izlenebilmektedir. Bu yöntemler içinde, oy çokluğu veya topluluk öğrenimi modeli oluşturan temeldeki modellerin, performans önemine göre ağırlıklandırılarak tahminler yapması yer almaktadır. Topluluk öğrenimi modellerinin avantajları arasında, çok farklı sorunları çözmeye yardımcı olmaları, büyük veri setlerinde çalışma performanslarının

geliştirilmiş olması, farklı alanlardan veri setleri ile çalışmaya uyumlu olmaları ve öğrenmesi zor olan hedeflerde yüksek performans göstermeleri yer almaktadır. Topluluk öğrenimi modelleri arasında, torbalama, yükseltme ve rassal orman modelleri yer almaktadır (Lantz, 2019).

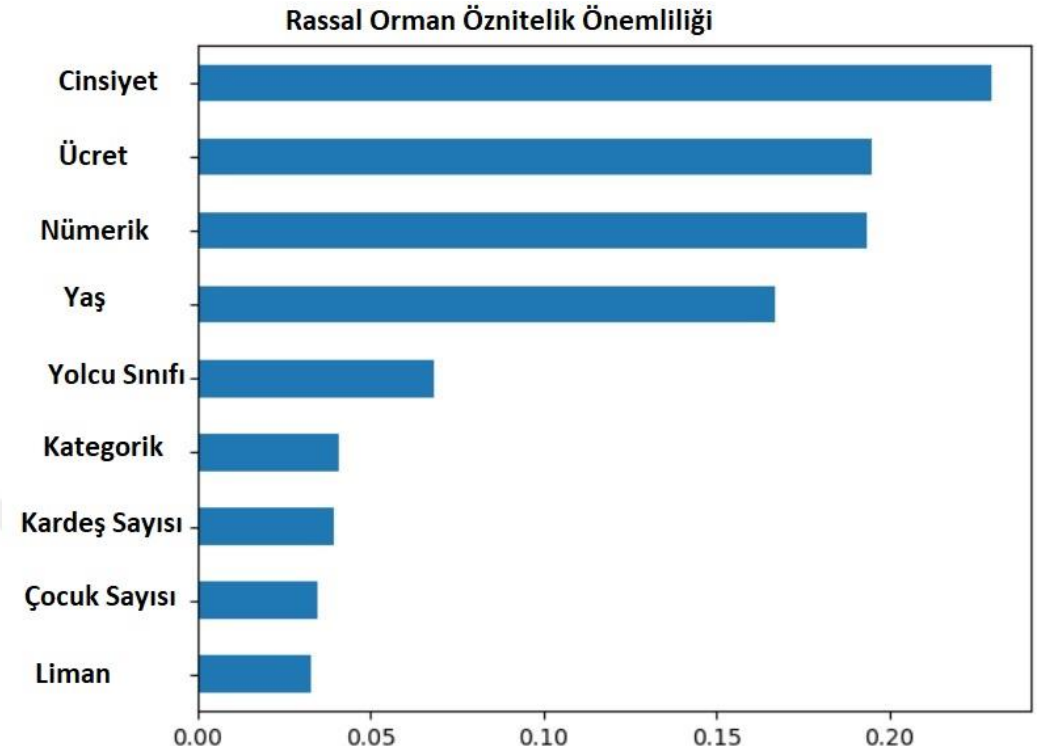
Rassal orman sınıflandırma algoritmaları; ağaç tabanlı sınıflandırma algoritmalarının, birlikte çalışarak yeni gözlemin sınıf tahminine yönelik oy kullanması şeklinde çalışmaktadırlar. Yeni gözlemin sınıfı, en fazla oya sahip olan sınıf olarak belirlenmektedir (Breiman, 2001).

Karar ağaçları sınıflandırma algoritmalarının karşılaştıkları güçlüklerden biri varyans problemleridir. Karar ağaçları modellerinde budama işlemi yapılmadığı takdirde, modellerin eğitim veri setinin üzerindeki performansı olumsuz etkilenmektedir. Bu ve benzeri durumlarda, modelin varyans değerini düşürmek için torbalama işlemi uygulanmaktadır. Rassal orman modelleri ile veri setlerinde sınıflandırma uygulaması gerçekleştirilirken, varyans değeri karar ağaçlarına göre düşük kalmaktadır. Bunun nedeni, karar ağacı modellerinin tek bir ağaç ile çalışmaktayken, rassal orman modellerinin çok fazla ağaç ile çalışıyor olmasıdır. Rassal orman sınıflandırma algoritmalarında budama işlemi gerçekleştirilmesine de ihtiyaç yoktur (Önder, 2020).

Rassal orman sınıflandırma algoritmalarında veri setinde bulunan tüm öznitelikler yerine, sadece model için en önemli olan öznitelikleri görebilmek önem taşımaktadır. Bunu yapmak için, modelin öznitelik önemliliği argümanı kullanılarak, model için en önemli öznitelikler görselleştirilebilmektedir. Öznitelik önemliliği, 0 ile 1 arasında değer almaktadır. 0 değeri, bir özniteliğin model için hiç önemli olmadığını belirtmekteyken, 1 değeri, bir özniteliğin model için çok önemli olduğunu belirtmektedir (Müller & Guido, 2017).

Grafik 1.7.'de rassal orman sınıflandırma modelinin, öznitelik önemliliği argümanının çıktı grafiği görülmektedir (Scikit-Learn, 2023).

Grafik 1.7.: Rassal Orman Sınıflandırma Modeline ait Öznitelik Önemliliği Grafiği



Rassal orman sınıflandırma algoritmalarının avantajları arasında, gürültü ve uç gözlemlere dirençli olması, torbalama yöntemlerinden daha hızlı çalışması, ayarlanabilecek hiperparametre sayısının az olması sebebiyle, modelin gerçek hayat verilerine uygulanmasına hızlı bir şekilde geçilebilmesi, eksik gözlem değerlerinden çok fazla etkilenmemesi ve kullanıcılara çeşitli ölçümlere göre çok sayıda bilgi vermesi yer almaktadır (Marins, ve diğerleri, 2021).

Rassal orman sınıflandırma algoritmalarının dezavantajları arasında ise, modellerde ağaç sayısının arttıkça modelin çalışma hızının düşüp, çalışma süresinin uzaması ve çok sayıda kategori içeren özniteliklere sahip veri setlerinde taraflı davranması yer almaktadır (Önder, 2020).

1.2.1.7 Gradyan Artırılmış Ağaçlar

Topluluk öğrenimi modelleri içinde yer alan bir diğer model, gradyan artırılmış ağaçlar modeli olarak adlandırılmaktadır.

Gradyan artırılmış ağaçlar modelleri, rassal orman modellerine zıt olarak, model kurulum aşamasında her bir tahmin edici ağacın, bir önceki ağacın hatasını

düzelterek yeni tahminler gerçekleştirmesi şeklinde çalışmaktadır. Gradyan artırılmış ağaçlar modellerinde, varsayılan olarak, rassallaştırma işlemi yerine güçlü ön-budama işlemi gerçekleştirilmektedir. Gradyan artırılmış ağaçlar modelleri, günümüzde çeşitli yarışmalarda elde ettikleri başarılarla popülerliklerini yükseltmektedirler. Genel olarak rassal orman modellerine göre hiperparametre ayarlamalarına daha duyarlı olmaktayken, hiperparametrelerin optimize edilmesi sonrasında, çok daha yüksek doğruluk oranıyla performans sergilemektedirler (Müller & Guido, 2017).

1.2.1.8 Ekstrem Gradyan Artırılmış Ağaçlar

2016 yılında Tianqi Chen ve Carlos Guestrin tarafından yayınlanan bir makale ile, gradyan artırılmış ağaçlar modeli üzerinde çeşitli düzenlemeler yapılmış ve daha gelişmiş şekilde çalışan, ekstrem gradyan artırma modelinin (XGBOOST) geliştirildiği duyurulmuştur.

Ekstrem gradyan artırma modelinin, ağaç modellerde yükseltme işlemi için uygun ve ölçeklenebilir bir model olmasının yanında, açık kaynak koda sahip olarak, geniş kullanıcı kitlesiyle, farklı uygulama alanlarında popülerliğini ve kullanımını artırdığı belirtilmektedir (Chen & Guestrin, 2016).

Paralel çalışması, dağıtık olması, çekirdeksiz, önbelleğe duyarlı işlemci yapısıyla, ekstrem gradyan artırma modeli, makine öğrenimi ve derin öğrenme modelleri arasındaki popüler modellerden çok daha iyi performans göstermektedir. Aynı zamanda ekstrem gradyan artırma modelinin kullanımının sağladığı avantajlardan biri de çok iyi optimize edilebilir ve ölçeklenebilir yapıya sahip olması olarak belirtilmektedir. Model, uygulanacağı veri setinde bulunan eksik gözlemlerin, yüksek frekansa sahip olan 0 değerlerinin ve uygulanan öznitelik mühendisliği tekniklerinin sonuçlarına göre çalışmasını güncellemektedir. Modelin uygulanmasında oluşan kayıp, gradyan azaltma algoritması ile en düşük seviyeye indirilmektedir (Paleczek, Grochala, & Rydosz, 2021).

1.2.1.9 Naive Bayes

Naive Bayes sınıflandırma algoritması, makine öğreniminde Bayes teoremine dayalı olarak çalışmaktadır.

Naive Bayes algoritması ile model oluşturulurken, modelin eğitim aşamasında, hedef öznitelikteki her sınıf için olasılıklar hesaplanmaktadır. Bu olasılıklar test verisinin hedef özniteliğinin sınıflarına yönelik tahminlemede kullanılmaktadır. Bayes Teoremi aşağıdaki şekilde formüle edilmektedir (Önder, 2020):

$$P(A / B) = \frac{P(B / A)P(A)}{P(B)}$$

Bu formülde P(A / B); B olayının gerçekleşmesi için A olayının koşullu olasılığını ifade etmekteyken, P(B / A); A olayının gerçekleşmesi için B olayının koşullu olasılığını ifade etmektedir. P(A); A olayının gerçekleşme olasılığını ifade etmekteyken, P(B); B olayının gerçekleşme olasılığını ifade etmektedir.

Naive Bayes sınıflandırma algoritmalarının avantajları arasında, küçük eğitim setleri için uygun olması, kolay uygulanması, eksik verilerden etkilenmemesi yer almaktadır.

Naive Bayes sınıflandırma algoritmalarının dezavantajları arasında, eğitim veri setinin büyümesiyle modelin yanlılığının artması, girdi verisinin hazırlanma şekline duyarlı olması ve veri setindeki tüm özniteliklerin bağımsız ve eşit öneme sahip olduğunu varsayması yer almaktadır (Önder, 2020).

1.2.1.10 Yapay Sinir Ağları

Yapay Sinir Ağları, 1944 yılında, Warren McCullough ve Walter Pitts tarafından geliştirilmiştir.

Yapay Sinir Ağları, insan beyninin çalışmasına benzer şekilde, insan beynini simüle ederek çalışmaktadır. Yapay Sinir Ağları'nda bilgisayar ortamında oluşturulan yapay düğümler kullanılmaktadır. Yapay Sinir Ağları modellerinde, girdiler önem derecelerine göre ağırlıklandırılmaktadır. Ağırlıklandırılmış girdi değerlerinin toplamı elde edilip, aktivasyon fonksiyonuna tabi tutulmaktadır. Modelin çalışmasının son bölümünde ise, çıktı değeri oluşturulmaktadır (Önder, 2020).

Yapay Sinir Ağları'nın temelinde üç katman bulunmaktadır. Bu katmanlar sırasıyla; girdi katmanı, çıktı katmanı ve gizli katmanlar olarak adlandırılmaktadır. Bu

katmanlardan, girdi katmanında ve çıktı katmanında yer alan nöron sayısı, bağımsız özniteliklerin ve hedef özniteliğinin sayısı ile belirlenmekte iken, gizli katmanlardaki nöron sayıları en iyi performansı sergileyecek şekilde kullanıcılar tarafından belirlenmektedir (Atalay, 2016).

Yapay Sinir Ağları algoritmasının avantajları sırasıyla, girdi ve çıktı öznitelikleri arasındaki ilişkileri kolayca tespit edebilmesi, gürültülü verilerle uyumlu çalışabilmesi olarak ifade edilmektedir (Önder, 2020).

Yapay Sinir Ağları algoritmasının dezavantajları sırasıyla, ezberleme sorununa dikkat etmeyi gerektirmesi, yorumunun zor olması, çalışmasının uzun süre gerektirmesi, uç değerlere duyarlı olması olarak ifade edilmektedir (Önder, 2020).

1.2.1.11 Denetimli Öğrenme Süreçleri

Denetimli öğrenimi süreci verinin, veri tabanlarından ve çeşitli kaynaklardan toplanmasıyla başlamaktadır. Toplanan ham veriler için öncelikle veri temizleme işlemleri olmak üzere çeşitli ön hazırlık işlemleri uygulanmaktadır. Ön hazırlık aşaması tamamlandıktan sonra, veri setinin eğitim seti ve test seti olarak ikiye bölünmesi ile eğitim seti üzerinde, model kurulumu için gerekli öznitelikleri seçmek ve bu öznitelikleri model için uygun hâle getirmek gerekmektedir. Sonrasında model için bir denetimli öğrenme algoritmasının seçimi ve eğitim setine modelin uygulanmasıyla eğitim işlemi gerçekleştirilmektedir. Eğitim işleminden sonra, modelin performansı beklenildiği düzeyde ise, model daha önce üzerinde hiç denenmemiş olduğu test setinin üzerine uygulanmaktadır. Böylece eğitim seti kullanılarak eğitilen modelin, daha önce hiç etkileşime girmediği test seti üzerinde çalıştırılarak performansı değerlendirilmektedir. Modelin performans değerleri bu aşamada da kabul edilecek düzeyde ise, model gerçek hayat problemlerinde kullanıma başlanmaktadır (Lewis, 2017).

Model kurulumundan sonra, modelin performansını değerlendirme yöntemi olarak k – kat çapraz doğrulama yöntemi sıklıkla kullanılmaktadır. Bu doğrulama yönteminde veri seti genel olarak k sayıda parçaya bölünmektedir. Veri setinden elde edilmiş k sayıda alt kümelerden biri test set olarak kullanılırken, $k - 1$ sayıda alt kümeler eğitim seti olarak kullanılmaktadır. Bu yöntemde veri setinden k sayıda elde

edilen parçalardaki ortalama hatalar hesaplanarak modelin performans değerlendirilmesi yapılmaktadır (Balaban & Kartal, 2018).

1.2.1.12 Denetimli Öğrenmede Performans Değerlendirme Ölçütleri

Sınıflandırma problemlerinde oluşturulan modelin performansını değerlendirirken çeşitli ölçütler kullanılmaktadır. Karmaşıklık matrisi oluşturularak elde edilen bu ölçütler, doğruluk, kesinlik, duyarlılık, özgüllük, F skoru ve duyarlılığın harmonik ortalaması olarak adlandırılmaktadır (Önder, 2020).

Karmaşıklık matrisi Grafik 1.8.'de görülmektedir

Grafik 1.8.: Karmaşıklık Matrisi

		Gerçek Sınıf	
		1	0
Tahmin Edilen Sınıf	1	Doğru Pozitif	Yanlış Pozitif
	0	Yanlış Negatif	Doğru Negatif

Bu matriste yer alan bölmelerin ifade ettikleri durumlar şu şekilde sıralanmaktadır:

Doğru Pozitif: Gözlemin gerçek sınıfının pozitif olduğu durumda, tahmin edilen sınıfının da pozitif olması durumu

Yanlış Pozitif: Gözlemin gerçek sınıfının negatif olmasına rağmen, tahmin edilen sınıfının pozitif olması durumu

Doğru Negatif: Gözlemin gerçek sınıfının negatif olduğu durumda, tahmin edilen sınıfının da negatif olması durumu

Yanlış Negatif: Gözlemin gerçek sınıfının pozitif olmasına rağmen, tahmin edilen sınıfın negatif olması durumu

Karmaşıklık matrisinde yer alan ölçütlerden biri olan, doğruluk; toplam sınıf tahminleri içindeki doğru sınıf tahminlerinin oranını vermektedir. Kesinlik ve duyarlılık sınıflandırma problemlerinde birlikte kullanılan ölçütlerdir. Kesinlik; doğru pozitif olarak tahmin edilen gözlem sayısının, pozitif olarak tahmin edilen toplam gözlem sayısına oranı iken, duyarlılık; doğru pozitif olarak tahmin edilen gözlem sayısının, doğru pozitif ve yanlış negatif olarak tahmin edilen gözlemlerin toplam sayısına oranıdır. Özgüllük ise, doğru negatif olarak tahmin edilen gözlemlerin, doğru negatif ve yanlış pozitif olarak tahmin edilen gözlemlerin toplam sayısına oranıdır. Makine öğreniminde, denetimli öğrenme modelinin performans değerlendirilmesinde kullanılan bir diğer ölçüt olan F skoru, kesinlik ve duyarlılık ölçütlerinin değerlerinin harmonik ortalamasının elde edilmesiyle bulunur (Önder, 2020).

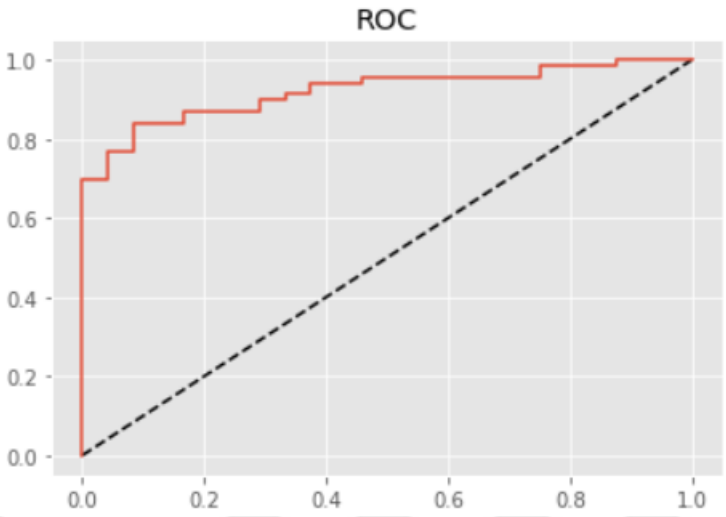
Denetimli öğrenme modelinin performansını değerlendirmede kullanılan başka bir ölçüt, eğrinin altında kalan alan olarak adlandırılmaktadır. Eğrinin altında kalan alan grafiği, X ekseninin yanlış pozitif oranlarını, Y ekseninin ise doğru pozitif oranlarını ifade ettiği iki boyutlu bir düzlemden oluşmaktadır. Doğru pozitif oranı artarken, yanlış pozitif oranının azalması hedeflenmektedir. Grafikte (0,1) noktası en iyi sınıflandırma durumunu ifade etmektedir (Balaban & Kartal, 2018).

Eğrinin altında kalan alan grafiğinde, modelin başarısı değerlendirilirken, eğrinin (0,1) noktasına yakınlığı dikkate alınır. Eğrinin (0,1) noktasına yakın olması, modelin başarısının yüksek olduğunu belirtmektedir.

Eğrinin altında kalan alan grafiğinin yanı sıra, eğrinin altında kalan alan değeri de makine öğreniminde modelin performansının değerlendirilmesi yapılırken kullanılan metrikler arasında yer almaktadır. Eğrinin altında kalan alan değeri 0 ile 1 arasında değer almaktadır.

Grafik 1.9.'de eğrinin altında kalan alan grafiği görülmektedir.

Grafik 1.9.: Eğrinin Altında Kalan Alan Grafiği



1.2.1.13 Hiperparametre Optimizasyonu

Hiperparametre kavramı; makine öğrenimi modellerinin dış kaynaklı yapılandırılmasında kullanılan değerleri ifade etmektedir. Modeller, hiperparametre değerlerini verilerden yola çıkarak tahmin edememektedirler. Modelin en yüksek seviyede başarı göstermesi için, hiperparametreler farklı yöntemlerle optimize edilmektedirler. Bu yöntemlerden en çok kullanılanları; GridSearch ve RandomSearch olarak adlandırılmaktadırlar.

GridSearch yönteminin temel çalışma prensibi; ayarlanacak olan hiperparametrelerin ve bu hiperparametrelerde denenmesi istenen değer aralıklarının tanımlanması ve sonrasında, bu değerlerin tüm olası kombinasyonları ile modelin eğitilerek, en iyi sonucu veren kombinasyonun, modelin hiperparametreleri olarak seçilmesi şeklindedir. GridSearch yöntemi, gerçekleştirilen işlem süreleri dikkate alındığında oldukça maliyetli bir yöntem olmaktadır. GridSearch yönteminin maliyetinden yola çıkarak, daha düşük bir maliyetli yöntem olan RandomSearch yöntemi de günlük hayatta sıklıkla kullanılmaktadır. RandomSearch yönteminde, sadece belirli sayıda ve önceden seçilmiş hiperparametre ve değer aralığı kombinasyonları denenmektedir. Bu durum, işlem süresinin kısalmasına ve maliyetin azalmasına yardımcı olmaktadır. Lâkin RandomSearch yönteminin, GridSearch yöntemine göre büyük bir dezavantajı olarak ta, denenecek kombinasyon sayısının azalmasından dolayı, en iyi kombinasyona ulaşma olasılığının düşüklüğü yer almaktadır (Troisi, Troisi, Scala, & Richards, 2022).

1.2.2 Denetimsiz Öğrenme

Denetimsiz öğrenme, veri setinde yer alan gözlemlerin çıktı değerlerinin bilinmediği durumlarda kullanılan algoritmaları içeren makine öğrenimi türüdür. Denetimsiz öğrenme, gözlemler arasındaki benzerlikleri veya uzaklıkları dikkate alarak çalışmaktadır. Denetimsiz öğrenmenin en yaygın kullanılan modeli kümeleme olmakla birlikte, boyut indirgeme de çok fazla kullanılmaktadır (Önder, 2020).

Denetimsiz öğrenme, veri setinde benzer girdiler içeren gözlemlerden oluşacak kümeleri tespit edecek şekilde çalışmaktadır. Denetimsiz öğrenme modelleri bu tespit işlemi sırasında, sınıflar oluşturup gözlemleri bu sınıflardan birine dahil etmek yerine, gözlemler arasındaki benzerlikleri ortaya çıkarmaya çalışmaktadır (Marsland, 2015).

Denetimsiz öğrenme, kümeleme, boyut indirgeme ve birliktelik kuralları olmak üzere üç farklı bölüme ayrılmaktadır.

Denetimsiz öğrenme modelleri oluşturulurken en çok kullanılan algoritmalar; K-Ortalamalar, Hiyerarşik Kümeleme, Temel Bileşenler Analizi, Doğrusal Ayırma Analizi, Apriori ve Eclat olarak sıralanmaktadır.

1.2.2.1 K-Ortalamalar

1967 yılında J. MacQueen tarafından geliştirilen K-Ortalamalar algoritması, makine öğreniminde kümeleme modelleri oluşturulurken, en çok kullanılan ve en popüler olan algoritmalardan biridir.

K-Ortalamalar algoritması ile model oluşturulurken, modelin uygulanacağı veri setinde, veri ön işleme adımlarından biri olan normalizasyon işleminin gerçekleştirilmesinin gerekli olduğu durumlar oluşabilmektedir. Normalizasyon, bir veri setindeki tüm nümerik özniteliklerde yer alan gözlemlerin, aynı değer aralıklarında ifade edilmesi için gerçekleştirilen bir dönüştürme işlemidir. Farklı özniteliklerin, farklı aralıklarda değerlere sahip oldukları durumlarda, kurulan ve veri setine uygulanan model, veri setindeki özniteliklerin çok yüksek ve çok düşük değerlerinden etkilenmektedir. Bu durumun engellenmesi ve veri setindeki tüm özniteliklere, modele etki açısından eşit önem verilmesi amacıyla normalizasyon

işlemi uygulanmaktadır. Normalizasyon işleminin sonucu olarak, birbirlerinden farklı değer aralıklarında dağılım gösteren nümerik özniteliklerin tamamı, aynı değer aralığında ifade edilmektedir. Bu şekilde normalizasyon işlemi uygulanmış özniteliklerin model için etkisi eşitlenmekte ve birbirleri ile karşılaştırılması olanaklı duruma gelmektedir. Normalizasyon işlemi için iki farklı yöntem uygulamalarda çok fazla kullanılmaktadır. Bunlar, minimum-maksimum normalizasyonu ve z-skoru olarak adlandırılmaktadır. Minimum-maksimum normalizasyonunda, normalizasyon işleminin uygulanacağı özniteliklerdeki nümerik değerler içerisinde en büyük ve en küçük değerler tespit edilerek, en büyük değer 1'e, en küçük değer 0'a eşitlenmektedir. Kalan değerler ise 0 ile 1 aralığında değerlere dönüştürülmektedir. Z-skoru yönteminde ise, normalizasyon işleminin uygulanmak istendiği öznitelikteki nümerik değerlerin ortalaması ve standart sapması kullanılmaktadır (Koçoğlu & Özcan, 2016).

K-Ortalamalar algoritması ile oluşturulan model ile üzerinde kümeleme işlemi gerçekleştirilecek veri setindeki gözlemler, kullanıcı tarafından belirlenen *k* sayıda kümelere paylaştırılmaktadır. Bir kümenin oluşması için öncelikle bir küme merkezinin belirlenmesi gerekmektedir. Bir gözlemin dahil olacağı küme, *k* sayıda kümeler içinden küme merkezine uzaklık olarak en yakın olan küme şeklinde belirlenmektedir (Davidson, 2002).

K-En Yakın Komşuluk sınıflandırma algoritmasına benzer şekilde, K-Ortalamalar algoritması da çoğunlukla Öklid uzaklık ölçüsünü kullanmakla birlikte, Manhattan ve Minkowski uzaklık ölçüleri de kullanılmaktadır (Önder, 2020).

K-Ortalamalar algoritması çalışırken sırasıyla; küme sayısı olan *k* değerinin belirlenmesi, kümelerin merkezlerinin hesaplanması, veri setindeki gözlemlerin, küme merkezlerine olan uzaklıklarının ölçülmesi, oluşturulan kümelerin toplam hata-kare değerlerinin hesaplanması, her bir gözlemin en yakın uzaklıkta yer alan kümeye dahil edilmesi işlemleri gerçekleştirilmektedir. Bu işlemler içinde yer alan küme merkezine olan uzaklıkların ölçümü, hata-kare değerlerinin hesaplanması ve gözlemlerin en yakın uzaklıktaki kümeye dahil edilmesi işlemi tekrarlanarak çalıştırılmaktadır (Koçoğlu & Özcan, 2016).

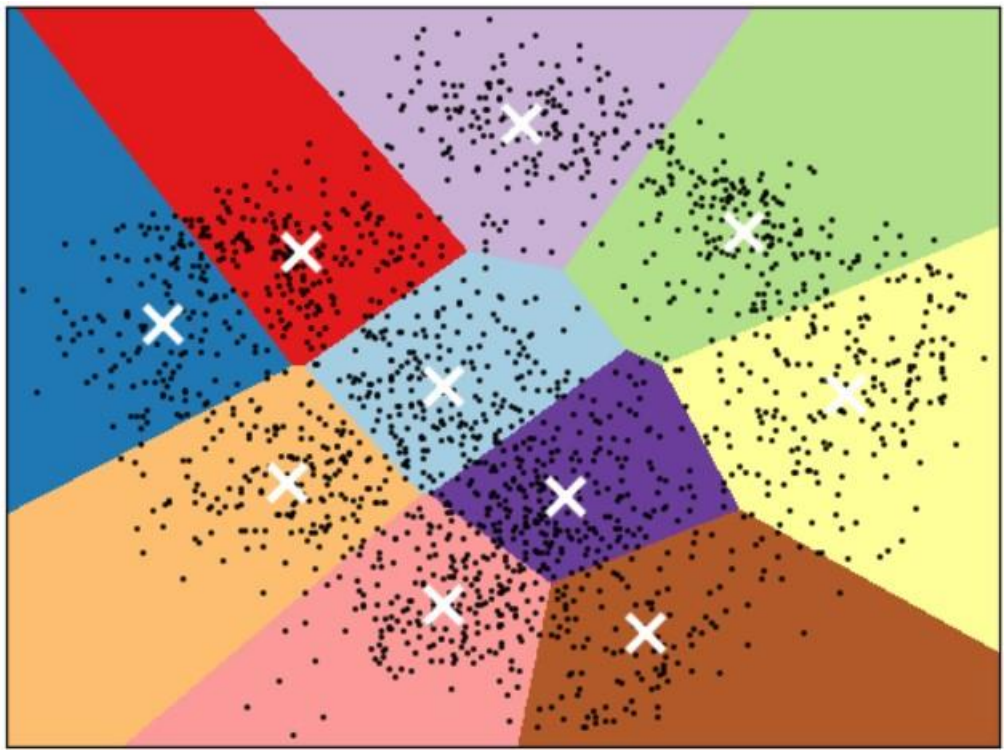
K-Ortalamalar algoritması ile kümeleme uygulamaları için model kurulurken, küme sayısının doğru seçimi önem taşımaktadır. Küme sayısını ifade eden *k* değeri çok büyük değerler aldığında, oluşan kümeler anlamlılık açısından yetersiz kalabilmekteyken, *k* değeri çok düşük değerler aldığında, oluşan kümeler fayda açısından yetersiz kalmaktadır. Küme sayısı olan *k* değerini en uygun şekilde seçmek için farklı yöntemler kullanılmaktadır. Bu yöntemler içinde en çok kullanılanları; dirsek yöntemi, siluet yöntemi ve gap istatistiği olarak adlandırılmaktadır. Dirsek yönteminde, her bir kümenin barındırdığı gözlemlerin homojenliği ile küme sayısı olarak seçilen *k* değerinin arasındaki ilişkinin belirlenmesi amaçlanmaktadır. Her bir kümenin içindeki gözlemlerin sahip olduğu homojenlik seviyesinin yüksek olduğu seviyeye karşılık gelen en düşük *k* değeri, en uygun küme sayısı olarak belirlenmektedir. Siluet yönteminde, bir kümenin içindeki gözlemlerin kendi kümelerine olan mesafeleriyle, komşu kümelere olan mesafeleri karşılaştırılmaktadır. Bu yöntemde en uygun *k* değerine sahip model, en büyük siluet değerine sahiptir. Gap istatistiği, rassal örnekleme metoduyla seçilen gözlemlerden oluşan veriler ile seçimde referans olan veriler arasındaki küme içi heterojenlik değeri toplamının, ne kadar farklı olabileceğini göstermektedir. Gap istatistiği ile küme sayısı seçilirken, büyük gap istatistik değerlerine karşılık olan küme sayısının en uygun şekilde olduğu görülmektedir (Önder, 2020).

Küme sayısını ifade eden *k* değerinin belirlenmesinde, dirsek yöntemi, siluet yöntemi ve gap istatistiği dışında, Dunn indeksi, Calinski-Harabasz indeksi, Davies-Bouldin indeksi ve Xie-Beni indeksi olmak üzere farklı ölçütler de kullanılmaktadır (Koçoğlu & Özcan, 2016).

K-Ortalamalar algoritmasının kümeleme problemlerinde uygulanmalarının avantajları arasında, kolay anlaşılan ve uygulanabilen algoritmalar olmaları, büyük veri setlerine kolaylıkla uygulanabilmeleri yer almaktadır. Bu avantajlarına karşılık olarak, K-Ortalamalar algoritmasının kümeleme problemlerinde uygulanmalarının dezavantajları arasında, küme sayısının tutarlı bir şekilde tahmin edilmesini veya alan bilgisine sahip olunarak doğrudan bilinmesini gerektirmesi, kategorik özniteliklerde performans sorunları yaşaması, uç gözlem değerlerinin ve gürültü verilerin modelin çalışmasını olumsuz etkilemesi yer almaktadır (Önder, 2020).

Grafik 1.10.'da K-Ortalamalar modeli ile gerçekleştirilen kümeleme uygulaması görülmektedir (Scikit-Learn, 2023).

Grafik 1.10.: K-Ortalamalar Modeli ile Kümeleme

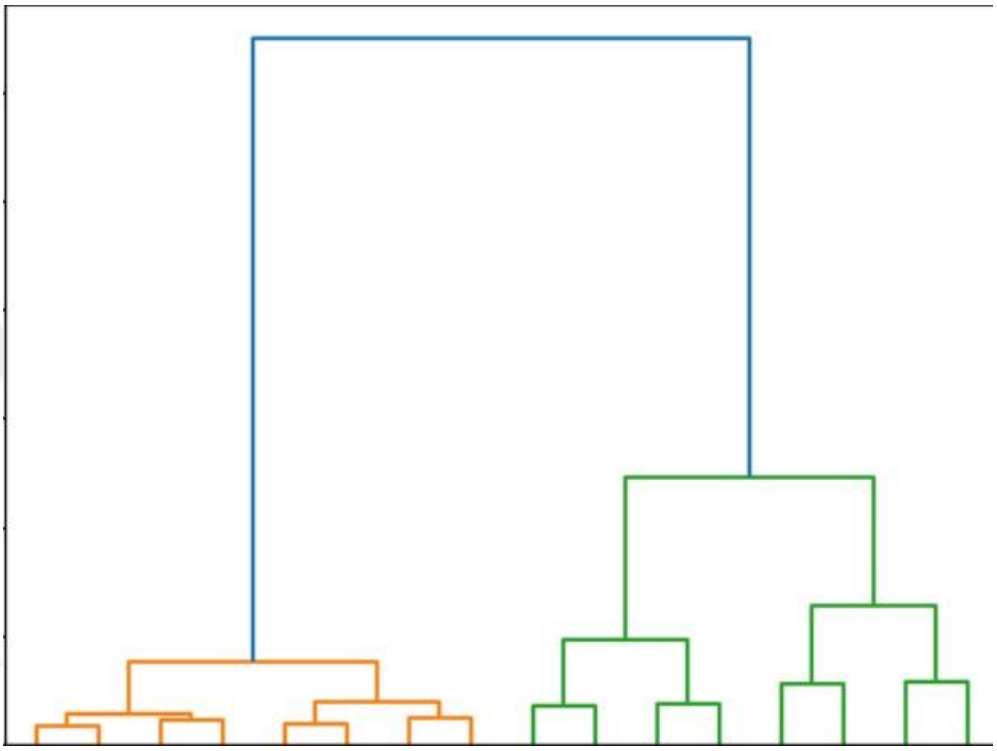


1.2.2.2 Hiyerarşik Kümeleme

Hiyerarşik kümeleme algoritması, veri setinde kümeleme işlemi uygulanıp kümeler elde edilirken, verilerin farklı gruplarda toplanması yoluyla, belirli bir hiyerarşiye sahip kümeler oluşturacak şekilde çalışmaktadır. Hiyerarşik kümeleme algoritması uygulanırken iki farklı strateji izlenmektedir. Bu stratejiler sırasıyla, toplanarak çalışma ve bölünerek çalışma olarak adlandırılmaktadır. Toplanarak çalışma stratejisinde, tabandan yani en küçük kümelerden, veri setinin tamamına doğru ilerlenirken, bölünerek çalışma stratejisinde, tepeden yani veri setinin tamamından, en küçük kümelere doğru olacak şekilde ilerlenmektedir. Hiyerarşik kümelerin görselleştirilmesi için özel bir görselleştirme türü olan dendrogramlar kullanılmaktadır. Ters U şeklinde görselleştirilen dendrogramlarda, her ters U şeklinin yüksekliği, birbiriyle bağlı olan iki veri noktası arasındaki uzaklığı ifade etmektedir (Caggiano, Napolitano, & Teti, 2021).

Grafik 1.11.'da Hiyerarşik Kümeleme modeli ile gerçekleştirilen kümeleme uygulaması görülmektedir (Scikit-Learn, 2023).

Grafik 1.11.: Hiyerarşik Kümeleme Modeli ile Kümeleme



1.2.2.3 Gürültülü Uygulamaların Yoğunluk Bazlı Mekânsal Kümelenmesi

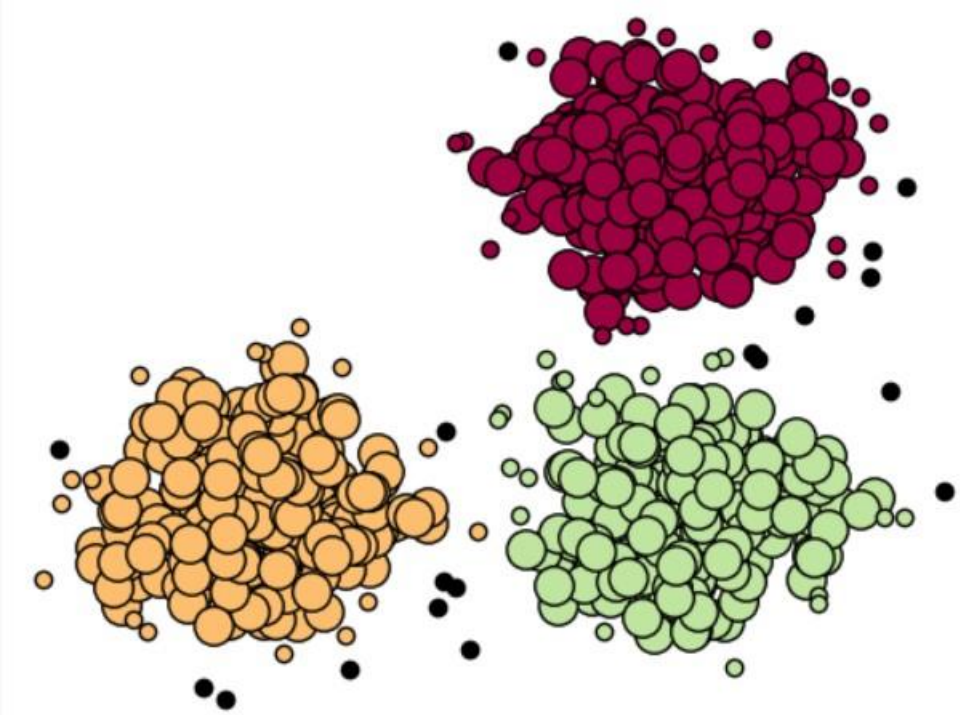
Gürültülü Uygulamaların Yoğunluk Bazlı Mekânsal Kümelenmesi modeli, Martin Ester, Hans-Peter Kriegel, Jörg Sander ve Xiaowei Xu tarafından, 1996 yılında literatüre sunulmuştur (Ester, Kriegel, Sander, & Xu, 1996).

Gürültülü Uygulamaların Yoğunluk Bazlı Mekânsal Kümelenmesi modelinin iki ana kullanım amacı vardır. Bu kullanım amaçları, rassal şekillere sahip kümeleri tespit etme işleminde gösterdiği yüksek performans ve büyük veri setleri için verimli çalışması şeklinde sıralanmaktadır. Gürültülü Uygulamaların Yoğunluk Bazlı Mekânsal Kümelenmesi modelinde, veri setinde yer alan her gözlemin yoğunluğu, o gözleme epsilon (ε) uzaklıkta olan gözlemlerin sayısı baz alınarak ölçülmektedir. Epsilon (ε) parametresi uzunluğunda yarıçapa sahip her çember minimum sayıda veri noktası içermektedir (Belyadi & Haghighat, 2021).

Gürültülü Uygulamaların Yoğunluk Bazlı Mekânsal Kümelenmesi modelinde, Eps ve MinPts değerleri giriş parametreleri olarak yer almaktadır. Veri setinde yer alan gözlemler üzerinde gürültülü uygulamaların yoğunluk bazlı mekânsal kümelenmesi modeli ile kümeleme işlemi gerçekleştirilirken, gözlemlerin hangi kümeye dahil olacağına yönelik bölge sorgulama işlemi gerçekleştirilmektedir. Bu sorgulama işleminin sonucunda elde edilen Eps komşu sayısı değeri, MinPts değerinden yüksek olursa, bu gözlem ve komşu gözlemler, oluşturulacak yeni bir kümeye dahil edilmektedir (Bilgin & Çamurcu, 2005).

Grafik 1.12.'de Gürültülü Uygulamaların Yoğunluk Bazlı Mekânsal Kümelenmesi modeli ile gerçekleştirilen kümeleme uygulaması görülmektedir (Scikit-Learn, 2023).

Grafik 1.12.: Gürültülü Uygulamaların Yoğunluk Bazlı Mekânsal Kümelenmesi Modeli ile Kümeleme



1.2.2.4 Temel Bileşenler Analizi

Denetimsiz öğrenme modelleri içinde yer alan temel bileşenler analizi ile doğrusal ayırma analizi, veri setlerinde boyut indirgeme uygulamaları gerçekleştirilirken sıklıkla kullanılmaktadır.

Boyut indirgeme; yüksek boyuta sahip veri setlerinde, gözlemlerin sınıflandırılmasının, görselleştirilmesinin ve birbirleri ile olan etkileşimlerinin kolaylaştırılması süreci olarak tanımlanmaktadır. Boyut indirgeme için basit ve çok kullanılan analizlerden biri olan temel bileşenler analizi veri setindeki en yüksek varyans değerinin yönlerini bulup, her bir gözlemi bu yönler doğrultusunda temsil etmektedir (Hinton & Salakhutdinov, 2006).

Temel bileşenler analizi, 1901 yılında Karl Pearson tarafından geliştirilmiştir.

Temel bileşenler analizi, yüksek boyuta sahip bir veri setini, en az bilgi kaybı ile daha düşük sayıda boyuta sahip olacak duruma indirgemektedir. Temel bileşenler analizinin uygulanması sonucunda birbirleriyle ilişkisi olmayan öznitelikler elde edilmektedir. Temel bileşenler analizi için verilerden oluşturulan kovaryans matrisinden yararlanılmaktadır. Kovaryans matrisinin özvektörleri hesaplanıp, en büyük özdeğerin karşılığındaki özvektörler, veri setindeki özniteliklerin varyans değerlerinin yapılandırılmasında kullanılmaktadır. Her ne kadar, bu süreçte veri kayıpları yaşansa da varyans, özvektörlerde tutulmaktadır. Temel bileşenler analizinin sağladığı avantajlara, gerekli olan depolama alanını düşürmesi, işlem sürelerini kısaltması ve ihtiyaç duyulmayan öznitelikleri elimine etmesi örnek verilmektedir. Temel bileşenler analizi, başka bir boyut indirgeme tekniği olan, doğrusal ayırma analizi ile karşılaştırıldığında, eğitim veri setinin küçük olması durumunda, daha iyi sonuçlar elde etmektedir (Önder, 2020).

1.2.2.5 Doğrusal Ayırma Analizi

Doğrusal ayırma analizi, 1936 yılında Ronald A. Fisher tarafından geliştirilmiştir.

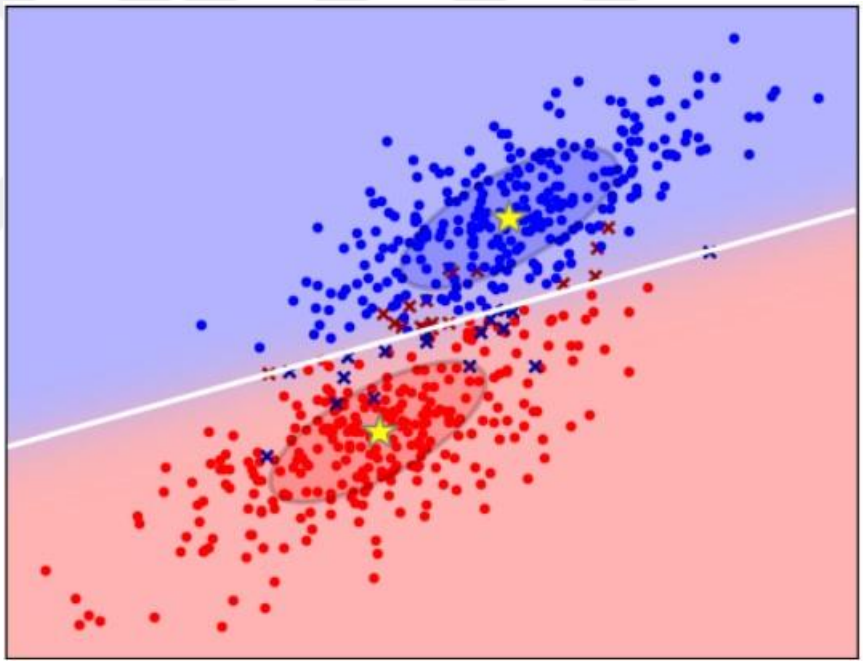
Doğrusal ayırma analizi, boyut indirgeme işlemlerinde kullanılmasının yanı sıra, sınıflandırma problemlerinde de uygulanmaktadır. Doğrusal ayırma analizi, bir veri setinde yer alan gruplar arasında, olabilecek en iyi ayrımı gerçekleştirecek

özniteliklerin doğrusal kombinasyonlarını hesaplamaktadır. Doğrusal ayırma analizi, veri setinde, hedef öznitelikte yer alan her bir sınıfa ait gözlemlerin, kendi sınıflarına olabildikleri kadar en yakın, farklı sınıftakilere olabildikleri kadar uzak olacağı şekilde sonuç verecek bir işlemi gerçekleştirmeye çalışmaktadır. Doğrusal ayırma analizi aşağıda görüldüğü gibi formüle edilmektedir (Önder, 2020).

$$\Delta = \frac{Maksimum\ (\ Ortalama\ A - Ortalama\ B\ )}{Minimum\ (\ Grup\ içi\ Varyans)}$$

Grafik 1.13.'de Doğrusal Ayırma Analizi'nin uygulaması görülmektedir (Scikit-Learn, 2023).

Grafik 1.13.: Doğrusal Ayırma Analizi



1.2.2.6 Apriori

Denetimsiz öğrenmenin modelleri içinde yer alan Apriori algoritması 1994 yılında R. Agrawal ve R. Srikant tarafından geliştirilmiştir.

Apriori algoritması birliktelik kuralları analizleri uygulamalarında, tüketici satın alma davranışlarının analizinde, ürün kampanyaları oluşturmada, pazarlama yönetimi ile ilgili kararlar alınmasında analitik yöntem olarak sıklıkla kullanılmaktadır.

Birliktelik kuralları veri madenciliğinde kullanılan bir yöntem olup, en çok uygulandığı problem ise Sepet Analizi olarak adlandırılmaktadır. Tüketici davranışlarına yönelik tahminlerde bulunmak, sepet analizinin temel amacını oluşturmaktadır. Sepet analizinin sonucunda elde edilen çıktılar değerlendirilerek, raf düzenlemeleri gerçekleştirilebilmekte, promosyonlar düzenlenebilmekte ve bir arada satılacak ürünlere karar verilebilmektedir (Şimşek Gürsoy, 2012).

Birliktelik kuralları analizlerinde farklı ölçütler kullanılmaktadır. Bu ölçütler, destek oranı, güven oranı ve kaldıraç oranı olarak adlandırılmakta ve analiz sonuçlarının yorumlanmasına yardımcı olmaktadırlar. Destek oranı; ikili ürün kümelerinin, veri setinin tamamında birlikte yer alma olasılığını yüzde olarak ifade ederken, güven oranı; ikili ürünlerden solda yer alanını içeren satın alma işlemlerinin içinde, sağda yer alanı da içerenlerin oranını yüzde olarak ifade etmektedir. Kaldıraç değeri ise, ikili ürünlerin birbirlerinden tamamen bağımsız olmaları durumunda, birlikte satın alınma oranlarını ifade etmektedir (Karahan Adalı, 2016).

Büyük veri tabanlarında ilişkiler aranırken en çok kullanılan algoritmalardan biri olan Apriori algoritması, veri setinde sıklık değeri yüksek olan bir nesnenin, dahil olduğu tüm alt kümelerde de sıklık değerinin yüksek olacağını varsayan basit bir şekilde çalışmaktadır. Apriori algoritması, sıklık değeri yüksek olan nesneleri güven oranı ile tespit etmektedir. Bir nesnenin güven oranı, o nesnenin veri seti içindeki sıklık ölçütüdür. Belirlenen güven oranından daha yüksek güven oranına sahip olan nesneler, sıklıkla satın alınan nesneler olarak adlandırılmaktadır (John & Shaiba, 2019).

Apriori algoritmasının yanında birliktelik kuralları için kullanılan Eclat algoritmasının, Apriori algoritmasından farkları; Eclat algoritmasının derin ilk arama tekniği kullanması ve satın alınan ürünlerin geçtiği işlemleri baz alması olarak sıralanmaktadır.

1.2.3 Pekiştirmeli Öğrenme

Pekiştirmeli öğrenmede, modeller eylemler üretmekte ve bu eylemlerin sonucunda nümerik ödül değerini en yüksek seviyeye çıkarmayı amaçlamaktadırlar. Modele gerçekleştireceği uygulamalara yönelik bilgiler kullanıcı tarafından öğretilmemekle birlikte, model en fazla ödül değerini elde etmek için çeşitli eylemler

geliştirmektedir. Denetimli öğrenmede, model kurulumunda, kullanıcının hedef özniteliğine ait etiket bilgisine sahip olması söz konusu iken, pekiştirmeli öğrenmede etiket bilgisi yer almamaktadır. Bir diğer öğrenme türü olan denetimsiz öğrenmede, model kurulumunda, veri setindeki örüntülerin bulunması amaçlanırken, pekiştirmeli öğrenmede ise, ödül değerinin en yüksek seviyeye çıkarılması amaçlanmaktadır (Sutton & Barto, 2018).

1.3 Makine Öğreniminin Kullanım Alanları

Günümüzde teknolojinin ilerlemesi, insanların hayatındaki payının artması, kullanımının ve erişiminin kolaylaşması her geçen gün makine öğreniminin hayatımıza daha fazla katkı sunmasını sağlamaktadır. Makine öğrenimi, günlük hayatın yanı sıra endüstriyel süreçler ve profesyonel hizmetlerde de geliştirilmiş olan farklı uygulamaların temelini oluşturmaktadır.

Makine öğreniminin günlük hayatta en çok kullanıldığı alanlardan biri, finansal işlem ve operasyonlardaki sahtecilik tespitidir. Finansal işlemlerde gerçekleşen sahteciliklerden dolayı, dünya çapında her yıl milyarlarca dolarlık kayıplar yaşanmaktadır. Kredi kartı harcamaları başta olmak üzere, sağlık sigortaları harcamaları, telekomünikasyon sektörü, internet üzerinden gerçekleştirilen satın alma işlemleri ve taşıt sigortaları işlemleri, günümüzde sahtecilik işlemlerinin en çok gerçekleştiği alanlar olarak sıralanmaktadır. Finans kurumları, sahtecilik işlemlerinin gerçekleştirilmesi için çok uygun yapıya sahip olmaları sebebiyle, faaliyetlerini güvenli bir şekilde gerçekleştirmek için teknoloji desteği kullanmaktadırlar. Günümüzde gerçekleşen finansal işlemlerin güvenliğini sağlamak amacıyla, sahtecilik tespiti için makine öğrenimi modellerinden yararlanılmaktadır (Błaszczyński, de Almeida Filho, Matuszyk, Szeląg, & Słowiński, 2021).

Makine öğreniminin günlük hayatta çok kullanıldığı alanlara başka bir örnek olarak sağlık uygulamaları gösterilmektedir. Her geçen gün sağlık sektöründeki kullanım payını artıran makine öğrenimi, ülkelerin sağlık sistemlerinin gelişmesinde de önemli rol oynamaktadır. Günümüzde makine öğrenimi modellerinden, genel olarak hastalıkların teşhis edilmesi başta olmak üzere, MR görüntülerinin sınıflandırılması, hastalıkların ilerleme veya gerilemelerine yönelik süreçlerin tahmin

edilmesi, kan değerlerinin tahmin edilmesi, kalp atış hızının ani değişimlerinden yola çıkarak genel değerlendirmeler yapılması, hastalıklara karşı uygulanan ilaç tedavilerinin takip edilmesi, anomali saptamaları, risk değerlendirmeleri, taburcu süresinden itibaren yeni başvuru riskinin belirlenmesi, hastaların kümelenmesi, görüntü işlenmesi amacıyla yararlanılmaktadır (Önder, 2020).

17 Kasım 2019 tarihinde Çin'in Hubei bölgesinin başkenti Vuhan şehrinde ortaya çıkan ve kısa süre içerisinde dünya geneline yayılıp pandemiye dönen SARS-CoV-2 virüsü kaynaklı COVID-19 pandemisi, aşı geliştirilmesi ve aşılama uygulamaları ile kontrol altına alınmıştır.

COVID-19 pandemisinde geliştirilen aşıların, uygulanmalarından sonra gerçekleşen olumsuz etkilerinin tespiti için, Twitter'da paylaşılan tweetler üzerinde metin madenciliği teknikleri ve doğal dil işleme modelleri kullanılarak anlam çıkarımları gerçekleştirilmiştir. Bu modellerin yardımıyla, Twitter kullanıcıların paylaşımlarından yola çıkarak, kullanıcıların, kaç doz aşı oldukları, hangi aşıyı tercih ettikleri ve aşının uygulanmasından sonra yaşadıkları yan etkiler tespit edilmiştir (Lian, Du, & Tang, 2022).

COVID-19 hastalığını atlatmasına rağmen, hastalığın atlatılmasından 4 hafta sonra yeni semptomlara ve uzun dönemli COVID olarak adlandırılan rahatsızlıklara sahip olması muhtemel insanların, sağlık durumlarının tahmin edilmesi için ekstrem gradyan artırma modellerinden yararlanılmıştır. Bunun yanında uzun dönemli COVID sorununu, aynı semptomlara sahip olarak yaşayan insanları, farklı özniteliklere göre kümelemek için makine öğrenimi modellerinden yararlanılmıştır (Pfaff, ve diğerleri, 2022).

Avrupa ülkelerindeki COVID-19 ölüm oranlarının, iklim ve hava kirliliği ile olan ilişkisinin modellenmesinde, farklı denetimli öğrenme modelleri kullanılmıştır. Araştırmada hava kirliliği kaynaklı olarak, azot dioksit gazının havadaki oranının arttığı tespit edilmiştir. Elde edilen 2 yıllık uydu verilerinden yola çıkılarak, geliştirilen makine öğrenimi modelleri ile tahminler gerçekleştirilmiştir. Bu tahminlerin sonucunda, azot dioksit kaynaklı hava kirliliği ile COVID-19 ölüm oranları arasında anlamlı bir ilişki olduğu görülürken, ülkelerdeki hemşire sayısı, hastane yatak sayıları

ve kişi başına düşen gayri safi yurtiçi hasıla gibi faktörlerin de COVID-19 ölüm oranları ile olan ilişkileri araştırılmıştır (Amoroso, ve diğerleri, 2022).

Makine öğrenimi modellerinin günümüzde spor alanında kullanımı yükselmektedir. Günümüzde, çocukluk ve gençlik dönemlerinde gerçekleşen spor bağlantılı sakatlıklar ve sarsıntıların, önlenmesi, sonuçlarının tahmini, tedavisi ve yönetimi, önemi artan çalışma alanları arasında yer almaktadır. Sportif faaliyetlerde gerçekleşen sakatlıklar ve sarsıntıların iyileşme sürelerine ve uzun süreli iyileşme dönemine ilişkin tahminler gerçekleştirmek için makine öğrenimi modellerinden yararlanılmaktadır. Bu alanda yapılan çalışmada, yaş, spor türü, spor faaliyetinde sakatlıklara ve sarsıntılara sebep olabilecek fiziki temas veya çarpışmanın yaşanıp yaşanmadığı ve halihazırda sporcularda mevcut olan ile gelecekte mevcut olması muhtemel olan risk faktörlerinden oluşan öznitelikler çalışmanın konusu olmaktadır. Bu öznitelikler bağlamında, yapılan çalışma incelendiğinde, veri setinde kategorik özniteliklerin de bulunduğu ve model kurulumunda, CatBoost başta olmak üzere, doğrusal regresyon, ekstrem gradyan artırma, Elastic Net, rassal orman algoritmalarının tercih edildiği görülmüştür (Chu, ve diğerleri, 2022).

Makine öğreniminden günümüzde, pazarlama yönetimi alanında da yararlanılmaktadır. Dijital içerik pazarlaması uygulamalarında, tüketicilerin bir ürünle veya hizmetle olan deneyimini ve etkileşimini optimize etmek için yapay zekadan yararlanılmaktadır. Günümüzde dijital pazarlama uygulamalarında tüketicilerle doğrudan iletişime geçmek, uzun süren ve maliyetli bir seçenek olarak değerlendirilmektedir. Bu sebeple pazarlama yöneticileri, tüketiciler ve hedef kitlelerine yönelik segmentasyon işlemleri gerçekleştirmektedirler. Bu işlemlerin sonucunda, bir ürünle veya hizmetle olan deneyimi ve etkileşimi olumlu yönde gerçekleşen tüketicileri tespit ederek, bu kitlede yer alan tüketicilere özellikle odaklanmaktadırlar. Günümüzde turizm sektörü, dijital içerik pazarlaması faaliyetlerinden en çok yararlanılan sektörlerden biri olarak tanımlanmaktadır. İnternet kullanıcılarının, turistik amaçla arama motorlarında kullandıkları kelimeler, internet siteleri ile etkileşimleri, sosyal ağları kullanımları, demografik özellikleri, davranış tercihleri, kişisel bilgileri hakkında çok fazla veri bulunmaktadır. Metin madenciliği teknikleri ve karar ağaçları modelleri kullanılarak, kullanıcıların profillerine uygun

içeriklerin ön plana çıkarılması, dijital içerik pazarlamasında çok fazla gerçekleştirilmektedir (Gregoriades, Pampaka, Herodotou, & Christodoulou, 2021).

Makine öğrenimi günümüzde müzik alanında da kullanımını yükseltmektedir. Genel olarak müzik alanında kullanımının yanında, müzik terapilerinin etki seviyesini tahminlemek için de kullanılmaktadır. Terapilerde kullanılacak ve etki seviyesi ölçülecek müziğin seçimi, terapistlerin tercihi veya rassal şekilde olmaktadır. Müzik dinlemenin terapi sürecinde, heyecan, endişe, anksiyete, stres ve acı gibi davranışsal ve psikolojik semptomlarda azalmaya yardımcı olması amaçlanmaktadır. Günümüzde gerçekleştirilen çok sayıda araştırmada, müzik türleri ile, kişilik ve bilişsel türler arasında bağlantı olduğu görülmektedir. Müziğin tedavi yöntemi olarak kullanılmasına ilişkin araştırmalarda, makine öğrenimi modellerinden yararlanılmaktadır. Müzik dinleme ile duygusal yargılar arasındaki ilişki, yapay sinir ağları, doğrusal regresyon ve rassal orman modelleri ile açıklanmıştır (Raglio, ve diğerleri, 2020).

Günümüzde COVID-19 pandemisinin küresel tedarik zinciri üzerinde oluşturduğu yıkıcı etkileri ve yakın zamanda başlayan Rusya-Ukrayna savaşı sebepleriyle, küresel boyutta bir enerji krizi yaşanmaktadır. Yeşil enerji olarak ta adlandırılmakta olan yenilenebilir enerji kaynaklarının öneminin çok fazla yükseldiği bu dönemde, makine öğrenimi modelleri ile enerji kaynaklarının kullanımlarının maliyetlerini tahmin etmeye yönelik modeller geliştirilmektedir.

Güneş enerjisi kaynaklarının hisse fiyatlarına yönelik tahminler gerçekleştirmek amacıyla makine öğrenimi modellerinden yararlanılmaktadır. Teknik indikatörler, gümüş madeni fiyatları, gümüş fiyatlarındaki volatilite ve petrol fiyatlarındaki volatilite özniteliklerine sahip veri setinde, rassal orman, torbalama, destek vektör makineleri ile oluşturulmuş model ile fiyat tahminleri gerçekleştirilmiştir (Sadorsky, 2022).

Avrupa enerji piyasalarında oldukça volatil şekilde seyreden elektrik enerjisi fiyatlarının tahminlenmesine yönelik gerçekleştirilen bir çalışmada, Almanya, Belçika ve Fransa verileri üzerinde modelleme yapılarak sonuçlar karşılaştırılmıştır. Çalışmada, rassal orman regresyonu, destek vektör regresyonu, derin sinir ağları ve

evrişimsel sinir ağları ile oluşturulan modellerin performansları karşılaştırılmıştır (Tschora, Pierre, Plantevit, & Robardet, 2022).

Makine öğrenimi, günümüzde tavsiye sistemlerinin oluşturulmasında çok fazla kullanılmaktadır. Tavsiye sistemleri temel olarak, farklı veri kaynaklarından yararlanarak, kullanıcılarının ilgisini çekecek sonuçları tespit edip, bu sonuçlara göre tavsiyelerde bulunmak şeklinde çalışmaktadır. Kullanıcılara tavsiyeler sağlanırken, tavsiye edilen ürün veya hizmet, nesne olarak adlandırılmaktadır. Tavsiye sistemleri, kullanıcılar ve nesneler arasındaki geçmişte olan etkileşimleri inceleyerek öğrenimlerini gerçekleştirmektedirler. Tavsiye sisteminin çalışma yapısına göre, geçmişteki etkileşimler, geleceğe yönelik tavsiyeleri tahmin etmek için kullanılabilecek en önemli göstergeler olarak yer almaktadırlar (Aggarwal, 2016).

Tavsiye sistemleri, günümüzde başta e-ticaret, perakendecilik, bankacılık ve eğlence sektörlerinde yoğun olarak kullanılmaktadır. Tavsiye sistemleri modeli, verileri toplayarak, kullanıcılara kişiselleştirilmiş tavsiyeler sunulması için analiz etmektedir. Tavsiye sistemlerinde en çok kullanılan yaklaşımlar; içerik bazlı filtreleme, işbirlikçi filtreleme ve hibrit filtreleme yaklaşımları olarak sıralanmaktadır. Bu yaklaşımlardan içerik bazlı yaklaşımı, her bir nesnenin içeriğini analiz edip, benzer karakteristik özellik gösteren nesneleri tavsiye etmesi için kullanılmaktadır. İşbirlikçi filtreleme yaklaşımı, kullanıcılar ve nesneler arasındaki benzerliklerden yola çıkarak tavsiyelerde bulunmaktadır. Hibrit filtreleme yaklaşımı ise, içerik bazlı filtreleme ve işbirlikçi filtreleme tekniklerinin birlikte kullanılmasıyla tavsiyelerde bulunmaktadır (Pavitha, ve diğerleri, 2022).

Günümüzde tavsiye sistemlerini, küresel boyutta faaliyet gösteren şirketler çeşitli yapımlarında kullanmaktadırlar. Amerikan teknoloji şirketlerinden biri olan Amazon, özellikle ticari kullanım için tavsiye sistemlerinin uygulanmasında öncü rolü oynamıştır. Amazon'un tavsiye sistemleri farklı kriterler üzerinden çalışabilmektedir. Bu kriterler, ürünlerin aldıkları puanlar, satın alma davranışları ve internet uygulamasında tarama verileri olarak sıralanmaktadırlar. Ürünlerin aldıkları puanlar üzerinden tavsiye sisteminin çalışma şekli, her bir ürüne yönelik, kullanıcıların 1 veya 5 olarak verdikleri değerlendirme puanlarını esas alarak gerçekleşmektedir. Satın alma

davranışları ve internet uygulamasında tarama verileri üzerinden tavsiye sisteminin çalışma şekli ise, Amazon'da üyelik kaydı oluşturulup elde edilen hesap üzerinden, sisteme bağlanıldığı andan itibaren toplanan verilerle gerçekleşmektedir (Aggarwal, 2016).

Günümüzde işletmeler mevcut müşterileri ve potansiyel kazanılabilecek olan müşterileri ile olan ilişkilerini bir sistem içinde yürütmektedirler. Müşteri ilişkileri yönetimi, bağlılık içeren ve uzun vadeli ilişkilerin oluşturulmasına, yönetimine ve güçlendirilmesine yönelik stratejileri ifade etmektedir. Müşteri ilişkileri yönetimi günümüzde, telekomünikasyon, bankacılık, sigortacılık ve perakendecilik sektöründe önemini artırmaktadır. Müşteri ilişkileri yönetiminin en önemli amaçlarından birisi, müşterilerin elde tutulmasını sağlayarak, müşteri kaybını engelleyecek çalışmaların gerçekleştirilmesidir. Günümüzde bu amaca yönelik, çok sayıda teknolojik uygulamadan yararlanılmaktadır. İşletmeler, günümüz şartlarında, çok farklı sebeplerden dolayı müşteri kaybı yaşayabilmektedir. Bu sebeplere sırasıyla, müşteri memnuniyetinde azalma, sert rekabet stratejileri, ürün geliştirmeleri ve hukuki düzenlemeler örnek verilebilmektedir. Makine öğrenimi uygulamalarından biri olan müşteri kaybı analizi ile model kurularak, farklı özniteliklere bağlı olarak, kaybedilme potansiyeli olan müşteriler tespit edilebilmektedir. Müşteri kaybı analizi, günümüzde başta telekomünikasyon, bankacılık, sigortacılık olmak üzere çoğu sektörde her geçen yıl işletmeler için önemini daha fazla artırmaktadır. Müşteri kaybı analizi için yapılan çalışmada, oluşturulan makine öğrenimi modellerinde, kullanılan algoritmalar, yapay sinir ağları, karar ağaçları, doğrusal regresyon, logistik regresyon, destek vektör makineleri, Naive Bayes, sepet analizi, doğrusal ayırma analizi olarak sıralanmaktadır (Vafeiadis, Diamantaras, Sarigiannidis, & Chatzisavvas, 2015).

Günümüzde makine öğreniminin kullanımının önem kazandığı alanlara madencilik ve jeoloji mühendisliği de örnek olarak verilebilmektedir. Bilgisayarlı görü ve makine öğreniminde denetimsiz öğrenme modelleri kullanılarak gerçekleştirilen çalışmalarda, kum tanelerinde yer alan mineraller farklı kümelere ayrılmıştır. Kum tanelerinde yer alan minerallerin içerikleri, optik mikroskop tarafından elde edilen görüntülerin, nümerik formatta bilgisayar ortamına aktarılması ve işlenmesi

sonrasında, bilgisayarlı görü modelleri ile süper pikseller değerlendirilerek mineraller tanımlanmıştır (Maitre, Bouchard, & Bédard, 2019).

Makine öğrenimi uygulamalarının günümüzde öneminin yükseldiği alanlardan biri de gıdadır. Gıda alanında, günümüzde yaşanan en büyük sorunlardan biri, soğuk zincir sistemlerinin performanslarında yaşanan düşüşler olarak belirtilmektedir. Soğuk zincir; bir gıda ürününün, üretim aşamasından başlanılarak, sevk edilme, depolanma ve son aşama olan tüketim aşamasına kadar geçen zaman aralığında, düşük sıcaklıklarda saklanılarak, korunmasını ifade etmektedir. Literatürde soğuk zincir sistemlerinde, gıda paletlerinin sıcaklıklarını tahmin etmeye yönelik makine öğrenimi modelleri kullanılarak yapılan çalışmalarda, sinir ağları algoritmalarından yararlanılmıştır (Loisel, ve diğerleri, 2022).

Makine öğreniminin gıda alanında uygulamalarına farklı bir örnek olarak, hiperspektral görüntülerin analiz edilerek, gıdaların kalitesinin belirlenmesi verilebilmektedir. Gıda ürünlerinin kalitesi, tüketicilerin tercihleri, işletmeciler ve sektörde yer alan diğer paydaşlar için büyük önem taşımaktadır. Makine öğrenimi, günümüzde tarım ve gıda ürünlerine yönelik çok fazla uygulamada kullanılmaktadır. Yapay sinir ağları, bulanık mantık, karar ağaçları, Naive Bayes, K-Ortalamalar, destek vektör makineleri, rassal orman ve K-En yakın komşuluk hem tarım hem de tarımla ilgili çok sayıda alandaki çalışmalarda kullanılan makine öğrenimi modellerindendir. Bunların yanında, makine öğreniminin bir alt dalı olan derin öğrenme de günümüzde kullanım alanını genişletmektedir. Günümüzde derin öğrenme algoritmalarından bilgisayarlı görü ve görüntü işleme amaçlı modellerde çok fazla yararlanılmaktadır. Literatürde, başta yapay sinir ağları olmak üzere farklı algoritmalar ile çok sayıda model kurulup gıdaların görüntülerinden yararlanılarak, kalitelerini tahmin etmeye yönelik çalışmalar yer almaktadır (Saha & Manickavasagan, 2021).

COVID-19 pandemisi sürecinde, dünya genelinde ülkelerin toplu yemek hizmetleri sistemlerinin büyük baskı altında kalması, sistemlerdeki zayıflıklar ve tüketicilerin hızla değişen tüketim kalıpları sebebiyle gıda sektöründe, atık gıdalar problemi yaşanmaya başlanmıştır. Kriz dönemlerinde kaynakların korunması ve en faydalı şekilde dağıtımı, büyük önem taşımaktadır.

2019 yılının Ağustos ayı ile 2020 yılının Haziran ayları arasındaki dönemde, İsveç'te halka açık yemek servisine katılacak insanların sayısını ve porsiyonları planlamak için gerçekleştirilen bir çalışmada, makine öğreniminde yer alan denetimli öğrenme modellerinden yararlanılmıştır. Çalışmada kullanılan verilerde dönüştürme işlemi uygulanmış olup, model kurulumunda rassal orman algoritması tercih edilmiştir (Malefors, Secondi, Marchetti, & Eriksson, 2022).

Makine öğreniminin günümüzde kullanıldığı alanlara farklı bir örnek olarak, e-ticaret uygulamaları yer almaktadır. Günümüzde internetin yaygınlaşması ile artan e-ticaret pazarları ve e-ticaret platformlarının yanı sıra, e-ticaret teknolojisini, geleneksel ticarete uygulayan işletmelerin sayısının da yükseldiği görülmektedir. Literatürde yer alan çalışmalara göre, e-ticaret alanında faaliyet gösteren işletmeler, stok takibi ve ürün-kategori eşleşmeleri için makine öğrenimi modellerinden yararlanmaktadırlar. İşletmeler, müşterilerinin siparişlerinde yer alan ürünlerine yönelik hem stok takibi hem de ürün-kategori eşleştirmeleri gerçekleştirmek amacıyla metin madenciliği tekniklerinden yararlanmaktadırlar. Literatürdeki çalışmalar incelendiğinde, doğal dil işleme modelinin, işletmelerin makine öğrenimi uygulamalarında en çok yararlandıkları modellerden biri olduğu görülmektedir (Kejriwal, Shen, Ni, & Torzec, 2021).

Makine öğrenimi türlerinden olan pekiştirmeli öğrenme de günümüzde kullanım alanını artırmaktadır. Pekiştirmeli öğrenme günümüzde sağlık sektöründe, insan kaynaklı hataları ve önyargıları etkisizleştirmek, eylem ve müdahale tavsiyeleri için en uygun politikaları oluşturmak amacıyla kullanılmaktadır. Pekiştirmeli öğrenme modelleri, insan benzeri öğrenme yaklaşımları sergilemektedirler. Modeller, elektronik ortamda bulunan sağlık kayıtları üzerinden tedavi oluşturup, tedaviyi uygulamakta başarılı sonuçlar elde etmektedirler. Pekiştirmeli öğrenme modelleri ayrıca, sağlık müdahalelerinde öneri oluşturabilmekte ve tavsiye sistemleri olarak çalışabilmektedirler. Pekiştirmeli öğrenme modelleri, sağlık görevlilerinin daha tutarlı kararlar almaları için destekte bulunarak, tedavi süreçlerinin gelişimine katkıda bulunmaktadırlar (Datta, ve diğerleri, 2021).

Pekiştirmeli öğrenmenin sağlık sektöründe uygulandığı alanlara örnek olarak, oftalmoloji verilmektedir. Tıpta görme yolları hastalıkları ve cerrahisiyle ilgilenen bilim dalı olan oftalmoloji alanında, pekiştirmeli öğrenme modelleri, kronik göz tutulumu hastalıklarının tedavisinde kullanılmaktadırlar. Bu kategorideki hastalıkların tedavisi, ekonomik anlamda fazla kaynak gerektirmesinin yanı sıra, uzmanların yoğun bilişsel emeklerini ve uzun zamanlarını da tüketmektedir. Bu sebeplerden dolayı, pekiştirmeli öğrenme modelleri, hastaların tedavisi ve klinik kaynakların değerlerinin eşitliğini korumak için kullanılmaktadırlar (Nath, ve diğerleri, 2022).

Pekiştirmeli öğrenme günümüzde tarım alanında, mahsul yönetimi için destek sağlaması amacıyla kullanılmaktadır. Pekiştirmeli öğrenme, belirsiz ve bilinmeyen durumlarda, ardışık karar verme işlemleri için pratik ile öğrenme metoduyla çalışmaktadır. Pekiştirmeli öğrenme mahsul yönetimine destek sağlamak için, gerçek dünya şartlarını simüle etmektedir. Her bir mahsul türünün kendine özgü sınırları olmaktadır. Pekiştirmeli öğrenme modellerinde, tasarımlarının tam tersi bir şekilde çalışma yolu izleyecek olan, simülasyondan öğrenilen modeli geliştirmek için, gerçek hayatta çiftçilerden elde edilecek olan, geri bildirimlerin değerlendirilmesi gerekmektedir (Gautron, Maillard, Preux, Corbeels, & Sabbadin, 2022).

Finansal piyasalarda yatırımcılar, çok sayıda farklı yatırım aracına yönelik, fiyat tahminleri ve fiyat hareket yönü tahminleri gerçekleştirirken makine öğrenimi modellerinden yararlanmaktadırlar. Finansal piyasalarda, makine öğrenimi modellerinden destek alarak işlemlerde bulunmanın zayıf noktalarından birisi, yanlış finansal işlem sinyalleridir. Yanlış finansal işlem sinyalleri, finansal piyasalarda yatırımcıların, yanlış kararlar almalarına neden olarak, zarar etmelerine sebep olmaktadır. Son dönemde yapılmış olan bir araştırmada, finansal piyasalardaki söz konusu olumsuz durumlara karşılık, tahminlemelerde bulunacak model geliştirilmiştir. Araştırmada geliştirilen model, finansal piyasalarda seyreden grafik modeller ile karşılaştırıldığında, çok farklı harekete sahip olan finansal hareketleri tespit edecek şekilde çalışmaktadır. Modelin oluşturulmasında, başta denetimli makine öğrenimi modelleri ve derin öğrenme modelleri olmak üzere, farklı otoregresif teknikler ile örüntü tanıma teknikleri kullanılmıştır. Araştırmada önerilen yaklaşım, yatırım getirisi açısından incelenerek değerlendirilmiştir (Cagliero, Fior, & Garza, 2023).

İKİNCİ BÖLÜM

DENGESİZ VERİ SETLERİ PROBLEMİ

2.1 Dengesiz Veri Setleri

Makine öğreniminde sınıflandırma algoritmaları ile kurulan modeller, veri setleri üzerinde çalıştırılıp, veri setinde yer alan gözlemlerin sınıflarına yönelik tahminler elde edilmektedir. Sınıflandırma algoritmaları genel olarak, veri setinde yer alan hedef özniteliğinin dağılımının dengeli olduğunu varsaymaktadır. Bu varsayımdan hareketle sınıflandırma algoritmaları ile kurulan modeller, performans ölçütlerinden biri olan doğruluk ölçütünü, en yüksek seviyede elde etmeyi amaçlamaktadır. Günlük hayatta elde edilen verilerden oluşan veri setlerinde, hedef özniteliğinin sınıf dağılımı çoğunlukla dengeli olmamaktadır. Dengesiz veri setleri problemi, sınıflandırma uygulamalarında, veri setindeki sınıf dağılımının eşit veya yakın olmadığı durumlarda gözlemlenmektedir. Dengesiz veri setleri problemi, gözlem sayısı olarak çoğunluğa sahip olan sınıfın, azınlık sınıfı baskılaması sonucu ortaya çıkmaktadır. Oluşturulan model çoğunluğa sahip sınıfın özniteliklerine daha fazla yanlılık göstermekte ve bu durum, modelin azınlık sınıfa yönelik sınıflandırma performansını olumsuz etkilemektedir.

Teknik olarak, hedef özniteliğindeki sınıf dağılımı eşit olmayan her veri seti, dengesiz veri seti olarak nitelendirilebilmektedir. Günlük hayatta olan gerçek uygulamalarda ise dengesiz veri seti kavramı daha çok, sınıf dağılımının anlamlı olarak farklılık gösterdiği, oransal olarak çok yüksek farkların belirgin bir şekilde gözlemlendiği veri setlerini tanımlamaktadır. Dengesiz veri setleri problemi ile günümüzde, makine öğreniminin kullanıldığı çok farklı uygulama alanlarında karşılaşılmaktadır. Bu alanlara örnek olarak, hata tespitleri, anomali tespitleri, yüz tanıma uygulamaları, e-posta klasörleme gösterilmektedir (Fernández, ve diğerleri, 2018).

Bu alanların dışında, dengesiz veri seti problemi ile karşılaşılan alanlara örnek olarak, başta kanser olmak üzere hastalık teşhisleri, bankacılık işlemlerindeki sahtekarlık tespitleri ve internet reklamlarının tıklama oranları verilebilmektedir.

2.2 Dengesiz Veri Setleri Problemini Çözüm Yöntemleri

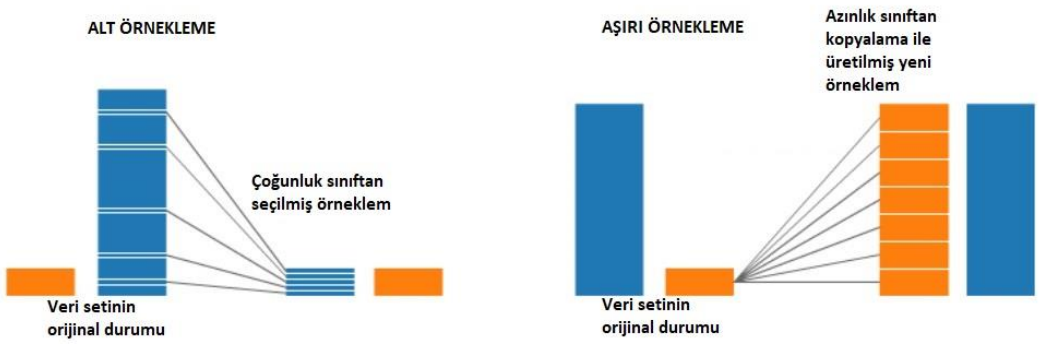
Dengesiz veri setleri problemini çözmek için günümüzde farklı yöntemlerden yararlanılmaktadır. Bu yöntemler, günlük hayatta yer alan uygulamalarda, veri seti seviyesinde çözüm yöntemleri ve maliyet duyarlı çözüm yöntemleri olarak iki bölüme ayrılmaktadırlar. Bu yöntemlerin dışında, algoritma seviyesinde çözüm yöntemlerinden de farklı uygulamalarda yararlanılmaktadır.

2.2.1 Veri Seti Seviyesinde Çözüm Yöntemleri

Dengesiz veri setlerinde, veri seti seviyesinde çözüm yöntemlerinin temel çalışma prensibi, veri setinde yeniden örnekleme işlemi uygulanarak, hedef özniteliğinin sınıf dağılımının dengelenmesi şeklindedir. Bu yöntemler; alt örnekleme ve aşırı örnekleme olarak ikiye ayrılmaktadır. Bu yöntemlere ek olarak, gerçek hayat uygulamalarında farklı gelişmiş yöntemler de kullanılmaktadır.

Alt örneklemenin ve aşırı örneklemenin çalışma şekli Grafik 2.1.'da görülmektedir (KDNuggets, 2023).

Grafik 2.1.: Alt Örnekleme ve Aşırı Örnekleme Yöntemlerinin Çalışma Şekli



Alt örnekleme yöntemi, kendi içinde farklı alt yöntemlere ayrılmaktadır. En çok kullanılan yöntemler; Rassal Alt Örnekleme, Küme Merkezleri, Tomek Bağlantıları, Düzenlenmiş En Yakın Komşular Kuralı ve Yakın Eksik Örnekleme

(Near Miss) olarak sıralanmaktadır. Bu yöntemlerin dışında, Yoğunlaştırılmış En Yakın Komşular Kuralı, Tek Taraflı Seçim, Bölge Temizleme Kuralı, Sınıf Saflığının Maksimizasyonu şeklinde farklı yöntemler de kullanılabilmektedir (Fernández, ve diğerleri, 2018).

Bu yöntemler içinde, Yakın Eksik Örnekleme ve Yoğunlaştırılmış En Yakın Komşu Kuralı yöntemleri, örneklem seçimlerini dönüştürülmüş veri setleri üzerinden yapmaktadırlar. Tomek Bağlantıları yöntemi bilgi kaybını en düşük seviyede tutacak şekilde çalışarak, gözlem sayısı olarak çoğunluğa sahip olan sınıfın, gözlem sayılarını azaltacak şekilde çalışmaktadır. Tek Taraflı Seçim ve Bölge Temizleme Kuralı, hibrit şekilde çalışmakta olan yöntemlerdir.

Bu yöntemler dışında, ileri seviye alt örnekleme yöntemleri olarak adlandırılan yöntemler de günümüzde kullanılmaktadır. İleri seviye alt örnekleme yöntemlerine örnek olarak; ACO, CBEUS, Ağırlıklı Örnekleme, Gözlem Sertlik Eşiği, IRUS, ClusterOSS ve DSUS olmak üzere farklı yöntemler verilmektedir.

Aşırı Örnekleme Yöntemi de Alt Örnekleme Yöntemi'ne benzer olarak, kendi içinde farklı alt yöntemlere ayrılmaktadır. En çok kullanılan yöntemler; Rassal Aşırı Örnekleme ve Sentetik Azınlık Aşırı Örnekleme (SMOTE) olarak sıralanmaktadır. SMOTE yönteminin geliştirilmesi ile; Sınırlandırılmış SMOTE, ADOMS, ADASYN, ROSE, Güvenli-Seviye SMOTE, DBSMOTE, MWMOTE, MDO gibi farklı ileri aşırı örnekleme yöntemleri elde edilmiştir. Bu yöntemler haricinde, günlük hayatta çok fazla kullanılmayan, LLE-SMOTE, CBSO, DSRBF, LVQ-SMOTE, BKS, PDFOS, OUPS, CURE-SMOTE gibi farklı aşırı örnekleme yöntemleri de literatürde yer almaktadır.

2.2.1.1 Alt Örnekleme Yöntemleri

Alt örnekleme yöntemlerinde uygulanan genel yaklaşım, çoğunluk sınıfa ait gözlemlerin veri setinden rassal veya belirli hesaplamalara ve işlemlere dayalı olarak çıkarılarak, hedef özniteliğindeki sınıf dağılımının dengelenmesidir. Alt örnekleme yaklaşımı, veri setinden gözlem silinmesi şekliyle çalışmaktadır.

2.2.1.1.1 Rassal Alt Örnekleme

Rassal Alt Örnekleme yönteminde, veri setinin hedef özniteliğinde, çoğunluk sınıfa ait olarak yer alan gözlemlerin, tamamen rassal seçilerek, veri setinden çıkarılması yoluyla, hedef özniteliğindeki sınıf dağılımının dengelenmesi amaçlanmaktadır.

2.2.1.1.2 Tomek Bağlantıları Yöntemi

Tomek Bağlantıları yöntemi, 1976 yılında, Ivan Tomek tarafından yayınlanan makalede, Yoğunlaştırılmış En Yakın Komşu Kuralı'nın değiştirilmiş bir sürümü olarak tanıtılmıştır. Tomek Bağlantıları yönteminde, öncelikle veri setinde yer alan gözlemler, hedef özniteliğinde yer alan her bir sınıftan en yakın Öklid uzaklığına sahip başka gözlemlerle ikililer hâline getirilmektedir. Sonrasında, Tomek bağlantıları, ikililer hâline gelen bu gözlem çiftlerini tespit ederek, çoğunluk sınıfına ait gözlemleri veri setinden silmekte ve bu şekilde hedef öznitelikteki sınıf dağılımının dengelenmesini amaçlamaktadır (Tomek, 1976).

2.2.1.1.3 Düzenlenmiş En Yakın Komşular Kuralı

Düzenlenmiş En Yakın Komşular Kuralı, 1972 yılında Dennis L. Wilson tarafından geliştirilen ve veri setinde yer alan, gürültülü gözlemleri silmek için kullanılan bir yöntemdir. Düzenlenmiş En Yakın Komşular Kuralı, veri setinden en düşük seviyede bilgi kaybı yaratacak şekilde, hedef özniteliğinde yer alan çoğunluk sınıfına dahil gözlemleri silerek veriden çıkartmak ve bu şekilde, hedef özniteliğindeki sınıf dağılımını dengeli hâle getirmek amacıyla çalışmaktadır. Düzenlenmiş En Yakın Komşular Kuralı için oluşturulan modelde k hiperparametresi genellikle 3 değerini almaktadır. Oluşturulan K-En Yakın Komşular sınıflandırma modeli, veri setinin hedef özniteliğinde yanlış sınıfa atanmış olan tüm gözlemleri silecek şekilde çalışmaktadır (Wilson, 1972).

2.2.1.1.4 Yoğunlaştırılmış En Yakın Komşu Kuralı

Yoğunlaştırılmış En Yakın Komşu Kuralı yöntemi, dengesiz T veri setinden, T veri seti ile tutarlı olan yani, en yakın 1 komşu modeline göre T'yi yüksek doğruluk

oranı ile sınıflayan S alt kümesini seçmek için kullanılmaktadır. Bu yöntem, gürültülü verilere çok duyarlıdır (Dal Pozzolo, Caelen, Waterschoot, & Bontempi, 2013).

2.2.1.1.5 Tek Taraflı Seçim

Tek Taraflı Seçim yöntemi, 1997 yılında Miroslav Kubat ve Stan Matwin tarafından hibrit bir yöntem olarak geliştirilmiştir. Tomek Bağlantıları ve Yoğunlaştırılmış En Yakın Komşu Kuralı yönteminin birleştirilmiş şekilde, sırayla uygulanmasıyla çalıştırılmaktadır. Tek Taraflı Seçim yönteminde, öncelikle Tomek Bağlantıları sınıflandırma için karar sınırlarında yer alan verileri ve gürültülü olarak belirlenen verileri veri setinden silmektedir. Sonrasında kalan verilerin oluşturduğu veri setinin yeni alt kümesinde, karar sınırlarından uzakta yer alan çoğunluk sınıfına ait gözlemlerin veri setinden Yoğunlaştırılmış En Yakın Komşu Kuralı yöntemi kullanılarak silinmesi şeklinde çalışmaktadır (Kubat & Matwin, 2000).

2.2.1.1.6 Bölge Temizleme Kuralı

Bölge Temizleme Kuralı, Tek Taraflı Seçim yöntemine göre, veri setinde gözlem azaltmaktansa, daha fazla veri temizleme işlemini öne çıkaran bir hibrit yöntemdir. Bu yöntemde, sınıflandırma işleminin sonuçlarının kalitesi, sınıfların boyutlarına bağımlı olmamaktadır. Bu nedenle, Bölge Temizleme Kuralı yönteminde, öncelikle hedef öznitelikteki sınıf dağılımı yerine, veri setinin başta gürültü olmak üzere, sınıflandırmayı olumsuz etkileyecek başka karakteristik özellikleri incelenmelidir. Veri setinde yapılacak olan veri temizleme işlemleri sırasında, eksik verilerin silinmesine karar verildiğinde, modelin doğruluk ölçütünün değerinin yüksek olması da amaçlanmalıdır.

Bölge Temizleme Kuralı'nda veri setinde yer alan her bir gözlem için, o gözleme en yakın üç komşu gözlem belirlenmektedir. Seçilen bir gözlem, hedef özniteliğinde yer alan çoğunluk sınıfına dahilse ve kendisine en yakın üç komşu gözlem tarafından yanlış sınıf tahmini gerçekleştirilerek yanlış sınıfa ait olduğu belirtilirse, o gözlem veri setinden çıkarılmaktadır. Seçilen gözlemin, hedef özniteliğinde yer alan azınlık sınıfına dahil olması ve kendisine en yakın üç komşu gözlem tarafından yanlış sınıflandırılması durumunda, komşularında yer alan hedef

özniteliğin çoğunluk sınıfına ait gözlemler veri setinden çıkarılmaktadır (Laurikkala, 2001).

2.2.1.2 Aşırı Örnekleme Yöntemleri

Aşırı örnekleme yöntemlerinde uygulanan genel yaklaşım, veri setinde azınlık sınıfa ait gözlemlerin sayısının farklı yöntemler kullanılarak artırılması ile hedef özniteliğindeki sınıf dağılımının dengelenmesidir.

2.2.1.2.1 Rassal Aşırı Örnekleme

Rassal Aşırı Örnekleme yönteminde, veri setinin hedef özniteliğinde, azınlık sınıfa ait olarak yer alan gözlemlerin, tamamen rassal seçilip, seçilmiş olan gözlemlere benzer çok sayıda yeni gözlemler oluşturulmasıyla, hedef özniteliğindeki sınıf dağılımının dengelenmesi amaçlanmaktadır. Bu yöntemin kullanılmasında, rassal durumun korunması amaçlanarak, gözlemler orijinal veri setinden seçilmektedir.

2.2.1.2.2 Sentetik Azınlık Aşırı Örnekleme

Sentetik Azınlık Aşırı Örnekleme yöntemi Nitesh V. Chawla tarafından 2002 yılında geliştirilmiştir. Bu yöntemde, azınlık sınıfa ait gözlemlerin sayısı sentetik olarak artırılarak, veri setinin hedef özniteliğindeki sınıf dağılımının dengelenmesi amaçlanmaktadır. (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

Sentetik Azınlık Aşırı Örnekleme yönteminin çalışma aşamaları; hedef özniteliğinde azınlık sınıfta yer alan gözlemlerin, K-En Yakın Komşuluk algoritması üzerine kurulan modelle, en yakın komşularının hesaplanması, sentetik gözlemlerin üretilmesi, sentetik birimlerin veri setine eklenmesi, örnekleme oranının sağlandığının kontrol edilmesi ve oran sağlandıysa işlemin sonlandırılması şeklinde sıralanmaktadır. Oran sağlanmadıysa ilk aşamaya dönülmektedir.

Sentetik Azınlık Aşırı Örnekleme yönteminde, veri setinde K-En Yakın Komşuluk algoritmasının yardımıyla, kendilerine en yakın komşu olan gözlemlerin öznitelikleri ve hedef özniteliğinde yer alan azınlık sınıfa ait gözlemlerin dikkate alınmasıyla, sentetik yeni gözlemler oluşturulmaktadır. Sentetik Azınlık Aşırı Örnekleme yönteminde, girdi veri setinin bütünü olarak gerçekleşmekteyken, çıktı

olarak sadece veri setinin hedef özniteliğinde yer alan azınlık sınıfa ait gözlemlerin sayısı artmaktadır.

Sentetik Azınlık Aşırı Örnekleme yönteminin kendi içinde çok sayıda farklı alt yöntemleri geliştirilmiştir. Bu alt yöntemlerin bazıları, bölümün başında örnek olarak verilmiştir. Verilen örneklere ek olarak olarak; Distance-SMOTE, Polynom-Fit-OS, Edge-Det-SMOTE, NRSBoundary-SMOTE, RWO-Sampling, NEATER, SMOTE-FRST-2T yöntemleri örnek verilebilmektedir (Fernández, ve diğerleri, 2018).

2.2.1.2.3 Uyarlanabilir Sentetik Örnekleme

Uyarlanabilir Sentetik Örnekleme yöntemi 2008 yılında Haibo He tarafından geliştirilmiştir. Sentetik Azınlık Aşırı Örnekleme yöntemini temel alarak çalışmaktadır (He, Bai, Garcia, & Li, 2008).

Uyarlanabilir Sentetik Örnekleme yönteminin, Sentetik Azınlık Aşırı Örnekleme yönteminden farkı; veri setinin hedef özniteliğinde yer alan azınlık sınıfına ait gözlemlerin düşük yoğunlukta olduğu alanlarda gözlemler oluşturarak, azınlık sınıfına ait gözlemlerin sayısını artırmak için, çoğunluk sınıfına ait gözlemlerin yüksek yoğunlukta olduğu alanlarda yer alan azınlık sınıfına ait gözlemleri saptayan bir yöntem kullanmasıdır.

Uyarlanabilir Sentetik Örnekleme yöntemi, veri setinin hedef özniteliğinde yer alan azınlık sınıfına ait gözlemler için öğrenmedeki zorluk derecelerine bağlı kalıp, ağırlıklı bir dağılım kullanarak, azınlık sınıfın gözlemlerine eklemek üzere sentetik gözlemler üretmektedir. Uyarlanabilir Sentetik Örnekleme yöntemine alternatif olarak, K-Ortalamalar Sentetik Azınlık Aşırı Örnekleme yöntemi de kullanılabilmektedir.

2.2.2 Maliyet Duyarlı Çözüm Yöntemleri

Dengesiz veri setlerinde, maliyet duyarlı çözüm yöntemlerinin temel çalışma prensibi, çözülmek istenen problemin belirli karakteristik özellikleri ile ilişkili olarak ortaya çıkan farklı maliyetlere duyarlı algoritmalar seti ile hedef özniteliğinin sınıf dağılımının dengelenmesidir. Maliyet duyarlı çözüm yöntemlerinde, söz konusu olan

maliyet kavramı; özniteliklerle ilişkili olan maliyetler ve veri setinde yer alan gözlemlerin, hedef özniteliğindeki sınıfları ile ilişkili olan maliyetler olmak üzere ikiye ayrılmaktadır. Özniteliklerle ilişkili olan maliyetler, finansal maliyetler, zaman maliyeti ve farklı zorluklardan kaynaklanabilmektedir. Maliyet duyarlı çözüm yöntemlerinin temel çalışma prensibi, maliyetleri mümkün olabilecek en düşük seviyede tutabilirken, aynı zamanda en iyi tahmin performansına sahip olan modeli oluşturmaktır. Veri setinde yer alan gözlemlerin, hedef özniteliğinde yer alan sınıfları ile ilişkili olan maliyetler söz konusu olduğunda ise, maliyet duyarlı çözüm yöntemlerinin varsayımı, hedef özniteliğinde belirli sınıflarda yer alan gözlemler ile hatalı model oluşturulmasının, yüksek maliyete yol açacağı şeklinde olduğu belirtilmektedir (Fernández, ve diğerleri, 2018).

Maliyete duyarlı öğrenme, modelin gerçekleştirdiği yanlış sınıflandırmaların maliyetlerini dikkate alan bir öğrenme türüdür. Modelin gerçekleştirdiği her yanlış sınıf tahminine bir maliyet değeri atanmaktadır. Bu tür öğrenmenin amacı, toplam maliyeti en düşük seviyeye indirgemektir (Sammut & Webb, 2010).

Maliyet duyarlı çözüm yöntemleri içinde en çok kullanılan modeller; başta maliyet duyarlı bölme ve gözlem ağırlıklandırma ile meta öğrenme yaklaşımı olarak iki farklı alt dala ayrılan, maliyet duyarlı karar ağaçları olmak üzere, destek vektör makineleri, yapay sinir ağları, en yakın komşuluk ve hibrit modeller olarak sıralanmaktadır (Fernández, ve diğerleri, 2018).

Maliyet duyarlı çözüm yöntemlerinde, gözlemlerin sınıf tahmini için gerçekleştirilen her bir yanlış tahminde, hataların türüne göre farklı maliyetler hesaplanmaktadır. Maliyetler, karmaşıklık matrisine benzer olan maliyet matrisi ile özetlenebilmektedir. Toplam maliyet; yanlış negatif ve yanlış pozitif değerlerinin, kendilerine ait maliyetleri ile çarpımlarının toplamı şeklinde hesaplanmaktadır.

2.3 Dengesiz Veri Setlerinde Performans Değerlendirme Ölçütleri

Hedef özniteliğinin sınıf dağılımı dengesiz olan veri setlerinde, modellerin performans değerlendirme ölçütlerinin doğru bir şekilde elde edilmesi ve bu ölçütlerden yola çıkılarak modelin doğru bir şekilde değerlendirilmesi kolaylıkla gerçekleştirilememektedir.

Sınıflandırma problemlerinde en çok kullanılan performans değerlendirme ölçütlerinden biri olan doğruluk oranı, dengesiz veri setlerinin hedef özniteliğinde, azınlık sınıfına dahil olarak yer alan gözlemlere yönelik tahminlerde çok zayıf kalmaktadır. Bu nedenle dengesiz veri setlerinde performans değerlendirme ölçütü olarak sadece doğruluk oranının kullanılması, modelin performansının değerlendirilmesini olumsuz etkilemektedir.

Dengesiz veri setleri üzerinde oluşturulup uygulanan makine öğrenimi modellerinin performansları değerlendirilirken, doğruluk oranı kullanışlı olmamaktadır. Bu duruma başta, performans değerlendirme ölçütlerinden biri olan doğruluk oranının, dengesiz veri setlerinde oluşturulup uygulanan modellerde çok kolay şekilde yüksek değere sahip olacak olması neden olmaktadır. Benzer durumda modelin hata oranı da çok düşük şekilde görülecektir. Doğruluk oranının kullanışlı olmama sebeplerinden bir diğeri de modelin, gerçekleştirdiği sınıf tahminlerinde, hatalı tahminlerin maliyetlerinin eşit olduğunu varsayması yer almaktadır. Lâkin dengesiz veri setlerinin sınıflandırılmasında, veri setinin hedef özniteliğinde yer alan azınlık sınıfa ait bir gözlemin sınıfının, hatalı bir şekilde tahmin edilerek sınıflandırılması, çoğu zaman, aynı veri setinin hedef özniteliğinde yer alan çoğunluk sınıfa ait bir gözlemin sınıfının, hatalı bir şekilde tahmin edilerek sınıflandırılmasından daha maliyetli sonuçlar ortaya çıkarmaktadır (Fernández, ve diğerleri, 2018).

Dengesiz veri setleri üzerinde oluşturulup uygulanan makine öğrenimi modellerinin performansları değerlendirilirken, kesinlik oranı, yanlış pozitif değerini en düşük seviyeye indirgemek amaçlandığında kullanışlı olmaktadır.

Model performans metriklerinden biri olan duyarlılık oranı, doğru pozitif olarak tahmin edilen gözlem sayısının, doğru pozitif ve yanlış negatif olarak tahmin edilen gözlemlerin toplam sayısına oranıdır. Yanlış negatif değerini en düşük seviyeye indirgemek amaçlandığında kullanışlı olmaktadır. Duyarlılık değeri, yanlış negatif olarak tahmin etme işleminin maliyetinin yüksek olduğu durumlarda kullanılabilmektedir.

Model performans değerlendirilmesinde kullanılan bir diğer ölçüt olan F skoru, kesinlik ve duyarlılık ölçütlerinin değerlerinin harmonik ortalamasının elde edilmesiyle bulunur.

F skoru hem kesinlik hem de duyarlılık oranlarını birleştirerek tek bir değer haline getirmektedir. Kesinlik değeri veya duyarlılık değeri, tek oldukları durumda, modelin performansı ile ilgili kullanışlı çıkarımlarda bulunulmasına yardımcı olmamaktadır. F skoru, iki farklı ölçütü, tek bir değer hâline getirerek sunmaktadır.

F skoru, veri setinin hedef özniteliğinde yer alan pozitif sınıfın analizine odaklanmaktadır. F skorunun hesaplanma amacı, pozitif gözlemlerin sınıflandırılmasında, doğruluk ve kapsayıcılık arasındaki ödünleşimleri analiz etmektir. Bu amaçla, kesinlik oranının ve duyarlılık oranının ağırlıklı harmonik ortalamasını kullanmaktadır (Fernández, ve diğerleri, 2018).

F skorunun en yüksek değeri olan 1 değeri, aynı zamanda modelin, en yüksek kesinlik oranı değerine ve en yüksek duyarlılık oranı değerine de sahip olduğunu göstermektedir. F skorunun en düşük değeri olan 0 değeri ise, aynı zamanda modelin, 0 kesinlik oranı değerine ve 0 duyarlılık oranı değerine de sahip olduğunu göstermektedir.

F skoru, kesinlik oranı ile duyarlılık oranını dengelemektedir. Bazı durumlarda, F skorunun kesinlik oranından ve duyarlılık oranından birine daha fazla ağırlık vermesi istenildiğinde, F Beta skoru kullanılabilmektedir. F Beta skoru, F skorunun, hesaplamalar gerçekleşirken harmonik ortalamada kesinlik oranının ve duyarlılık oranının bir beta (β) katsayısı tarafından kontrol edildiği versiyonudur.

Dengesiz veri setleri üzerinde oluşturulup uygulanan makine öğrenimi modellerinin performanslarının değerlendirilmesinde, yukarıda açıklanan performans metriklerinin dışında, Matthews Korelasyon Katsayısı, ROC eğrisi, eğrinin altında kalan alan, Kappa değeri gibi performans metrikleri de kullanılmaktadır.

2.4 Literatür Taraması

Bir veri setinde yer alan gözlemlerin, hedef özniteliklerindeki sınıf dağılımlarının eşit olmaması durumunda, söz konusu veri seti, dengesiz olarak

nitelendirilebilmektedir. Günlük hayatta gerçekleştirilen uygulamalarda ise, gözlemlerinin sınıf dağılımlarının, oransal olarak çok yüksek seviyede farklılık gösterdiği veri setleri dengesiz veri setleri olarak adlandırılmaktadır.

Günlük hayatta dengesiz veri setleri ile hemen her alanda karşılaşılmaktadır. Bu alanlara; finans, bankacılık, mühendislik, bilgi teknolojileri, biyoinformatik, tıp, işletme, yönetim, güvenlik, eğitim gibi örnekler verilebilmektedir.

Dengesiz veri seti problemleri, finans sektöründe, sahtecilik tespiti için üzerinde çalışılan veri setlerinde sıklıkla görülmektedir. 2018 yılının Ocak ile Haziran ayları arasında 33305 kredi kartı sahtecilik işlemi raporlanmıştır. 2019 yılında, kredi kartlarındaki sahtecilik işlemlerini tespit etmeye yönelik geliştirilen uygulamanın yer aldığı bir araştırmada kullanılan veri setinde, kredi kartı işlemlerini temsil eden gözlemlerin, yüzde 5.96'sının hedef öznitelikteki değeri sahte olarak görülmüştür. Bu durumda dengesiz veri seti probleminin varlığı ortaya çıkarılmıştır. Dengesiz veri seti problemini çözmek için, Rassal Aşırı Örnekleme yöntemi ile Maliyet Duyarlı C5.0 ve Maliyet Duyarlı Destek Vektör Makineleri modellerinden yararlanılmıştır (Makki, ve diğerleri, 2019).

Dengesiz veri seti problemleri, işletmelerin müşteri kaybına yönelik olan veri setlerinde de görülmektedir. 2017 yılında yayınlanan araştırmada, kullanılan müşteri kaybı veri setinin hedef özniteliğinin, sınıf dağılımı açısından dengesiz olduğu görülmüştür. Hedef özniteliğinin sınıf dağılımındaki bu dengesiz durumun çözülmesine yönelik olarak, örnekleme yöntemi bazında; Uyarlanabilir Sentetik Örnekleme yöntemi, Küme Bazlı Alt Örnekleme yöntemi, Rassal Alt Örnekleme, Rassal Aşırı Örnekleme, Sentetik Azınlık Aşırı Örnekleme, Tomek Bağlantıları ile Sentetik Azınlık Aşırı Örnekleme yöntemleri ile maliyet duyarlı yöntemler bazında; Maliyet Duyarlı Karar Ağaçları modelleri ve Maliyet Duyarlı Destek Vektör Makineleri modelleri olmak üzere çok sayıda yöntem kullanılmıştır (Zhu, Baesens, & vanden Broucke, 2017).

Dengesiz veri seti problemine sahip olan gerçek hayat verileri ve problemi çözmeye yönelik uygulamalara farklı bir örnek olarak, 1998 yılında Kanada'da yapılan bir çalışmada, deniz yüzeyinde, gemi kaynaklı petrol sızıntılarının tespiti için

RADARSAT ve ERS-1 uyduları tarafından çekilmiş deniz yüzeyi fotoğraflarından yola çıkılarak model kurulmuştur. Modelde veri olarak kullanılan uydu görüntülerinde petrol sızıntılarının, deniz yüzeyinde daha az yansımaya sahip olduğu ve bu nedenle kolayca tespit edildiği görülmüştür. Lâkin deniz yüzeyindeki yansımanın düşük seviyede olması, yağmur, rüzgâr ve yosun birikmesi gibi petrol sızıntısı dışında kalan nedenlerle de ilgili olabilmektedir. Petrol sızıntısı dışında kalan nedenlerle ilişkili olduğu düşünülen bu görüntülerin temsil edildiği gözlemlerin, veri setinde, hedef özniteliklerindeki değerleri sızıntıya benzer olarak yer almaktadır. Geriye kalan görüntülerin temsil edildiği gözlemlerin de hedef özniteliklerindeki değerleri sızıntı değeri olarak yer almakta olup, veri setinin hedef öznitelik dağılımı, 896 sızıntıya benzer, 41 sızıntı şeklinde gerçekleşmiştir. Bu durumda veri setinde dengesiz veri seti problemi görülmüş ve problemin çözümü için Tomek Bağlantıları'nın değiştirilmiş şekli olan Tek Taraflı Seçim yöntemi uygulanıp, sonrasında K-En Yakın Komşuluk modeli ile sınıflandırma gerçekleştirilmiştir (Kubat, Holte, & Matwin, 1998).

Sağlık sektöründe toplanan verilerde de dengesiz veri seti problemi görülmektedir. 2014 yılında akciğerlerde bulunan nodüllerin teşhisi ve tedavisi için hastalar üzerinde tıbbi görüntüleme ve teşhis edici radyoloji işlemleri ile toplanan görüntülerden oluşturulan veri setinde, dengesiz veri seti problemi görülmüştür. Dengesiz veri seti problemini çözmek için hem alt örnekleme hem de aşırı örnekleme yöntemlerinden yararlanılmıştır (Cao, Yang, Li, Zhao, & Zaiane, 2014).

Günümüzde makine öğrenimi ve alt dallarından olan derin öğrenmenin, hayatlarımızdaki kullanım payı artmaktadır. Makine öğreniminin alt dalı olan derin öğrenmenin kullanıldığı alanlardan birisi de yüz tanıma uygulamalarıdır. Video gözetiminde, yüz tanıma uygulaması için 2014 yılında Kanada'da yapılan bir araştırmada, kullanılan veri setinin hedef özniteliğindeki sınıf dağılımının dengesiz olduğu görülmüştür. Dengesiz veri seti problemini çözmeye yönelik olarak, Eğim Duyarlı Mantıki Kombinasyonlar yöntemi geliştirilmiştir (Radtke, Granger, Sabourin, & Gorodnichy, 2014).

ÜÇÜNCÜ BÖLÜM

UYGULAMA

3.1 Uygulamaların Kapsamları ve Amaçları

Uygulama bölümünde, 3 farklı veri seti ile yapılmış çalışmalara ait bilgiler ve elde edilmiş çıktılar ile bu çıktılara yönelik yorumlar yer almaktadır. Bu çalışmada, Brezilya'da faaliyet gösteren araç ve konut sigorta şirketine ait veri seti, beyin felci hastalarının bilgilerini içeren veri seti ve felçli askeri gaziler için bir bağış toplama talebinde, bağış yapan donörlerin bilgilerini içeren veri seti kullanılmıştır. Tüm veri setlerinin, dengesiz veri seti olduğu görülmüştür.

3.1.1 Porto Seguro Sigorta Şirketi Verisine ait Dengesiz Veri Seti Uygulaması

Uygulama kısmı için ilk olarak, Brezilya menşeli Porto Seguro adlı sigorta ve reasürans şirketinin paylaşmış olduğu veri seti kullanılmıştır. Porto Seguro sigorta ve reasürans şirketi, Brezilya'nın en büyük şehri olan São Paulo şehrinde, 1945 yılında kurulmuştur. Porto Seguro sigorta ve reasürans şirketi, günümüzde 14000'den fazla çalışan ve 100'den fazla şube ile 18 milyondan fazla müşterisine hizmet vermektedir. Şirketin sigortacılık faaliyet alanları arasında; otomobil sigortaları, şirket sigortaları, hayat sigortaları, miras sigortaları ve ulaşım sigortaları yer almaktadır.

Literatür incelemesi yapıldığında, sigortacılık sektöründe makine öğrenimi uygulamalarının, farklı kullanım alanlarına sahip oldukları görülmüştür. Bu kullanım alanlarından öncelikle müşteri kaybı tahmini olmak üzere, ödeme tutarları tahmini, dolandırıcılık tespitleri gibi çok sayıda farklı örneklerin, literatürdeki çalışmalarda yer aldığı görülmüştür. Bu çalışmaların büyük bölümünde, makine öğrenimi modelleri kullanılarak, sigorta şirketlerinin gerçekleştirmiş oldukları faaliyetlere yönelik sınıflandırma uygulamalarının geliştirilmiş olduğu görülmüştür.

Çalışmada kullanılan veri seti 2017 yılında bir veri bilimi yarışmasında açık erişime sunulmuştur. Veri seti incelendiğinde, temel olarak veri setinin, eğitim seti ve test seti olarak iki farklı veri seti şeklinde sunulduğu görülmüştür. Veri setinde yer alan

her bir gözlem, farklı bir poliçe sahibine ait verileri içermektedir. Veri setinin eğitim setinde 595212 gözlem bulunmaktayken, test setinde 892816 gözlem bulunmaktadır.

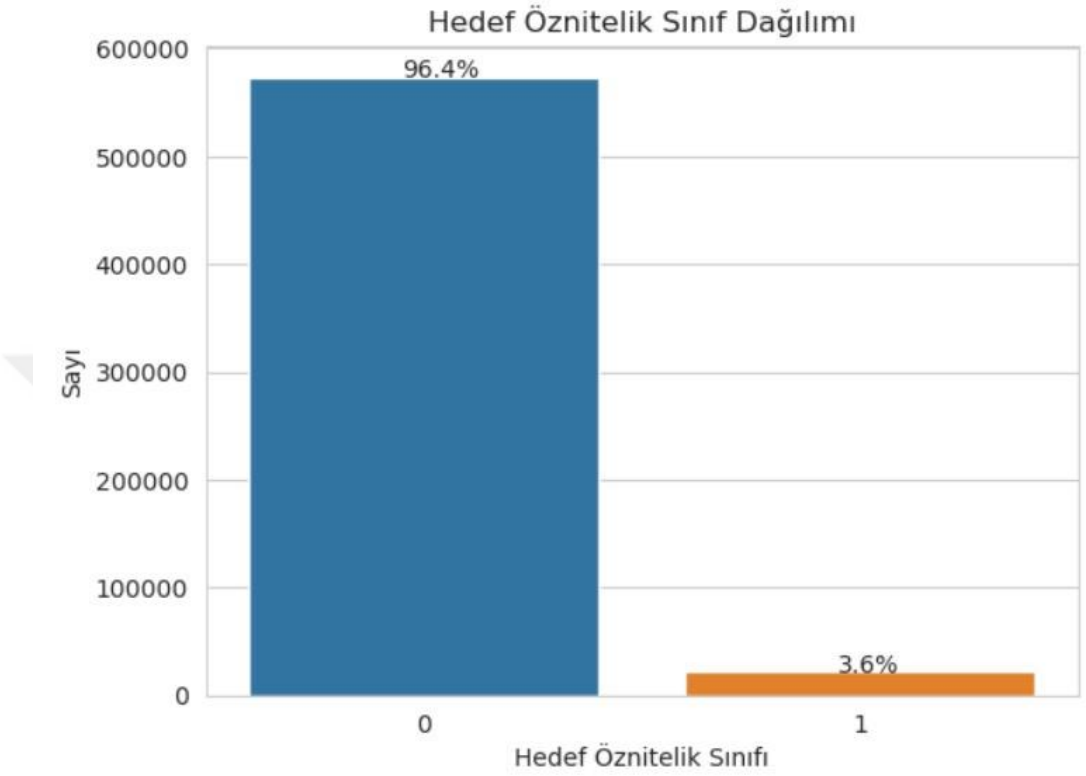
Veri setinde toplamda 1488028 gözlem bulunmaktadır. Bunun yanında eğitim veri setinde 59 öznitelik bulunmaktayken, test veri setinde 58 öznitelik bulunmaktadır. Test veri setinde yer almayan öznitelik, oluşturulacak olan makine öğrenimi modellerinin tahmin edeceği "target" özniteliğidir.

Yarışma veri setlerinin formatı gereğince, test veri setinde "target" özniteliğinin yer almadığı görülmüştür. Bu nedenle, çalışmada eğitim veri seti; ana veri seti olarak kullanılmıştır. Ana veri seti olarak kabul edilmiş veri setinin, hedef özniteliğindeki sınıf dağılımı incelenmiştir. Bu incelemenin sonucunda, veri setinde yer alan 595212 gözlemin içinden, 573518 gözlemin 0 etiketli sınıfa dahil olduğu ve 21694 gözlemin 1 etiketli sınıfa dahil olduğu görülmüştür. Böylece veri setinde yer alan gözlemlerin, yaklaşık olarak %96.4'ünün 0 etiketli sınıfa dahil olduğu ve geriye kalan yaklaşık olarak %3.6'sının 1 etiketli sınıfa dahil olduğu görülmüştür.

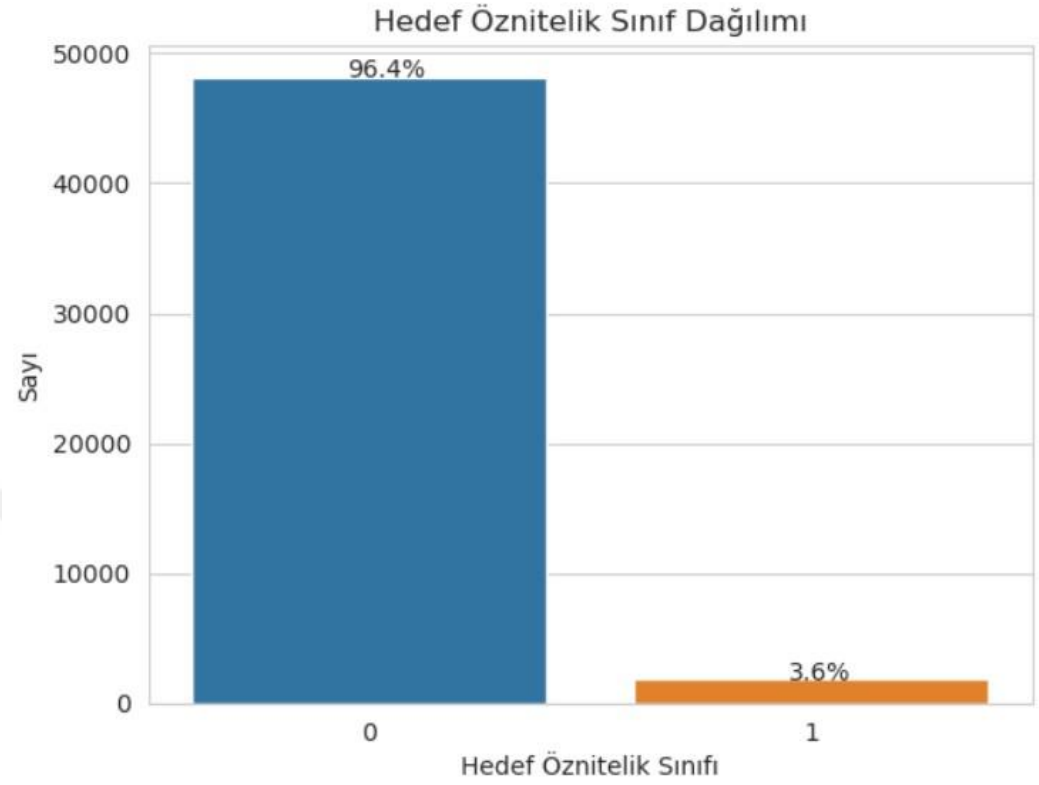
Veri setinde yer alan gözlem sayısının çok fazla olması, uygulamanın analizler ve modellemeler aşamalarında performans kayıplarına neden olmuştur. Bu performans kayıplarını engellemek amacıyla; veri setinden, sınıf dağılımını tamamen koruyarak ve rassal şekilde gözlemler seçilip, bu çalışmada kullanılacak olan daha küçük boyutta bir veri seti oluşturulmuştur. Oluşturulan yeni veri setinde toplam 50000 gözlem yer almıştır. Yeni oluşturulan veri setinde yer alan 50000 gözlemin, veri setinin hedef özniteliğinde karşılık geldiği sınıf dağılımı incelenmiştir. Bu inceleme sonucunca, 48178 gözlemin 0 etiketli sınıfa dahil olduğu ve 1822 gözlemin 1 etiketli sınıfa dahil olduğu görülmüştür. Rassal örnekleme ile elde edilmiş yeni veri setinde yer alan gözlemlerin, yaklaşık olarak %96.4'ünün 0 etiketli sınıfa dahil olduğu ve geriye kalan yaklaşık olarak %3.6'sının 1 etiketli sınıfa dahil olduğu görülerek, sınıf dağılımının, veri setinin asıl durumundaki dağılımıyla aynı olduğu görülmüştür.

Veri setinde, rassal örnekleme işlemi uygulanmasından önceki durumda ve rassal örnekleme işlemi uygulanmasından sonraki durumda, hedef özniteliğindeki sınıf dağılımları aşağıda yer alan Grafik 3.1. ve Grafik 3.2.'de görülmektedir.

Grafik 3.1.: Rassal Örnekleme İşlemi Uygulanmasından Önceki Durumda Veri Setinin Hedef Özniteliğinin Sınıf Dağılımı



Grafik 3.2.: Rassal Örnekleme İşlemi Uygulanmasından Sonraki Durumda Veri Setinin Hedef Özniteliğinin Sınıf Dağılımı



Veri setinde yer alan özniteliklerin, Python programlama dilinin, nümerik veri tipleri olan "int64" ve "float64" formatında oldukları görülmüştür. Veri setinde yer alan "id" özniteliği yapılacak olan analizlerde ve kurulacak olan modellerde kullanılmayacağı için hem eğitim veri setinden hem de test veri setinden çıkarılmıştır. Veri setinde yer alan "ps_ind_14" özniteliği de bağımsız özniteliklerle arasında yüksek korelasyon değerine sahip olduğu için hem eğitim veri setinden hem de test veri setinden çıkarılmıştır.

Çalışmada kullanılan veri setinde öncelikle keşifsel veri analizi işlemleri gerçekleştirilmiştir. Bu bağlamda, veri setinin yapısının incelenmesi, özniteliklerin incelenmesi, özniteliklerin dağılımlarının görselleştirilmesi, aykırı değerlere sahip gözlemler ile eksik gözlemlerin oluşturduğu sorunların çözülmesi, aralarında yüksek korelasyon olan özniteliklerin incelenmesi ve bağımsız öznitelikler ile hedef özniteliğinin arasındaki ilişkinin incelenmesi adımları uygulanmıştır.

Keşifsel veri analizi aşamasının sonunda, modelleme için kullanımı gerekmeyen öznitelikler, veri setinden çıkarılmıştır. Veri setinde öznitelik

mühendisliği işlemleri uygulanarak, veri setinde yer alan öznitelikler modelleme için uygun şekle dönüştürülmüştür.

Çalışmada; Logistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları, Rassal Orman, Gradyan Artırılmış Ağaçlar ve Ekstrem Gradyan Artırma sınıflandırma algoritmaları ile oluşturulan modeller kullanılmıştır. Bu modeller ilk olarak, dengesiz veri seti problemini çözme yaklaşımları uygulanmamış veri seti üzerinde eğitilip, aynı veri seti üzerinde test edilmiştir. Sonrasında aynı modeller, dengesiz veri seti problemini çözme yaklaşımları uygulanmış veri seti üzerinde eğitilip, yine dengesiz veri seti problemini çözme yaklaşımları uygulanmış veri seti üzerinde test edilmiştir. Sonrasında dengesiz veri seti problemini çözme yaklaşımları uygulanmadan oluşturulan modellerin performans metrikleri ile, dengesiz veri seti problemini çözme yaklaşımları uygulandıktan sonra oluşturulan modellerin performans metrikleri karşılaştırılmıştır. Dengesiz veri seti problemini çözme yaklaşımları uygulandıktan sonra oluşturulan modellerin, F-Skoru performans metrikleri, Tablo 3.1.'de görülmektedir.

Tablo 3.1.: Porto Seguro Sigorta Şirketi Veri Setinde Dengesiz Veri Seti Problemini Çözme Yaklaşımları Uygulandıktan Sonra Oluşturulan Modellerin F-Skoru Performans Metriklerinin Değerleri

Yöntem \ Model	Logistik Regresyon	Destek Vektör Makineleri	Karar Ağaçları	Rassal Orman	Gradyan Artırılmış Ağaçlar	Ekstrem Gradyan Artırılmış Ağaçlar
Rassal Alt Örnekleme	0.42	0.42	0.37	0.40	0.41	0.39
Tomek Bağlantıları	0.49	0.49	0.50	0.49	0.49	0.49
Küme Merkezleri	0.42	0.44	0.04	0.04	0.04	0.04
Tek Taraflı Seçim	0.49	0.49	0.50	0.49	0.49	0.49
Düzenlenmiş K-En Yakın Komşular Kuralı	0.49	0.49	0.50	0.49	0.49	0.49
Bölge Temizleme Kuralı	0.49	0.49	0.97	1.00	0.50	0.68
Rassal Aşırı Örnekleme	0.43	0.45	0.50	0.49	0.45	0.52
SMOTE	0.43	0.41	0.51	0.49	0.49	0.49

Dengesiz veri seti problemini çözmek için veri seti üzerinde kullanılan farklı yöntemlerin çıktılarına, altı farklı sınıflandırma algoritması ile kurulan modeller uygulanmıştır. Bu modellerin uygulanmalarından sonra, modellere ait performans metrikleri ile karmaşıklık matrisleri incelenmiştir. En başarılı performans

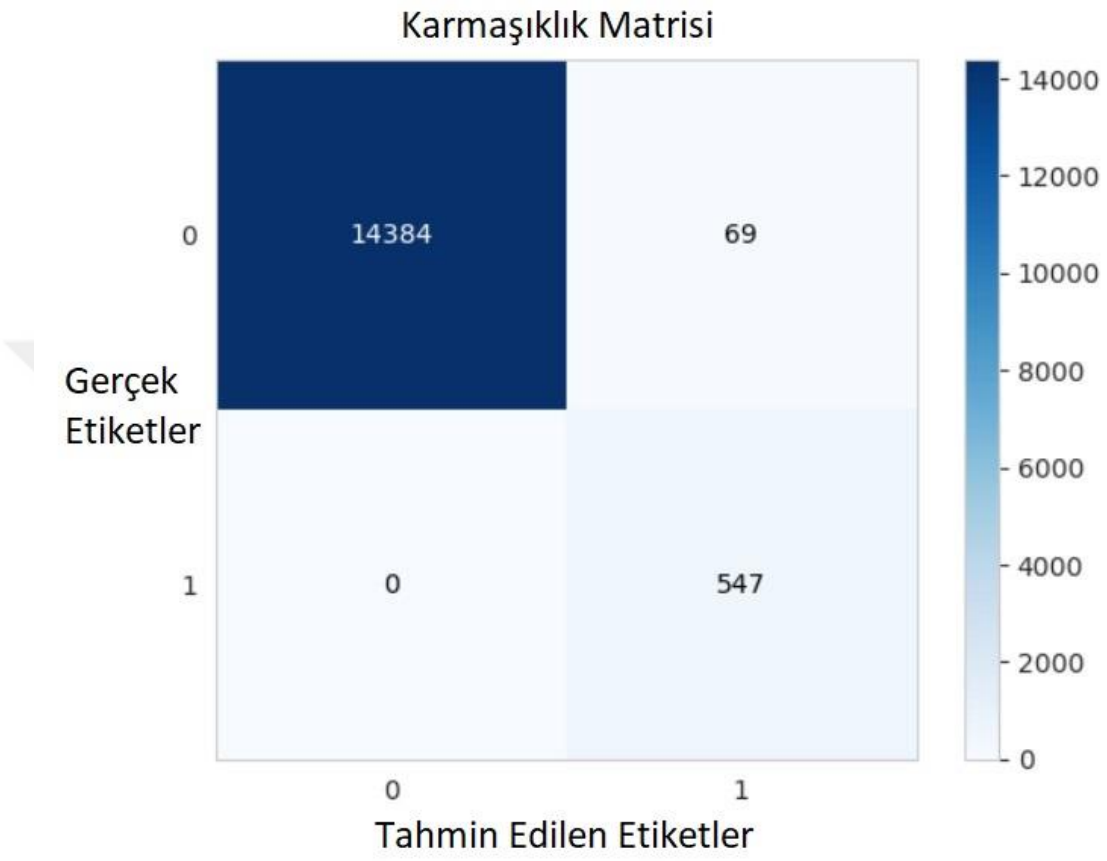
metriklerinin, alt örnekleme yöntemi olan Bölge Temizleme Kuralı'nın uygulanmasından sonra elde edilen çıktılara uygulanan Karar Ağaçları ve Rassal Orman modelleri ile elde edildiği görülmüştür. Her iki model de Bölge Temizleme Kuralı'nın uygulanmasından sonra, veri setinin hedef özniteliğinde yer alan azınlık sınıfa ait gözlemleri, çok başarılı bir şekilde tahmin etmiştir.

Bölge Temizleme Kuralı'nın uygulanmasından sonra Karar Ağaçları modelinin performans metrikleri incelendiğinde, doğruluk ölçütü değerinin 0.9954 olarak çok yüksek olduğu görülmüştür. Bölge Temizleme Kuralı'nın uygulanmasından sonra Karar Ağaçları modeline ait performans metrikleri tablosu Tablo 3.2.'de ve karmaşıklık matrisi Grafik 3.3.'te görülmektedir.

Tablo 3.2.: Bölge Temizleme Kuralı Uygulanmasından Sonra Karar Ağaçları Modeline ait Performans Metrikleri

	Kesinlik	**Duyarlılık**	**F1 Skoru**	**Destek**
0	1.00	1.00	1.00	14453
1	0.89	1.00	0.94	547
Doğruluk			1.00	15000
Makro Ortalama	0.94	1.00	0.97	15000
Ağırlıklı Ortalama	1.00	1.00	1.00	15000

Grafik 3.3.: Bölge Temizleme Kuralı Uygulanmasından Sonra Karar Ağaçları Modeline ait Karmaşıklık Matrisi

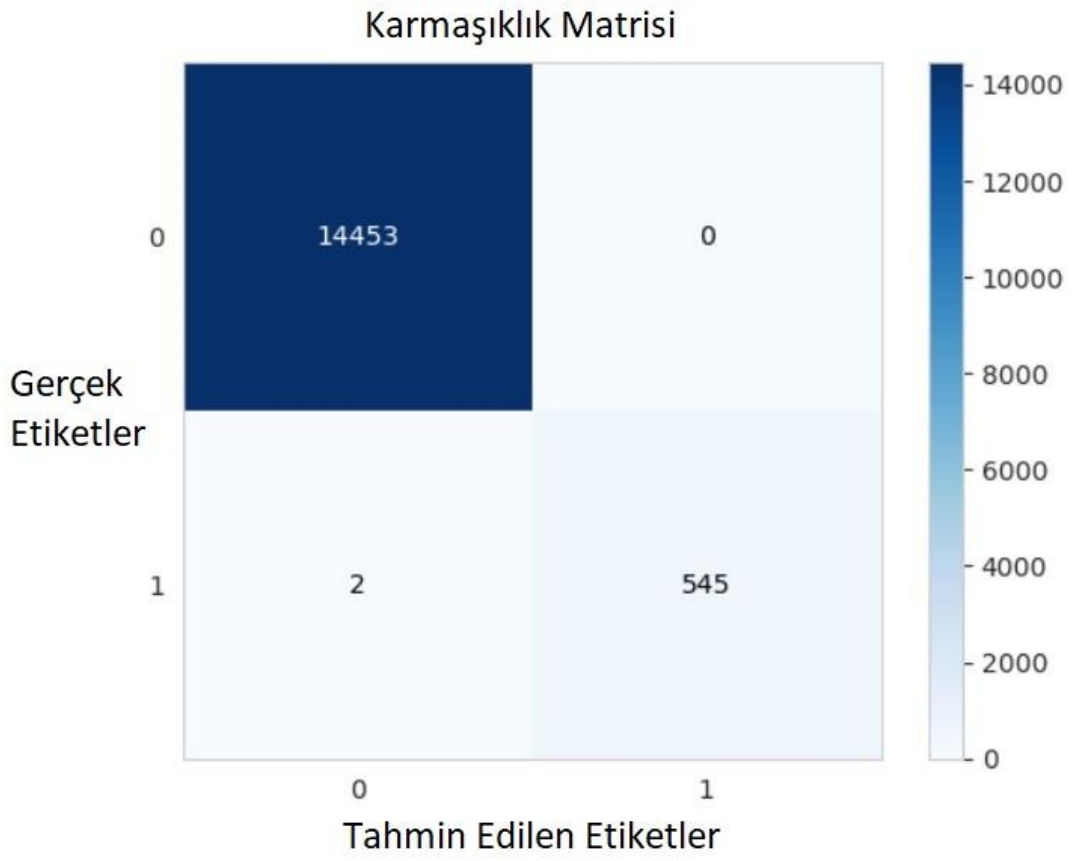


Bölge Temizleme Kuralı'nın uygulanmasından sonra Rassal Orman modelinin performans metrikleri incelendiğinde, doğruluk ölçütü değerinin 0.9998 olarak çok yüksek olduğu görülmüştür. Bölge Temizleme Kuralı'nın uygulanmasından sonra Rassal Orman modeline ait performans metrikleri tablosu Tablo 3.3.'te ve karmaşıklık matrisi Grafik 3.4.'te görülmektedir.

Tablo 3.3.: Bölge Temizleme Kuralı Uygulanmasından Sonra Rassal Orman Modeline ait Performans Metrikleri

	Kesinlik	**Duyarlılık**	**F1 Skoru**	**Destek**
0	1.00	1.00	1.00	14453
1	1.00	1.00	1.00	547
Doğruluk			1.00	15000
Makro Ortalama	1.00	1.00	1.00	15000
Ağırlıklı Ortalama	1.00	1.00	1.00	15000

Grafik 3.4.: Bölge Temizleme Kuralı Uygulanmasından Sonra Rassal Orman Modeline ait Karmaşıklık Matrisi



3.1.2 Beyin Felci Hastaları Verisine ait Dengesiz Veri Seti Uygulaması

Uygulamada ikinci sırada, Kaggle isimli veri bilimi platformundan elde edilen, beyin felci hastalarına ait bilgileri içeren veri seti kullanılmıştır.

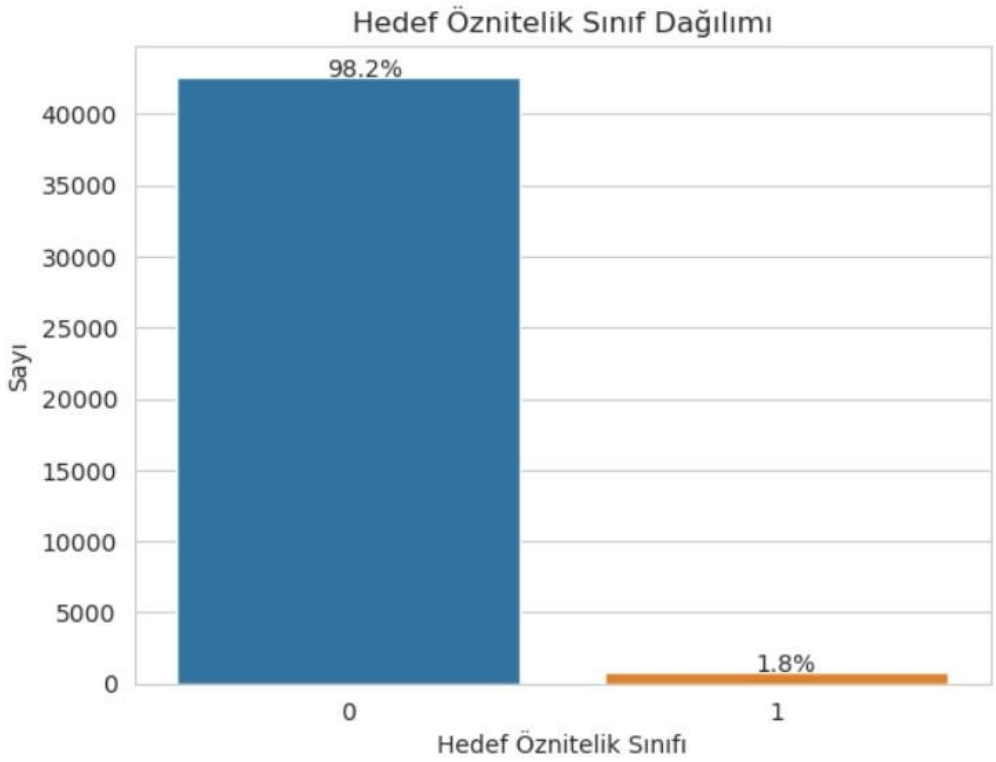
Literatür incelemesi yapıldığında, sağlık sektöründe makine öğrenimi uygulamalarının, kullanım alanlarının ve öneminin gün geçtikçe yükseldiği görülmüştür. Bu kullanım alanlarına, hastalık teşhisleri, tanı ve teşhis aşamalarında karar destek sistemleri olarak kullanımı, tedavi planlama ve uzun vadeli hasta takibi gibi farklı örnekler verilebilmektedir.

Çalışmada kullanılan veri seti, Amerika Birleşik Devletleri'nin sağlık alanında çok sayıda veri seti ve kaynak paylaşımı yaptığı www.healthdata.gov internet adresinde yer almakla birlikte, Kaggle'da da yarışma verisi olarak geçmişte kullanılmıştır. Günümüzde Kaggle'da açık kullanıma sunulmaya devam etmektedir. Veri seti incelendiğinde, temel olarak veri setinin 11 öznitelik ve 43400 gözlem içerdiği görülmektedir. Veri setinde yer alan her bir gözlem, farklı bir hastaya ait verileri içermektedir. Veri setinin hedef özniteliği, hastalarda beyin felci rahatsızlığının varlığını ifade eden "stroke" özniteliğidir. Veri setinde yer alan "smoking_status" adlı öznitelikte çok fazla eksik gözlem olması, bu eksik gözlemlerin, veri setinde oluşturulacak olan modellerin performanslarını olumsuz etkileyecek olması ve bu eksik gözlemleri tamamlamak için herhangi bir güvenilir referans bulunamaması sebebiyle, söz konusu öznitelik, veri setinden çıkarılarak analiz ve modelleme aşamalarına dahil edilmemiştir. Veri setinde yer alan "id" adlı öznitelik, analiz ve modelleme aşamalarında kullanılmayacağı için veri setinden çıkarılarak analiz ve modelleme aşamalarına dahil edilmemiştir. Veri setinde yer alan "bmi" adlı öznitelikte düşük sayıda eksik gözlem olduğu görülmüş olup, bu özniteliğin dağılımı incelenerek, eksik gözlemlere özniteliğin medyan değeri atanmıştır.

Veri setinin hedef özniteliğindeki sınıf dağılımının incelenmesi sonucunda, veri setinde yer alan 43400 gözlemin içinden, 42617 gözlemin 0 etiketli sınıfa dahil olduğu ve 783 gözlemin 1 etiketli sınıfa dahil olduğu görülmüştür. Böylece veri setinde yer alan gözlemlerin, yaklaşık olarak %98.2'sinin 0 etiketli sınıfa dahil olduğu ve geriye kalan yaklaşık olarak %1.8'inin 1 etiketli sınıfa dahil olduğu görülmüştür.

Veri setinin hedef özniteliği olan “stroke” özniteliğinin sınıf dağılımı aşağıda yer alan Grafik 3.5.’de görülmektedir.

Grafik 3.5.: Beyin Felci Veri Setinin Hedef Özniteliğinin Sınıf Dağılımı



Veri setinde yer alan özniteliklerin, Python programlama dilinin, nümerik veri tipleri olan “int64” ve “float64” ve kategorik veri tipi olan “object” formatlarına sahip oldukları görülmüştür.

Veri setinde gerçekleştirilen keşifsel veri analizi ile eksik gözlemler ve aykırı gözlemler incelenmiş, bu gözlemlerle ilgili olarak gerekli işlemler yapılmıştır. Özniteliklerin dağılımları incelenmiş ve gerekli olan veri dönüştürme işlemleri yapılmıştır. Bağımsız özniteliklerin hem kendi aralarında hem de hedef öznitelik ile olan korelasyon değerleri incelenmiştir. Son olarak gerekli öznitelik mühendisliği işlemleri uygulanarak veri setinde yer alan öznitelikler, modelleme için uygun duruma getirilmiştir.

Çalışmada; Logistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları, Rassal Orman, Gradyan Artırılmış Ağaçlar ve Ekstrem Gradyan Artırma sınıflandırma

algoritmaları ile oluşturulan modeller kullanılmıştır. Bu modeller önce, dengesiz veri seti problemini çözme yaklaşımları uygulanmayarak, veri seti üzerinde eğitilip test edilmiştir. Sonrasında aynı modeller, dengesiz veri seti problemini çözme yaklaşımları uygulanmış veri seti üzerinde eğitilip test edilmiştir.

Dengesiz veri seti problemini çözme yaklaşımları uygulanmadan oluşturulan modellerin performans metrikleri ile, dengesiz veri seti problemini çözme yaklaşımları uygulandıktan sonra oluşturulan modellerin performans metrikleri karşılaştırılmıştır. Dengesiz veri seti problemini çözme yaklaşımları uygulandıktan sonra oluşturulan modellerin, F-Skoru performans metrikleri, Tablo 3.4.'te görülmektedir.

Tablo 3.4.: Beyin Felci Hastaları Veri Setinde, Dengesiz Veri Seti Problemini Çözme Yaklaşımları Uygulandıktan Sonra Oluşturulan Modellerin F-Skoru Performans Metrikleri

Yöntem \ Model	Logistik Regresyon	Destek Vektör Makineleri	Karar Ağaçları	Rassal Orman	Gradyan Artırılmış Ağaçlar	Ekstrem Gradyan Artırılmış Ağaçlar
Rassal Alt Örnekleme	0.48	0.47	0.44	0.46	0.47	0.46
Tomek Bağlantıları	0.50	0.50	0.53	0.50	0.50	0.50
Küme Merkezleri	0.40	0.40	0.39	0.41	0.39	0.40
Tek Taraflı Seçim	0.50	0.50	0.54	0.50	0.50	0.50
Düzenlenmiş K-En Yakın Komşular Kuralı	0.50	0.50	0.54	0.52	0.50	0.49
Bölge Temizleme Kuralı	0.50	0.50	0.91	0.98	0.53	0.62
Rassal Aşırı Örnekleme	0.47	0.47	0.52	0.51	0.49	0.54
SMOTE	0.47	0.48	0.51	0.53	0.50	0.52

Beyin felci hastalarına ait verileri içeren veri setinde, dengesiz veri seti problemini çözmek için kullanılan yaklaşımların sonucunda elde edilen dönüştürülmüş veri setlerine, altı farklı sınıflandırma algoritması ile kurulan modeller uygulanarak, modellere ait performans metrikleri ve karmaşıklık matrisleri elde edilmiştir. Modellere ait performans metriklerinin ve karmaşıklık matrislerinin

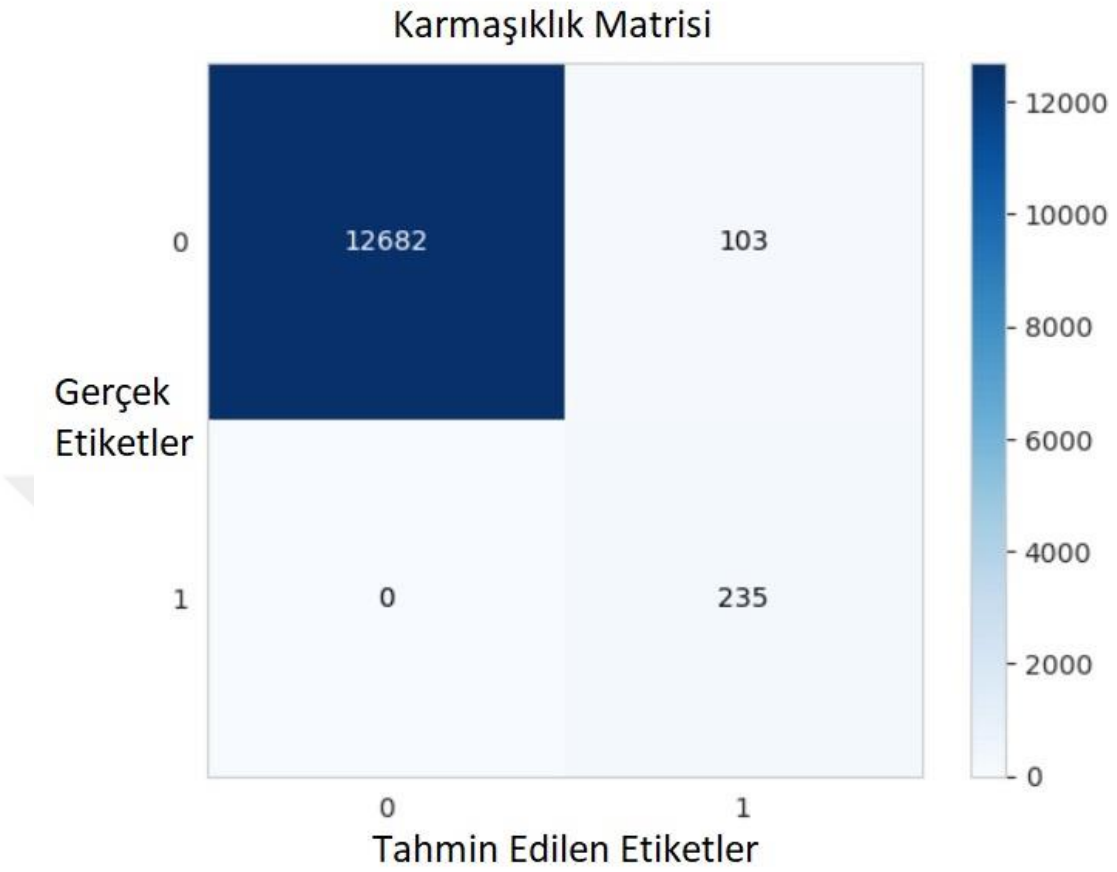
incelenmesi sonucunda, en başarılı performans metriklerinin, alt örnekleme yöntemi olan Bölge Temizleme Kuralı'nın uygulanmasıyla elde edilen, dönüştürülmüş veri setine uygulanmış Karar Ağaçları ve Rassal Orman sınıflandırma modelleri ile elde edildiği görülmüştür.

Bölge Temizleme Kuralı'nın uygulanmasıyla dönüştürülmüş veri setinde, Karar Ağaçları sınıflandırma modelinin performans metrikleri incelendiğinde, doğruluk ölçütü değerinin 0.9920 olarak çok yüksek olduğu görülmüştür. Bölge Temizleme Kuralı'nın uygulanmasından sonra Karar Ağaçları modeline ait performans metrikleri tablosu Tablo 3.3.'de ve karmaşıklık matrisi Grafik 3.6.'te görülmektedir.

Tablo 3.5.: Bölge Temizleme Kuralı Uygulanmasından Sonra Karar Ağaçları Modeline ait Performans Metrikleri

	Kesinlik	**Duyarlılık**	**F1 Skoru**	**Destek**
0	1.00	0.99	1.00	12785
1	0.70	1.00	0.82	235
Doğruluk			0.99	13020
Makro Ortalama	0.85	1.00	0.91	13020
Ağırlıklı Ortalama	0.99	0.99	0.99	13020

Grafik 3.6.: Bölge Temizleme Kuralı Uygulanmasından Sonra Karar Ağaçları Modeline ait Karmaşıklık Matrisi



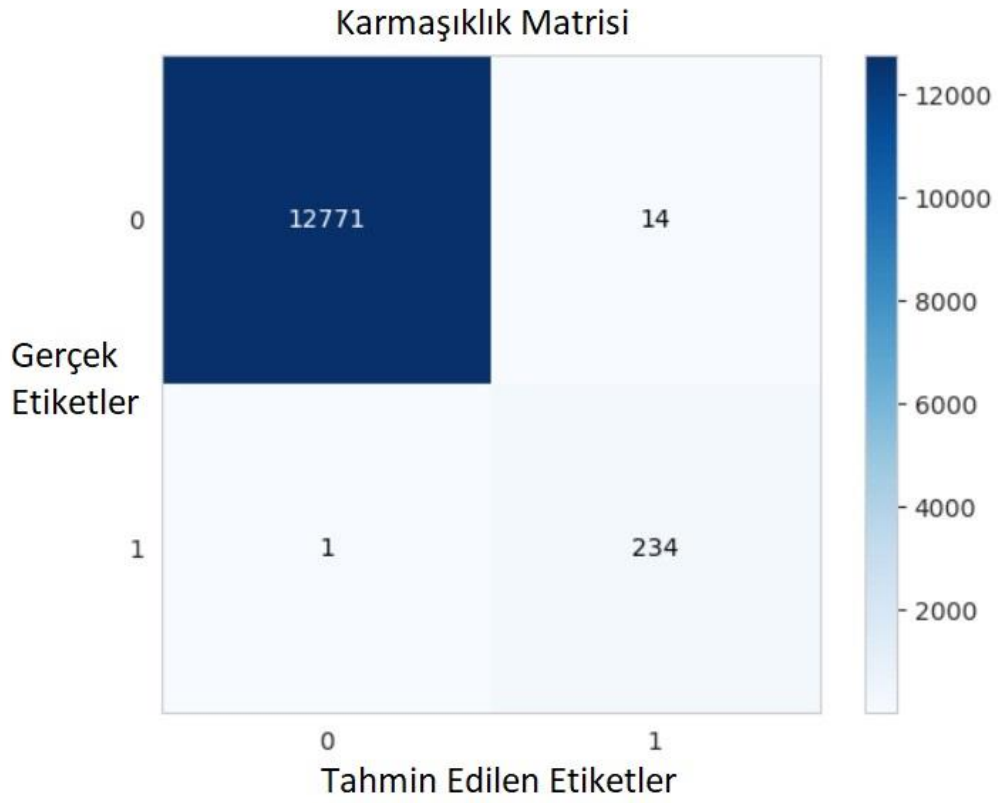
Grafik 3.6.'da yer alan karmaşıklık matrisinde görüldüğü gibi karar ağaçları modeli, azınlık sınıf olan 1 sınıfına ait gözlemlerin tamamını doğru tahmin etmiştir.

Bölge Temizleme Kuralı'nın uygulanmasıyla dönüştürülmüş veri setinde, Rassal Orman sınıflandırma modeli de uygulanmış olup, performans metrikleri incelendiğinde, doğruluk ölçütü değerinin 0.9988 olarak Karar Ağaçları sınıflandırma modeline göre daha yüksek olduğu görülmüştür. Bölge Temizleme Kuralı'nın uygulanmasından sonra Rassal Orman modeline ait performans metrikleri tablosu Tablo 3.6.'de ve karmaşıklık matrisi Grafik 3.7.'te görülmektedir.

Tablo 3.6.: Bölge Temizleme Kuralı Uygulanmasından Sonra Rassal Orman Modeline ait Performans Metrikleri

	Kesinlik	**Duyarlılık**	**F1 Skoru**	**Destek**
0	1.00	1.00	1.00	12785
1	0.94	1.00	0.97	235
Doğruluk			1.00	13020
Makro Ortalama	0.97	1.00	0.98	13020
Ağırlıklı Ortalama	1.00	1.00	1.00	13020

Grafik 3.7.: Bölge Temizleme Kuralı Uygulanmasından Sonra Rassal Orman Modeline ait Karmaşıklık Matrisi



Beyin felci hastalarına ait verileri içeren veri setinde yaşanan dengesiz veri seti problemini çözmek için Bölge Temizleme Kuralı yaklaşımı uygulanıp, sonrasında

dönüştürülmüş veri setine Karar Ağaçları sınıflandırma modeli ve Rassal Orman sınıflandırma modeli uygulanarak elde edilen performans metrikleri ve karmaşıklık matrisleri incelendiğinde, dikkat çeken bir durum görülmektedir. Rassal Orman sınıflandırma modelinin doğruluk oranı, Karar Ağaçları sınıflandırma modelinin doğruluk oranından 0.0068 değerinde daha yüksek olarak hesaplanmıştır. Sadece doğruluk oranına bakılarak Rassal Orman modelinin daha başarılı olduğu şüphesiz bir şekilde söylenebilmektedir. Rassal Orman sınıflandırma modelinin karmaşıklık matrisindeki değerler ile Karar Ağacı sınıflandırma modelinin karmaşıklık matrisindeki değerler karşılaştırıldığında, Rassal Orman sınıflandırma modelinin, beyin felci rahatsızlığı olmayan hastaları temsil eden, yani veri setinin hedef özniteliğinde çoğunluk sınıfı olan 0 sınıfını daha başarılı bir şekilde tahmin ettiği görülmektedir. Lâkin aynı karşılaştırmada Karar Ağaçları sınıflandırma modelinin, beyin felci rahatsızlığı olan hastaları temsil eden, yani veri setinin hedef özniteliğinde azınlık sınıf olan 1 sınıfını, Rassal Orman sınıflandırma modeline göre daha başarılı bir şekilde tahmin ederek, Rassal Orman sınıflandırma modelinden daha iyi performans gösterdiği görülmektedir. Bu ve benzeri durumlarda, işletmelerin iş planları, veri setinde modelleme ile elde edilmek istenen başarımlar ve veri setinin bağlı olduğu alan hakkında arka planda sahip olunan bilgiler dikkate alınarak maliyet kavramının da hesaplanması önem arz etmektedir.

3.1.3 Bağış Yapan Donörlerin Verisine ait Dengesiz Veri Seti Uygulaması

Uygulamada üçüncü sırada, felçli askeri gaziler için düzenlenen bir bağış kampanyasında bağış yapan donörlere ait bilgileri içeren veri seti kullanılmıştır.

Veri setinde yapılan incelemeler sonucunda, veri setinde 13 özniteliğin ve 93462 gözlemin yer aldığı görülmüştür. Veri setinde yer alan her bir gözlem, farklı bir donöre ait verileri içermektedir. Veri setinin hedef özniteliği, donörlerin bağış yapma durumunu ifade eden “donated” özniteliğidir. Veri setinde yer alan “age” özniteliğinde çok fazla sayıda eksik gözlem olması ve bu eksik gözlemlerin, veri setinde oluşturulacak olan modellerin performanslarını etkileyecek olmasından dolayı söz konusu öznitelikte yer alan eksik gözlemlerin tamamlanmasına karar verilmiştir. Söz konusu öznitelikte eksik gözlemler tamamlanırken, özniteliğin dağılımının mümkün

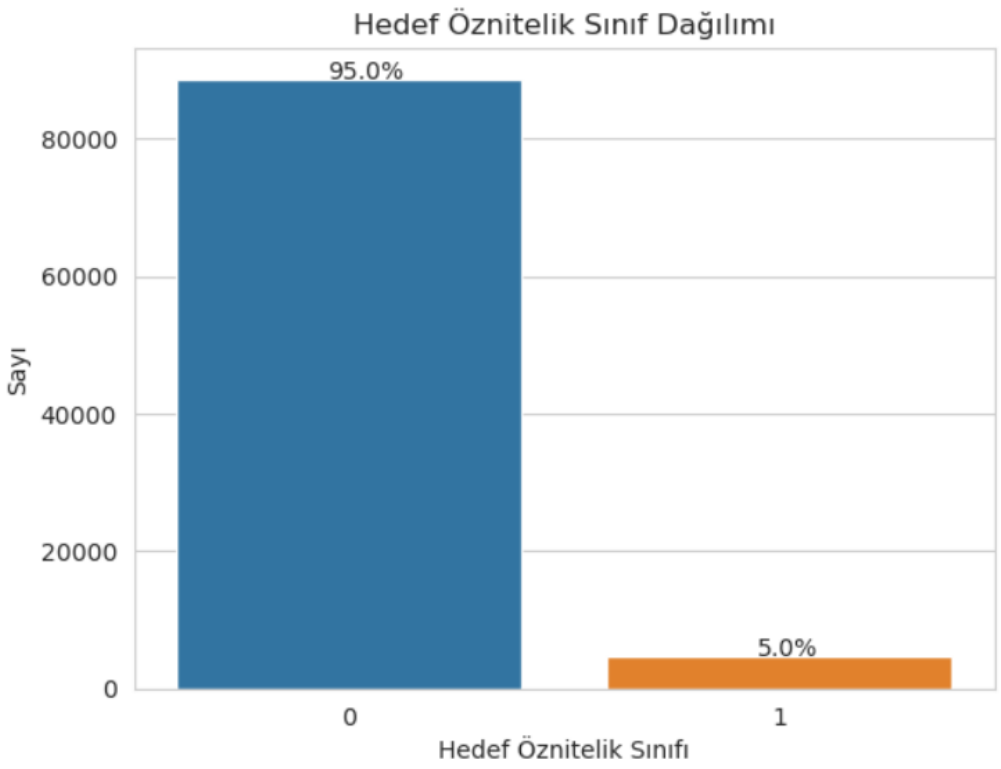
olduğu kadar korunması amaçlanmıştır. Bu nedenle, öncelikle özniteliğin dağılımı incelenmiştir. Sonrasında özniteliğin ortalama değeri ve standart sapma değeri hesaplanarak, belirlenmiş aralıklarda, rassal şekilde üretilmiş sayılar eksik gözlemlere atanarak, söz konusu öznitelikte bulunan eksik gözlemler tamamlanmıştır.

Veri setinde yer alan kategorik değişkenler için gerekli olan dönüştürme işlemleri gerçekleştirilmiştir.

Veri setinin hedef özniteliği olan, "donated" özniteliğinde sınıf dağılımının incelenmesi sonucu, veri setinde bulunan 93462 gözlemden, 88751 gözlemin, 0 etiketli sınıfa dahil olduğu ve geriye kalan 4711 gözlemin, 1 etiketli sınıfa dahil olduğu görülmüştür. Böylece veri setinde yer alan gözlemlerin, yaklaşık olarak %95'inin 0 etiketli sınıfa dahil olduğu ve geriye kalan yaklaşık olarak %5'inin 1 etiketli sınıfa dahil olduğu görülmüştür.

Veri setinde hedef öznitelik olarak yer alan "donated" özniteliğinin sınıf dağılımı aşağıda yer alan Grafik 3.8.'de görülmektedir.

Grafik 3.8.: Bağış Yapan Donörler Veri Setinin Hedef Özniteliğinin Sınıf Dağılımı



Veri setinde yer alan kategorik özniteliklere yönelik gerekli dönüştürme işlemlerinin uygulanmasından sonra veri setinin yapısı incelenmiştir. Bu incelemede, veri setinin özniteliklerinin Python programlama dilinin nümerik veri tipleri olan, "int64", "float64" veri tiplerine sahip olmalarının yanında, "uint8" veri tipine sahip olan iki öznitelik olduğu da görülmüştür. Söz konusu iki özniteliğin, kategorik özniteliklerin dönüştürülme işleminden sonra elde edilmiş olan yeni öznitelikler oldukları görülmüştür.

Veri setinde gerçekleştirilmiş olan keşifsel veri analizi ile özniteliklerin dağılımları incelenmiş ve gerekli olduğuna karar verilen çeşitli veri dönüştürme işlemleri gerçekleştirilmiştir. Veri setinde yer alan bağımsız özniteliklerin hem kendi aralarında hem de hedef öznitelik ile olan korelasyon değerleri incelenmiştir. Tüm bu incelemelerin sonucunda, öznitelik mühendisliği işlemleri uygulanarak veri setinde yer alan öznitelikler, modelleme için uygun duruma getirilmiştir.

Çalışmada; Logistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları, Rassal Orman, Gradyan Artırılmış Ağaçlar ve Ekstrem Gradyan Artırma sınıflandırma

algoritmaları ile oluşturulan modeller kullanılmıştır. Bu modeller uygulama bölümünde daha önce anlatılan 2 uygulamada da olduğu gibi, öncelikle dengesiz veri seti problemini çözme yaklaşımları uygulanmayarak, veri setlerine uygulanıp test edilmiştir. Sonrasında aynı modeller, dengesiz veri seti problemini çözme yaklaşımları uygulanmış veri seti üzerinde eğitilip test edilmiştir. Bu modellerin uygulanıp test edilme aşamalarının sonucunda, dengesiz veri seti problemini çözme yaklaşımları uygulanmadan oluşturulan modellerin performans metrikleri ile, dengesiz veri seti problemini çözme yaklaşımları uygulandıktan sonra oluşturulan modellerin performans metrikleri karşılaştırılmıştır. Dengesiz veri seti problemini çözme yaklaşımları uygulandıktan sonra oluşturulan modellerin, F-Skoru performans metrikleri, Tablo 3.7.'te görülmektedir.

Tablo 3.7.: Bağış Yapan Donörler Veri Setinde, Dengesiz Veri Seti Problemini Çözme Yaklaşımları Uygulandıktan Sonra Oluşturulan Modellerin F-Skoru Performans Metrikleri

Yöntem \ Model	Logistik Regresyon	Destek Vektör Makineleri	Karar Ağaçları	Rassal Orman	Gradyan Artırılmış Ağaçlar	Ekstrem Gradyan Artırılmış Ağaçlar
Rassal Alt Örnekleme	0.41	0.30	0.40	0.38	0.42	0.39
Tomek Bağlantıları	0.49	0.49	0.49	0.49	0.49	0.49
Küme Merkezleri	0.31	0.28	0.17	0.16	0.09	0.06
Tek Taraflı Seçim	0.49	0.49	0.49	0.50	0.49	0.49
Düzenlenmiş K-En Yakın Komşular Kuralı	0.49	0.49	0.50	0.50	0.49	0.49
Bölge Temizleme Kuralı	0.49	0.49	0.59	0.59	0.49	0.49
Rassal Aşırı Örnekleme	0.41	0.39	0.44	0.44	0.43	0.43
SMOTE	0.44	0.39	0.42	0.42	0.43	0.42

Bağış yapan donörlere ait verileri içeren veri setinden, dengesiz veri seti problemini çözmek için kullanılan yaklaşımların sonucunda elde edilen dönüştürülmüş veri setlerine, altı farklı sınıflandırma algoritması ile kurulan modeller uygulanarak, modellere ait performans metrikleri ve karmaşıklık matrisleri elde edilmiştir. Modellere ait performans metriklerinin ve karmaşıklık matrislerinin

incelenmesi sonucunda, en başarılı performans metriklerinin, aşırı örnekleme yöntemi olan Rassal Aşırı Örnekleme'nin uygulanmasıyla elde edilen, dönüştürülmüş veri setine uygulanmış Gradyan Artırılmış Ağaçlar sınıflandırma modelleri ile elde edildiği görülmüştür.

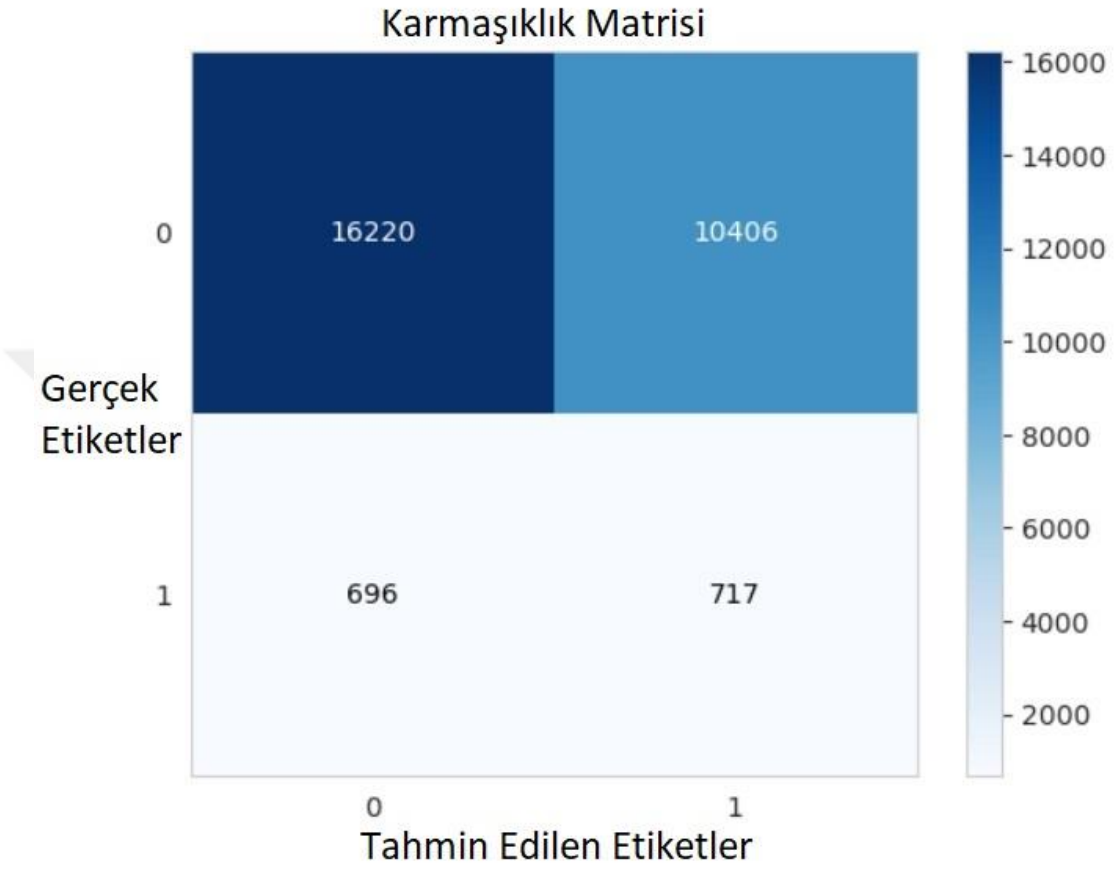
Rassal Aşırı Örnekleme yaklaşımının uygulanmasıyla dönüştürülmüş veri setinde, Gradyan Artırılmış Ağaçlar sınıflandırma modelinin uygulanması sonucunda elde edilen performans metrikleri incelendiğinde, doğruluk ölçütü değerinin 0.6040 olarak ortalama seviyede olduğu görülmüştür. Modele ait doğruluk ölçütü değerinin, uygulama bölümünde yer alan önceki uygulamalarla kıyaslandığında daha düşük olarak elde edildiği görülmesine karşın, dengesiz veri seti problemine sahip olan veri setinde uygulanmış olan çözüm yaklaşımları ve sınıflandırma modelleri içinde en iyi sonucu verdiği görülmüştür. Buna ek olarak dengesiz veri setleri ile sınıflandırma modellerinin kullanıldığı çalışmalarda performans metrikleri incelenirken, doğruluk ölçütünün tek başına yeterli olmadığı ve diğer ölçütlerin de incelenmesi gerektiği, dengesiz veri setleri bölümünde açıklanmıştır.

Rassal Aşırı Örnekleme yaklaşımının uygulamasından sonra Gradyan Artırılmış Ağaçlar modeline ait performans metrikleri tablosu Tablo 3.8.'de ve karmaşıklık matrisi Grafik 3.9.'da görülmektedir.

Tablo 3.8.: Rassal Aşırı Örnekleme Yaklaşımının Uygulamasından Sonra Gradyan Artırılmış Ağaçlar Modeline ait Performans Metrikleri

	Kesinlik	**Duyarlılık**	**F1 Skoru**	**Destek**
0	0.96	0.61	0.75	26626
1	0.06	0.51	0.11	1413
Doğruluk			0.60	28039
Makro Ortalama	0.51	0.56	0.43	28039
Ağırlıklı Ortalama	0.91	0.60	0.71	28039

Grafik 3.9.: Rassal Aşırı Örnekleme Yaklaşımının Uygulanmasından Sonra Gradyan Artırılmış Ağaçlar Modeline ait Karmaşıklık Matrisi



SONUÇ

Bu çalışmada dengesiz veri seti problemi, teorik olarak tanıtılarak çözüm yaklaşımları örneklerine uygulama bölümünde yer verilmiştir. Uygulama bölümünde 3 farklı veri setinin incelenmesiyle söz konusu veri setlerinin, dengesiz veri seti olarak nitelendirilebileceği görülmüştür. Dengesiz veri setlerinin her biri için, farklı çözüm yaklaşımları uygulanmıştır. Söz konusu çözüm yaklaşımlarının uygulanmalarından sonra, veri setlerinin her birinde, 6 farklı sınıflandırma algoritması ile kurulmuş modeller eğitilerek test edilmiş ve modellerin her birine ait performans metrikleri ve karmaşıklık matrisleri elde edilerek incelenmiştir.

Uygulama bölümünde yer alan ilk uygulamada, Brezilya menşeli Porto Seguro sigorta ve reasürans şirketine ait veri seti kullanılmıştır. Veri setinde yer alan gözlem sayısının çok fazla olmasından dolayı, analiz ve modelleme aşamalarında performans kayıpları yaşanmış ve bu nedenle veri setinden hedef özniteliğindeki sınıf dağılımının korunacağı şekilde, rassal olarak gözlemler seçilerek daha küçük boyutta bir veri seti elde edilmiştir. Elde edilen bu yeni veri setinde, dengesiz veri seti problemini çözmek için alt örnekleme yöntemlerinden, Bölge Temizleme Kuralı yaklaşımı uygulanmıştır. Bu yaklaşımın uygulanmasından sonra elde edilen çıktılara uygulanan Karar Ağaçları ve Rassal Orman sınıflandırma modellerinin performans metrikleri ve karmaşıklık matrisleri incelendiğinde, çok başarılı sonuçlar elde edildiği görülmüştür.

Uygulama bölümünde yer alan ikinci uygulamada, beyin felci hastalarının bilgilerini içeren veri seti kullanılmıştır. Veri seti, hedef özniteliğinin incelenmesi sonucu, dengesiz veri seti olarak nitelendirilmiştir. Dengesiz veri seti problemini çözmek için alt örnekleme yöntemlerinden Bölge Temizleme Kuralı yaklaşımı uygulanan veri setinde, elde edilen çıktılara Karar Ağaçları ve Rassal Orman sınıflandırma modelleri uygulanmıştır. Bu modellere ait performans metrikleri ve karmaşıklık matrisleri incelendiğinde, çok başarılı sonuçlar elde edildiği görülmüştür.

Uygulama bölümünde yer alan üçüncü ve son uygulamada, felçli askeri gaziler için düzenlenen bağış toplama etkinliğinde, bağış yapan donörlere ait bilgileri içeren veri seti kullanılmıştır. Veri seti, hedef özniteliğinin incelenmesi sonucunda

dengesiz veri seti olarak nitelendirilmiştir. Dengesiz veri seti problemini çözmek için aşırı örnekleme yöntemlerinden biri olan Rassal Aşırı Örnekleme yaklaşımı uygulanmıştır. Bu işlemin sonucunda elde edilen çıktılara Gradyan Artırılmış Ağaçlar sınıflandırma modeli uygulanmıştır. Gradyan Artırılmış Ağaçlar sınıflandırma modeline ait performans metrikleri ve karmaşıklık matrisi incelendiğinde, Rassal Aşırı Örnekleme yaklaşımından sonra elde edilmiş çıktılara uygulanan sınıflandırma modelleri içinde en iyi performans metriklerinin ve en başarılı sonuçları gösteren karmaşıklık matrisinin elde edildiği görülmüştür.

Her bir veri setinde gerçekleştirilen analiz ve modelleme uygulamalarından sonra elde edilen performans metriklerinin ve karmaşıklık matrislerinin incelenmesi sonucunda kullanılan dengesiz veri setleri çözüm yöntemleri içinden, Bölge Temizleme Kuralı'nın en başarılı çıktıları verdiği görülmüştür. Bölge Temizleme Kuralı'nın başarılı çıktılar vermesinin nedenleri incelendiğinde, Bölge Temizleme Kuralı'nın çalışma şekli dikkat çekmektedir. Bölge Temizleme Kuralı'nın çalışma şekli diğer alt örnekleme yaklaşımlarına göre farklılık göstermektedir. Bölge Temizleme Kuralı, veri setinin hedef özniteliğinde sınıf dengesini sağlamak için doğrudan çoğunluk sınıfa ait gözlem sayısını azaltmak yerine, veri temizleme işlemini ön plana çıkarmaktadır. Bu şekilde, Bölge Temizleme Kuralı'nın aslında hibrit bir yöntem olarak çalıştığı söylenebilmektedir. Veri temizleme işlemi sonucunda veri kalitesinin artması, modelin başarısının artmasını sağlamaktadır.

Dengesiz veri setleri ile ileride yapılacak olan benzer çalışmalarda, bu çalışmada kullanılan veri seti seviyesinde çözüm yöntemlerinin yanında, modellerin hatalı veya yanlış sınıflandırma maliyetlerinin, veri setinin ait olduğu alana yönelik alan bilgisi kullanılarak hesaplanması prensibiyle çalışan, maliyet duyarlı çözüm yöntemleri de kullanılabilir. Yine veri seti seviyesinde çözüm yöntemleri yaklaşımları ile maliyet duyarlı çözüm yöntemleri yaklaşımlarının birlikte çalıştığı, hibrit yaklaşımlar kullanılabilir. Literatürde kısmen kullanılmış olsa da algoritma seviyesinde çözüm yöntemleri de dengesiz veri setleri problemini çözmek için kullanılabilir.

KAYNAKÇA

Aggarwal, C. C. **Recommender Systems The Textbook**. Switzerland: Springer., 2016, doi:10.1007/978-3-319-29659-3.

Akpınar, H. Veri tabanlarında bilgi keşfi ve veri madenciliği. **İ.Ü. İşletme Fakültesi Dergisi**, 2000, 29(1), 1-22.

Amoroso, N., Cilli, R., Maggipinto, T., Monaco, A., Tangaro, S., Bellotti, R. Satellite data and machine learning reveal a significant correlation between NO2 and COVID-19 mortality. **Environmental Research**, 2022, 204, 1-10. doi: 10.1016/j.envres.2021.111970.

Atalay, M. Bulanık Zaman Serilerinde Yapay Sinir Ağları ile Tahmin. M. E. Balaban, & E. Kartal içinde, **R ile Veri Madenciliği Uygulamaları**, 2016, (s. 103-105). İstanbul: Çağlayan Kitabevi.

Ayodele, T. O. Machine Learning Overview. In Y. Zhang, **New Advances in Machine Learning**, 2010, (pp. 9-16). Vukovar, Croatia: IntechOpen. doi:10.5772/9374.

Aytekin, Ç., Sütcü, C. S., Özfidan, U. Karar Ağacı Algoritması ile Metin Sınıflandırma: Müşteri Yorumları Örneği. **Uluslararası Sosyal Araştırmalar Dergisi**, 2018, 11(55), 782-792. doi:10.17719/jisr.20185537249.

Balaban, M. E., Kartal, E. **Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları**, 2018, (2. b.). İstanbul: Çağlayan.

Belyadi, H., Haghighat, A. **Machine Learning Guide for Oil and Gas Using Python A Step-by-Step Breakdown with Data, Algorithms, Codes, and Applications.**, 2021, United States: Gulf Professional Publishing, Elsevier.

Bilgin, T. T., Çamurcu, Y. DBSCAN, OPTICS ve K-Means Kümeleme Algoritmalarının Uygulamalı Karşılaştırılması. 2005, **Politeknik Dergisi**, 8(2), 139-145.

Błaszczyński, J., de Almeida Filho, A. T., Matuszyk, A., Szeląg, M., Słowiński, R. Auto loan fraud detection using dominance-based rough set approach versus machine learning methods. 2021, **Expert Systems with Applications**, (163), 1-2. doi: 10.1016/j.eswa.2020.113740.

Boehmke, B., Greenwell, B. **Hands-On Machine Learning with R**. Boca Raton: CRC Press., 2020.

Breiman, L. Random Forests. **Machine Learning**, (45), 2001, 5-32. doi:10.1023/A:1010933404324.

Caggiano, A., Napolitano, F., Teti, R. Hierarchical cluster analysis for pattern recognition of process conditions in die sinking EDM process monitoring. **14th CIRP Conference on Intelligent Computation in Manufacturing Engineering**, 2021, CIRP ICME '20 (pp. 514-519). Naples: Procedia CIRP.

Cagliero, L., Fior, J., Garza, P. Shortlisting machine learning-based stock trading recommendations using candlestick pattern recognition. 2023, **Expert Systems with Applications**, (216), 1-15., doi: 10.1016/j.eswa.2022.119493

Cao, P., Yang, J., Li, W., Zhao, D., Zaiane, O. Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD. **Computerized Medical Imaging and Graphics**, 2014, 38(3), 137-150., doi: 10.1016/j.compmedimag.2013.12.003.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, (16), 2002, 321-357.

Chen, T., Guestrin, C. XGBoost: A Scalable Tree Boosting System. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 2016, 785-794. doi:10.1145/2939672.2939785.

Chu, Y., Knell, G., Brayton, R. P., Burkhart, S. O., Jiang, X., Shams, S. Machine learning to predict sports-related concussion recovery using clinical data. **Annals of Physical and Rehabilitation Medicine**, 2022, 65(4), 1-9. doi: 10.1016/j.rehab.2021.101626.

Cunningham, P., Delany, S. J. k-Nearest Neighbour Classifiers - A Tutorial. **ACM Computing Surveys**, 2022, 54(6), 1-25. doi:10.1145/3459665.

Dal Pozzolo, A., Caelen, O., Waterschoot, S., Bontempi, G. Racing for unbalanced methods selection. **Intelligent Data Engineering and Automated Learning**, 2013, 24-31. doi:10.1007/978-3-642-41278-3_4.

Datta, S., Li, Y., Ruppert, M. M., Ren, Y., Shickel, B., Ozrazgat-Baslanti, T., . . . Bihorac, A. Reinforcement learning in surgery. **Surgery**, 2021, 170(1), 329-332. doi: 10.1016/j.surg.2020.11.040.

Davidson, I. Understanding K-means non-hierarchical clustering. **SUNY Albany - Technical Report**, 2002, 02-2, 2.

Dietterich, T. G. Machine-Learning Research. **AI MAGAZINE**, 1997, 18(4), 97-136. doi:10.1609/aimag.v18i4.1324.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. A Density-Based Algorithm for Discovering Clusters. A Density-Based Algorithm for Discovering Clusters (pp. 226-231). 1996, **Portland: KDD-96 Proceedings.**

Favero, L. P., Belfiore, P. **Data Science for Business and Decision Making**. 2019, Sao Paulo: Academic Press**.**

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., Herrera, F. **Learning from Imbalanced Data Sets**. Cham: 2018, Springer. doi:10.1007/978-3-319-98074-4.

Fix, E., Hodges, J. L. Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. **Randolf Field Texas: USAF School of Aviation Medicine**, 1951.

Gautron, R., Maillard, O.-A., Preux, P., Corbeels, M., Sabbadin, R. Reinforcement learning for crop management support: Review, prospects and challenges. **Computers and Electronics in Agriculture**, (200), 2022, 1-14. doi: 10.1016/j.compag.2022.107182.

Géron, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2 ed.).** 2019, Sebastepol: O'Reilly.

Gregoriades, A., Pampaka, M., Herodotou, H., Christodoulou, E. Supporting digital content marketing and messaging through topic modelling and decision trees. **Expert Systems with Applications**, 2021, 184, 1-16. doi: 10.1016/j.eswa.2021.115546.

Hastie, T., Tibshirani, R., Friedman, J. **The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2 b.)**, Stanford: Springer. 2017.

He, H., Bai, Y., Garcia, E. A., Li, S. ADASYN: Adaptive Synthetic Sampling Approach For Imbalanced Learning. 2008, **IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)** (pp. 1322-1328). Hong Kong: IEEE. doi:10.1109/IJCNN.2008.4633969.

Hinton, G. E., Salakhutdinov, R. Reducing the Dimensionality of Data with Neural Networks. **Science**, 2006, 313(5786), 504-507. doi:10.1126/science.1127647.

Hoffman, J. I. **Basic Biostatistics for Medical and Biomedical Practitioners** (2 ed.). 2019, San Diego: Academic Press.

John, M., Shaiba, H. Apriori-Based Algorithm for Dubai Road Accident Analysis. 2019,**16th International Learning & Technology Conference**, 2019 (pp. 218-227). Suudi Arabistan: Elsevier.

Karahan Adalı, G. **Birliktelik Kuralları ile Veri Madenciliğinde Bir Pazarlama Uygulaması Gerçekleştirilmesi**. M. E. Balaban, & E. Kartal

içinde, R ile Veri Madenciliği Uygulamaları (s. 221-223). 2016, İstanbul: Çağlayan Kitabevi.

KDNuggets.com "The 5 Most Useful Techniques to Handle Imbalanced Datasets.", 2020, (Çevrimiçi), https://www.kdnuggets.com/2020/01/5-most-useful-techniques-handle-imbalanced-datasets.html, 25 Mart 2023.

Kejriwal, M., Shen, K., Ni, C.-C., Torzec, N. An evaluation and annotation methodology for product category matching in e-commerce. **Computers in Industry**, 2021, 131, 1-12. doi: 10.1016/j.compind.2021.103497

Koçoğlu, F. Ö., Özcan, T. **Veri Madenciliğinde Kümeleme Algoritmaları ile Müşteri Segmentasyonu**. M. E. Balaban, E. Kartal, M. E. Balaban, & E. Kartal (Dü) içinde, R ile Veri Madenciliği Uygulamaları (s. 193-200). İstanbul: Çağlayan Kitabevi. 2016.

Kubat, M., Matwin, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. 2000, **Fourteenth International Conference on Machine Learning**, 1-8.

Kubat, M., Holte, R. C., Matwin, S. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. **Machine Learning**, (30), 1998, 195-215. doi:10.1023/A:1007452223027.

Lantz, B. **Machine Learning with R Expert Techniques for Predictive Modeling**. Birmingham: Packt Publishing. 2019.

Laurikkala, J. Improving Identification of Difficult Small Classes by Balancing Class Distribution. **Artificial Intelligence in Medicine**, 2001, 63-66. doi:10.1007/3-540-48229-6_9.

Lewis, N. D. **Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners**. United States: CreateSpace Independent Publishing Platform. 2017.

Lian, A. T., Du, J., Tang, L. Using a Machine Learning Approach to Monitor COVID-19 Vaccine Adverse Events (VAE) from Twitter Data. **Vaccines** (Basel), 2022, 10(1), 1-11. doi:10.3390/vaccines10010103.

Loh, W.-Y. Fifty Years of Classification and Regression Trees. **International Statistical Review**, 2014, 82(3), 329-348. doi:10.1111/insr.12016.

Loisel, J., Cornuéjols, A., Laguerre, O., Tardet, M., Cagnon, D., de Lamotte, O. D., Duret, S. Machine learning for temperature prediction in food pallet along a cold chain: Comparison between synthetic and experimental training dataset. **Journal of Food Engineering**, 2022, 335, 1-11. doi: 10.1016/j.jfoodeng.2022.111156.

Maitre, J., Bouchard, K., Bédard, L. P. Mineral grains recognition using computer vision and machine learning. **Computers & Geosciences**, 2019, 130, 84-93. doi: 10.1016/j.cageo.2019.05.009.

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., Zeineddine, H. An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. **IEEE Access**, 2019, 7, 93010-93022. doi:10.1109/ACCESS.2019.2927266.

Malefors, C., Secondi, L., Marchetti, S., Eriksson, M. Food waste reduction and economic savings in times of crisis: The potential of machine learning methods to plan guest attendance in Swedish public catering during the Covid-19 pandemic. **Socio-Economic Planning Sciences**, 2022, 82, 1-11. doi: 10.1016/j.seps.2021.101041.

Marins, M. A., Barros, B. D., Santos, I. H., Barrionuevo, D. C., Vargas, R. E., Prego, T. d., . . . Netto, S. L. Fault detection and classification in oil wells and production/service lines using random forest. **Journal of Petroleum Science and Engineering** (197), 2021, 1-10. doi: 10.1016/j.petrol.2020.107879.

Marsland, S. **Machine Learning An Algorithmic Perspective (2 ed.)**. Boca Raton: CRC Press. 2015.

Morgan, J. N., Sonquist, J. A. Problems in the Analysis of Survey Data, and a Proposal. **Journal of the American Statistical Association**, 1963, 58(302), 415-434.

Müller, A. C., Guido, S. **Introduction to Machine Learning with Python A Guide for Data Scientists (1 ed.)**. California: O'Reilly. 2017.

Nath, S., Korot, E., Fu, D. J., Zhang, G., Mishra, K., Lee, A. Y., Keane, P. A. Reinforcement learning in ophthalmology: potential applications and challenges to implementation. **The Lancet Digital Health**, 2022, 4(9), e692-e697. doi:10.1016/S2589-7500(22)00128-5.

Önder, E. **Sağlıkta Gelişmekte Olan Teknolojiler (1. b.)**. Bursa: Dora. 2020.

Özdemir, S. **Principles of Data Science**. Birmingham: Packt Publishing. 2016.

Paleczek, A., Grochala, D., Rydosz, A. Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection. **Sensors**, 21(12), 2021. doi:10.3390/s21124187

Pavitha, N., Pungliya, V., Raut, A., Bhonsle, R., Purohit, A., Patel, A., Shashidhar, R. Movie recommendation and sentiment analysis using machine learning. **Global Transitions Proceedings**, 2022, 3(1), 279-284. doi: 10.1016/j.gltp.2022.03.012.

Pfaff, E. R., Girvin, A. T., Bennett, T. D., Bhatia, A., Brooks, I. M., Deer, R. R., . . . Niehaus, E. Identifying who has long COVID in the USA: a machine learning approach using N3C data. **The Lancet Digital Health**, 2022, 4(7), e532-e542. doi:10.1016/S2589-7500(22)00048-6.

Radtke, P. V., Granger, E., Sabourin, R., Gorodnichy, D. O. Skew-sensitive boolean combination for adaptive ensembles – An application to face recognition in video surveillance. **Information Fusion**, (20), 2014, 31-48. doi: 10.1016/j.inffus.2013.11.001.

Raglio, A., Imbriani, M., Imbriani, C., Baiardi, P., Manzoni, S., Machine learning techniques to predict the effectiveness of music therapy: A randomized controlled trial. **Computer Methods and Programs in Biomedicine**, 2020, 185, 1-6. doi: 10.1016/j.cmpb.2019.105160.

Gianotti, M., . . . Manzoni, L.

Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. **Encyclopedia of Bioinformatics and Computational Biology (Vol. 1)**. (M. Cannataro, B. Gaeta, A. M. Khan, Eds.) Cambridge: Elsevier. 2019.

Riffenburgh, R. H., Gillen, D. L. **Statistics in Medicine (4 ed.)**. Kaliforniya: Elsevier. 2020.

Sadorsky, P. Forecasting solar stock prices using tree-based machine learning classification: How important are silver prices? **The North American Journal of Economics and Finance**, 2022, 61, 1-15. doi: 10.1016/j.najef.2022.101705.

Saha, D., Manickavasagan, A. Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. **Current Research in Food Science**, 2021, 4, 24-44. doi: 10.1016/j.crfs.2021.01.002.

Sammut, C., Webb, G. I. **Encyclopedia of Machine Learning**. New York: Springer. 2010. doi:10.1007/978-0-387-30164-8

Schapire, R. E. The Boosting Approach to Machine Learning: An Overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, B. Yu, D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), **Nonlinear Estimation and Classification**, 2003, (Vol. 171, pp. 149-171). New York: Springer. doi: 10.1007/978-0-387-21579-2

Scikit-Learn "A demo of K-Means clustering on the handwritten digits data.", 2023, (Çevrimiçi), https://scikitlearn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py, 24 Mart 2023.

Scikit-Learn "Clustering.", 2023, (Çevrimiçi), https://scikit-learn.org/stable/modules/clustering.html#dbscan, 24 Mart 2023.

Scikit-Learn "Linear and Quadratic Discriminant Analysis with covariance ellipsoid.", 2023, (Çevrimiçi), https://scikitlearn.org/stable/auto_examples/classification/plot_lda_qda.html#sphx-glr-auto-examples-classification-plot-lda-qda-py, 24 Mart 2023.

Scikit-Learn "Linear Regression Example.", 2023, (Çevrimiçi), https://scikitlearn.org/stable/auto_examples/linear_model/plot_ols.html, 21 Mart 2023.

Scikit-Learn "Nearest Neighbors Classification.", 2023, (Çevrimiçi), https://scikitlearn.org/stable/auto_examples/neighbors/plot_classification.html#sphx-glr-auto-examples-neighbors-plot-classification-py, 21 Mart 2023.

Scikit-Learn "Permutation Importance vs Random Forest Feature Importance (MDI).", 2023, (Çevrimiçi), https://scikitlearn.org/stable/auto_examples/inspection/plot_permutation_importance.html#sphx-glr-auto-examples-inspection-plot-permutation-importance-py, 25 Mart 2023.

Scikit-Learn "Plot Hierarchical Clustering Dendrogram.", 2023, (Çevrimiçi), https://scikitlearn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html, 24 Mart 2023.

Scikit-Learn "Support Vector Machines.", 2023, (Çevrimiçi), https://scikitlearn.org/stable/_images/sphx_glr_plot_separating_hyperplane_001.png, 21 Mart 2023.

Scikit-Learn "SVM-Kernels.", 2023, (Çevrimiçi), https://scikitlearn.org/stable/auto_examples/svm/plot_svm_kernels.html#sphx-glr-auto-examples-svm-plot-svm-kernels-py, 21 Mart 2023.

Siegel, A. F. **Practical Business Statistics**. Seattle: Academic Press. 2016.

Sutton, R. S., Barto, A. G. **Reinforcement Learning An Introduction (2 ed.)**. Kaliforniya: The MIT Press. 2018.

Şimşek Gürsoy, U. T. **Uygulamalı Veri Madenciliği Sektörel Analizler (3 b.)**. Ankara: Pegem Akademi. 2012.

Tolun, S. **Destek Vektör Makineleri: Banka Başarısızlığının Tahmini Üzerine Bir Uygulama**. İstanbul: İktisadi Araştırmalar Vakfı. 2008.

Tomek, I. Two Modifications of CNN. **IEEE Transactions on Systems, Man, and Cybernetics**, 1976, 769-772. doi:10.1109/TSMC.1976.4309452

Troisi, J., Troisi, G., Scala, G., Richards, S. M. Data analysis in metabolomics: from information to knowledge. In J. Troisi, J. Troisi (Ed.), **Metabolomics Perspectives** (pp. 287-381). Salerno: Academic Press. 2022. doi:10.1016/B978-0-323-85062-9.00009-X.

Tschora, L., Pierre, E., Plantevit, M., Robardet, C. Electricity price forecasting on the day-ahead market using machine learning. **Applied Energy**, 313, 1-14., 2022, doi: 10.1016/j.apenergy.2022.118752

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., Chatzisavvas, K. C. A comparison of machine learning techniques for customer churn prediction. **Simulation Modelling Practice and Theory**, 55, 1-9., 2015, doi: 10.1016/j.simpat.2015.03.003

Wilson, D. L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. **IEEE Transactions on Systems, Man, and Cybernetics, SMC-2(3)**, 408-421., 1972, doi:10.1109/TSMC.1972.4309137

Zhu, B., Baesens, B., vanden Broucke, S. K. An empirical comparison of techniques for the class imbalance problem in churn prediction. **Information Sciences**, (408), 84-99., 2017, doi: 10.1016/j.ins.2017.04.015