AYKIRI DEĞERLER VARLIĞINDA SINIFLANDIRMA YÖNTEMLERİ

CLASSIFICATION METHODS IN THE PRESENCE OF OUTLIERS

CEMİLE AŞLAR KIRMIZI

DR. ÖĞR. ÜYESİ ONUR TOKA

Tez Danışmanı

Hacettepe Üniversitesi

Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin

İstatistik Anabilim Dalı için Öngördüğü

YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2023

ÖZET

AYKIRI DEĞERLER VARLIĞINDA SINIFLANDIRMA YÖNTEMLERİ

Cemile AŞLAR KIRMIZI

Yüksek Lisans, İstatistik Bölümü
Tez Danışmanı: Dr. Öğr. Üyesi Onur TOKA
Ocak 2023, 70 sayfa

Veri madenciliğinde denetimli öğrenme başlığı altında yer alan sınıflandırma tekniklerinin önemi, sürekli değişen, çeşitlenen ve çoğalan verilerin hızıyla beraber artmaktadır. Verilerdeki bu değişkenlik, sınıflandırma tekniklerinin de değişim ve gelişimi ihtiyacını doğurmaktadır. Temel sınıflandırma teknikleri altında önerilen sağlam (robust) sınıflandırma teknikleri, yanlış sınıflandırma oranlarının düşmesi için geliştirilmektedir. Verilerdeki gelişim ve değişimlerin meydana getirdiği aykırı değerlerin varlığında, doğru sınıflandırma yönteminin bulunabilmesi gün geçtikçe önemini arttırmaktadır.

Bu çalışmada makine öğrenmesi başlığı altında toplanan bazı sınıflandırma teknikleri incelenmiştir. Literatürde en çok kullanılan ve kaynaklarda başarılı olarak nitelendirilen algoritmalar ile benzetim ve gerçek veri kümeleri üzerinde analizler yapılmış ve yorumlanmıştır. Sınıflandırma algoritmalarının tahmin hatalarını sayısallaştırarak yorumlayabilmek için Eşik Değerleri (Threshold Metrics) kullanılmıştır. Sınıflandırma algoritmalarının başarıları duyarlılık (sensitivity), belirlilik/özgüllük (specificity), genel doğruluk oranı ve F1-değerileri hesaplanarak, bu veriler üzerinden değerlendirilmiştir.

Değerlendirme çeşitlendirilerek 4 tip benzetim veri kümesi ve 2 farklı gerçek veri üzerinden analizler yapılmıştır. Yapılan analiz sonuçları tablolaştırılmış, yorumlanmış ve F1-değerlerinin grafiksel gösterimlerinden faydalanılmıştır.

Benzetim veri kümeleri ile analizlerde lojistik regresyon, benzer özelliğe sahip olan sağlam lojistik regresyon, tanjantboost, gudermannianboost algoritmaları, sağlam doğrusal ayrıştırıcı (Robust Linear Discriminant Analysis - RLDA) ve sağlam karesel ayrıştırıcı ((Robust Quadratic Discriminant Analysis - RQDA), OGK (Ortogonolize Gnanadesikan- Kettenring) kestiricili sağlam doğrusal ayrıştırıcının (RLDA-OGK) başarıları ön plana çıkmıştır. Gerçek veri kümelerinin çalışıldığı analiz sonuçlarında ise lojistik regresyon, sağlam lojistik regresyon, tanjantboost, gudermannianboost algoritmaları, duyarlılık (sensitivity), belirlilik/özgüllük (specificity), genel doğruluk oranı ve F1-değerlerinin tümü göz önüne alındığında önemli bir farkla başarılı bulunmuşlardır.

Anahtar Kelimeler: Eşik Değerleri, F1-değerleri, Makine Öğrenmesi, Sağlam Sınıflandırma Algoritmaları, Sınıflandırma Algoritmaları

ABSTRACT

CLASSIFICATION METHODS IN THE PRESENCE OF OUTLINERS

Cemile AŞLAR KIRMIZI

Master of Science, Department of Statistics
Supervisor: Asst. Prof. Dr. Onur TOKA
January 2023, 70 pages

In the field of data mining, the importance of classification techniques categorized under supervised learning is increasing with the constantly changing, diversifying and multiplying data. This variety in data creates the need for change and advancement of classification techniques. Robust classification techniques under basic classification techniques are being developed in order to reduce misclassification rates. In the presence of outliers caused by advancements and changes in the data, finding the right classification method increases its importance day by day.

In this study, some classification techniques gathered under machine learning were examined. Analyzes were made and interpreted on simulation and real data sets with algorithms that are most used in the literature and described as successful in the sources. Threshold Metrics were used to interpret the prediction errors of classification algorithms by digitizing them. The success of classification algorithms was evaluated based on these data by calculating sensitivity, specificity, overall accuracy and F1-scores. By diversifying the evaluation, analyzes were made on 4 types of simulation data sets and 2

different real data. The results of the analysis were charted, interpreted and graphical representations of the F1-scores were used.

In the analyzes using simulation datasets, the successes of logistic regression, robust logistic regression with similar features, tangentboost, gudermannianboost algorithms, robust linear discriminant analysis (RLDA) and robust quadratic discriminant analysis (RQDA), robust linear discriminant analysis with OGK estimator (RLDA-OGK) came forward. In the analysis results where real datasets were studied, logistic regression, robust logistic regression, tangentboost, gudermannianboost algorithms, sensitivity, specificity, overall accuracy and F1-scores were all found to be successful with a significant margin.

Keywords: Threshold Metrics, F1-score, Machine Learning, Robust Classification Algorithms, Classification Algorithms

İÇİNDEKİLER

ÇİZELGELER

ŞEKİLLER

KISALTMALAR

AI	Artificial Intelligence
CART	Sınıflandırma ve Regresyon Ağaçları
DBN	Derin İnanç Bağı (Deep Belief Network)
DNN	Derin Sinir Ağları (Deep Neural Network)
DVM (SVM)	Destek Vektör Makineleri (Support Vector Machine)
K-NN	K-En Yakın Komşu (K-Nearest Neighbours)
LDA	Doğrusal Ayrıştırıcı Analizi (Linear Discriminant Analysis)
QDA	İkinci Dereceden Ayrıştırıcı Analizi
PR	Kesinlik Geri çağırma (Precision Recall)
RLDA	Sağlam Doğrusal Ayrıştırıcı Analizi
ROC	Alıcı İşletim Karakteristiği (Receiver Operating Characteristic)
RQDA	Sağlam Karesel Ayrıştırıcı Analizi
VM (DM)	Veri Madenciliği (Data Mining)

1. GİRİŞ

Veri, istatistiksel yöntemlerinin, istatistiksel öğrenmenin, makine öğrenmesinin ve yapay zekanın geliştirilmesinde ve güncellenmesindeki en önemli girdidir. Bilgisayar ve ardından internetin hayatlarımızın son on yıllarına çok hızlıca entegre olması ile dünya çapında kontrolsüzce büyüyen veri yığınları, veri temelli olan tüm öğrenme yöntemlerinde çeşitli algoritmalarının gelişmesine hız katmıştır. Veri yığınlarını okuyabilmek, anlamlı modeller oluşturabilmek, modeller ile tahminler ve sonuçlar elde edebilmek için veri madenciliği yani verinin kıymetli olanını ortaya çıkarma işlemini istatistik, matematik ve bilgisayar bilimleri ile harmanlayan yöntemler kullanılır. Bu yöntemler, analistler yardımıyla bir veri tabanı üzerinden kullanılırken istatistiksel öğrenme yöntemleri, bilgisayar ve yazılımlar yardımıyla kendini iteratif ve veriye uygun şekilde modelleyen, algoritmasını geliştirebilecek otomatik sistemler üzerinden yapıldığında ise makine öğrenme yöntemleri adını alır. Benzer şekilde, ortaya çıkacak bilgiyi insan kaynağı olmadan eyleme dönüştürebilen yapılar haline getirdiğinde ise yapay zeka yöntemleri olarak adlandırılabilir. Yöntemler temel olarak farklı alanlarda kullanılıyor olsa bile altındaki algoritmalar ve mantıklar benzer sistemlerle ortaya çıkartılmaktadır. Verinin hızlıca artması, veriyi anlamlandırma isteği ve teknolojinin bu isteğe cevap verme arayışı, bilgisayar ve veri tabanında devasa gelişimler yaratmış ve bunun sonucunda yukarıda bahsedilen tüm yöntemlerin altında kullanılabilen matematik, istatistik ve bilgisayar temelli algoritmalar geliştirilmiştir.

Teknolojinin artışı, verilerin fazlalaşması ile birleşince, önceden kullanılan birçok algoritma zayıf yanlarını tamamlayan başka yöntemlerle güçlendirilmiş, eksiklikler oluşunca yeni algoritmalar ve yöntemler geliştirilmiştir. Tüm bu yöntemler, tahmin etmeye çalıştığı bilgiye elindeki veride sahipse denetimli öğrenme ve tahmin etmeye çalıştığı bilgi verinin içinde yoksa denetimsiz öğrenme olarak adlandırılan iki genel isim altında toplanmaktadır. Denetimli öğrenme regresyon ve sınıflandırma algoritmalarını kullanarak öğrenmeyi amaçlar. Veriden öğrenerek çalışan algoritmalar sayesinde çözümlemeler yapılması kolaylaşır. Veri madenciliği algoritmalarından birisi olan sınıflandırma algoritmaları birkaç aşamadan oluşur. Öncelikle problemin çözümüne giden yolu öğrenir ve öğrendiği bu yolu yeni problemlerde kullanarak çözümlemeye gitmeyi amaçlar. Dünya üzerinde tarımsal kalkınmadan bankacılığa, medya alanından

sağlığa, internet ve sosyal ağlardan market alışverişlerine kadar her alanda her sektörde durum analizleri, modern gelişmeler ve kalkınma stratejileri geliştirilmektedir. Bu çalışmalar için temelde güncel veri kümeleri üzerinden yapılan analizler ile veriden öğrenilen bilgiler anlamlandırılır. Veriden öğrenme, eğitim verileri üzerinden oluşturulan modelin, test verisi üzerinde sınanmasını konu alır.

Güncel veri kümeleri çoğalırken içindeki hatalı, verinin genel durumundan uzaklaşabilen, çok olası olmamasına rağmen ortaya çıkabilecek veriler de artmaktadır. Bu yapılar aykırı değer (outlier) olarak adlandırılmaktadır. Var olan aykırı değerler analizlerin doğru sonuçlara ulaşmasını manipüle edebilmektedir. Bu nedenle "çalışılan verideki" aykırı değerlerin sınıflandırma çalışması üzerindeki etkisi ve bu etkiyi arındırarak "daha başarılı" sınıflandırma elde edebilmek için var olan yöntemlere geliştirmeler yapmak ve alternatif yöntemler önermek, son yıllardaki literatürde oldukça sık rastlanan çalışmalardır.

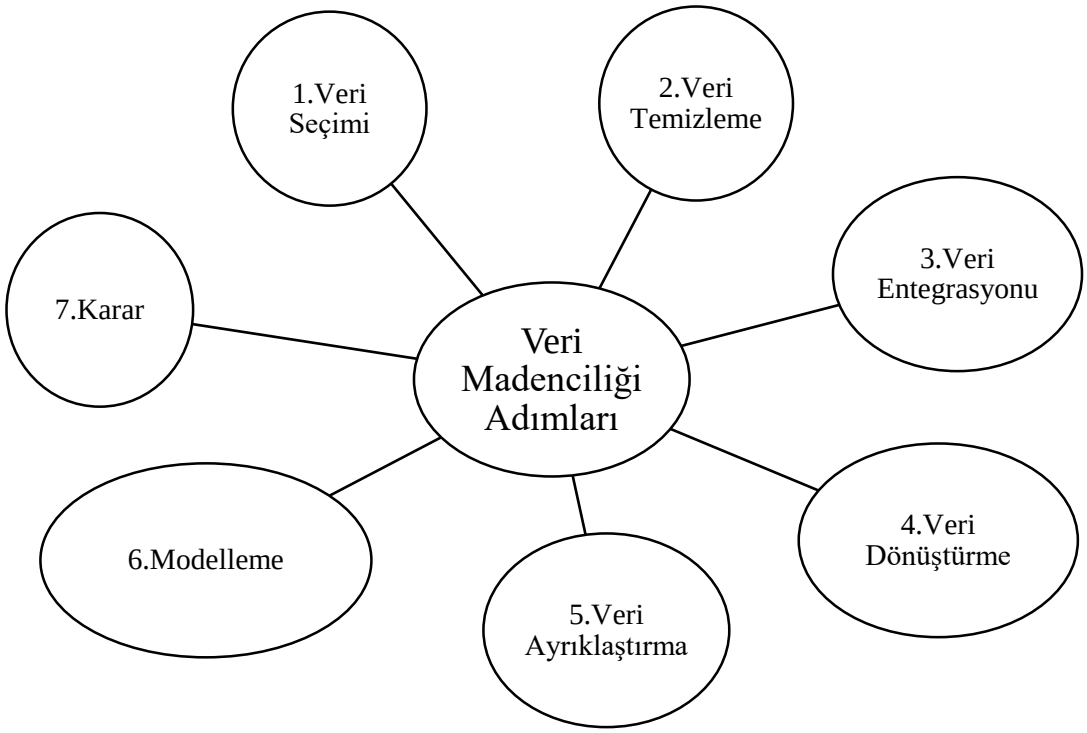
Bu çalışmada, bazı önemli tanımlamalardan sonra denetimli öğrenme yöntemleri içinde yer alan sınıflandırma yöntemlerinden eğitim kümesinde aykırı değer bulunması durumunda kullanılması önerilen sağlam (robust) sınıflandırma özelliğine sahip yöntemler incelenmiştir. Bu yöntemlerden son yıllarda önerilen ve birbirlerine karşı alternatif olarak gösterilenler, klasik bazı yöntemler de dahil edilerek aykırı değer oranı, değişken sayısı, gözlem sayısı, eğitim, test ve geçerlilik oranlarına göre daha genel hale getirilmiş geniş bir benzetim (simülasyon) ile karşılaştırılmıştır. İlgili yöntemler genelleştirilmiş benzetimin kısıtlarına göre karşılaştırılmış ve yorumlanmıştır.

2. GENEL BİLGİLER

2.1. Veri Madenciliği

Veri bilimi (data science) altında değerlendirilen veri madenciliği (data mining) (VM-DM), verinin ham halinden yola çıkarak, veri üzerindeki bilgileri, birbiri ile ilişkileri ve veri yapılarını öğrenmek üzere geliştirilmiş algoritmalar kullanır. Tarihsel olarak VM mevcut haliyle, son birkaç on yılda klasik istatistik ve yapay zekanın (artificial intelligence - AI) etkileşiminden evrimleşmiştir [1].

Literatürde veri madenciliği için birçok tanımlama yapılmıştır. Veri madenciliğini, Fayyad, bir veri tabanında saklanan verilerden örtük, daha önce bilinmeyen ve potansiyel olarak yararlı bilgilerin bir şekilde çıkarılması süreci olarak tanımlarken; Giudici de eldeki büyük miktarlardaki veriden net ve faydalı sonuçlar elde etmek amacıyla başlangıçta bilinmeyen düzenlilikleri veya ilişkileri keşfetmek için seçme, keşfetme ve modelleme süreci olarak tanımlamıştır [2].
Veri madenciliği adımlarına kısaca değinmek gerekirse Şekil 2.1.'de 7 başlık altında incelenebilir.



Şekil 2.1. Veri Madenciliği Adımları

1. Veri Seçimi (Data Selection): Birincil verileri elde etmek maliyetli ve zaman alıcı bir süreçtir. Veri toplamanın ana tekniklerinden başlıcaları gözlem, görüşmeler, anketler, programlardır. Toplanan veriler bilgisayar ortamlarında değerlendirilmek üzere yapılandırılırlar. Çalışma konusuna göre toplanan veriler derlenir ve veri yığınları çalışılmak üzere seçilir.

2. Veri Temizleme (Data Cleaning): Bir dizi tablo, veri tabanı ve kayıt kümesinden yanlış ve aldatıcı verilerin tanımlanması ile mevcut verilerin kaldırılması veya eksik bilgilerin ortalamalar gibi genel sabitler ile doldurulması veya değiştirilmesi gibi çözümlerin kullanıldığı bir adımdır.

3. Veri Entegrasyonu (Data Integration): Bazı çalışmalarda eldeki mevcut veri setinin yeni veri setleri ile birleştirilmesi adımıdır. Ancak yeni veriler mevcut veri seti ile tam olarak uyumlu olmayabileceği için farklı uygulamalar yardımı ile entegrasyon/birleştirme işlemi yapılabilir.

4. Veri Dönüştürme (Data Transformation): Çalışmada kullanılacak kaynak sistem verilerinin hedef sisteme formatlanırken genellikle dönüştürülmesi gerekir. Bunun için kullanılan tekniklerden bazıları yumuşatma (smoothing), toplama (aggregation), normalleştirme (normalization) ve genelleştirme (generalization) olarak bilinmektedir.

5. Veri Ayrıklaştırma (Data Discretization): Sürekli verilerin bir nevi sınıflandırılarak veri kümesinin küçültülmesini ve küçük alt kümeler haline getirilmesini sağlar. Bu şekilde veriler üzerinde çalışmak daha kolay bir hale getirilir. Veri ayıklama işlemleri Veri dönüştürme başlığı altında da incelenebilir. Ancak dönüştürme için bazı özel teknikler kullanılmakta olup veri dönüşümü daha sistematik yapıldığı için ayrıklaştırmayla karıştırılmamalıdır.

6. Modelleme (Modelling): Veri madenciliği modelleri, veri madenciliğinin çekirdeğini oluşturur. Veri tablolarına benzerdir ancak tablolar gerçek verileri temsil etmek için kullanılırken, madencilik modelleri bu verilerin olay/durum (cases) olarak literatürde adlandırılan yorumlarıdır. Model mevcut bilgiler ve tahminlerle oluşturulur. Veri madenciliği modeli, yapısal olarak modelin girdi, çıktıları, model düğümleri ve veri madenciliği algoritmasından oluşturularak eğitilir [3].

7. Karar (Decision): Modelde elde edilen sonuçlar tercih edilen yöntemler kullanılarak karşılaştırılabilir. Tablolaştırılarak, grafiksel gösterimlerle veya modelin yapısal görselleri kullanılarak sonuçlar üzerinden değerlendirmeler yapılabilir.

Kapsamlı analizler büyük verilerle çalışmayı gerektirir. Büyük verilerle çalışırken çoğunlukla aykırı değerlerle (hatalı veriler/gürültülü etiketler) karşılaşılır. Gerçek dünya senaryosunda büyük verilerle çalışıldıkça verilerin aykırı değerler içermesi kaçınılmazdır. Bu durum bazı veri madenciliği yöntemlerinde genelleme performansını ciddi şekilde düşürmektedir.Bu durumda veri madenciliğinde, aykırı değer varlığında geliştirilen ve literatürde sağlam öğrenme (robust learning) olarak bilinen sınıflama algoritmaları sayesinde veriden başarılı öğrenmeler gerçekleşir.
Tahminlerin test kümelerinde başarılı olabilmesi için öğrenmenin gerçekleştiği eğitim veri kümesinin süreci başarılı şekilde açıklayabilecek yapıda kurulmuş olması gerekir. Gerçek sürece uygun olmayan verilerin eğitim kümesinde bulunması, sınıflandırıcı fonksiyona etki ederek modelin performansını düşürebilir. Bu şekilde gerçek verilere uygun olmayan veriler varlığında en iyi sınıflamanın ve tahminin yapılabilmesi için sağlam sınıflandırma yöntemleri kullanılması önerilmektedir.

2.1.1. Makine Öğrenmesi

Yapay zekanın bir alt dalı olan makine öğrenmesinin (machine learning) kullanım alanlarını hayatımızın her anında görebiliyoruz. Teknolojinin takip edilmekte zorlanılan gelişimi ile bazen farkında bile olmadan kullanıyor ve faydalanıyoruz. Sosyal hayatımızdan, iş hayatımıza, sağlık alanlarından eğitim alanlarına, kısaca yaşamın her alanına yayılmış durumda ve teknolojinin sunduğu “kolaylıklar” ile artık otomatik olarak kullandığımız boyuta gelmiştir. Örneğin gören, duyan ve okuyan akıllı telefonlar ile oluşturulan veriler işlenerek teknolojinin her an daha ileri bir boyuta taşınması sağlanmaktadır. Sadece internette ekran karşısında gezilen sayfalar değil, attığımız adımlardan uyku saatlerine, içtiğimiz sulardan, gezdiğimiz yerlere, kalp atışlarımızdan oturma süremize, baktığımız sitelerden konuştuğumuz konulara kadar biz hiç farkında bile değilken veri oluşturuyoruz. Bu verilerin işlenerek farklı alanlarda kullanılmasının veya bir şekilde karşımıza çıktığının çoğu zaman farkına bile varmadan yaşamaya devam ediyoruz.

Makine Öğrenmesi son yarım yüzyılda birçok alanda kullanılmaktadır. Tarihte Yapay Zekanın ilk defa 1950'lerde bir bilim dalı olarak ele alınmaya başlanmasından itibaren, yapay zekanın önemli bir konusu olarak Makine Öğrenmesi de araştırmalara konu olmuştur. Bu beraber işleyen mekanizmayı Quinlan [2] şu şekilde ifade eder, Öğrenme yeteneği, akıllı davranışın bir özelliğidir, bu nedenle zekayı bir fenomen olarak anlamaya yönelik herhangi bir girişim, bir öğrenme anlayışını içermelidir. Öğrenme ve işleme birlikte yürüyen ayaklar gibi, biri dururken diğerinin yol kat etmesi beklenemez. Öğrenme tabanlı yapay zeka, bilgiyi içeren veriden öğrenerek algoritmaların çalışmasını sağlıyor. Verilere dayanarak hızlı otomatik kararlar veren programlar geliştirmeyi amaçlamıştır.

Kuralların modele tek tek verilmesi yerine, modelin kuralları öğrenerek karar verebilir olması makine öğrenmesidir. Klasik bir örnek olan 'mail süzme' işlemini düşünelim, spam/çöp mail olarak adlandırılan maillerde olabilecek kelimeleri baştan belirtmek, istenen performansı vermeyeceği gibi ayrıca çok zaman alan yorucu bir iştir. Ancak bir model oluşturulup makinenin kelimeleri öğrenip ona göre mailleri sınıflandırması kolaylaşan bir yöntemdir. Model eğitildikçe öğrenme algoritmasındaki matematiksel yapılar gelişir, iyileşir, algoritmanın becerisi artar, beklenen performansı yükselir.

Verilerle çalışırken, istatistik anlamında sınıflandırma denildiğinde, çok fazla gözlemle çalışılması durumunda, gözlemlerin kolay ölçülebilir ve analiz edilebilir olması için sınıflandırdığımız ve frekanslarıyla tablolar oluşturduğumuz durumlar vardır. Makine öğrenmesinde kısaca sınıflandırma, kategori tahmini diye nitelendirilir [4]. Bu çalışmaya konu olan sınıflandırma yöntemleri, daha önceden belirlenmiş sınıflar üzerinden yapılmaktadır. Kalabalık sınıflar varlığında veri analizi sırasında faydalı bir katkısı olmayan sınıflar çıkartılır. Analiz faydalı, etkili olabileceği düşünülen sınıflar üzerinden devam eder. Doğrusal ayrıştırıcı yöntemlerinde (lineer diskriminantlarda) gereksiz değişkenleri kaldırmak için yerleşik prosedürler (örneğin, ileriye doğru adım adım seçim) vardır [5].

Makine öğrenme teknikleri 1990'larda, denetimli (supervised), denetimsiz (unsupervised) ve yarı denetimli olarak üç ana başlıkta incelenmekteyken, zamanla farklı isimlerle anılan teknikler literatüre girmiştir. Ayrıca bunlardan başka Topluluk (Ensemble) öğrenimi, Takviyeli/Güçlendirilmiş (Reinforcement) öğrenme ve Aktif

(Active) öğrenme teknikleriyle de karşılaşılabilmektedir. Farklı özellikler ve uygulama yöntemlerine sahip bu teknikler literatür çalışmalarında birbirleri ile iç içe geçmişçesine çalışmalara yön veriyor. Topluluk öğrenmesi, dört kategoride gruplandırılabilir: denetimli topluluk sınıflandırması, yarı denetimli topluluk sınıflandırması, kümeleme topluluğu ve yarı denetimli kümeleme topluluğu [6].

2.1.2. Makine Öğrenme Teknikleri

Makine Öğrenmesi veriden bilgi elde etme bilimidir. Verideki bilginin işlenmesini ve o bilgiyi açığa çıkaran algoritmaların geliştirilmesidir. Verilerden öğrenen programlar da, öğrenme sürecinin temeli olarak mevcut örneklerden yola çıkar ve girdiler üzerinden çıktılara ulaşmak için bir haritalandırma yapar. Öncelikle eldeki bilinen veriler öğrenilir, daha sonra bilinmeyen bir sonuç için tahmin yapılabilir. Bu tahminler süreklilik içeren sayısal sonuçlarla ilgili ise regresyon, nitelik olarak etiketleme çalışmaları ile ilgili ise sınıflandırma veya kümeleme olarak değerlendirilebilir.

Verilerde bir dizi gözlem içerisinde sınıfların veya kümelerin varlığını belirlemek amacıyla denetimsiz öğrenme (kümeleme) kullanılır; ne kadar sınıf olduğunu kesin olarak bilindiği durumlarda amaç, yeni bir gözlemi mevcut sınıflardan birine sınıflandırabilecek bir kural oluşturmak ise, denetimli öğrenme tekniklerinin kullanılması uygundur [7].

Makine öğrenmesinde tahminler, devamlılık içeriyorsa, çıktı somutlaştırılabilen bir değer niteliğindeyse regresyon olarak tanımlanabilir; tahminler kategorik yapıda olacaksa ve var olan etiketler göz önünde bulundurularak bir etiketleme işleminin yapılmasını kapsıyor ise sınıflandırma; etiketlerin önceden belirlenmemiş olması durumu ise kümeleme olarak tanımlanabilir. Denetimli ve Denetimsiz öğrenmeler aslında önceden sınıfların analizciler tarafından belirlenmesi ve belirlenmemiş olması durumlarını ifade eder. Denetimli öğrenmede sınıflar bellidir. Verilerin yer aldığı sınıflar bilindiğinden, analiz sınıflarda yer alan verilerin ışığında sınıfların özellikleri üzerinden yapılır. Girdi olarak belli sınıflara ait veri bilgilerinden öğrenerek, yeni verilerin hangi sınıflara ait olabileceği tespit edilir.

1) Denetimli Öğrenme (Supervised Learning): Gerçek sınıfların bilindiği bir dizi veriden bir sınıflandırma prosedürünün oluşturulması, çeşitli şekillerde isimlendirilir:

örüntü tanıma, ayırt etme veya denetimli öğrenme [5]. Etiketlenmiş eğitim verilerinden model öğrenilir. Etiketler kesikli, kategorik veya sürekli olabilir. Girdi ve çıktıların bilindiği durumlardır. Girdi-çıktı ilişkisinin öğrenilerek analizlerin bunun üzerinden yürütülmesini sağlayan yöntemlerdir. Neyin nasıl öğrenileceği baştan belirlenmiş kurallar üzerinden gidilir. Daha önceden oluşmuş, geçmiş verilere dayanarak geleceğe yönelik tahminlerin yapılabilmesi de bu öğrenme tekniği ile mümkün görülür. Bağımlı değişken hakkında bilginin varlığında, bağımsız değişkenleri hangi sınıfa gireceğini öğrenme çalışmalarıdır.

2) Denetimsiz Öğrenme (Unsupervised Learning): Sınıfların verilerden çıkarıldığı öğrenme tipidir. Verilerin sahip olduğu özellikler ve taşıdıkları bilgilere göre sınıflamalar yapılır. Veri madenciliğinde sınıflamanın veriden yapıldığı bilindiğine göre, en doğru sonuçlar için çalışılan verinin de en doğru bilgiye ulaştıracak bir kaynak olması istenir.

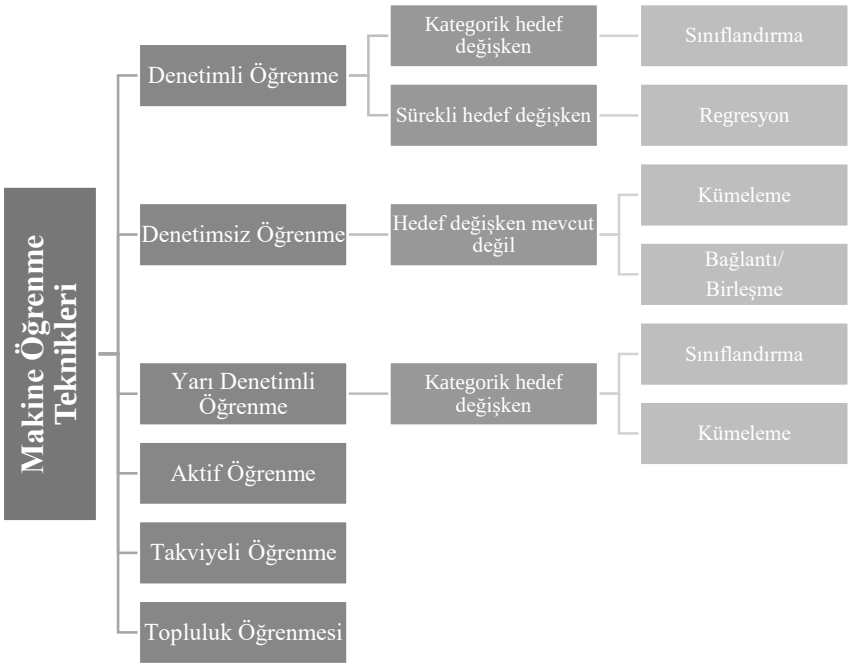
Mevsimlere ilişkin görüntü veri seti üzerinde denetimsiz algoritmanın çalıştırıldığını varsayalım. Algoritma verilen veriler üzerinden eğitilmemiş ve veri kümesinin özellikleri hakkında bilgisi olmayan bir yapıdadır. Görevi, görüntü özelliklerini kendi başına tanıyıp, görüntüler arası benzer özelliklere göre kümeleme yapmaktır. Sınıfların verilerden çıkarıldığı bu öğrenme tipi, "denetimsiz öğrenme" veya "kümeleme" olarak kaynaklarda adlandırılmıştır.

3) Yarı Denetimli Öğrenme (Semi Supervised Learning): Sınıflandırmanın başarısını arttırmak için daha ucuz ve etkili bir yol olarak tercih edilir. Aynı zamanda insan çabası gerektiren etiketleme işlemi azaltılmış olması, sınıflandırmanın daha kolay ve ucuz olmasında ve doğruluk oranının artmasında etkilidir.

4) Aktif Öğrenme (Active Learning): Aktif öğrenme, eğitim verilerinin aktif olarak ve ekstra esneklikle seçilmesine olanak tanımasıyla bilinir. Bu, eğitim için gerekli veri seçimini etkileyerek büyük miktarda etiketlenmiş veri ihtiyacını azaltır [8].

5) Takviyeli Öğrenme (Reinforcement Learning): Görevin nasıl yerine getirileceğini belirtmeye gerek kalmadan ödül ve ceza yoluyla programlama ajanlarını kandırmaktır [9].

6) Topluluk Öğrenmesi (Ensemble Learning): Topluluk öğrenme yöntemleri, veriler üzerinde çeşitli projeksiyonlar yoluyla çıkarılan özelliklere dayalı zayıf tahmine dayalı sonuçlar üretmek için çoklu makine öğrenimi algoritmalarından yararlanır ve tek başına herhangi bir kurucu algoritmadan elde edilenden daha iyi performans elde etmek için sonuçları çeşitli oylama mekanizmalarıyla birleştirir [6]. Genel olarak literatürde 4 çeşidi yer almaktadır; Denetimli topluluk sınıflandırması, yarı denetimli topluluk sınıflandırması, kümeleme topluluğu ve yarı denetimli kümeleme topluluğudur [6]. Başarıyı arttırmak için birden fazla tekniği aynı anda kullanarak daha iyi sonuçlar elde etmeyi amaçlar.



Şekil 2.2. Makine Öğrenme Teknikleri [10]

Veriden öğrenen yapıdaki algoritmalar aşırı uyum/ezber (overfitting) sıkıntılarından kurtulduğunda sağlıklı sınıflamalar yapabilirler. Çünkü verilerin ezberlenmiş özellikleri için ağ tahminleri de belirsizlik içerecektir.

Aykırı değerler varlığında veriden öğrenmenin literatürdeki karşılığı ise sağlam eğitimdir (robust training). Aykırı değerlerin olumsuz etkilerinin azaltılarak doğru sınıflama başarısını arttırmayı hedefleyen sağlam eğitimler aykırı değerlerden faydalanarak gelişir. Aslında her veri gibi aykırı değer de modeli eğitir.

Son yıllardaki çalışmalarda küçük gruplar olarak veride var olan aykırı değerlerden faydalanmanın yöntemlerini daha sonra değineceğimiz modellemeler arasında yer alan Derin Sinir Ağları (DNN) modelleri ışığında araştırılmaktadır. Araştırmalar farklı yöntemler sunarken, bu çalışmada kullanılan yöntemler içinde teze konu olan veri seti için hangi yöntemin daha iyi sonuç verdiği araştırılmıştır.

Ezberden kaçınmanın en temel yolu büyük veriyle çalışmaktır. Büyük veri olmadığı durumlarda başarıyı arttırmak için sentetik veri üretme yöntemlerine başvurmak gibi farklı çözümler üretilmektedir. Bu çalışmada da aynı şekilde büyük veriyle çalışarak ezberin en aza indirilmesi de temel alınmıştır.

2.2. Sınıflandırma

Denetimli öğrenmenin bir alt kategorisi olan sınıflandırma çalışmaları, ilk olarak "klasik" aşama, Fisher'in doğrusal ayrımcılık teorisi üzerinden, ikinci olarak "modern" aşama, daha esnek model sınıflarından yararlanılarak oluşturulmaya başlamıştır. Günümüzün gelişen konularından biri olan sınıflandırma için sayısız kaynak ve bilgiye ulaşılıyor. Literatür taramalarında makine öğrenmesinde sınıflandırma başlığı altında daha derinlere inildiğinde Marcello Pelillo'ya ait makaleye göre, bir sınıflandırma tekniği olan K-En Yakın Komşu (KNN), İslam ile bilimin iç içe olduğu ve literatürde İslam'ın altın çağı olarak nitelendirilen yıllara dayanmakta olduğunu gösteriyor. Dünyanın ismi bilinen ilk bilim insanlarından, "İbnü'l-Heysem", Batı dünyasındaki adıyla "Alhazen", Irak'ta doğmuştur, İbn Sînâ ve Bîrûnî ile aynı dönemde yaşamıştır. İbnü'l Heysem'in deney yapmanın inceliklerini bilim dünyasına göstermesi ile deneye dayalı bilimlerin temel taşını oluşturduğu belirtilir. Marcello Pelillo [11] makalesinde, birçok bilim için öncü olan İbnü'l Heysem'in optik ile ilgili yaptığı çalışmasında fazla öne çıkarılmasa da temel bir sınıflama tekniğinin varlığını göstermiştir. Cover ve Hart'ın NN kuralıyla temelde aynı bir sınıflandırma mekanizmasını kullanmış olduğu ortaya çıkarılmıştır [11].

En iyi sınıflandırma, temiz yansız verinin kaybını en aza indirmelidir. Eldeki doğru bilgi içeren verilerin sınıflandırmasını yaparken ise olası aykırı değer içeren verilerin de doğru sınıflandırılabilmesi çok önemlidir. Çünkü analizi yapılmak istenen veri kümeleri veya üzerinde çalışılan konu istatistiklerinde hiçbir veri hatasız değildir. Gün geçtikçe verilerin taşıdığı bilgiler artmakta, her alanda her konuda veri yoğunluğunu getirmektedir. "Çok

söz yalansız, çok para haramsız olmaz" ata sözümüz aslında durumu çok güzel özetlemiş; "Çok/büyük veri hatasız olmaz." Çalışılan veya toplanan her veri grubunda yanlışlıklar, hatalar oluşması engellenememektedir. Ancak büyük verinin kıymetinin artmasıyla, aykırı değerler varlığında en iyi sınıflandırmayı yapabilmek ve en iyi sonuçlara ulaşabilmek de hem zorlaşmış hem önem kazanmıştır.

Veri madenciliğinde sınıflandırma iki çeşit ile genelleştirilmiştir. İlki ikili sınıflandırma (binary classification); tek bir hedef değişken kullanılarak yapılan analizler bu kategoride yer alır [12]. Bu tarz sınıflandırma problemlerinde örneğin hastalığın varlığı veya yokluğu sonuçlarına ulaşılmak istenebilir. İkinci problem çeşidi, birden fazla hedef değişkenin varlığında yapılan analizlerdir. Bu tarz sınıflandırma problemlerine örnek olarak da hastalığın çeşidi üzerinden değerlendirmeler, analizler yapılır. Örneğin kişinin burada bir hastalığının varlığından ziyade çeşidi ile ilgilenilebilir. Kişinin grip, alerjik, astım gibi çeşitlerden hangisinin sınıfında yer alacağı araştırma sonuçları ile gösterilir.

Sınıflandırmada etiketler kesikli veya kategorik olabilir. Görsel bir analizde arabanın resimde olup olmadığı, belirtiler ve mevcut veriler ile hastalığın olup olmadığı, mailin spam mail olması gibi durumlar denetimli makine öğrenme teknikleri ile model kurularak analiz edilebilir. Bu şekilde kesikli veya sıralı olmayan kategorik değişkenler üzerinden ikili sınıflandırma yapılabilir. Ayrıca algoritmalar kurallar kümesi, ikiden fazla yani çok kategorili sınıflandırmalar için de oluşturularak (örneğin sayılar, harfler vb.) çalıştırılabilir. Temel olarak, sınıflandırma bir etiketi tahmin etmekle ilgilidir ve regresyon bir miktarı tahmin etmekle ilgilidir ve her ikisi de tahmine dayalı modellemedir ki bu, girdilerden çıktılara yapılan bir yolculuk şeklidir [13]. Denetimli makine öğrenimi şemsiyesi altında yer edinmiştir [13]. İstatistik, Makine Öğrenimi ve Sinir Ağı ana başlıkları büyük oranda farklı akademik gruplardır ancak "sınıflandırma" alanında ortak amaçları vardır: Bu ortak amacın girdilerden çıktılara ulaşma yolunda tahmin içermesidir. Her bir sınıf içindeki özelliklerin ortak dağılımının bir tahmini üzerinden bir sınıflandırma kuralı elde edilmeye çalışılmıştır. Michie, Spiegelhalter ve Taylor [5] aşağıdaki gibi özetlemektedirler:

- Karar verici olarak insan ile eşit düzeyde, tutarlı ve belli oranda değişkenliğe sahip olmak,
- Çok çeşitli sorunları ele almak ve yeterli veri ile son derece genel sonuçlar vermek,
- Kanıtlanmış başarı ile pratik ortamlarda kullanılabilir olmak.

İstatistiksel yaklaşımlar genellikle, basit bir sınıflandırmadan ziyade her bir sınıfta olma olasılığını sağlayan açık bir temel olasılık modeline sahip olmaları ile karakterize edilir. Ek olarak, genellikle tekniklerin istatistikçiler tarafından kullanılacağı varsayılır ve bu nedenle değişken seçimi ve dönüşümü ve sorunun genel yapılandırılması ile ilgili olarak bazı insan müdahalesi olduğu varsayılır [5].

Verilerde sınıfların veya kümelerin varlığını belirlemek amacıyla bize bir dizi gözlem verilebilir, ya da ne kadar sınıf olduğunu kesin olarak biliyor olabiliriz. Amaç, yeni bir gözlemi mevcut sınıflardan birine sınıflandırabileceğimiz bir kural oluşturmaktır. İlk tür Denetimsiz Öğrenme (veya Kümeleme), ikincisi Denetimli Öğrenme olarak 1994'te Henery, Michie, Spiegelhalter ve Taylor tarafından verilmiştir [7]. Gelişimler devam ederken öğrenme algoritmalarının da arttığına Makine Öğrenme teknikleri başlığı altında değinilmişti. Şekil 2.3.'te makine öğrenme teknikleri 3 ana başlık altında, araştırmalarda genel olarak kullanılan algoritmaları ile gösterilmiştir.

1.Denetimli Öğrenme **Genel olarak bilinen Denetimli Öğrenme Algoritmaları:** Karar Ağaçları Rasgele Ağaçlar K-Enyakın Komşu Yapay Sinir Ağları Destek Vektör Makineleri Lineer Regresyon Ragele Orman Sınıflandırma ve Regresyon Ağaçları Gradyan Destekli Regresyon Ağaçları	Kategorik hedef değişken	Sınıflandırma
	Sürekli hedef değişken	Regresyon
2.Denetimsiz Öğrenme **Genel olarak bilinen Denetimsiz Öğrenme Algoritmaları:** k-means kümeleme ve sınıflandırma Birleşme/Bağlantı Kuralları	Hedef değişken mevcut değil	Kümeleme Bağlantı/ Birleşme
3.Yarı Denetimli Öğrenme **Genel olarak bilinen Yarı Denetimli Öğrenme Algoritmaları:** Lineer Regresyon Lojistik Regresyon	Kategorik hedef değişken	Kümeleme Sınıflandırma

Şekil 2.3. Denetimli, Denetimsiz ve Yarı Denetimli Öğrenme Teknikleri [10]

Denetimli Öğrenmenin iki ana dalı regresyon ve sınıflandırma arasındaki en dikkat çekici fark çıktı verilerinin yapısıdır. Analiz sonucu elde edilen çıktı verisi kategorik ise sınıflandırma, çıktı verisi sürekliyse regresyon yöntemi kullanılır.

2.2.1. Denetimli Öğrenme - Eğitim Test ve Geçerlilik

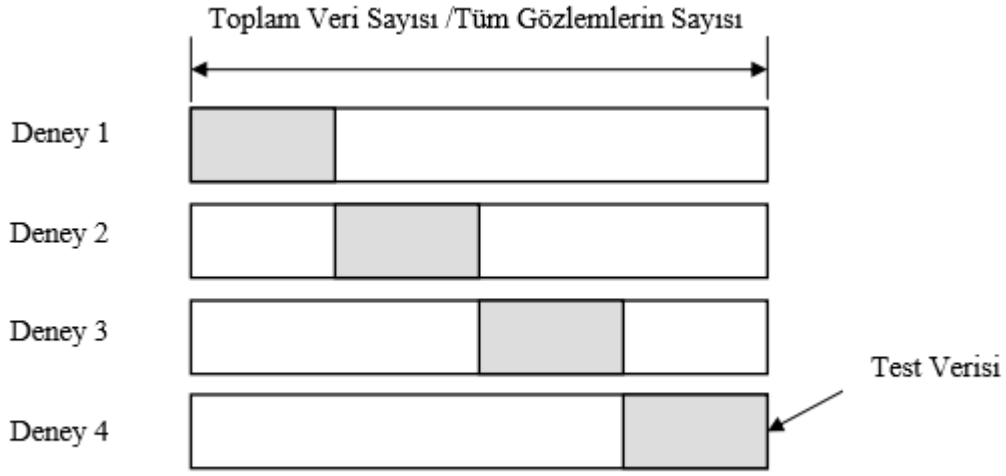
Sınıflandırma algoritması, eğitim verilerinin işlenmesi ve öğrenilmesi teknikleri ile bir algoritma elde edilir. Yeni gözlemlerin kategorisini belirlemek için kullanılan sınıflandırma algoritmaları bir "Denetimli Öğrenme" tekniğidir. Denetimli öğrenmede gerçek sınıfların bilindiği bir çalışmada, var olan sınıfların yapısı veriler yardımıyla analiz edilerek, bir yapı oluşturulur. Veri modellemede girdiler ve çıktılar vardır. Modelleri eğitebilmek için yapıtaşları olarak girdiler çıktılara ulaşmak için yol haritası oluşturur. Girdilerin ağ, ağaç, örüntü tarzında haritalandırılarak çıktılara ulaşma yolculuğu modellerin eğitilmesini sağlar. Eldeki verilere girdi, elde edilen sonuçlara çıktı dersek, girdi ve çıktılar arasında çok karmaşık ilişkiler kurularak tahmin veya sınıflamalar yapılır. Var olan sınıflar analiz edilerek yeni girilen verinin, hangi sınıfa ait olacağının belirlenmesinde eğitim ve test verileri kullanılır. Eğitim ve test verileri ile model eğitilir, test edilir ve modelin geçerliliği (validity) sınanabilir. Verilerin yer aldığı sınıflar bilindiğinde, analiz; sınıflarda yer alan verilerin ışığında sınıfların özellikleri üzerinden yapılır. Girdi olarak belli sınıflara ait veri bilgilerinden öğrenerek, yeni verilerin hangi sınıflara ait olabileceği tespit edilir.

a) Eğitim: Veriler arasındaki gizli örüntülerin analizi ile oluşturulan modeller ve bu modellerin test edilmesi amaçlanır. Veriden öğrenen modelin geçerliliğini inceleyebilmek için veriyi sadece eğitim ve test olarak ayırmakla kalmayıp, gerekli görülen durumlarda aynı veri seti üzerinden çapraz geçerleme yöntemini de kullanarak daha sağlam bir model elde edilmeye çalışılır. Denetimli öğrenmede eğitim verileri insan faktörü ile oluşturulur. Pyton, R gibi programlama dilleri ile verilerimizi eğitim ve test verileri olarak ayırabilmek mümkündür.

b) Geçerlilik (Validity): Eğitim setiyle model eğitilir, modelin görmediği veri seti ile test işlemine geçilir. Eğitim verilerinden bir alt küme oluşturularak testte karşılaşılması muhtemel hatayı ölçme işlemine geçerlilik tespiti denir. Veri eğitilir, test edilir, artık doğruluğunu sınamak amacıyla geçerlilik kontrolü yapılır. Oluşturulan modelin doğruluğunu sınamak amacıyla kullanılır. Girdi ve çıktılar arasında çok karmaşık ilişkiler

kurularak tahmin veya sınıflamalar yapılır. Bu ilişkileri eğitim verileri ile öğrenen modellerde aşırı öğrenme/ezber (overfitting) gibi problemler ortaya çıkabilir. Modelin performansı hakkında tahmin oluşturur. Eğitim setlerinden, belirlenen büyüklükte farklı ölçüm kümeleri alınarak, eğitim setinden hiç eğitime katılmamış veriler üzerinden geçerlilik ölçülür. Burada hiç karşılaşılmamış verilerin eğitime alınmaması gibi bir risk oluşmasına karşı her seferinde farklı kümeler geçerliliğin ölçülmesi için seçiliyor ki bu yöntem literatürde çapraz doğrulama (cross validation) olarak adlandırılır.

Modelin Çapraz Doğrulaması, K katlı çapraz geçerleme yöntemini kullanmak eğitim ve test verilerindeki dengesiz bir temsili ortadan kaldırmak için ilkel bir koruma olarak görülmektedir [14]. Yapılan çalışmada eğitim ve test olarak ayrılan kısımların bir kerelik belirlenmesi yerine, k. eğitim test seçimleri yapılırken her seferinde farklı bir test seti seçilerek gerçekleştirilen dengeli bir eğitim ile daha iyi bir model elde edilmesi amaçlanır.



Şekil 2.4. Çapraz Doğrulama [15]

Elimizdeki verinin %80 eğitim ve %20 test olarak ayrılması ile 1000 veride 200 test için ayrılmış olarak düşünülebilir. Eğitim setine yeterince güvenilmediği zamanlarda, sınıflandırma sonuçlarını değerlendirmek için bazı ölçüm değerlerine başvurulur. Aşağıda verilen ölçüm araçları, literatürde çalışmalarda rastlanan ve sıklıkla kullanılanlardır:

- Karışıklık Matrisi,
- ROC Eğrisi,

- Cohen'in κ puanı,
- Gini katsayısı.

Ayrıca çapraz doğrulama, farklı modelleme özelliklerinin performansını karşılaştırmak için kullanılabilir (yani, etkileşimli ve etkileşimsiz modeller, polinom terimlerinin hariç tutulması, sınırlı kübik eğrilere sahip düğüm sayısı, vb.) [16].

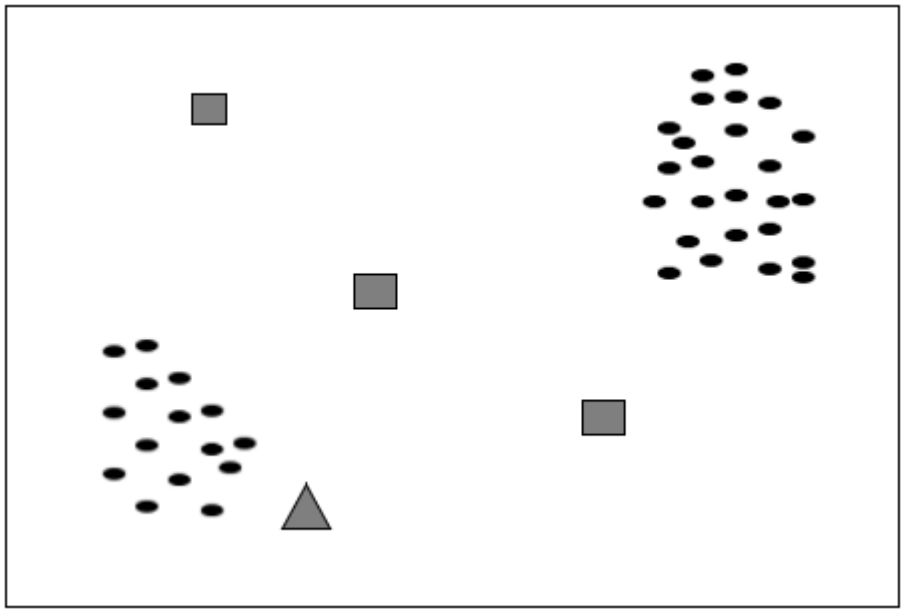
Genel olarak denetimli makine öğrenmesi algoritmaları Sınıflandırma ve Regresyon olarak iki türe ayrılır [17] [29]. Bu çalışmada denetimli öğrenmenin sınıflandırma tekniği ile araştırmalar yapılmaktadır.

2.3. Denetimli Öğrenmede Aykırı Değer

Aykırı değer, veri içerisinde diğer gözlemlerden farklı olması, başka bir kaynaktan alınmış olabileceği veya başka bir kaynak tarafından oluşturulduğuna dair şüphe uyandırması nedeniyle dağılımdan sapan bir gözlem olarak tanımlanabilir.

Şüphe uyandıran aykırı değerlerin meydana gelişindeki yanlış etiketlenme ile yanlış sınıflandırma yapılmasının verideki oranını araştırmak için veri madenciliğinde farklı sınıflandırma yöntemleri kullanılabilir.

Bazı araştırmalarda "aykırı değer puanı" olarak nitelendirilen bir değer hesaplanarak, verinin yanlış etiketlenme riski araştırılır. Sınıflandırma teknikleri ile aykırı değer tespiti genellikler genel (global) ve yerel (local) aykırı değer üzerinde çalışır. Genel aykırı değerlerin belirlenmesi sorunu, kabaca küresel ve yerel aykırı değer modelleri olarak sınıflandırılabilecek çok farklı yaklaşımlarla ele alınabilir [15]. Genel aykırı değer, çalışılan verilerin hepsi dikkate alındığında, normal aralığın içinde kalmayan verilere denir. Yerel aykırı değer ise, tüm veri kümesi içerisinde normal aralıkta yer alırlar ancak, çevreleyen veriler bazında incelendiğinde normal aralıkta yer almayan veriler olarak tanımlanırlar. Şekil 2.5.'te iki çeşit aykırı değerin veri kümesi içerisindeki gösterimi verilmiştir.



Şekil 2.5. Genel ve Yerel Aykırı Değerler

Dörtgen şekilli 3 adet nokta tüm veri içerisindeki genel aykırı değerleri gösterirken, sol alttaki üçgen nokta, çevre veriler dikkate alınarak yerel aykırı değeri göstermektedir.

Veri analizi için oluşturulan algoritma 2 aykırı değerden birini bulmada daha iyi sonuçlar verecek şekilde ayarlanırsa, diğer aykırı değer çeşidini tespit etmede yetersiz kalacaktır. Çünkü verilerin ya geneline ya da ayrıntıda kalan kümelerine bakacaktır. Bu nedenle, etkili bir algoritma aykırı değerlerin iki türünün de tespitinde başarılı olmalıdır.

Veri madenciliğinde yapılan analizlerde eğitim verilerinde ezberleme hemen hemen her çalışmada karşılaşılabilecek bir sorun haline gelmiştir. Konu üzerine yapılan son araştırmalar, önce temiz etiketlerin eğitim verilerini ve ardından gürültülü etiketlerin eğitim verilerini ezberleyeceklerini gösteriyor. Son yıllarda bu problemi çözmek için bilim insanları farklı yöntemler deneyip geliştirmiştir. Literatürde “robustness” olarak karşımıza çıkan sağlamlık, genel anlamda “sistemin işlevsel gövdesini etkileyebilecek bozulmaları tolere etme yeteneği” olarak ifade edilir. Mühendislikte, hata direncini belirtmek için kullanılırlar, örneğin sağlam yöntemler, küçük hatalar varlığından fazla etkilenmeden yöntem oluşturmaya çalışırlar [1]. İstatistikte ise, model varsayımları tam olarak doğru olmadığında kullanılan yöntemlerin bütününü ifade eder [1].

Aykırı değerler içeren araştırmalarda birçok yöntem kullanılır. Temiz veri olarak adlandırdığımız aykırı değer içermeyen veriler gerçek zamanlı çalışmalarda rastlanmayan bir durum haline gelmiştir. Acuna ve Rodriguez'in [18] araştırmalarına göre, aykırı değerler varlığında sınıflandırma yapılırken, bazı yöntemlerin aykırı değerlerden daha az etkilendiği görülmüştür. Bu çalışmada eğitimi yapacak olan veri kümesinde aykırı değer bulunması durumunda bu aykırı değerlerden etkilenmeden sınıflandırma yöntemi önerebilecek algoritmalar incelenecektir.

2.4. Eğitim Kümesinde Aykırı Değer Olması Durumunda Kullanılabilecek Algoritmalar

2.4.1. Lojistik Regresyon

İki veya daha fazla kategoriden oluşan bağımlı değişkenin bağımsız değişkenler ile ilişkisinin ölçümünü sağlayan bir tekniktir. Lojistik regresyonda (logistic regression) bağımsız değişkenler kategorik veya sürekli olabilirler. Bağımlı değişken kategori sayısı iki ise ikili lojistik regresyon, ikiden fazla kategori içeriyorsa çoklu lojistik regresyon kullanılır. Yarı denetimli bir öğrenme tekniği olarak kategorilendirilen bir algoritmadır.

Regresyonda olduğu gibi bağımsız değişkenlerin bağımlı değişkenler üzerindeki etkisini araştıran bu yöntem regresyonu kullanarak sınıflama yapmakta başarılıdır. Sınıflandırma ve regresyon arasındaki temel fark ise sınıflandırma çıktı değişkenlerinin kesikli olmasıdır. Regresyonda sürekli değişken yapıları mevcuttur.

Aykırı değerler varlığında yapılan araştırmalardan birinde, lojistik regresyonun aykırı değerler varlığında etkili sınıflandırma yapmasının pek mümkün olmadığı sonucu üzerinde durulmuştur. Sasaki, Takenouchi, Monti ve Hyvarinen araştırmalarında lojistik regresyonun aykırı değerlere karşı savunmasız olduğunu ve bu nedenle performansın aykırı değerler tarafından güçlü bir şekilde zayıflatılabileceğini vurgulamışlardır [20].

Lojistik Regresyon (LR) çalışma mantığında bir örneklemde pozitif ve negatif olma durumlarının olasılıkları ile ilgilenme durumu mevcuttur. Bu özelliği ile sınıflandırmada yaygın olarak kullanılmaktadır. Daha başarılı sınıflandırmalar elde edilebilmesi için Sağlam Lojistik Regresyon (RLR) modelleri geliştirilmiştir. Aykırı değerler varlığında 2014 yılında önerilen RoLR metodu ve 2020 yılında mevcut RLR lerden daha başarılı

olduğu öne sürülen ağırlıklı maksimum olabilirlik tahmin edicili Mallows tipi tahmin edicileri (WMLE) bu sınıfta yer alır. RoLR, çalışma mantığında önce aykırı değerleri örneklemden çıkarır, sonra tahmin edilmiş LR ile kalan örneklerin düzenlenmiş korelasyon değerlerinin maksimumunun elde edilmesini sağlar [21]. Sağlam lojistik tahmin edicilerden Mallows ve Maksimum Tahmin Edici (Maximum Likelihood Estimator-MLE) ile geliştirilmiş WMLE önerilen bir diğer sağlam lojistik regresyon olarak gösterilmektedir [22].

2.4.2. Karar Ağaçları

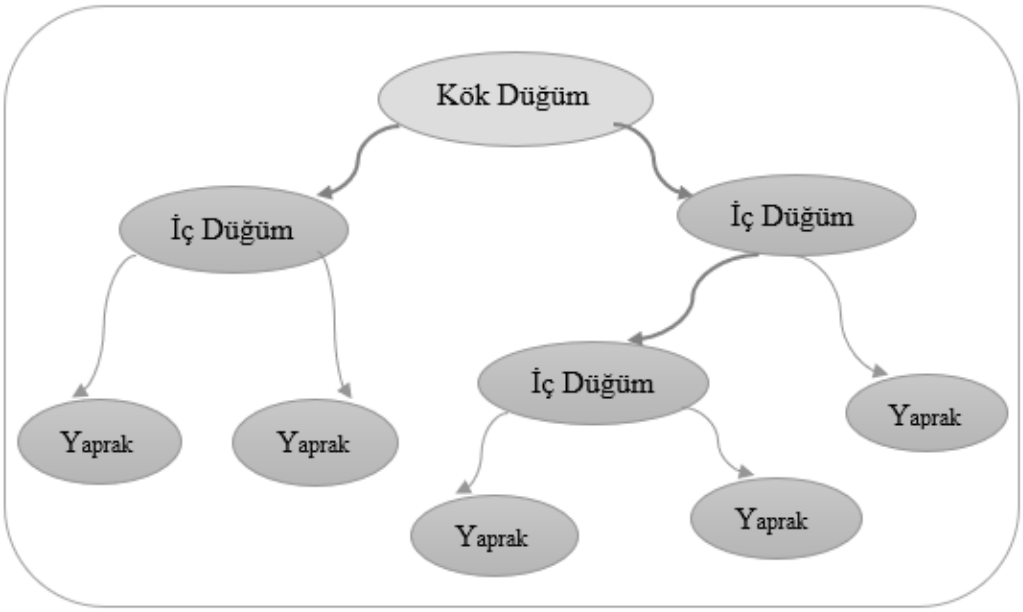
Hem kategorik hem de sayısal verilerden oluşabilir. Koşullu denetim ifadeleri içeren bu algoritmalar dallanarak ilerlediği için ağaca benzetilmiştir. Eksik veya hatalı veriler varlığında da başarılı sonuçlar verebilen bir algoritmadır.

Karar Ağaçları (Decision Trees), kurgulanmasının, yorumlanmasının ve veri tabanları ile entegrasyonun kolaylığı nedeniyle en yaygın kullanılan sınıflandırma tekniklerinden biri olarak bilinir. Ayrıca gürültü toleransı, düşük hesaplama ihtiyaçları, uygulama kolaylığı ve sağladığı görsellik sayesinde sınıflama teknikleri içerisinde tercih görmektedir [23]. İstatistiksel algoritmalar arasında en sık kullanılan tahmin ve sınıflandırma mekanizmalardan biridir.

Araştırmalarda karar ağaçları güçlü bir sınıflandırma tekniği olarak bilinmektedir. Bir sınıflandırma yöntemleri ailesi olan karar ağaçları, alanı özyinelemeli olarak bölmeyi ve noktaların düştüğü bölgeye göre tahminler yapmayı amaçlar [24]. Karar ağaçlarının çalışma düzeneğini Barry de Ville [24] şu şekilde açıklıyor: Ağaç, ilk olarak kök düğümün, bir düğüm içinde birbirine benzeyen ancak ağacın herhangi bir seviyesindeki diğer düğümlerle karşılaştırıldığında farklı olan gözlem kümelerini oluşturan alt yaprakları (veya düğümleri) tanımlayan dalları oluşturmak üzere bölümlenmesiyle oluşturulur.

Karar ağacı, her düğümde bazı amaç fonksiyonlarını en aza indirerek bir bölme kuralının öğrenildiği yukarıdan aşağıya akışı sağlayacak bir şekilde öğrenir [25]. Karar ağaçları oluşturmak için kullanılan algoritmalar içerisinde C4.5, hata oranı ve hız kombinasyonuna bakıldığında literatürde en iyi bilinen algoritma olarak tanımlanmıştır [26]. Özellikle öznitelikleri farklı olan veri setleriyle çalışılırken ağaç tabanlı modeller

daha başarılı sonuçlar verebilmektedir. Şekil 2.6.'da, karar ağaçlarının temel yapısının net ve en basit hali şematik olarak gösterilmiştir.



Şekil 2.6. Karar Ağacı Temel Yapısının Basit Bir Gösterimi

Karar ağaçlarının hedefi, bağımlı değişkendeki farklılıkları en ayırt edici biçimde saptayarak, sıralı bir biçimde dallarına (farklı gruplara) ayırmaktır. Karar ağacındaki her düğüm, sınıflandırılacak bir örnekteki bir özelliği temsil eder ve her dal, düğümün üstlenebileceği bir değeri temsil eder [27]. Eğitim veri kümelerinden karar ağaçları oluşturmak için çeşitli algoritmalar mevcuttur. Geliştirilen algoritmalardan bazıları; ID3 (Iterative Dichotomizer 3), CART (Sınıflandırma ve Regresyon Ağacı) ve C4.5 [28], CHAID, C5.0 ve QUEST' tir. Sınıflandırma ve Regresyon Ağaçları (CART) kategorik değişkenlerle de çalışırken, QUEST algoritması kategorik verilerde uygulanmamaktadır. Yaygın bir görüşe göre karar ağaçları oluşturmak için kullanılan algoritmalar içerisinde C4.5, hata oranı ve hız kombinasyonuna bakıldığında literatürde en iyi bilinen algoritma olarak tanımlanmıştır [26].

2.4.3. Rasgele Ağaçlar

Makine Öğrenme başlığı altında tanımlandığında, rasgele ağaçlar temelde tek model ağaçlar ile rasgele orman fikirlerinin yani çoklu öğrenme modellerinin birleşimi olarak değerlendirilmektedir. Model ağaçlar, her bir yaprağın bu yaprak tarafından açıklanan veri kümesi için optimize edilmiş doğrusal bir modeli içerdiği karar ağaçlarıdır. Gini

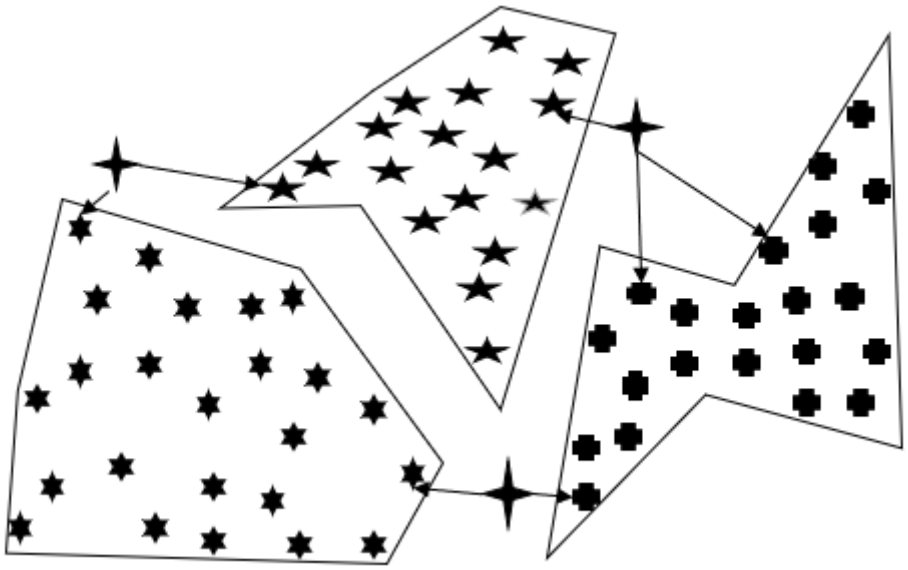
indeksini kullanarak çalışan bir algoritmadır. Bir dizi zayıf sınıflandırıcı olarak görülen karar ağaçlarının bir araya gelmesiyle meydana gelen öğrenme ise rasgele orman sınıflandırıcısı olarak adlandırılmıştır [18].

Hatalı verilerin varlığında karar ağacı öğrenmesinin sağlamlığı üzerine yapılan bir çalışmada karar ağaçları ve rasgele orman algoritmaları kullanılmış, simetrik hatalı veri ile sınıflandırma yapıldığında yanlış sınıflandırma olasılığı, hatasız veri ile yapılan sınıflandırmadaki yanlış sınıflandırma olasılığı ile kıyaslanmıştır. Elde edilen sonuca göre, yanlış sınıflandırma olasılıklarının aynı olması karar ağaçları ve rasgele orman algoritmalarının öğrenmedeki sağlamlığına bir örnek olmuştur [25]. Karar ağaçlarının sağlamlıkları araştırılmıştır. Bu sonuçlar simetrik hatalı veri üzerinden yapılmış ve farklı veri setlerinde de denenebileceği görülmüştür.

2.4.4. K- En Yakın Komşu Algoritması

K-En yakın komşu, veri madenciliğinde sıkça kullanılan sınıflama tekniklerinden biridir. Yapısı basit işleyişi daha kolaydır. Yeni veri öznitelik değerlerine/özelliklerine göre eğitim veri setinde en yakın komşu sınıfa atanır. Bu yöntemin işleyişinde, sınıf bilgisi uyarınca n boyutlu uzay içinde örneklem tasvir edilmektedir [23]. Sürekli ve kategorik değişken tahminlerinde kullanılır, parametrik olmayan yöntemler arasındadır. (İngilizcesi K-Nearest Neighbor olan bu algoritma literatürde K-NN olarak da bilinmektedir.)

K-En yakın komşu algoritması, büyük eğitim setlerinin varlığında, oldukça etkin sonuçlar verebilmektedir. Bu algoritma, en iyi bilinen, eski, basit ve etkili örüntü sınıflandırma yöntemlerinden biridir ve makine öğrenme algoritmaları arasında popüler olarak kullanılmaktadır. K en yakın komşu sınıflandırma yöntemi, benzer sınıflar belirli özellik uzayları etrafında kümelendiğinde iyi çalışır [29]. Değişkenler üzerinde çalışırken, örnekler arası uzaklıkları ile ilgilenir. Algoritmanın temelinde komşu sayısı ve uzaklık parametreleri olduğu için bu işlemler çalışılan ekipmanda yüksek bellek ihtiyacı oluşturur. Bu nedenle çok boyutlu veri setleri üzerinde çalışılırken tercih edilmez. K-NN algoritmasını diğer sınıflandırma yöntemlerinden ayıran en çarpıcı özelliği için, algoritmanın çalışma mantığının, veriyi test ve öğrenme (training) olarak ayırmadan inceleyip analiz ediyor olmasıdır. Şekil 2.7.'de analize giren yeni verilerin en yakın sınıfına atanması ile ilgili görsel verilmiştir.



Şekil 2.7. K En Yakın Komşu Algoritması ile Sınıflandırma, k=3

Yeni gelen verileri mevcut sınıflara atayabilmek ve en yakın sınıfın hesaplanabilmesi için farklı uzaklık ölçümleri hesaplanabilir. Bu ölçüm sonuçlarına göre en kısa mesafeyle atanacak yeni verinin en yakın olduğu verinin bağlı olduğu sınıf seçilir. Mesafenin ölçülmesinde kullanılan birçok ölçüm ailesi vardır ve en sık rastlananları aşağıdaki gibidir:

1) L_p Minkowski Ailesi (p-norm uzaklığı), vektörler arası uzaklık hesaplaması daha basit ve genel yöntemlerden faydalanılarak bulunmaktadır.

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

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

2) İç Çarpım Ailesi (Inner Product Family), mevcut vektörlerden bazı vektörlerin çarpılması ile hesaplanır. Bu alanda 2 hesaplama vardır:

a) Kosinüs Benzerlik Ölçüsü:

$$\mathrm{S(x,y)} = \frac{x.y}{||x||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

b) Jaccard Mesafesi/Uzaklığı:

$$d(x,y) = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{(x_i)^2 + (y_i)^2 - [(x_i)(y_i)]}$$

3) L_1 Uzaklık ailesinden olan mutlak farka göre hesaplanan Canberra Mesafesi/Uzaklığı:

$$d(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

'N' sayıda eğitim veri boyutu, 'P' sayıda özellik alanı boyutu ve 'R' özellik alanında doğrusal olan bir mesafe metriği varlığında, Neeb ve Kurrus, tek bir veri noktasının en yakın k komşusunu belirlemek için R(NP) hesaplamaları yapılmasını öneriyor [29]. Ancak Cover ve Hart'ın da dikkat çektiği üzere sonsuz sayıda sınıflandırılmış örnek koleksiyonundaki mevcut bilgilerin yarısı en yakın komşuda bulunur [30], ki bu durum, Neeb ve Kurrus'un önerdiği hesaplamanın, tahmin edilmesi istenen veriler için yinelenmesi hesap yoğunluğuna neden olacaktır. Bu tarz bir işlem yoğunluğu büyük verilerle çalışılırken büyük bir dezavantaj oluşturmaktadır.

2.4.5. Yapay Sinir Ağları

Birçok bilim alanında yapılan icatlarda doğanın izlerini bulmak mümkündür. Japonya'nın hızlı trenlerini tasarlayan Eiji Nakatsu adlı mühendis, aynı zamanda bir kuş bilimcidir. Balıkçıl kuşların daha iyi hızlanabilmesini sağlayan gaga yapılarını, tasarladığı hızlı trenlere yansıtmıştır. Benzer olarak sadece gemiler değil, 1960'larda yapılan ses hızını aşan Concorde uçakların burun yapısının da yunusların burun şeklinden esinlendiği iddia edilir. Çağımızın gelişen robot teknolojisi, sineklerden ve böceklerden esinlenerek böcekler kadar ufak, her yere erişimi olan, uçabilen, tavanlarda gezebilen robotlar geliştirilmektedir. Günümüzün en ilgi çekici çalışma alanlarından biri olan yapay zeka da yine bu şekilde doğa ve yaşamdan örneklerle doludur. Yapay sinir ağları (ANN) da yine bu şekilde ortaya çıkmıştır. İnsan beyninin çalışma şeklinden esinlenen ve geliştirilen bu ağlar, hayatımıza hızla giriş yaptı. Literatür araştırmaları yapıldığında 1940 yıllarında aktif bir çalışma grubunun sinir ağları üzerinde matematiksel olarak çalıştığı bilgisine ulaşılmaktadır. Ancak sinir ağları üzerinde matematiksel ve mantıksal çalışmaların tarihi 1943 yılında S. McCulloch ve Walter H. Pitt'in "A Logical Calculus of the Ideas in

Manent in Nervous Activity" adlı makalesi ile başlar. 1950 de artık bir bilim dalı olarak kabul edilmiştir.

Yapay sinir ağı, geleneksel yapay zeka ve bilgi işleme teknolojilerinden tamamen farklı bir mekanizmayı benimseyen, geleneksel mantık tabanlı yapay zekanın sezgi ve yapılandırılmamış bilgiyi işlemedeki kusurlarını ortadan kaldıran, uyarlanabilir, organize etme ve gerçek zamanlı öğrenme özelliklerine sahiptir [31]. Bilgilerden öğrenerek yeni bilgiler oluşturabilme yeteneğine sahiptir.

Yapay sinir ağları, ismini insan beynindeki benzer nöron sisteminden almıştır. Algoritmanın işleyiş şekli, tıpkı insan beyni gibi topladığı bilgileri bir nörondan diğerine aktarması gibidir. Bilgileri aktarırken oluşan katmanlar ise paralel işlem yapan nöronlar gibi değerlendirilmektedir. Nöronların bilgi alışverişleri aktivasyon fonksiyonu değerleri ile sağlanır. Öğretilen verilerden elde edilen bu bilgilerle genelleme yaparak yeni sayısal veri seti tahminlerini verir. Yapay sinir ağı modeli girdi katmanı, gizli katman ve çıktı katmanı olmak üzere üç tabakadan oluşturulabilir. Her katmanda düğümler(nöronlar) vardır. Bir düğüm birden fazla düğüm tarafından etkilendiği için, etkileyici düğümlerdeki verinin yetersiz veya bozuk olması, sistemin performansını etkilemektedir.

Yapay Sinir Ağları, tahmin/öngörü çalışmalarında, Sınıflandırma ve Kümeleme, ayrıca uyarı niteliği taşıyan Kontrol çalışmalarında ve diğer birçok analiz yan işlemlerinde (veri filtreleme, veri birleştirme vb.) kullanılır. Kompleks modeller kurarak büyük veri setlerindeki bilgilere ulaşılabilir. Özellikle verilerin homojen niteliklere sahip olmasıyla analiz daha etkili sonuçlar verebilmektedir.

Sınıflandırma çalışmalarında veriden öğrenme gerçekleşirken oluşabilen ezber (aşırı uyum) durumundan kurtulabilmek için ise bagging, boosting ve rasgele orman gibi popüler ve başarılı topluluk öğrenme yöntemleri kullanılmaktadır.

2.4.6. Derin Sinir Ağları

Yapay zekanın alt segmenti olan makine öğrenmesinde, büyük verilerle çalışıldığında derin öğrenme söz konusu olmaktadır. Büyük veriler artık sadece yapısal veriler üzerinde çalışmaz, yani satır sütunlarda analiz edilebilen verilerle değil; söz konusu resim, ses veya video gibi yapısal olmayan veriler de analiz edilebilir.

Yapay Sinir Ağlarını “derin” yapan, sinir ağındaki katman sayısının artmasıdır. Katman sayısı arttıkça derinlik artacaktır. Özellikle artan katmana bağlı olarak, katmanın kendi içinde bağlantı olmaması da durumu daha karmaşık hale getiren bir etkendir. İşte bu ve bunun gibi katmanların kompleks yapıları derin sinir ağlarına olan ihtiyacı arttırmıştır.

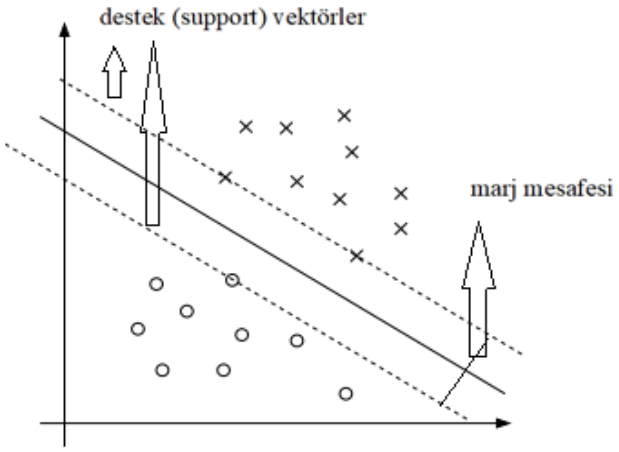
Klasik yapay sinir ağlarının yetersiz kaldığı daha karmaşık işlemlerde derin sinir ağları (DNN/DSA) tercih edilir. 2006 yılına kadar, birkaç özel problem dışında, daha geleneksel yaklaşımları aşmak için sinir ağlarını nasıl eğitilebileceği tam olarak bilinmezken, 2006'da, derin sinir ağları olarak adlandırılan öğrenme tekniklerinin keşfiyle bu tarz çalışmalara yol açılmıştır [32]. Verilerin tanımlanmasında özniteliklerinin ayarlanması derin öğrenmede ihtiyaç duyulmayan bir durumdur. Bu da derin öğrenme algoritmalarının çalışma mantığının açıklanamamasına neden olur.

Derin sinir ağının, eğitim açısından kolay olduğu söylenemez. Özellikle denetimsiz öğrenme altında kullanıldığında karmaşık işlerin varlığında etkilidir. Ayrıca derin ağlar, öğrenme kapasitelerinin yüksek olması nedeniyle, verilerin tamamen rasgeleliğinde veya aykırı değerlerin varlığında da tüm verileri ezberleyebilirler. DNN alanında ilk önemli katkı, birçok gizli katmana sahip ancak katmanın kendi içinde gözlemlerle bağlantı kurulamayan bir sinir ağı olarak nitelendirilen Derin İnanç Ağı(DBN) için yapıldığı belirtilmektedir [33]. Derin öğrenme nitelikleri gereği büyük veriler ile çalışırken hem çok hem kompleks yapıdaki verilerle çalışmak eğitim süresinin uzamasına neden olur. Bu nedenle Derin Öğrenme Algoritmalarının çalışabilmesi için gerekli donanımı sağlayacak daha üstün makinelere ihtiyaç duyulmaktadır.

2.4.7. Destek Vektör Makineleri

Destek vektör makinesi (DVM/SVM), diğer sınıflandırma algoritmalarına benzer olarak nesnelere etiket atamayı örneklerin ışığında öğrenen bir başka bilgisayar algoritmasıdır. DVM'ler bir dışbükey optimizasyon problemi olarak tanımlanabilir ve denetimli öğrenme algoritmalarının geliştirilmesi için doğrusal programlama yöntemini kullanabilir [1]. Biyomedikal araştırmalarda da yaygın olarak kullanılan bir model olmuştur. Teorik olarak, bir DVM, bir tümör numunesinden türetilen gen ekspresyon profilini inceleyebilir ve bir teşhise ulaşabilir [34]. Binlerce sahte ve sahte olmayan hesap bilgilerini inceleyerek sahte hesapları tanımayı öğrenebilir.

DVM verileri sınıflara ayırmada hiper düzlemleri kullanır. Bunlar 2 boyutlu çizgiler veya 3 boyutlu düzlemler gibi düşünülebilir. Üç boyutlu levha gibi bir düzlemden bahsediliyorsa veya doğrusal olarak 2 boyutlu bir düzlemle sınıflandırılamayan veriler söz konusu olduğunda orada DVM'nin çekirdekleme yöntemine yani kernel trick işlemine başvurulur. En iyi düzlemi seçerken, sınıflara ait gözlemlerin içerisinde düzleme en yakın olan noktaların, düzlemle olan uzaklığı dikkate alınır. Bu noktalarla düzlemin arasındaki maksimum uzaklığa marj mesafesi denir ve bu mesafenin optimumunu veren düzlem DVM tarafından seçilir. Bu düzleme en yakın noktalar yani düzlemin belirleyicileri diyebileceğimiz gözlemler, destek (support) vektörler olarak adlandırılır.



Şekil 2.8. Destek Vektörler Grafiksel Gösterimi

Şekil 2.8.' de DVM' lere ilişkin grafiksel gösterim verilmiştir. Destek vektör makinesi, verileri doğrusal veya doğrusal olmayan bir ayırma hiperdüzlemiyle ayırarak sınıflandırma yapan denetimli bir öğrenme yöntemidir [35].

Verimiz üzerinde sınıf dengesizliklerine neden olan aykırı değerlerin, eğitim setleri üzerindeki negatif etkisini çözümlemek için en başarılı sınıflama tekniği araştırılmaktadır. Denenen her modelin farklı bir metodolojisi vardır. Yapay sinir ağlarına benzer olarak DVMler de homojen öznitelikli verilerle daha etkili ve başarılı sonuçlar vermektedir.

2.4.8. Sağlam (Robust) Kayıp Fonksiyonu

Matematiksel kökenli uzaklık ölçümünü ifade eden fonksiyon yapısındadır. Hedef değişkenle tahmin edilen değer arasındaki uzaklıktan yola çıkarak, hedef değişkene ne kadar yaklaştığı sonucu ile ilgilenir. En sık kullanılan çeşidi hata kareler ortalaması (mean squared error) olarak bilinir. Hata kareler ortalaması (MSE) ve ortalama mutlak hata (MAE) gibi geleneksel öklid mesafe ölçümlerine dayalı kayıp fonksiyonlar, sıralı veri karakteristiğini veya tahmin edilen dizilerin sürekliliğini hesaba katmaz. Dizileri bir dağılımdan örneklenen bir dizi değer olarak ele alarak, MSE ve MAE sırasıyla bu dağılımın ortalamasını ve medyanını belirlemeyi amaçlar [36].

Denetimli makine öğreniminde eğitim verileri ile çalışılırken amaç kayıp fonksiyonunun (loss function) minimize edilmesidir. Makine öğrenimi yöntemlerinin eğitimi için kayıp işlevi olarak belirli bir hata metriğinin seçimi çok önemlidir [36]. Kayıp fonksiyonu birçok önemli parametreyi etkiler, gradyanı, dolayısıyla optimizasyonun yakınsama davranışını ve potansiyel olarak yanlı tahminleri şekillendirir [36].

Sınıflandırma çalışmalarında verilerle beraber aykırı değerlerdeki artışlar kayıp fonksiyonların sağlamlaştırılması ihtiyacını ortaya çıkarmıştır. Literatürde Boosting yöntemler olarak karşılaşılmaya başlanmıştır. Bu yöntemler, birçok zayıf sınıflandırıcının birlikte vereceği çıktı sonucundaki kararın, bir sınıflandırıcının kararından daha mantıklı olacağından hareketle önerilmiştir [37]. Toka [37] sağlam kayıp fonksiyon özelliklerine sahip olan TanjantBoost algoritması üzerinde çalışmış ve istatistiksel tutarlılık için algoritmada bir düzeltme yaparak, sınıflandırıcıya yakın gözlemler üzerindeki etkiyi başarılı hale getirmiştir. Ceza miktarının önemli olduğu kayıp fonksiyonlar için daha büyük ceza veren kayıp fonksiyonlarına bir alternatif de Gudermannian kayıp fonksiyonudur. Yanlış sınıflandırma ve doğru sınıflandırmada farklı cezalandırma değerleri için trigonometrik fonksiyonların kullanılması önerilebilir [37].

Toka [37], ceza katsayısının büyütülme başarısını, Gudermannian kayıp fonksiyonu ile boosting algoritması olan GudermannianBoost ikili sınıflandırma yöntemi önerisiyle gerçek veri kümeleri üzerinde göstermiştir. Cezalar sınırlandırılmaktadır ve doğru sınıflandırma, yanlış sınıflandırmadan daha az şekilde ceza almaktadır. Gudermannian kayıp fonksiyonunun çalışma mantığında, kayıp fonksiyonun sınır değerlerine fazla ceza

vermesi vardır, bu şekilde sınıflandırıcıya yakın yanlış sınıflandırılmış gözlemlerin sınıflandırıcının kararlılığını etkilemesinin önüne geçilmiştir [37].

2.4.9. Doğrusal Sınıflandırma Analizi

Denetimli öğrenme tekniklerinden biri olan Doğrusal Ayrıştırıcı Analizi (Linear Discriminant Analysis-LDA), istatistik bilimi ışığında ayrıştırıcı analizi kullanılarak sınıflandırma işlemini yapan temel algoritmalardandır. Ayrıştırıcı analizi istatistikte grupların ortalamalarına ilişkin bağımsız değişkenlerin doğrusal fonksiyonlarını bulmak için kullanılır. Bu şekilde gruplar birbirlerinden ayrılması sağlanır. Bulunan foksiyonlar ile sınıflardaki farklılıklar göz önüne alınarak değişkenlerin doğrusal kombinasyonları belirlenir ve bu doğrusal kombinasyonlar yeni gözlemlerin sınıflara atanabilmesi için kullanılır. Fonksiyonun temeli grup ortalamaları arasındaki farka dayanır. Farkın en büyük olması önemlidir.

Grup sayısına göre LDA, çoklu grup ve ikili grup olarak adlandırılır. LDA'nın çalışma prensibinde bazı varsayımların sağlanması gerekir. Öncelikle verilerin normal dağıldığı varsayımı sağlanmalıdır. Ayrıca gruplar için varyans-kovaryans matrislerinin homojen olduğu varsayımı altında LDA kullanılabilir. Varyans-kovaryans homojenliği sağlanmıyorsa lojistik regresyona geçilebilir. Aynı şekilde normallik ve varyans-kovaryans matrisinin homojenliği varsayımı sağlamadığında makine öğrenmesinde denetimli sınıflandırma yöntemlerinde LDA' nın alternatifi olarak kullanılmak üzere sağlam doğrusal ayrıştırıcı analiz (RLDA) yöntemleri geliştirilmiştir.

Aykırı değerler varlığında RLDA'nın sağlam bir sınıflandırıcı olarak araştırmalarda kullanıldığı görülmektedir. Simülasyon çalışmaları ve gerçek hayat problemi sonuçları, önerilen RLDA modellerinin LDA' da karşılaştırılabilir bir performans veya daha iyi performans sağladığını savunulmuştur [38]. Özellikle kullanılan RLDA'ların gerçek veri sonuçlarında, aykırı değerleri tespit edebildiği ve klasik LDA'dan daha küçük hata oranları ile sonuçlara ulaştığı belirtilmiştir [38].

Todorov ve Pires [39] 2007 yılındaki çalışmalarında ele aldıkları grup ortalamaları ve ortak kovaryans matrisi için RLDA sınıflandırma tekniklerini karşılaştırmışlardır. İkili tahmin edici yöntemi olarak değerlendirmeye aldıkları Orthogonalized Gnanadesikan-Kettering (OGK)'i en iyi performans gösteren yöntemlerden biri olarak belirtmişlerdir

[39]. Benzer araştırmalarda OGK'ya dayalı olarak yeni yöntemler geliştirilmiş ve başarılı sonuçlar alınmıştır. Bu tezde de OGK'ya dayalı RLD algoritması RLD-OGK olarak analizlerde kullanılmıştır.

Bir diğer sınıflandırma tekniği olan Karesel (İkinci Dereceden) Diskriminant Analizi (Quadratic Discriminant Analysis - QDA) LDA gibi normal dağılımlı gözlemlerin sınıflandırılması için kullanılır. Ancak yine LDA'daki gibi normal dağılmayan veriler için de çalışabilmektedirler. Normallikten sınırlı sapmalar olması durumunda ve dağılımların şekli veri kümeleri arasında benzer olduğu durumlarda LDA'nın, artan örneklem boyutuna sahip bir dizi normal olmayan dağılım için QDA'nın daha iyi performans gösterdiği belirtilmiştir [40].

Gerçek dünyada, aykırı değerlere karşı oldukça savunmasız olan verilerle yüzleşmek oldukça yaygındır ve bu nedenle, ortalama vektörün ve dağılım matrisinin klasik tahmin edicilerinin sağlamlığının olmaması, genelleştirilmiş karesel diskriminant analizi sınıflandırıcısının etkinliğini önemli ölçüde azaltmakta ve yanlış sınıflandırma hatalarını artırmaktadır [40]. Normallik sağlanmadığında da sınıflandırma yöntemlerinde QDA kullanılabilmekte ancak sınıflandırmadaki performansını yükseltmek için alternatifi olarak kullanılmak üzere sağlam karesel ayrıştırıcı analizi yöntemleri tercih edilmektedir.

2.4.10. Temel Bileşenler Analizi

Makine öğrenmesinde sınıflandırma amaçlı kullanılan algoritmaların çalışma prensiplerine etki eden bir istatistiksel analizdir. Düşük boyutlu veriler için sınıflandırma kuralı oluşturabilir her gruba ayrı ayrı uygulanabilir özelliktedir [41]. Temel Bileşen Analizine (PCA) dayalı popüler bir istatistiksel sınıflandırma model önermesi Sınıf Analojisine Göre Yumuşak Bağımsız Modelleme kısaca SIMCA (Soft Independent Modelling Of Class Analogies) olarak bilinir. Gözlem seti yüksek boyutlu verilerde uygulanabilirdir. PCA'nın aksine SIMCA ile farklı gruplar hakkında farklı değişkenlerin ilişkisi hakkında ek bilgiler elde edilebilir. Tek sınıflı bir sınıflandırıcı olarak kullanılabilmesinin yanında aykırı değerlerin tespiti için verilerin açıklayıcı analizlerinde de kullanılan bir yöntemdir [41].

Büyük verilerle çalışılması ve aykırı değerlerin varlığında da sınıflandırmanın doğruluğunu arttırabilmek için yüksek boyutlu veriler için temel bileşenler analizine

dayanan yani literatürde karşılaşılan kısaltmasıyla ROBPCA tabanlı sağlam bir ikili sınıflandırma analizi olan RSIMCA yöntemi önerilmektedir [42]. ROB kısaltması konunun yüksek boyutlu veriler hakkında olduğu bilgisini içerir.

2.5. Eğitim Veri Kümesinde Aykırı Değer Bulunması Durumunda Önerilen Bazı Algoritmalar

Denetimli öğrenme algoritmalarının tahmin performansı, veri etiketlerinin kalitesiyle doğrudan ilgilidir. Tipik bir etiket toplama sürecinde, çoklu ek açıklamalar, değişen beceri düzeylerinin ve önyargılarının etkisi altında "gerçeğin" öznel gürültülü tahminlerini sağlar [43]. Aykırı değerleri doğru bildirilmiş veriler olarak ele almak, öğrenme algoritmalarının doğruluğunu sınırlar. Bu tarz sorunlar, hem ek açıklama maliyetinin hem de gözlemciler arası değişkenliğin yüksek olduğu tıbbi görüntüleme gibi alanlardaki uygulamalar için kritik öneme sahiptir [43]. Literatürde oldukça karşılaşılan tanım veriler içerisinde yer alan gürültülerdir. Gürültü, yanlış etiketlenmiş örnekler (sınıf/etiket gürültüsü) veya öznitelik değerlerindeki hatalar (nitelik gürültüsü) olarak tanımlanabilirken, aykırı değer, yalnızca hataları değil aynı zamanda popülasyon veya süreç içindeki doğal varyasyondan kaynaklanabilecek uyumsuz verileri de içeren daha geniş bir kavramdır [44]. Birçok büyük veri probleminde, etiketli numuneler genellikle kitle kaynak kullanımı yoluyla elde edilir ve bu tür etiketlerin güvenilmezliği, etiket gürültüsünün başka bir nedenidir [25]. Aykırı değerlerin oluşmasında insan faktörü dendiğinde veriyi karşılayacak yapıda olmaması, etiketlemenin nesnellikten uzaklaşması, ölçüm hataları gibi durumlar düşünülebilir.

C. Pelletier, S. Valero, J. Inglada, G. Dedieu ve N. Champion [28], çalışmalarında Random Forest (Rasgele Orman) sınıflandırma tekniğini kullanmışlardır. Rasgele Orman sınıflandırıcısı, güçlü bir karar kuralına sahip bir sınıflandırıcı oluşturmak için bir dizi zayıf sınıflandırıcı olarak görülen karar ağaçlarının öğrenilmesinden oluşan bir topluluk öğrenme yöntemi olarak tanımlanmıştır [18]. Aykırı değer tespit yöntemlerini nicel olarak değerlendirmek için, eğitim veri setlerine yapay yanlış etiketlenmiş veriler enjekte edilir [18].

Herhangi bir problem için hangi prosedürün en iyi veya sadece yeterli performans göstereceği önceden bilinen bir durum değildir [46]. "Off the shelf" (standart/hazır) denilen yöntem veri madenciliğinde hızlı, pratik olan, çok zaman almamakla beraber,

fazla öğrenme prosedürleri ile öğrenmek zorunda bırakılmayan ve direk veriye uygulanabilen yöntemlerdir [46]. Tüm bilinen sınıflandırma yöntemleri arasında, karar ağacının, aykırı değerlere sağlamlık da dahil olmak üzere istenen tüm özelliklere sahip olan 'off-the-shelf' olarak adlandırılan şekilde, sonuca en yakın değeri veren bir yöntem olduğu kabul edilmektedir [25].

Literatürde yapılan bazı çalışmalar incelenmiştir. Hatalı verileri içeren veri setleriyle çalışılarak, sağlam fonksiyonları ile farklı algoritmaların karşılaştırıldığı ve yorumlandığı görülmüştür. Aykırı değerlerle ve ezberle mücadele için 2018 yılında StudentNet'in eğitimini denetlemek için MetorNet [47] adı verilen yeni bir teknik geliştirilmiş. Aynı yıl MetorNet ile kıyaslanarak co-teaching (birlikte öğretme) adıyla derin öğrenme paradigması önerilmiş ve özellikle aykırı değerler varlığında daha iyi sonuçlar verdiği öne sürülmüştür. Co-teaching çalışma şeklini araştırmacılar şu şekilde ifade etmişlerdir; iki derin sinir ağı aynı anda eğitiliyor ve her mini partide birbirlerine öğretmelerine izin veriliyor: ilk olarak, her ağ tüm verileri iletir ve muhtemelen temiz etiketlerden bazı verileri seçer; ikinci olarak, iki ağ, bu mini partide hangi verilerin eğitim için kullanılması gerektiği hakkında birbirleriyle iletişim kurar; son olarak, her ağ, kendi eş ağı tarafından seçilen verileri geri yayar ve kendini günceller [48]. Co-teaching özellikle rakam tanıma olarak bilinen MNIST veri setleri, CIFAR-10 görüntü koleksiyonları ve daha küçük görsellerden oluşan CIFAR-100 görüntü koleksiyon verileri üzerinde çalışılmış bir teknik olup ve başarılı sonuçlara ulaşıldığı araştırmacılar tarafından savunulmuştur.

Büyük verilerle çalışan araştırmacılar 3 ana sınıflandırma yöntemi üzerinde durmuşlardır. Bunlar karar ağaçları, destek vektör makineleri (DVM) ve lojistik regresyon olarak karşımıza çıkıyor. Karar ağaçlarında CART yöntemi en sık rastlanılan yöntemler arasındadır. Optimal karar ağacı probleminin sağlam karşılıklarını, nominal optimal karar ağacı probleminden ziyade CART heuristik (sezgisel/buluşsal) yöntemiyle karşılaştırılması mümkündür [24]. Sınıflandırma ve Regresyon Ağaçları (CART) yöntemini cazip hale getiren bir diğer özelliği ise kategorik değişkenlerle de çalışabilme özelliğidir.

Bu tez çalışmasında karşılaştırılacak sınıflandırma teknikleri; karar ağaçları, rasgele ağaçlar, destek vektör makineleri (DVM), doğrusal ayrıştırıcı analizi ve ikinci dereceden ayrıştırıcı analizi, sağlam kayıp fonksiyonu tabanlı gudermannianboost algoritması ve tanjantboost

algoritması, temel bileşenler analizi tabanlı RSIMCA, K-En yakın komşu algoritmaları ve lojistik regresyon üzerinden çalışılmış, ilişkin sonuçlar değerlendirilerek tablolaştırılmıştır.

2.6. Sınıflandırıcıların Değerlendirilmesi için Kullanılan Ölçüm Değerleri

Sınıflandırma analizleri sonucunda elde edilen sonuçları, yapılan sınıflandırma ölçümlerini değerlendirebilmek için bazı metrikler kullanılır. Daha önceden değinilen karışıklık matrisi, ROC eğrisi gibi yöntemler, bu konu başlığı altında ayrıntılı bir şekilde açıklanacaktır. Sınıflandırıcıların değerlendirilmesi ile ilgili aşağıdaki ölçüm değerlerinden bahsedilecektir:

1. Eşik Değerleri (Threshold Metrics)
2. Sıralama Değerleri (Ranking Metrics)
3. Olasılık Değerleri (Probability Metrics)

Bu konu başlıklarının daha iyi anlaşılabilmesi için ölçüm değerleri açıklanmadan önce genel terimlerin ne demek olduğu üzerinde durmak gerekir. Sıklıkla karşılaşılan "Sınıflandırma Doğruluğu (accuracy)", bir veri kümesi analizi sonucunda elde edilen toplam doğru sayısının toplam tahmin sayısına bölünmesi ile hesaplanır. Performans ölçüsü olarak bilinen doğruluk büyük verilerde ve sonuç olarak aykırı değerler içeren veri kümelerinin sınıflandırılması sonucunda karşılaştırmalar için çoğu zaman yeterli değildir.

Karışıklık matrisi, bir model tarafından sınıflandırma görevleri üzerine yapılan tahminlerin bir özetidir. Her bir örneğin ait olduğu gerçek sınıfa göre ayrılmış, her sınıf için yapılan tahmin sayısını özetleyen bir tablodur. Genel gösterimi şu şekildedir:

Çizelge 2.1. Karışıklık Matrisi

	Gerçek Negatif	Gerçek Pozitif
Tahmin Edilen Negatif	Doğru Negatif	Yanlış Negatif
Tahmin Edilen Pozitif	Yanlış Pozitif	Doğru Pozitif

Çizelge 2.1. Gerçek Pozitif, tahmin edilen Pozitifler kesiştiği yer yani doğru pozitiflerin (true positive) tahmin edilen pozitiflere oranı duyarlılığı (sensitivity) verirken, doğru

negatiflerin tahmin edilen negatiflere oranı ise belirlilik/özgüllük (specificity) olarak tanımlanır. Konu ile ilgili ayrıntılara girilmesi sınıflandırma yöntemlerinin değerlendirilirken neye göre metriklerin seçildiğinin anlaşılmasında yol gösterici olacağından genel başlıklar ve bilgilerle tanımlamalar, hesaplamalar ve örnekler verilmiştir.

2.6.1. Eşik Değerleri

Eşik metrikleri, sınıflandırmadaki tahmin hatalarını sayısallaştıran metriklerdir. Tahmin edilen sınıfın çalışılan veri setinde beklenen sınıfla eşleşmediği durumların kesirli, oranlı veya orantılı özetlenmesini sağlamak için eşik değerleri kullanılır. Bu ölçümler bir eşik değerine ve niteliksel hata anlayışına dayanırlar ve bir modelin hata sayısını en aza indirecek şekilde model oluşturmak istendiğinde kullanılan ölçümlerdir.

En çok kullanılan eşik metriği sınıflandırma doğruluğudur (accuracy) ve şu şekilde formüle edilir:

$$Doğruluk = \frac{Doğru\ Tahminlerin\ Sayısı}{Tüm\ Tahminlerin\ Sayısı}$$

Doğruluğun tamamlayıcısı olarak da sınıflandırma hatası olarak tanımlanan değer gelir. O da yanlış tahminlerin tüm tahminlere oranıdır.

$$Hata = \frac{Yanlış\ Tahminlerin\ Sayısı}{Tüm\ Tahminlerin\ Sayısı}$$

Bu şekilde oranlara bakıldığında eşik ölçütleri/metrikleri genelde ikili sınıflandırma (binary) problemi için hesaplanan oranlarla anlaşılabilir. Çünkü sınıflandırmada kullanılan karışıklık matrisi tahmine dayalı model performansına bakarken, hangi sınıfların doğru, hangi sınıfların yanlış tahmin edildiğine ve ne tip hataların yapıldığına dair bilgilere de ulaşılmasını sağlayabilir.

Aşağıda sınıflandırma ile ilgili verilen ikili karışıklık matrisinde sütunlar tahminleri yani sınıflandırma sonucunda verilen kararları, satırlar da gerçekte olan durumu/gerçekleşen sonucu göstermektedir.

Çizelge 2.2. İkili Karışıklık Matrisi

	Pozitif Tahmin	Negatif Tahmin
Pozitif Sınıf	Doğru Pozitif	Yanlış Negatif
Negatif Sınıf	Yanlış Pozitif	Doğru Negatif

Burada da aynı şekilde duyarlılık (sensitivity) hesaplanmasında, doğru olan sınıfın doğru tahmini yani doğru pozitif (true positive) oranlarına bakılırken, belirlilik (specificity) ise bazı kaynaklarda, yanlış pozitif (false positive) oranlarını 1'den çıkararak hesaplanır [49]. Yanlış pozitif, bir şeyi aslında yanlış iken doğru kabul etme durumunu kapsar ve en istenmeyen durumlardandır.

Eşik değerleri anlaşılması ve hesaplamadaki kolaylığı ile sık kullanılırlar. Duyarlılık ve belirlilik eşik değerlerindendir ve yine kolay anlaşılan metrikleridir. Duyarlılık pozitif sınıfın ne kadar iyi tahmin edildiğinin bir özetidir, doğru/gerçek pozitif oranını ifade eder. Belirlilik ise, duyarlılığın tamamlayıcısı olarak veya doğru/gerçek negatif oranını gösterir. Negatif sınıfın ne kadar iyi tahmin edildiğini özetler:

$$Duyarlılık = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + Yanlış\ Negatif}$$

$$Belirlilik = \frac{Doğru\ Negatif}{Doğru\ Negatif + Yanlış\ Pozitif}$$

Araştırmalarda analizlerin sonuç değerleri için tablo gösterimlerinde ve yorumlamalarda eşik değerlerine bakıldığında kesinlik (precision) ve recall metrikleri yer alır. Recall tam Türkçe ifade edilmese de geri çekme, hatırlama gibi kelime anlamları vardır. Ancak hesaplamada duyarlılıkla yani sensitivity ile bir farkı olmadığı için Türkçe kaynaklarda recall, duyarlılık olarak yer almaktadır.

Kesinlik, pozitif sınıfa ait olan ve pozitif sınıfa atanan örneklerin oranı olarak tanımlanabilir:

$$Kesinlik = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + Yanlış\ Pozitif}$$

Recall, pozitif sınıfın ne kadar iyi tahmin edildiğini özetler. Temelinde duyarlılıkla aynı anlama gelir (Konular anlatılırken Recall terimi yerine Duyarlılık kelimesi kullanılacaktır.).

$$Recall(Duyarlılık) = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + Yanlış\ Negatif}$$

Literatür çalışmalarında bu iki önemli metriğin (kesinlik ve duyarlılık) dengelenmesiyle oluşan tek bir hesaplama ölçeği sıklıkla kullanılır; F-değeri (F-measure) veya diğer adıyla F-skoru (F-score) [49]. Daha çok F1-skoru olarak isimlendirilmektedir. Bu şekliyle tek bir ölçekle tek bir değerin karşılaştırılarak sınıflandırmaların değerlendirilmesi için sıklıkla kullanılır.

$$F1 - skoru = \frac{2 * Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık}$$

Araştırmalarda, analiz sonuçlarının gösterildiği tablolarda her bir sınıflandırma tekniği için F1 skorları kıyaslanır. F1 Skoru yüksek olan sınıflandırma tekniğinin daha etkin bir sınıflandırma becerisine sahip olduğu söylenebilir.

Konuyu bir örnekle daha anlaşılır kılabiliriz:

- 10.000 adet verisi olan majör bir sınıf ile az sayıda veriye sahip 100 verili minör bir sınıf, 1:100 oranı ile gösterilmektedir. Pozitif sınıf için 150 örnek tahmin eden bir model olsun, bunun 95 tanesi doğru/gerçek pozitiflerden, 5 tanesi ise yanlış negatiflerden oluşan bir sınıf için, 55 tanesi de yanlış pozitiflerden oluşuyor.

Bu tarz bir örnekte Kesinlik hesaplaması şu şekilde;

$$Kesinlik = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + Yanlış\ Pozitif}$$

$$Kesinlik = \frac{95}{95 + 55} = 0{,}633$$

Duyarlılık hesaplaması şu şekilde:

$$Duyarlılık = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + Yanlış\ Negatif}$$

$$Duyarlılık = \frac{95}{95 + 5} = 0{,}95$$

F1 skoru hesaplanması ise şu şekilde yapılmaktadır:

$$F1\ skoru = \frac{2 * Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık}$$

$$F1\ skoru = \frac{2 * 0{,}633 * 0{,}95}{0{,}633 + 0{,}95}$$

$$F1\ skoru = \frac{2 * 0{,}601}{1{,}583} = 0{,}759$$

2.6.2. Sıralama Değerleri/Metrikleri

Sıralama metrikleri (ranking metrics), sınıflandırıcıların sınıfları ayırmadaki etkinliklerinin değerlendirilmesi için kullanılırlar. Bir sınıflandırıcının tahmin puanı/skoru veya sınıf üyeliği olasılığını tahmin etmesiyle kullanılabilen ölçümlerdir. Burada tahmin puanı ile sınıflandırıcının etkinliğini test etmek için eşik değerlerinden faydalanılır. Farklı eşik değerleriyle bir dizi oluşturulur, başarılı görülen modeller iyi bir sınıf ayrımını sağlayacak ve böylece daha etkili bulunacaktır.

Literatürde en sık kullanılan sıralama metriği ROC eğrisi, diğer bir adıyla ROC analizidir. ROC İngilizce olarak açılımı Receiver Operating Characteristic yani Alıcı/Toplama Operasyon Karakteristiği olarak çevrilebilir. Amacı ikili sınıflandırıcıların sınıfları ayırt etme yeteneklerine dayanan, sınıfları ayırt etme yeteneklerini analiz etmek için bir

çalışma alanı oluşturmaktır. Oluşturulan bu çalışma alanı doğru pozitif oranı ve yanlış pozitif oranı gibi eşik değerlerinden faydalanarak oluşturulan bir modeldir ve grafiksel bir özet niteliğindedir.

Model oluşturulması için gerçek/doğru pozitif oranları(true positive rate) ve yanlış pozitif oranlarına (false positive rate) ihtiyaç vardır. Gerçek pozitif oranı burada da recall yani duyarlılıkla aynı şekilde hesaplanır:

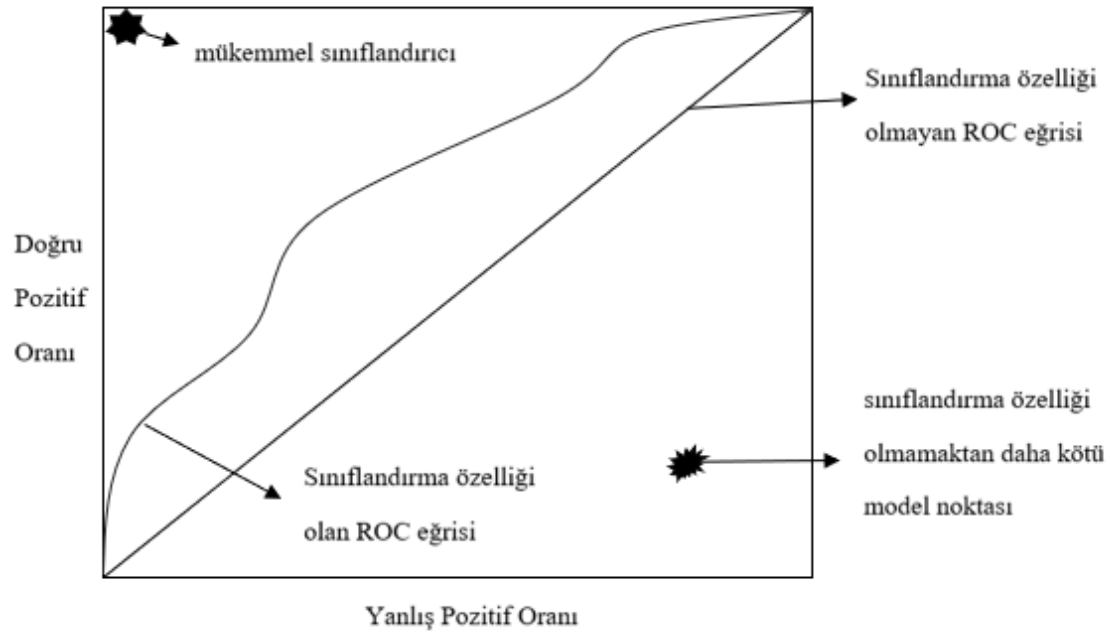
$$Doğru\ Pozitif\ Oranı = \frac{Doğru\ Pozitif}{Doğru\ Pozitif + Yanlış\ Negatif}$$

Yanlış pozitif oranı hesaplanması:

$$Yanlış\ Pozitif\ Oranı = \frac{Yanlış\ Pozitif}{Yanlış\ Pozitif + Doğru\ Negatif}$$

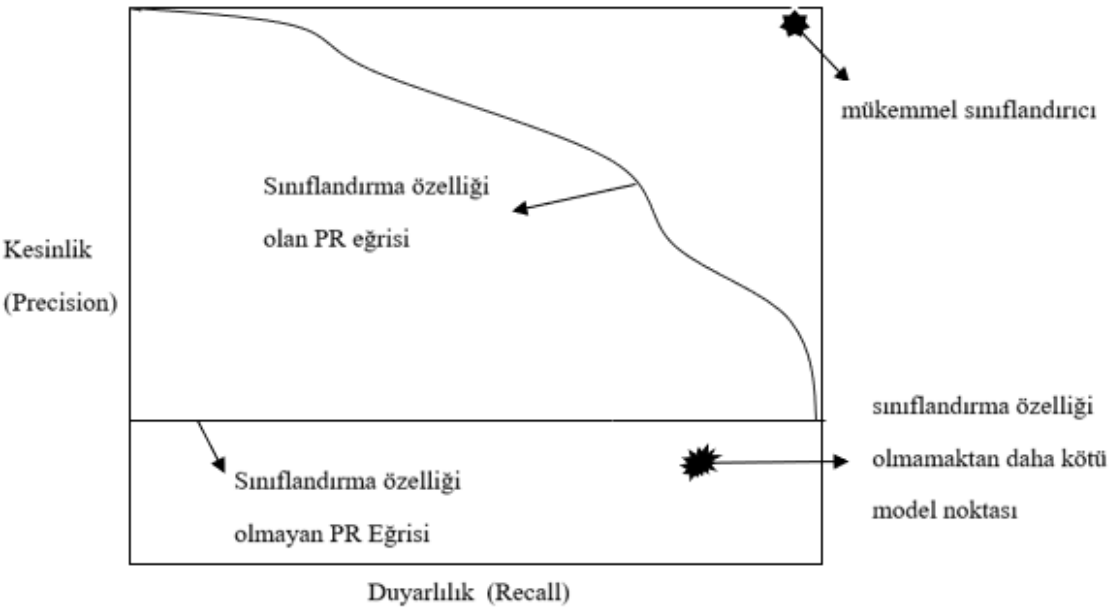
Grafiksel gösterimi oluştururken eşik değerleri hesaplanır ve tüm eşik değerleri noktaları bir eğri oluşturacak şekilde birleştirilir. Bu yüzden de ROC eğrisi (ROC curve) olarak bilinir. Tabloda sınıflandırıcı özelliğinin olmadığını gösteren bir eğri çizilir. Bu çizginin altında kalan nokta sınıflandırıcı özelliği olmamasından daha kötü bir model olarak tanımlanır. Eğrinin üzerinde kalan nokta ise sınıflandırma yeteneği kuvvetli olan modeli verir ve doğru pozitif rate oranı yükseldikçe mükemmel bir modele yaklaşır.

ROC eğrisi ile sınıflandırıcı model değerlendirilmesindeki şematik gösterim şu şekilde ifade edilebilir:



Şekil 2.9. ROC Eğrisi Şeması

Kesinlik ve duyarlılık metrikleri PR (Precision-Recall) eğrisi için şema çizilmektedir. Yeni grafikte sınıflandırma yeteneği olmayan eğri yatayda gösterilir ve yine onun altında kalan nokta sınıflandırılamayan modelden daha da kötüsü olarak tanımlanır. Burada mükemmel sınıflandırıcı model noktası artık sağ üst köşede yer alacaktır. PR Eğrisi çizimi aşağıdaki gibi gösterilir:



Şekil 2.10. PR Eğrisi Şeması

Farklı eşik metriklerini kullanarak oluşturulan ROC eğrisi ve PR eğrisi tek bir sınıflandırıcının değerlendirilmesi için kullanılan metriklerdir [49]. Analizdeki sınıflandırıcıları karşılaştırmak için kullanılmamaktadır.

3. UYGULAMA

3.1. Sınıflandırma Teknikleri için Benzetim Çalışması

Sınıflandırma algoritmalarında sınıflandırmanın başarılı olması için eğitim veri kümesindeki öğrenim sürecinin başarılı olması oldukça önemlidir. Eğitim veri kümesinde yanlış sınıflandırılmış gözlemlerin bulunması durumunda bu süreç sekteye uğrayabilir. Bu durumu incelemek için R 4.5.0 [50] programında klasik ve sağlam yöntemlerin bulunduğu algoritmalar kullanılarak çeşitli karşılaştırmalar yapılmıştır. Programdaki library() fonksiyonunda yer alan library(e1071), library(caret), library(ROCR), library(robust), library(rrcov), library(caTools), library(class), library(ggplot2) ve library(rrcovHD) paketleri kullanılmıştır. Çalışmanın temel amacı, klasik yöntemler yanlış sınıflandırılmış gözlemlerden dolayı ne kadar sınıflandırma başarısızlığına uğruyor ve istatistik temelli sağlam yöntemler hangi yanlış sınıflandırma süreçlerinde başarılı oluyor sorularına cevap vermektir. Bu yöntemlerin karşılaştırılmasının amacı ile Toka'nın [37] üretmiş olduğu 4 farklı benzetim verisi [37] gözlem sayısı sırasıyla 100, 250 ve 1000 olacak şekilde ve eğitim veri kümesinde ise sırasıyla 100 ve 250 gözlem için %5, %10 ve 1000 gözlem için %10 ve %20 aykırı değer olacak şekilde üretilmiştir. Benzetim durumları, kullanılan algoritmalar ve karşılaştırma için kullanılan sınıflandırma başarı ölçütleri Çizelge 3.1.'de verilmiştir.

Çizelge 3.1. Benzetim Çalışması Temel Bilgiler

<table>
<tr><th>Eğitim ve Test Kümesi Gözlem Sayısı</th><th>Eğitim Kümesindeki Yanlış Sınıflandırma Yüzdesi</th><th>Kullanılan Algoritmalar</th><th>Karşılaştırma Ölçütleri</th></tr>
<tr><td>100;250</td><td>%5; %10</td><td rowspan="2">TS-DVM
RSIMCA
RLDA
RLDA-OGK
RLogReg
GudBoost
TanBoost
LogReg
RQDA
KNN
RegTree</td><td rowspan="2">Duyarlılık
Belirlilik/Özgüllük
Genel Doğruluk Oranı
F1-Skoru</td></tr>
<tr><td>1000</td><td>%10;%20</td></tr>
</table>

Benzetim çalışması için farklı yapıya sahip dört benzetim verisi üretilmiştir. Benzetim veri kümelerinde sınıflandırıcıya yakın tek bir yönden, sınıflandırıcıya yakın iki yönden de yanlış sınıflandırma bulunmasıdır. Diğer taraftan aynı iki grup için sınıflandırıcıya uzak tek ve iki taraflı olarak veri üretilmiştir. Veri kümeleri aşağıdaki gibi özetlenebilir. Bu yapı Toka [37] çalışmasındaki yapıya benzer farklı sayılarda elde edilmiştir.

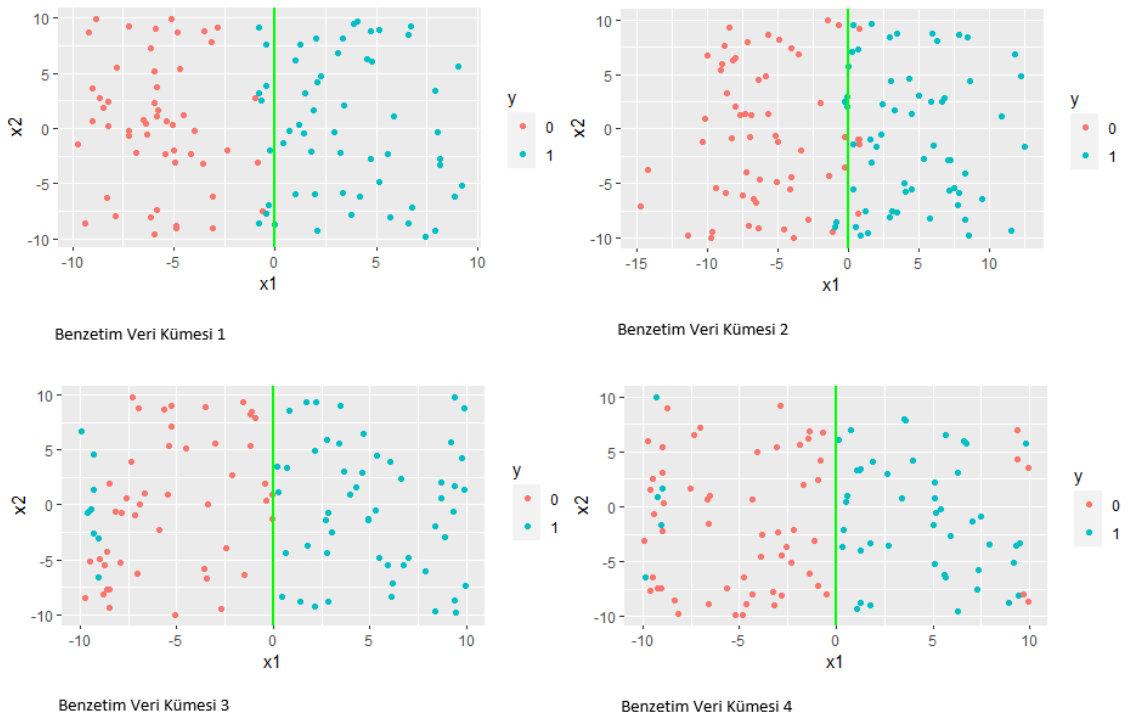
Benzetim verisi 1: Sınıflandırıcıya yakın tek bir yönden yanlış sınıflandırılmış verilerin olduğu durumdur.

Benzetim verisi 2: Sınıflandırıcıya yakın iki yönden yanlış sınıflandırılmış verilerin olduğu durumdur.

Benzetim verisi 3: Sınıflandırıcıya uzak tek yönden yanlış sınıflandırılmış verilerin olduğu durumdur.

Benzetim verisi 4: Sınıflandırıcıya uzak iki grupta da, uzak iki yönlü yanlış sınıflandırılmanın olduğu durumdur.

Üretilmiş olan benzetim veri kümeleri için birer örnek olarak Şekil 11'de görselleştirilmiştir. İki değişkenli yapı ile görsellik kolaylığı sağlanmış ve $x_1 = 0$ doğrusu gerçek sınıflandırıcı olarak kabul edilmiştir. Benzetim çalışması için yukarıda belirtilen özelliklere sahip dört veri kümesi için yukarıda belirtilen kısıtlarla çoğaltılmıştır. 100, 250 ve 1000 gözleme sahip eğitim kümeleri oluşturulmuş, 1000 tekrar sonucunda elde edilen ortalama karşılaştırma ölçütleri, ortalama ve standart sapma olarak tablolaştırılmış ve yorumlanmıştır. Ayrıca tüm ölçüt değerleri kullanılarak tablo ve grafiksel gösterimler sunulmuştur.



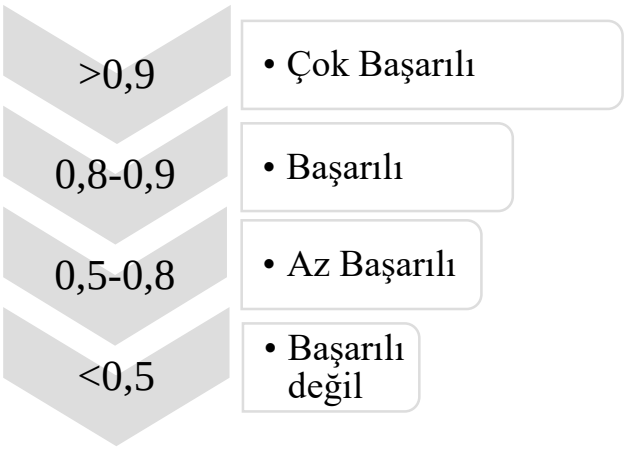
Şekil 3.1. Benzetim Veri Kümeleri

3.1.1. Uygulamada Skor Puanları Yorumları

Benzetim veri kümeleri üzerinden yapılan karşılaştırmalarda duyarlılık, belirlilik, genel doğruluk ve F1-skor (F-ölçüsü veya dengelenmiş F-skoru olarak da bilinir) değerleri kullanılmıştır. Duyarlılık, belirlilik ve doğruluk oranları aritmetik hesaplanan ortalamalar iken, F1-skoru, duyarlılık ve kesinlik gibi aritmetik ortalamaların harmonik ortalamasını hesaplayarak model performansını ölçer. Bu şekilde harmonik ortalama ile daha hassas bir ölçüm elde edilir. F1 puanı, ikili sınıflandırmanın preformansını ölçmek için kullanılır,

0 ile 1 arasında değişir; burada en iyi puan 1, olabilecek en kötü puan da 0'dır. F1-skoru 1 değerini alıyor ise modelin her gözlemi doğru tahmin ettiği söylenebilir.

Çalışılan veri tipine göre F1-skor değerlerinin de performans yorumları değişkenlik gösterir. Örneğin müşteri kaybı sorunu ile ilgilenen bir modelde 0,8-0,9 skor değerleri iyi bir başarı skoru olarak kabul edilebilirken, hastalığın ortaya çıkışı ile ilgilenen veya yine hayati riskler içeren durumlarda 0,9-0,95 skor değeri iyi bir başarı olarak sınıflandırılabilir. Genel olarak F1 skoru için yorumlama Şekil 3.2.'deki gibi verilir:



Şekil 3.2. F1-Skoru Başarı Sıralaması

1) Benzetim Veri Kümesi 1 için:

Benzetim veri kümesi 1 ile yani tek bir tarafta sınıflandırıcıya yakın yanlış sınıflandırılmış eğitim gözlemleri ile çalışıldığında oluşturulan, Çizelge 3.2. ve Çizelge 3.3.'te gözlem sayısı, yanlış sınıflandırma oranı verilen gözlemlerin aynı gözlem sayısına sahip test kümesindeki duyarlılık ve belirlilik değerlerinin 1000 tekrar için ortalama ve standart sapma değerleri verilmiştir. Elde edilen duyarlılık değerlerine göre inceleme yapıldığında tek örneklem destek vektör makinesi olarak adlandırılan algoritma (TS-DVM) en başarısız sonuçları elde etmiştir. Temel olarak eğitim kümesi içindeki aykırı kısmı tespit etme amacını taşımasından dolayı odaklandığı yapıda sınıflandırma anlamında yanlılık oluştuğu görülmüştür. Diğer algoritmalarda benzetim kümesi veri 1 için yani tek bir tarafta sınıflandırıcıya yakın yanlış sınıflandırılmış eğitim gözlemleri olduğunda, sonuçlar duyarlılık için başarı göstermektedir.

KNN algoritması diğer yöntemlere göre sınıflandırıcı yanındaki yanlış sınıflandırılmış gözlemlerin etkisiyle duyarlılık skorlarında düşme göstermiştir. Ama yanlış sınıflandırma

oranı arttığında sınıflandırma için baz alınan komşular nedeniyle duyarlılık skorunun arttığı görülmüştür. Lojistik regresyon sınıflandırma yakınındaki hatalı sınıflı eğitim kümelerinden az etkilendiği için duyarlılık skorunda bu veri kümesi için yeterli başarıyı göstermiştir, benzer özelliğe sahip olan sağlam lojistik regresyon, tanjantboost ve gudermannianboost algoritmalarında da aynı başarı söz konusudur. Gözlem sayısı arttıkça başarının daha da yükseldiği, şimdilik hatalı sınıflandırma oranının başarıda bir etki yaratmadığı tespit edilmiştir. Aynı veri kümesi için belirlilik ölçüm tablosu incelendiğinde Çizelge 3.2.'de doğru sınıflandırma durumuna göre belirlilik skorlarında genel olarak bir düşme eğilimi görülmüştür. Ayrıca gözlem sayısı arttıkça belirlilik değerlerindeki artış yanlış sınıflandırma oranı olan eğitim kümesindeki aykırı değerler için yine düşme eğilimi göstermiştir.

Sağlam yöntemlerin, klasik yöntemlere göre daha iyi olduğu görülmüş ancak lojistik regresyonun özellikle RSIMCA, tanjantboost ve gudermannianboost algoritmalarına göre daha yüksek belirlilik skoruna sahip olduğu görülmüştür. Tek bir taraftan sınıflandırıcıya yakın yanlış sınıflandırılmış gözlemlerin eğitim kümesindeki yanlılığı lojistik regresyonun kestirim değerlerini minimum karesel artık hata düzeyindeki az etkisinden dolayı bozmadığı görülmüştür. RLDA ve RQDA belirlilik skorunda en başarılı algoritmalar olmuştur. Ancak duyarlılık ve belirlilik değerleri sınıflandırma başarılarını tek başlarına ölçmek yerine birlikte hesaplandığında karşılaştırma açısından daha mantıklı olacağından Çizelge 3.4. ve Çizelge 3.5.'te sırasıyla dengelenmiş genel doğruluk oranı ve F1-skoru incelenmiştir. Genel doğruluk oranıyla incelendiğinde özellikle OGK kestiricili sağlam doğrusal ayrıştırıcının (RLDA-OGK) benzetim verisi 1 için oldukça başarılı hesaplamalara sahiptir. TS-DVM algoritması bu veri kümesinde en başarısız sınıflandırma becerisine sahipken RLDA, tanjantboost ve gudermannianboost algoritmaları özellikle gözlem sayısının ve yanlış sınıflandırma oranının en yüksek olduğu durumda başarısız genel doğrulama oranına sahiptir. Yanlış sınıflandırma oranının arttığı durumlarda tüm algoritmalarda genel doğruluk oranında düşmeler görülmüştür.

Çizelge 3.2. Benzetim Veri Kümesi 1 - Duyarlılık İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.81932	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.98061	1.00000
		St.Sapma	0.01879	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00113	0.00000
	%10	Ortalama	0.83163	0.99770	0.98789	0.99727	0.99986	0.99646	0.99659	0.99986	0.99110	0.98932	1.00000
		St.Sapma	0.07409	0.00994	0.02544	0.00967	0.00236	0.01219	0.01185	0.00236	0.02788	0.01933	0.00000
250	%5	Ortalama	0.88649	0.99860	0.98800	0.99716	0.99998	0.99496	0.99495	0.99998	0.99599	0.98997	0.99947
		St.Sapma	0.04128	0.00596	0.01955	0.00755	0.00037	0.01299	0.01308	0.00037	0.00969	0.01194	0.00318
	%10	Ortalama	0.88613	0.99981	0.99606	0.99967	1.00000	0.99953	0.99953	1.00000	0.99928	0.99632	0.99986
		St.Sapma	0.03963	0.00193	0.01040	0.00215	0.00000	0.00298	0.00291	0.00000	0.00348	0.00692	0.00135
1000	%10	Ortalama	0.91417	1.00000	0.99958	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.99834	1.00000
		St.Sapma	0.01687	0.00000	0.00207	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00230	0.00000
	%20	Ortalama	0.90632	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.99956	1.00000
		St.Sapma	0.01879	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00113	0.00000

Çizelge 3.3. Benzetim Veri Kümesi 1 - Belirlilik İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.94601	0.91782	0.94224	0.95660	0.94716	0.92835	0.92780	0.94704	0.92794	0.92923	0.95016
		St.Sapma	0.01376	0.03417	0.02459	0.01758	0.01363	0.02670	0.02682	0.01363	0.02037	0.01623	0.01343
	%10	Ortalama	0.92716	0.88230	0.90981	0.92798	0.92698	0.88543	0.88404	0.92698	0.88892	0.89018	0.90936
		St.Sapma	0.04033	0.07171	0.06834	0.04869	0.04009	0.07541	0.07618	0.04009	0.07802	0.05710	0.04619
250	%5	Ortalama	0.94040	0.92366	0.95700	0.96533	0.94976	0.94021	0.93960	0.94976	0.95023	0.94262	0.94740
		St.Sapma	0.02472	0.04638	0.03498	0.02513	0.02302	0.04281	0.04303	0.02302	0.03311	0.02703	0.03410
	%10	Ortalama	0.92548	0.89109	0.92204	0.93396	0.92796	0.89036	0.88912	0.92796	0.91054	0.90067	0.91610
		St.Sapma	0.02544	0.04753	0.04237	0.02946	0.02513	0.04954	0.04994	0.02513	0.03749	0.03211	0.03045
1000	%10	Ortalama	0.92141	0.89505	0.92122	0.93466	0.92790	0.89191	0.89069	0.92790	0.91294	0.90810	0.90243
		St.Sapma	0.01349	0.03241	0.02852	0.01495	0.01295	0.02550	0.02575	0.01295	0.01826	0.01502	0.01430
	%20	Ortalama	0.90743	0.84901	0.78807	0.87661	0.90454	0.80005	0.79787	0.90454	0.84609	0.87495	0.89952
		St.Sapma	0.01376	0.03417	0.02459	0.01758	0.01363	0.02670	0.02682	0.01363	0.02037	0.01623	0.01343

Çizelge 3.4. Benzetim Veri Kümesi 1 - Genel Doğruluk Oranı İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.87281	0.95739	0.96125	0.97458	0.97385	0.95943	0.95920	0.97380	0.95530	0.95610	0.97533
		St.Sapma	0.01229	0.01549	0.01193	0.00878	0.00685	0.01254	0.01260	0.00685	0.01006	0.00805	0.00681
	%10	Ortalama	0.87118	0.94338	0.95101	0.96383	0.96439	0.94445	0.94393	0.96439	0.94320	0.94221	0.95636
		St.Sapma	0.04823	0.03284	0.03081	0.02328	0.01989	0.03334	0.03357	0.01989	0.03711	0.02681	0.02300
250	%5	Ortalama	0.91058	0.96274	0.97311	0.98168	0.97551	0.96874	0.96844	0.97551	0.97388	0.96704	0.97425
		St.Sapma	0.02583	0.02135	0.01681	0.01182	0.01129	0.01898	0.01908	0.01129	0.01549	0.01324	0.01631
	%10	Ortalama	0.90322	0.94805	0.96041	0.96767	0.96494	0.94774	0.94720	0.96494	0.95660	0.95049	0.95940
		St.Sapma	0.02519	0.02273	0.01993	0.01450	0.01275	0.02309	0.02322	0.01275	0.01817	0.01584	0.01507
1000	%10	Ortalama	0.91741	0.95035	0.96204	0.96838	0.96520	0.94884	0.94829	0.96520	0.95832	0.95525	0.95349
		St.Sapma	0.01126	0.01486	0.01309	0.00724	0.00640	0.01180	0.01191	0.00640	0.00872	0.00734	0.00703
	%20	Ortalama	0.90672	0.92994	0.90417	0.94186	0.95443	0.90913	0.90823	0.95443	0.92852	0.94092	0.95214
		St.Sapma	0.01229	0.01549	0.01193	0.00878	0.00685	0.01254	0.01260	0.00685	0.01006	0.00805	0.00681

Çizelge 3.5. Benzetim Veri Kümesi 1 - F1-Skoru İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.88267	0.95532	0.96054	0.97414	0.97334	0.95781	0.95754	0.97328	0.95398	0.95492	0.97453
		St.Sapma	0.01163	0.01709	0.01229	0.00879	0.00682	0.01335	0.01341	0.00682	0.01018	0.00810	0.00672
	%10	Ortalama	0.87939	0.94000	0.94885	0.96263	0.96342	0.94095	0.94032	0.96342	0.94001	0.93975	0.95468
		St.Sapma	0.04128	0.03510	0.03308	0.02436	0.02007	0.03664	0.03700	0.02007	0.04006	0.02777	0.02310
250	%5	Ortalama	0.91344	0.96113	0.97250	0.98125	0.97487	0.96758	0.96728	0.97487	0.97311	0.96630	0.97343
		St.Sapma	0.02353	0.02248	0.01758	0.01225	0.01150	0.02010	0.02019	0.01150	0.01609	0.01341	0.01697
	%10	Ortalama	0.90581	0.94545	0.95905	0.96682	0.96398	0.94494	0.94433	0.96398	0.95491	0.94850	0.95798
		St.Sapma	0.02306	0.02365	0.02060	0.01464	0.01257	0.02451	0.02471	0.01257	0.01858	0.01587	0.01517
1000	%10	Ortalama	0.91779	0.94753	0.96040	0.96733	0.96395	0.94596	0.94534	0.96395	0.95647	0.95322	0.95121
		St.Sapma	0.01070	0.01621	0.01412	0.00748	0.00647	0.01275	0.01288	0.00647	0.00913	0.00746	0.00715
	%20	Ortalama	0.90688	0.92450	0.89403	0.93831	0.95227	0.90002	0.89894	0.95227	0.92305	0.93725	0.94976
		St.Sapma	0.01163	0.01709	0.01229	0.00879	0.00682	0.01335	0.01341	0.00682	0.01018	0.00810	0.00672

2) Benzetim Veri Kümesi 2 için Sonuçlar:

Benzetim veri kümesi 2 için yani iki taraftan sınıflandırıcıya yakın yanlış sınıflandırılmış eğitim gözlemlerine göre oluşturulan, Çizelge 3.6. ve Çizelge 3.7.'de gözlem sayısı, yanlış sınıflandırma oranı verilen gözlemlerin aynı gözlem sayısına sahip test kümesindeki duyarlılık ve belirlilik değerlerinin 1000 tekrar için ortalama ve standart sapma değerleri verilmiştir.

Algoritmalarda iki taraftan sınıflandırıcıya yakın yanlış sınıflandırılmış eğitim gözlemleri olduğunda elde edilen duyarlılık değerlerine göre inceleme yapıldığında TS-DVM burada da en başarısız sonuçları elde etmiştir. Diğer algoritmalarda sonuçlar duyarlılık için benzer başarılar göstermektedirler. KNN algoritması diğer yöntemlere göre sınıflandırıcı yanındaki yanlış sınıflandırılmış gözlemlerin etkisiyle duyarlılık ve belirlilik skorlarında düşme göstermiştir. Yanlış sınıflandırma oranı arttığında sınıflandırma için baz alınan komşular nedeniyle duyarlılık skorunda çok az farklarla giderek düştüğü görülmüştür. Lojistik regresyon sınıflandırma yakınındaki hatalı sınıflı eğitim kümelerinden az etkilendiği için duyarlılık skorunda bu veri kümesi için yeterli başarıyı gösterdiği gibi benzer özelliğe sahip olan sağlam lojistik regresyon, RSIMCA, RLDA, RLDA-OGK, tanjantboost ve gudermannianboost algoritmalarında da benzer başarılar söz konusudur. Gözlem sayısı arttıkça başarılı bulunan bu algoritmalar için, başarının daha da yükseldiği, artan gözlem sayısı ve hatalı sınıflandırma oranının da artışı başarıda artış gösterme yönünde bir etki yarattığı tespit edilmiştir.

Aynı veri kümesi için belirlilik ölçüm tablosu incelendiğinde Çizelge 3.7. etiketinin doğru sınıflandırma durumuna göre belirlilik skorlarında genel olarak hafif bir düşme eğilimi görülmüştür. Ayrıca gözlem sayısı arttıkça belirlilik değerlerindeki artış yanlış sınıflandırma oranı olan eğitim kümesindeki aykırı değerler için yükselme eğilimi göstermektedir. Sağlam yöntemlerin, klasik yöntemlere göre daha iyi olduğu görülmüş ancak sağlam lojistik regresyon ve onunla aynı sonuçlara yakın olan lojistik regresyonun, özellikle RSIMCA, RLDA, RLDA-OGK, tanjantboost ve gudermannianboost algoritmalarına göre daha yüksek belirlilik skoruna sahip olduğu görülmüştür. İki taraftan sınıflandırıcıya yakın yanlış sınıflandırılmış eğitim gözlemleri kümesindeki yanlılığı lojistik regresyonun kestirim değerlerini minimum karesel artık hataya olan az etkisinden dolayı bozmadığı görülmüştür.

Duyarlılık ve belirlilik değerlerinin yanında sınıflandırma başarılarını ölçmek için Çizelge 3.8. ve Çizelge 3.9.'da sırasıyla genel doğruluk oranı ve F1-skoru incelenmiştir. Genel doğruluk oranıyla incelendiğinde özellikle en başarılı skor RLDA, RQDA ile elde edilmiş, RLDA-OGK' nın benzetim verisi 2 için de oldukça başarılı hesaplamalara sahip olduğu gözlenmiştir. TS-DVM algoritması bu veri kümesinde en başarısız sınıflandırma becerisine sahipken tanjantboost ve gudermannianboost algoritmaları ile yaklaşık bir F1 skoru ile RLDA da özellikle gözlem sayısının ve yanlış sınıflandırma oranının en yüksek olduğu durumdan etkilenmeyerek başarılı genel doğrulama oranına sahiptir. Benzetim verisi 2 için yanlış sınıflandırma oranının arttığı durumlarda tüm algoritmalarda genel doğruluk oranında önemsiz düşmeler görülmüştür. RSIMCA, RQDA, tanjantboost ve gudermannianboost benzer, RLDA-OGK ve lojistik regresyonun sınıflandırma başarıları çok az bir farkla daha önde denilebilir.

Çizelge 3.6. Benzetim Veri Kümesi 2 - Duyarlılık İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.80493	0.96789	0.96445	0.97918	0.98869	0.96502	0.96479	0.98864	0.96396	0.96165	0.98793
		St.Sapma	0.07883	0.04623	0.04522	0.02977	0.01835	0.04885	0.04935	0.01839	0.04869	0.03727	0.02655
	%10	Ortalama	0.82968	0.96924	0.96333	0.97753	0.98764	0.96200	0.96165	0.98764	0.96224	0.96115	0.97943
		St.Sapma	0.07300	0.04397	0.04488	0.03019	0.01894	0.04953	0.05017	0.01894	0.05219	0.03740	0.03495
250	%5	Ortalama	0.89102	0.97949	0.97792	0.98832	0.99315	0.97921	0.97890	0.99315	0.98338	0.97538	0.99424
		St.Sapma	0.04063	0.02820	0.02683	0.01606	0.00974	0.02818	0.02846	0.00974	0.02132	0.01890	0.01340
	%10	Ortalama	0.90598	0.98425	0.97658	0.98700	0.99266	0.97570	0.97551	0.99266	0.98237	0.97106	0.98460
		St.Sapma	0.03669	0.02105	0.02699	0.01727	0.01000	0.03164	0.03201	0.01000	0.02144	0.01996	0.02459
1000	%10	Ortalama	0.93630	0.99314	0.98764	0.99379	0.99645	0.98886	0.98876	0.99645	0.99210	0.97725	0.99509
		St.Sapma	0.01601	0.00804	0.01303	0.00750	0.00428	0.01332	0.01346	0.00428	0.00888	0.00867	0.00943
	%20	Ortalama	0.94421	0.99204	0.98664	0.99303	0.99611	0.98723	0.98715	0.99611	0.99151	0.95571	0.95381
		St.Sapma	0.01424	0.00957	0.01477	0.00818	0.00421	0.01534	0.01547	0.00421	0.00922	0.01137	0.03912

Çizelge 3.7. Benzetim Veri Kümesi 2 - Algoritmaların Belirlilik İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.98090	0.96341	0.96220	0.97736	0.98679	0.96172	0.96156	0.98676	0.95859	0.95937	0.98620
		St.Sapma	0.02263	0.05073	0.04532	0.03098	0.02066	0.05036	0.05067	0.02068	0.05385	0.04024	0.02745
	%10	Ortalama	0.96996	0.96941	0.96275	0.97703	0.98671	0.96467	0.96418	0.98671	0.96189	0.96029	0.98151
		St.Sapma	0.02785	0.04457	0.04782	0.03146	0.01990	0.05068	0.05107	0.01990	0.05302	0.03831	0.03263
250	%5	Ortalama	0.97876	0.98111	0.97692	0.98671	0.99294	0.97652	0.97633	0.99294	0.98202	0.97467	0.99311
		St.Sapma	0.01731	0.02858	0.02649	0.01742	0.00990	0.02938	0.02965	0.00990	0.02236	0.01927	0.01493
	%10	Ortalama	0.95881	0.98432	0.97660	0.98691	0.99244	0.97685	0.97666	0.99244	0.98290	0.97090	0.98646
		St.Sapma	0.02184	0.02097	0.02580	0.01656	0.01017	0.02947	0.02967	0.01017	0.02037	0.02058	0.02265
1000	%10	Ortalama	0.95354	0.99268	0.98728	0.99328	0.99610	0.98787	0.98779	0.99610	0.99133	0.97664	0.99552
		St.Sapma	0.01220	0.00785	0.01407	0.00752	0.00415	0.01459	0.01470	0.00415	0.00916	0.00891	0.00892
	%20	Ortalama	0.93454	0.99138	0.98685	0.99323	0.99609	0.98849	0.98842	0.99609	0.99183	0.95598	0.95534
		St.Sapma	0.01235	0.00936	0.01393	0.00758	0.00434	0.01403	0.01412	0.00434	0.00869	0.01243	0.03835

Çizelge 3.8. Benzetim Veri Kümesi 2 - Genel Doğruluk Oranı İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.88037	0.96535	0.96287	0.97807	0.98757	0.96307	0.96284	0.98753	0.96103	0.96009	0.98695
		St.Sapma	0.05036	0.02713	0.02873	0.01958	0.01411	0.02918	0.02950	0.01414	0.03245	0.02454	0.01728
	%10	Ortalama	0.89034	0.96866	0.96250	0.97690	0.98699	0.96263	0.96221	0.98699	0.96132	0.96020	0.98021
		St.Sapma	0.04562	0.02709	0.03038	0.02055	0.01410	0.02941	0.02965	0.01410	0.03701	0.02466	0.02007
250	%5	Ortalama	0.93152	0.98033	0.97743	0.98758	0.99304	0.97792	0.97767	0.99304	0.98277	0.97506	0.99364
		St.Sapma	0.02385	0.01574	0.01606	0.01011	0.00679	0.01566	0.01573	0.00679	0.01291	0.01195	0.00916
	%10	Ortalama	0.93014	0.98424	0.97647	0.98692	0.99256	0.97620	0.97600	0.99256	0.98257	0.97085	0.98540
		St.Sapma	0.02182	0.01233	0.01615	0.01015	0.00682	0.01644	0.01659	0.00682	0.01260	0.01250	0.01582
1000	%10	Ortalama	0.94436	0.99291	0.98748	0.99355	0.99628	0.98839	0.98830	0.99628	0.99172	0.97693	0.99530
		St.Sapma	0.00989	0.00445	0.00772	0.00426	0.00291	0.00710	0.00715	0.00291	0.00511	0.00565	0.00562
	%20	Ortalama	0.93956	0.99170	0.98672	0.99312	0.99609	0.98781	0.98773	0.99609	0.99165	0.95577	0.95435
		St.Sapma	0.00953	0.00505	0.00823	0.00440	0.00298	0.00760	0.00766	0.00298	0.00503	0.00812	0.01554

Çizelge 3.9. Benzetim Veri Kümesi 2 - F1-Skoru İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.89292	0.96565	0.96333	0.97827	0.98774	0.96337	0.96317	0.98770	0.96128	0.96051	0.98707
		St.Sapma	0.04029	0.02652	0.02748	0.01919	0.01383	0.02811	0.02834	0.01386	0.03191	0.02394	0.01677
	%10	Ortalama	0.89982	0.96932	0.96304	0.97728	0.98717	0.96334	0.96292	0.98717	0.96207	0.96072	0.98047
		St.Sapma	0.03804	0.02585	0.02914	0.01974	0.01375	0.02811	0.02832	0.01375	0.03461	0.02360	0.01953
250	%5	Ortalama	0.93489	0.98030	0.97742	0.98752	0.99305	0.97787	0.97762	0.99305	0.98270	0.97502	0.99367
		St.Sapma	0.02125	0.01573	0.01591	0.01011	0.00677	0.01553	0.01561	0.00677	0.01297	0.01178	0.00898
	%10	Ortalama	0.93240	0.98429	0.97659	0.98695	0.99255	0.97628	0.97609	0.99255	0.98263	0.97098	0.98553
		St.Sapma	0.02017	0.01208	0.01576	0.00999	0.00678	0.01611	0.01625	0.00678	0.01240	0.01223	0.01551
1000	%10	Ortalama	0.94492	0.99291	0.98746	0.99354	0.99627	0.98836	0.98827	0.99627	0.99171	0.97695	0.99530
		St.Sapma	0.00953	0.00443	0.00774	0.00426	0.00290	0.00713	0.00718	0.00290	0.00510	0.00562	0.00561
	%20	Ortalama	0.93938	0.99171	0.98674	0.99313	0.99610	0.98786	0.98778	0.99610	0.99167	0.95584	0.95458
		St.Sapma	0.00931	0.00503	0.00818	0.00439	0.00298	0.00751	0.00757	0.00298	0.00500	0.00800	0.01525

3) Benzetim Veri Kümesi 3 için Sonuçlar:

Benzetim verisi 3 ile sınıflandırıcıya uzak tek yönden yanlış sınıflandırılmış verilerin olduğu durumlar için, Çizelge 3.10. ve Çizelge 3.11.'de gözlem sayısı, yanlış sınıflandırma oranı verilen gözlemlerin aynı gözlem sayısına sahip test kümesindeki duyarlılık ve belirlilik değerlerinin 1000 tekrar için ortalama ve standart sapma değerleri verilmiştir. Elde edilen duyarlılık değerlerine göre sınıflandırıcıdan uzaklaşılması TS-DVM başarısını daha da düşürmüş, yanlılığını daha belirgin hale getirmiştir. Diğer algoritmalarda benzetim kümesi veri 3 için yani tek bir tarafta sınıflandırıcıya uzak yanlış sınıflandırılmış eğitim gözlemleri olduğunda, sonuçlar duyarlılık için başarı göstermektedir. Lojistik regresyon sınıflandırma yakınındaki hatalı sınıflı eğitim kümelerinden az etkilendiği için duyarlılık skorunda bu veri kümesi için yeterli başarıyı gösterdiği gibi benzer özelliğe sahip olan sağlam lojistik regresyonla neredeyse aynı skoru paylaşıyorlar, yaklaşık değerlerle tanjantboost ve gudermannianboost algoritmalarında da yanlış sınıflama oranının ve n'in artması ile lojistik algoritmalarla aynı başarı düzeyine çıktıkları görülüyor. Gözlem sayısı ve yanlış sınıflama oranlarındaki artışın benzetim verisi 3 için sağlam lojistik regresyon, lojistik regresyonun tanjantboost ve gudermannianboost algoritmaların başarısında yükseltici bir etki yarattığı söylenebilir.

Aynı veri kümesi için belirlilik ölçüm tablosu incelendiğinde Çizelge 3.11.'de etiketinin doğru sınıflandırma durumuna göre belirlilik skorlarında genel olarak bir düşme eğilimi görülmüştür. Ayrıca gözlem sayısı arttıkça belirlilik değerlerindeki artış yanlış sınıflandırma oranı olan eğitim kümesindeki aykırı değerler için keskin bir düşme eğilimi göstermiştir. Sağlam yöntemlerin, klasik yöntemlere göre daha iyi olduğu görülmüş ancak RLDA, RLDA-OGK, RQDA, özellikle RSIMCA, tanjantboost ve gudermannianboost algoritmalarına göre daha yüksek belirlilik skoruna sahip olduğu gözlenmiştir. RLDA ve RQDA belirlilik skorunda n artışından etkilenmeyerek en başarılı algoritmalar olmuştur.

Hesaplandığında karşılaştırma açısından daha mantıklı olması için Çizelge 3.12. ve Çizelge 3.13.'te sırasıyla dengelenmiş genel doğruluk oranı ve F1-skoru incelenmiştir. Genel doğruluk oranıyla incelendiğinde özellikle OGK kestiricili RLDA-OGK benzetim verisi 3 için oldukça başarılı hesaplamalara sahiptir. Ancak gözlem sayısının ve yanlış sınıflandırma oranının yükseldiği durumlarda RLDA-OGK'nın başarısı düşmüş, yerini RLDA ve RQDA'ya bırakmıştır. TS-DVM algoritması bu veri kümesinde de en başarısız

sınıflandırma becerisine sahipken lojistik regresyon, RSIMCA, tanjantboost ve gudermannianboost algoritmaları özellikle gözlem sayısının ve yanlış sınıflandırma oranının en yüksek olduğu durumda başarısız genel doğrulama oranına sahiptir. Yanlış sınıflandırma oranının arttığı durumlarda tüm algoritmalarda genel doğruluk oranında düşmeler görülmüştür. F1-skoru için RQDA, RLDA artan gözlem değerlerine ve yanlış sınıflandırma oranlarına rağmen başarılarını korumuşlardır. Ancak diğer algoritmaların başarılarında ciddi düşüşler gözlenmiştir. F1-skoruna göre tek taraftan yanlış sınıflandırmalı da başarısı biraz düşen RLDA-OGK iki taraflı sınıflandırıcıya uzakta yanlış sınıflandırmaların olduğu (benzetim veri kümesi 4) başarısını yükseltmiştir. Artan veriler ve aykırı değer oranları ile en başarılı algoritmalar, RQDA, RLDA-OGK ve RLDA olarak tespit edilmiş ve artan gözlem değerlerine ve yanlış sınıflandırma oranlarına rağmen başarıları yükselmiştir. Lojistik Regresyon, RSIMCA, Tanjantboost ve Gudermannianboost algoritmaları birbirlerine yakın ve başarılı sonuçlar vermişlerdir.

Çizelge 3.10. Benzetim Veri Kümesi 3 - Duyarlılık İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.84616	0.98385	0.97509	0.98498	0.99647	0.98909	0.98912	0.99649	0.97705	0.96433	0.99442
		St.Sapma	0.07278	0.03451	0.03803	0.02684	0.01239	0.02866	0.02877	0.01251	0.03904	0.03603	0.01359
	%10	Ortalama	0.85110	0.99257	0.98006	0.98627	0.99767	0.99621	0.99621	0.99764	0.98167	0.96207	0.99566
		St.Sapma	0.07276	0.02254	0.03158	0.02412	0.01105	0.01670	0.01670	0.01107	0.03241	0.03910	0.01140
250	%5	Ortalama	0.89845	0.99176	0.98501	0.99183	0.99917	0.99713	0.99713	0.99919	0.98985	0.97712	0.99799
		St.Sapma	0.03798	0.01875	0.02021	0.01256	0.00347	0.01049	0.01050	0.00343	0.01583	0.01975	0.00507
	%10	Ortalama	0.89634	0.99789	0.99148	0.99366	0.99989	0.99964	0.99966	0.99989	0.99444	0.97668	0.99816
		St.Sapma	0.03972	0.00757	0.01449	0.01123	0.00148	0.00359	0.00352	0.00148	0.01134	0.01936	0.00494
1000	%10	Ortalama	0.91961	0.99962	0.99830	0.99885	1.00000	1.00000	1.00000	1.00000	0.99958	0.98924	0.99951
		St.Sapma	0.01694	0.00201	0.00380	0.00290	0.00000	0.00000	0.00000	0.00000	0.00189	0.00641	0.00140
	%20	Ortalama	0.91417	1.00000	0.99991	0.99996	1.00000	1.00000	1.00000	1.00000	1.00000	0.98874	0.99957
		St.Sapma	0.01730	0.00000	0.00085	0.00044	0.00000	0.00000	0.00000	0.00000	0.00009	0.00646	0.00115

Çizelge 3.11. Benzetim Veri Kümesi 3 - Belirlilik İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.64478	0.93954	0.95848	0.97373	0.89271	0.88034	0.87914	0.89043	0.94964	0.95799	0.94034
		St.Sapma	0.16679	0.06660	0.04683	0.03286	0.06384	0.09115	0.09188	0.06427	0.06105	0.04128	0.05741
	%10	Ortalama	0.34164	0.91272	0.95528	0.97200	0.79286	0.76569	0.76430	0.79159	0.94247	0.91584	0.89756
		St.Sapma	0.15047	0.07291	0.04951	0.03189	0.08421	0.12593	0.12630	0.08418	0.06222	0.05433	0.04803
250	%5	Ortalama	0.61238	0.94937	0.97602	0.98424	0.90418	0.89786	0.89675	0.90284	0.97442	0.96098	0.94723
		St.Sapma	0.13923	0.04739	0.02511	0.01820	0.03753	0.05774	0.05822	0.03787	0.02594	0.02464	0.03343
	%10	Ortalama	0.25311	0.92396	0.97035	0.97939	0.79847	0.77215	0.77027	0.79781	0.96351	0.90008	0.91478
		St.Sapma	0.09333	0.05299	0.02679	0.02068	0.05187	0.07955	0.08034	0.05204	0.03016	0.03332	0.03031
1000	%10	Ortalama	0.18783	0.92750	0.97860	0.98371	0.79648	0.76878	0.76701	0.79633	0.96930	0.89927	0.90187
		St.Sapma	0.04053	0.04296	0.01269	0.01072	0.02609	0.03986	0.04016	0.02607	0.01525	0.01561	0.01424
	%20	Ortalama	0.09075	0.44399	0.95661	0.87060	0.58281	0.45857	0.45613	0.58266	0.94399	0.86515	0.89959
		St.Sapma	0.02254	0.14910	0.01492	0.07865	0.03318	0.05546	0.05563	0.03313	0.01729	0.01655	0.01309

Çizelge 3.12. Benzetim Veri Kümesi 3 - Genel Doğruluk Oranı İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.76963	0.96229	0.96669	0.97931	0.94708	0.93863	0.93814	0.94608	0.96361	0.96080	0.96842
		St.Sapma	0.06783	0.03123	0.02663	0.01887	0.03007	0.03846	0.03866	0.03041	0.03206	0.02267	0.02864
	%10	Ortalama	0.67736	0.95434	0.96794	0.97914	0.90472	0.89447	0.89394	0.90419	0.96282	0.93968	0.94852
		St.Sapma	0.05964	0.03305	0.02612	0.01810	0.03868	0.05247	0.05255	0.03860	0.03065	0.03063	0.02460
250	%5	Ortalama	0.78756	0.97142	0.98060	0.98810	0.95380	0.95036	0.94986	0.95319	0.98229	0.96921	0.97324
		St.Sapma	0.04825	0.02049	0.01381	0.00943	0.01786	0.02547	0.02565	0.01807	0.01301	0.01399	0.01665
	%10	Ortalama	0.67862	0.96286	0.98118	0.98667	0.90867	0.89845	0.89771	0.90839	0.97942	0.94065	0.95829
		St.Sapma	0.03599	0.02383	0.01308	0.00996	0.02365	0.03382	0.03411	0.02371	0.01394	0.01801	0.01486
1000	%10	Ortalama	0.67279	0.96514	0.98854	0.99133	0.90743	0.89629	0.89558	0.90737	0.98465	0.94655	0.95287
		St.Sapma	0.01675	0.01993	0.00618	0.00516	0.01232	0.01734	0.01744	0.01232	0.00736	0.00798	0.00701
	%20	Ortalama	0.64695	0.78462	0.97867	0.94023	0.82707	0.78679	0.78603	0.82702	0.97269	0.93102	0.95185
		St.Sapma	0.01599	0.04835	0.00728	0.03463	0.01450	0.02006	0.02009	0.01448	0.00841	0.00839	0.00648

Çizelge 3.13. Benzetim Veri Kümesi 3 - F1-Skoru İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.74547	0.96170	0.96679	0.97935	0.94459	0.93471	0.93413	0.94346	0.96335	0.96116	0.96738
		St.Sapma	0.08318	0.03128	0.02598	0.01857	0.03146	0.04250	0.04279	0.03179	0.03245	0.02232	0.02940
	%10	Ortalama	0.59637	0.95265	0.96767	0.97914	0.89526	0.88095	0.88025	0.89462	0.96207	0.93895	0.94661
		St.Sapma	0.07338	0.03416	0.02624	0.01777	0.04164	0.06124	0.06142	0.04163	0.03176	0.03034	0.02467
250	%5	Ortalama	0.75541	0.97056	0.98051	0.98804	0.95168	0.94750	0.94694	0.95102	0.98213	0.96905	0.97261
		St.Sapma	0.06632	0.02142	0.01382	0.00947	0.01860	0.02768	0.02790	0.01879	0.01315	0.01400	0.01685
	%10	Ortalama	0.57473	0.96093	0.98091	0.98653	0.89918	0.88590	0.88497	0.89885	0.97898	0.93838	0.95647
		St.Sapma	0.04489	0.02541	0.01341	0.01015	0.02590	0.03953	0.03993	0.02598	0.01441	0.01816	0.01526
1000	%10	Ortalama	0.55372	0.96356	0.98845	0.99128	0.89824	0.88439	0.88350	0.89817	0.98444	0.94425	0.95069
		St.Sapma	0.01927	0.02122	0.00623	0.00518	0.01304	0.01993	0.02008	0.01304	0.00747	0.00830	0.00715
	%20	Ortalama	0.50246	0.72200	0.97826	0.93528	0.79140	0.72929	0.72806	0.79133	0.97199	0.92695	0.94958
		St.Sapma	0.01250	0.07455	0.00744	0.03931	0.01659	0.02773	0.02781	0.01657	0.00864	0.00854	0.00653

4) Benzetim Veri Kümesi 4 için Sonuçlar:

Benzetim verisi 4 ile sınıflandırıcıya iki grupta da uzak iki yönlü yanlış sınıflandırılmanın olduğu durumlar için, Çizelge 3.14. ve Çizelge 3.15.'te duyarlılık ve belirlilik değerlerine göre algoritmaların başarıları verilmiştir. Test kümesindeki duyarlılık ve belirlilik değerlerinin 1000 tekrar için ortalama ve standart sapma değerleri göz önüne alınmıştır. RLDA-OGK, lojistik regresyon benzetim kümesi veri 4 için eğitim gözlemleri olduğunda, sonuçlar duyarlılık için en iyi olduklarını gösteriyor. Genel tabloda RSIMCA, RLDA, RQDA, tanjantboost ve gudermannianboost algoritmaları duyarlılık skorunda benzer sonuçlar vermişlerdir. Çok yaklaşık değerlerle tanjantboost ve gudermannianboost algoritmalarında da yanlış sınıflama oranının ve n'in artması ile lojistik algoritmalarla aynı başarı düzeyine yaklaştıkları görülüyor. Ayrıca gözlem sayısındaki artış da başarıyı yükseltmiştir. Ayrıca gözlem sayısı ve yanlış sınıflama oranlarındaki artışın benzetim verisi 4 için RLDA ve RQDA algoritmaların başarısında yükseltici bir etki yarattığı söylenebilir. Aynı veri kümesi için belirlilik ölçüm tablosu incelendiğinde Çizelge 3.15.'te gözlem sayısı arttıkça belirlilik değerlerindeki artış yanlış sınıflandırma oranı olan eğitim kümesindeki aykırı değerler için yine düşme eğilimi göstermiştir. Sağlam yöntemlerin, klasik yöntemlere göre daha iyi olduğu görülmüş ancak RLDA, RLDA-OGK, RQDA, özellikle RSIMCA, tanjantboost ve gudermannianboost algoritmalarına göre daha yüksek belirlilik skoruna sahip olduğu görülmüştür. RLDA-OGK en yüksek belirlilik skorunda en başarılı algoritma olduğu gözlenirken, RLDA ve RQDA belirlilik skorunda RLDA-OGK ile aynı başarı düzeylerinde olduğu söylenebilir. Ayrıca tanjantboost ve gudermannianboost başarıları lojistik regresyon ile aynı düzeydedir.

Çizelge 3.16. Genel doğruluk oranları incelendiğinde özellikle RLDA-OGK benzetim verisi 4 için oldukça başarılı hesaplamalara sahiptir. Ancak gözlem sayısının ve yanlış sınıflandırma oranının yükseldiği durumlarda RLDA-OGK'nın başarısına, RLDA ve RQDA algoritmaları çok yaklaştığı görülmektedir. TS-DVM algoritması bu veri kümesinde de yine en başarısız sınıflandırma becerisine sahipken lojistik regresyon, tanjantboost ve gudermannianboost algoritmaları özellikle gözlem sayısının ve yanlış sınıflandırma oranının en yüksek olduğu durumda benzer doğrulama oranlarına sahiptir. RLDA-OGK, RLDA ve RQDA hariç yanlış sınıflandırma oranının arttığı durumlarda tüm algoritmalarda genel doğruluk oranında düşmeler görülmüştür.

Çizelge 3.17.'de F1-skoruna göre tek taraftan yanlış sınıflandırmalı da başarısı düşen RLDA-OGK iki taraflıda başarısını yükseltmiştir. En başarılı algoritmalar, RQDA, RLDA-OGK ve RLDA olarak tespit edilmiş ve artan gözlem değerlerine ve yanlış sınıflandırma oranlarına rağmen başarıları yükselmiştir. Lojistik Regresyon, RSIMCA, Tanjantboost ve Gudermannianboost algoritmaları birbirlerine yakın ve başarılı sonuçlar vermişlerdir.

Çizelge 3.14. Benzetim Veri Kümesi 4 - Duyarlılık İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.82745	0.96706	0.96748	0.97995	0.97452	0.95959	0.95909	0.97352	0.96580	0.96341	0.99505
		St.Sapma	0.07632	0.04794	0.04324	0.02845	0.03336	0.05530	0.05584	0.03426	0.04638	0.03678	0.01314
	%10	Ortalama	0.84411	0.96534	0.97147	0.98068	0.95655	0.94005	0.93984	0.95576	0.96862	0.95985	0.93698
		St.Sapma	0.07516	0.04892	0.03888	0.02839	0.05000	0.07505	0.07542	0.05029	0.04365	0.03919	0.05831
250	%5	Ortalama	0.88886	0.97446	0.98029	0.98776	0.98080	0.97044	0.97026	0.98009	0.98247	0.97524	0.99308
		St.Sapma	0.04127	0.03515	0.02323	0.01547	0.02079	0.03523	0.03549	0.02127	0.02059	0.01876	0.01885
	%10	Ortalama	0.89602	0.97526	0.98460	0.98900	0.97282	0.96150	0.96127	0.97271	0.98435	0.95382	0.94353
		St.Sapma	0.03935	0.03480	0.01992	0.01463	0.02835	0.04705	0.04739	0.02864	0.01955	0.02744	0.03443
1000	%10	Ortalama	0.92116	0.98260	0.99283	0.99462	0.98694	0.98146	0.98139	0.98685	0.99271	0.94750	0.95596
		St.Sapma	0.01740	0.02650	0.00842	0.00671	0.01337	0.02247	0.02257	0.01345	0.00863	0.01352	0.03698
	%20	Ortalama	0.92032	0.98125	0.99375	0.99445	0.97785	0.96964	0.96958	0.97785	0.99288	0.90011	0.90109
		St.Sapma	0.01718	0.02684	0.00764	0.00668	0.02180	0.03584	0.03595	0.02180	0.00854	0.01619	0.01462

Çizelge 3.15. Benzetim Veri Kümesi 4 - Belirlilik İstatistikleri

n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.92570	0.96334	0.96681	0.97816	0.97121	0.95684	0.95651	0.97010	0.96275	0.96255	0.99561
		St.Sapma	0.06583	0.05055	0.04296	0.03050	0.03530	0.05578	0.05622	0.03581	0.05317	0.03653	0.01245
	%10	Ortalama	0.65164	0.96813	0.96979	0.97940	0.95722	0.94027	0.93996	0.95654	0.96764	0.96048	0.93614
		St.Sapma	0.15349	0.04729	0.04033	0.02992	0.04965	0.07436	0.07490	0.05010	0.04339	0.03957	0.05857
250	%5	Ortalama	0.90105	0.97733	0.98153	0.98897	0.98284	0.97295	0.97273	0.98251	0.98395	0.97700	0.99327
		St.Sapma	0.05394	0.03251	0.02262	0.01468	0.01981	0.03581	0.03603	0.02009	0.02010	0.01834	0.01787
	%10	Ortalama	0.56350	0.97737	0.98344	0.98800	0.97254	0.96083	0.96065	0.97226	0.98375	0.95656	0.94304
		St.Sapma	0.14408	0.03266	0.02001	0.01566	0.02970	0.04891	0.04922	0.03004	0.01960	0.02463	0.03451
1000	%10	Ortalama	0.56002	0.98016	0.99209	0.99409	0.98625	0.98055	0.98046	0.98613	0.99188	0.94647	0.95803
		St.Sapma	0.09077	0.02737	0.00867	0.00693	0.01363	0.02332	0.02346	0.01367	0.00916	0.01328	0.03602
	%20	Ortalama	0.18882	0.98186	0.99322	0.99386	0.97726	0.96880	0.96873	0.97726	0.99216	0.90024	0.90195
		St.Sapma	0.03979	0.02652	0.00816	0.00714	0.02227	0.03738	0.03748	0.02227	0.00913	0.01658	0.01452

Çizelge 3.16. Benzetim Veri Kümesi 4 - Genel Doğruluk Oranı İstatistikleri

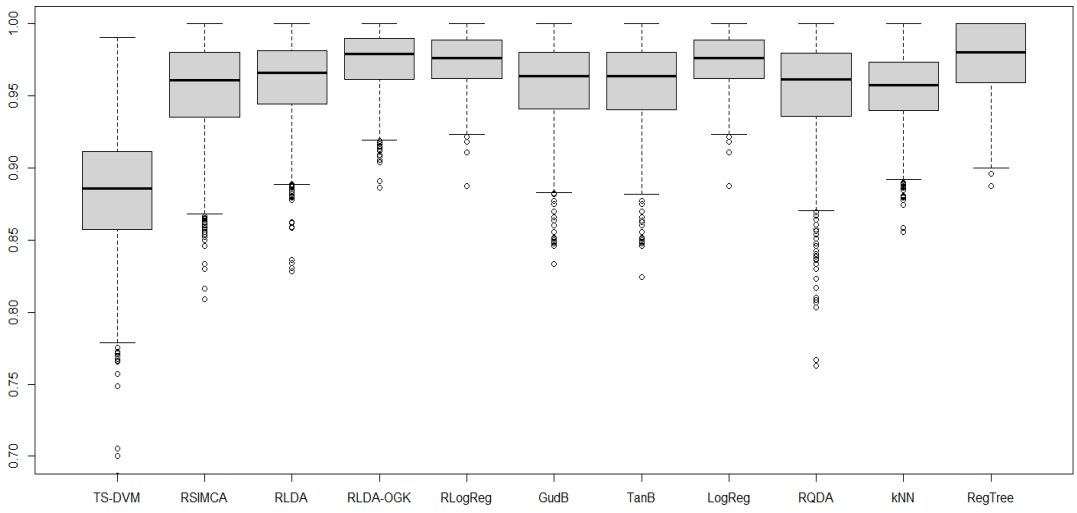
n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.86864	0.96482	0.96675	0.97890	0.97261	0.95745	0.95701	0.97153	0.96404	0.96263	0.99527
		St.Sapma	0.05262	0.02726	0.02687	0.01876	0.02241	0.03214	0.03239	0.02306	0.03178	0.02236	0.00853
	%10	Ortalama	0.76968	0.96615	0.97022	0.97984	0.95637	0.93915	0.93888	0.95563	0.96765	0.95967	0.93587
		St.Sapma	0.06491	0.02715	0.02491	0.01849	0.03286	0.04116	0.04143	0.03294	0.02821	0.02413	0.04180
250	%5	Ortalama	0.89412	0.97570	0.98078	0.98830	0.98173	0.97153	0.97133	0.98119	0.98315	0.97597	0.99316
		St.Sapma	0.03156	0.01798	0.01463	0.00921	0.01291	0.01892	0.01904	0.01317	0.01193	0.01160	0.01283
	%10	Ortalama	0.77066	0.97615	0.98394	0.98841	0.97259	0.96106	0.96085	0.97240	0.98396	0.95495	0.94312
		St.Sapma	0.04921	0.01788	0.01250	0.00919	0.01822	0.02501	0.02516	0.01838	0.01160	0.01727	0.02490
1000	%10	Ortalama	0.78071	0.98134	0.99245	0.99434	0.98656	0.98095	0.98087	0.98646	0.99228	0.94687	0.95683
		St.Sapma	0.02959	0.01397	0.00527	0.00390	0.00828	0.01172	0.01177	0.00833	0.00498	0.00919	0.02564
	%20	Ortalama	0.67403	0.98152	0.99347	0.99414	0.97755	0.96919	0.96913	0.97755	0.99251	0.90014	0.90145
		St.Sapma	0.01629	0.01378	0.00433	0.00380	0.01280	0.01805	0.01809	0.01280	0.00474	0.01114	0.01054

Çizelge 3.17. Benzetim Veri Kümesi 4 - F1-Skoru İstatistikleri

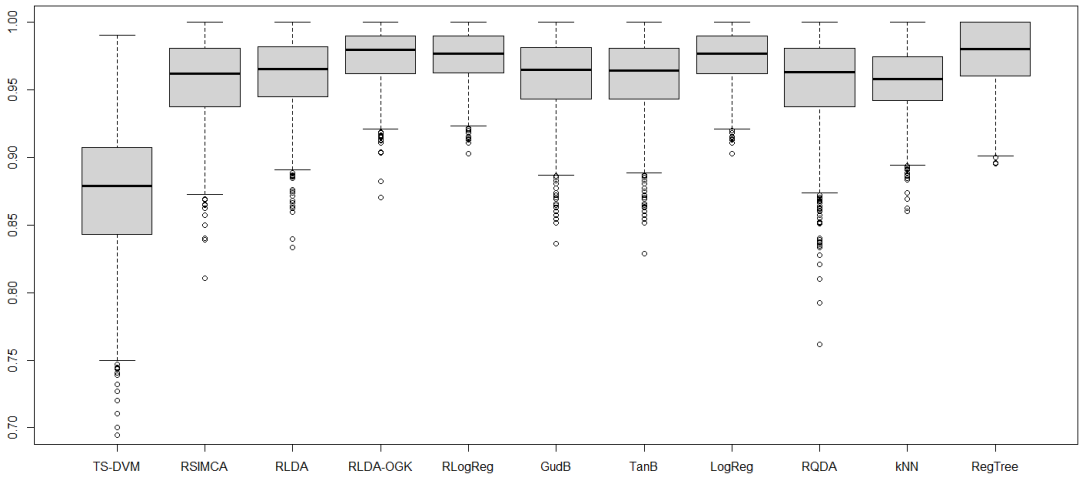
n	YSO(%)		TS-DVM	RSIMCA	RLDA	RLDA-OGK	RLogReg	GudB	TanB	LogReg	RQDA	KNN	RegTree
100	%5	Ortalama	0.87658	0.96520	0.96714	0.97905	0.97287	0.95821	0.95780	0.97181	0.96427	0.96298	0.99533
		St.Sapma	0.04669	0.02636	0.02600	0.01863	0.02197	0.03074	0.03095	0.02258	0.03153	0.02166	0.00843
	%10	Ortalama	0.74788	0.96673	0.97063	0.98004	0.95688	0.94016	0.93990	0.95615	0.96813	0.96017	0.93656
		St.Sapma	0.07698	0.02591	0.02425	0.01809	0.03197	0.03924	0.03946	0.03207	0.02693	0.02360	0.04048
250	%5	Ortalama	0.89496	0.97590	0.98091	0.98836	0.98182	0.97169	0.97149	0.98130	0.98321	0.97612	0.99317
		St.Sapma	0.03141	0.01745	0.01437	0.00912	0.01270	0.01862	0.01872	0.01296	0.01184	0.01142	0.01273
	%10	Ortalama	0.72976	0.97631	0.98402	0.98850	0.97268	0.96117	0.96096	0.97248	0.98405	0.95519	0.94329
		St.Sapma	0.06965	0.01743	0.01228	0.00901	0.01801	0.02472	0.02486	0.01819	0.01135	0.01661	0.02433
1000	%10	Ortalama	0.74059	0.98138	0.99246	0.99435	0.98659	0.98100	0.98092	0.98649	0.99229	0.94698	0.95700
		St.Sapma	0.04329	0.01389	0.00526	0.00389	0.00827	0.01166	0.01171	0.00831	0.00497	0.00901	0.02551
	%20	Ortalama	0.55457	0.98156	0.99349	0.99415	0.97756	0.96922	0.96915	0.97756	0.99252	0.90017	0.90152
		St.Sapma	0.01892	0.01375	0.00430	0.00377	0.01276	0.01790	0.01794	0.01276	0.00471	0.01081	0.01017

3.1.2.Uygulamada F1-Skor Değerlerinin Yorumu ve Grafiksel Gösterimleri

1000 tekrar sonucunda elde edilen ortalama ve standart sapma ile ilgili bilgiler yorumlandığı gibi tüm elde edilen doğru sınıflandırma oranı ve F1 skorları için bazı benzetim veri kümelerinin bazı durumları için boxplot grafikleri çizdirilmiştir. Bu durumlarla algoritmaların genel doğru sınıflandırma oranı ve F1-skorları değerlerine olan etkisi ve yine algoritmaların birbirlerine göre farkları yorumlanmaya çalışılmıştır. Regresyon ve sınıflandırma ağacı, lojistik regresyon, sağlam doğrusal ayrıştırıcı – OGK kestirici ve sağlam lojistik regresyon yöntemlerinde ortanca olarak yüksek başarı olduğu gibi doğru sınıflandırma oranı yayılımında daha az olduğu görülmüştür. Aynı durumun benzetim verisi 1 için elde edilen F1-skorlarında da elde edilmiştir. Yöntemlerde doğru sınıflandırma oranı ve F1-skor değerlerinin sonuçlarının benzer olması gözlemlerin eşit sayıda sınıflara ayrılmış olmasıdır. Bu nedenle bundan sonraki grafiklerde sadece F1-skor değerlerinin sonuçları verilmiştir. Benzetim Verisi 1; 100 gözlem ve %5 yanlış sınıflandırmaya sahip Eğitim kümesinde 1000 tekrar sonucu elde edilen test kümesi dengelenmiş doğru sınıflandırma oranları Boxplot grafikleri Şekil 3.3. ve F1-Skor grafikleri 3.4.'te gösterilmiştir.

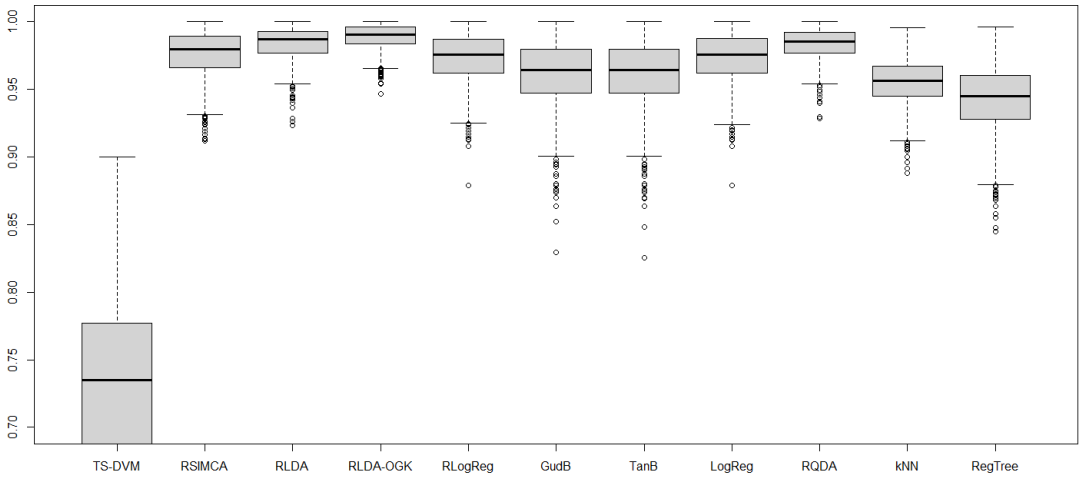


Şekil 3.3. Benzetim Verisi 1; Doğru Sınıflandırma Oranları Boxplot Grafikleri



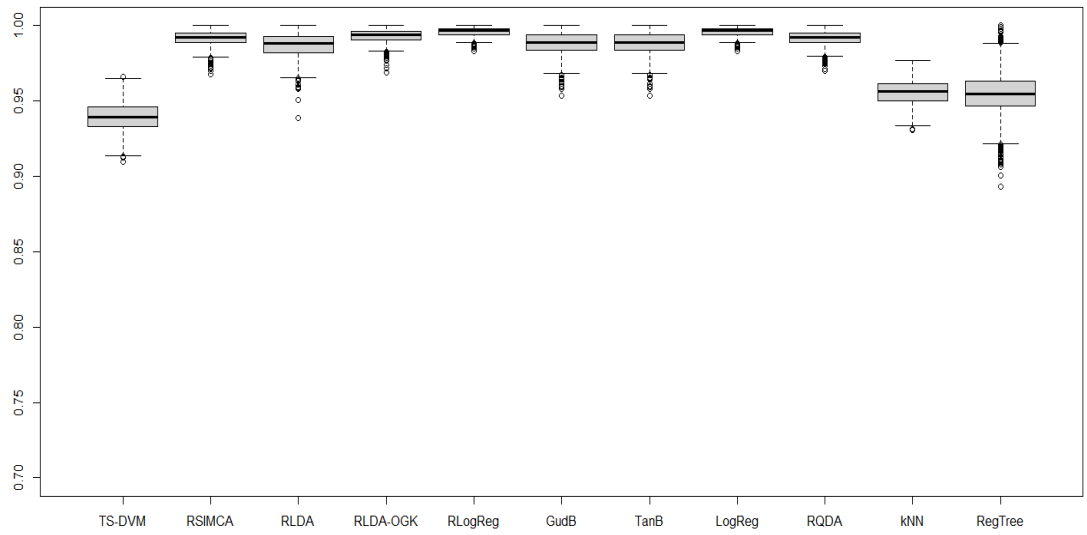
Şekil 3.4. Benzetim Verisi 1; F1-Skoru Boxplot Grafikleri

Benzetim verisi 2 için gözlem sayısının biraz artmış olması ve yanlış sınıflandırılmış gözlem oranındaki artış da birlikte incelendiğinde iki tarafta da yanlış sınıflandırma bulunma durumunun yöntemleri çok etkilemediği görülmüştür. Elde edilen değerlerde en olumsuz etkiyi TS-DVM'nin, sağlam lojistik regresyon ile regresyon ve sınıflandırma ağacı algoritmasının daha başarısız F1-skorları elde ettiği görülmüştür. Sağlam doğrusal ayrıştırıcı ve sağlam karesel ayrıştırıcı algoritmalarının performanslarında diğer veri kümeleri, gözlem sayısı ve yanlış sınıflandırılmış gözlem değerleri performanslarından daha iyi sonuçlar elde edilmiştir. Benzetim Verisi 2 için, 1000 gözlem ve %20 yanlış sınıflandırmaya sahip Eğitim kümesinde 1000 tekrar sonucu elde edilen test kümesi F1-skorları Boxplot grafikleri Şekil 3.5.'te verilmiştir.



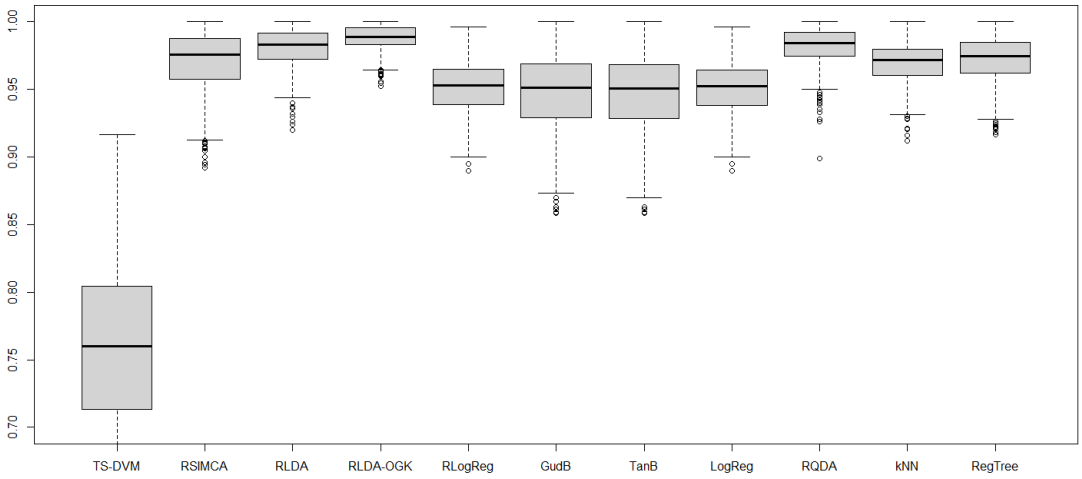
Şekil 3.5. Benzetim Verisi 2; F1-skorları Boxplot Grafikleri

Benzetim verisi 3 için 250 eğitim veri kümeli gözlem sayısı ve %5 yanlış sınıflandırılmış gözlem oranı durumunda elde edilen bilgiler Şekil 3.6.'da verilmiştir. Yöntemlerin en başarılı olduğu benzetim veri kümesidir ve TS-DVM, KNN, RegTree algoritmalarının haricinde başarılı sonuçlar elde edilmiştir. F1-skorları incelendiğinde lojistik ve sağlam lojistik regresyonun başarısı göze çarparken RSIMCA'da da başarılı sonuçlar elde edilmiştir. Benzetim veri kümelerinin yanlış sınıflandırılmış gözlem değerlerinin olması ve sınıflandırıcının doğrusal olmasının lojistik regresyon ve doğrusal ayrıştırıcı gibi algoritmalarının tahmin değerlerini etkilemediği bu şekilde de görülmüştür. Benzetim Verisi 3 250 gözlem ve %5 yanlış sınıflandırmaya sahip eğitim kümesinde 1000 tekrar sonucu elde edilen test kümesi F1-skorları Boxplot grafikleri Şekil 3.6.' da yer almaktadır.



Şekil 3.6. Benzetim Verisi 3; F1-Skorları Boxplot Grafikleri

Benzetim veri kümesi 4 için elde edilen örnek saçılımda ise doğrusal ayrıştırıcıdan sağlam olan ve aynı zamanda sağlam OGK kestirimine sahip olan RLDA-OGK algoritmasının oldukça başarılı olduğu gözlenmiştir. Benzetim Verisi 4 250 gözlem ve %10 yanlış sınıflandırmaya sahip Eğitim kümesinde 1000 tekrar sonucu elde edilen test kümesi F1-skorları Boxplot grafikleri Şekil 3.7.'de verilmiştir.



Şekil 3.7. Benzetim Verisi 4; F1-Skorları Boxplot Grafikleri

3.2. Aykırı Değerler Varlığında Sınıflandırma için Gerçek Veri Uygulamaları

Benzetim veri kümelerinin haricinde gerçek veri kümeleri üzerinde de algoritmalar karşılaştırılmıştır. Benzetim veri kümesi ile gerçek veri kümelerindeki farklılık, benzetimde sadece yanlış sınıflandırmalar ve kesin sınıflandırmalar bulunmaktadır. Ancak gerçek veri kümesinde yanlış sınıflandırmalar, aykırı değerler vb. tüm gerçek yapı olduğu için sağlam yöntemlerin biraz daha ön plana çıkacağı öngörülebilir. Ancak bu durumu açıklayabilmek adına sağlam sınıflandırma yöntemlerinde daha önce kullanılan bazı gerçek veri kümelerine yer verilmiştir.

Wdbc veri kümesi, meme kanseri tanısı için 569 gözlemde 32 değişkende kötü huylu/ iyi huylu tümör taramasının olduğu bir veri kümesidir [51]. 569 gözlemin 300'ü eğitim veri kümesine rasgele alınarak eğitim gerçekleştirilmiş ve geride kalan tüm gözlemler test veri kümesinde kullanılmıştır. Verilerde kötü huy ve iyi huy dağılımının dengesiz olmasından dolayı F1-skoru karşılaştırma yapılabilecek en iyi ölçüt olarak görülebilir. Bu nedenle F1-skor değerleri incelendiğinde sırasıyla GudermannianBoost, sağlam lojistik regresyon, lojistik regresyon ve TanjantBoost algoritmalarının başarılı sonuçlar verdiği görülmüştür. Elde edilen sonuçlar Çizelge 3.18.'de duyarlılık, belirlilik, dengelenmiş genel doğruluk oranı ve F1-skoru ile verilmiştir.

Çizelge 3.18. Wdbc Veri Kümesi Uygulama Sonucu Algoritma Başarıları

Yöntemler	**Duyarlılık**	**Belirlilik**	**Dengelenmiş Genel Doğruluk Oranı**	**F1-Skoru**
TS-DVM	0.12048	0.81553	0.46801	0.19512
RSIMCA	0.74699	0.93204	0.83951	0.83501
RLDA	**0.97591**	0.70874	0.84230	0.90503
RLDA-OGK	**0.99397**	0.72815	0.86106	0.91922
RLogReg	**0.96592**	**0.95132**	**0.95355**	**0.97058**
GudB	**0.97590**	**0.95145**	**0.96368**	**0.97207**
TanB	**0.98193**	0.89320	0.93756	**0.95882**
LogReg	0.95783	0.94174	0.94979	**0.96072**
RQDA	0.92771	0.94175	0.93473	0.94479
KNN	0.90361	0.35923	0.63141	0.78534
RegTree	0.95181	0.89232	0.92251	0.94329

Duyarlılık, belirlilik, dengelenmiş genel doğruluk oranı ve F1-Skoru incelendiğinde 4 değer için %95 in üzerinde kalarak, sağlam lojistik regresyon ve gudermannianboost algoritmaları genel olarak en başarılı algoritmalar olmuşlardır. Sağlam boosting algoritma olan GudermannianBoost ve TanjantBoost algoritmaları, benzetim veri kümesinde sadece yanlış sınıflandırmanın vermiş olduğu etkiyi yeterince ayıklayamamış olmasına rağmen gerçek veri kümesinde oldukça başarılı sınıflandırma sağlamıştır. En başarısız algoritma, duyarlılık, doğruluk oranı ve F1-skoru için TS-DVM; belirlilik için KNN olmuştur. Lojistik regresyon ve RQDA bu algoritmalara yakın bir sonuç vermiştir. Ayrıca duyarlılık ve F1-skoru için incelendiğinde sağlamlaştırılmış algoritmalardan, RLDA ve RLDA-OGK birbirine benzer yüksek başarılar elde etmiştir.

İkinci gerçek veri kümesi üzerinde yapılan analizler ise parkinsons veri kümesi kullanılarak karşılaştırılmıştır [52]. Veri kümesi 24 değişken ve 195 gözlemden oluşmaktadır. Değişkenlere göre hastaların Parkinson tanısı alıp almadıkları sınıflandırılmaya çalışılmıştır. 100 gözlem rasgele alınarak eğitim kümesi olarak kullanılırken geriye kalan 95 gözlem test kümesi olarak kullanılmıştır. Elde edilen sonuçlar aşağıdaki Çizelge 3.19.'da verilmiştir. Sağlam karesel ayrıştırıcı ve OGK kestirimine sahip sağlam doğrusal ayrıştırıcı da dahil olmak üzere bir önceki gerçek veri

kümesindeki sonuçlardaki yöntemler yine başarı anlamında birbirleriyle karşılaştırılabilir sonuçlar elde etmiştir.

Çizelge 3.19. Parkinsons Veri Kümesi için Algoritma Başarıları

Yöntemler	**Duyarlılık**	**Belirlilik**	**Dengelenmiş Genel Doğruluk Oranı**	**F1-Skoru**
TS-DVM	0.70588	0.29629	0.50109	0.71111
RSIMCA	0.63235	0.51852	0.57544	0.69354
RLDA	**0.95588**	0.18518	0.57053	0.83873
RLDA-OGK	0.89705	0.74074	**0.81889**	**0.89706**
RLogReg	0.90248	0.73127	**0.81266**	**0.89704**
GudB	**0.95588**	0.5556	0.75572	**0.89655**
TanB	**1.00000**	0.48148	0.74074	**0.90667**
LogReg	**0.89706**	0.74074	**0.81889**	**0.89700**
RQDA	**0.97098**	0.59296	0.78159	**0.91034**
KNN	**0.97060**	0.29629	0.63344	0.86275
RegTree	0.85294	0.59259	0.72277	0.84691

Sonuçlar her bir skor değerinde farklı algoritmaları öne çıkarıyor. En başarısız sonuç veren algoritma, TS-DVM algoritmasına aittir. En başarılı sonuçlar, duyarlılık için tanjantboost, belirlilik ve genel doğruluk skorları için, RLDA-OGK ve lojistik regresyondur. F1-Skoru incelendiğinde 0,9 değeri üzerinde kalan skorlarla, tanjantboost ve RQDA algoritmaları en başarılı sonucu vermişlerdir.

4. SONUÇ

Sınıflandırma algoritmalarının ikili sınıflandırma alanındaki başarıları üzerinde değerlendirmeler yapılmıştır. Karşılaştırılan yöntemler; TS-DVM, RSIMCA, RLDA, RLDA-OGK, sağlam lojistik regresyon, sağlam boosting algoritma olan GudermannianBoost ve TanjantBoost algoritmaları, RQDA, KNN, regresyon ve sınıflandırma ağacı algoritmasıdır. Bu on bir algoritma için R 4.5.0 programı kullanılmıştır. Analizler için dört farklı benzetim veri kümesi ve iki farklı gerçek veri kümesi kullanılmıştır.

Çalışmanın başında %5, %10 ve %20 yanlış sınıflandırma oranları ile oluşturulan 4 farklı benzetim veri kümesi kullanılmıştır. Kullanılan algoritmalar, duyarlılık, belirlilik, genel doğruluk oranı ve F-skoru sonuçlarına göre değerlendirilmiştir. Duyarlılık ve kesinlik değerlerinin harmonik ortalaması kullanılarak hesaplanan F1-skoru, bu yönüyle daha hassas bir eşik değeridir. Bu nedenle de karşılaştırma yorumlarında diğer eşik değerlerine göre daha etkilidir. TS-DVM başarısı diğer algoritmalara göre düşük kalmıştır. Burada TS-DVM'nin eğitim kümesindeki aykırı değerleri tespit etme özelliğinden ötürü, sınıflandırma anlamında yanlılığı görülmüştür.

İlkinde sınıflandırıcıya yakın tek taraflı yanlış sınıflandırılan verilerin varlığında yapılan analizde, TS-DVM başarısı diğer algoritmalara göre düşük, diğer algoritmalar birbirlerine yakın başarı değerleri elde etmiştir.

İkinci benzetim veri kümesinde sınıflandırıcıya yakın iki taraflı yine %5, %10 ve %20 oranlarla yanlış sınıflandırılmış veriler varlığında analizler yapılmıştır. Yanlış sınıflandırma oranının ve n sayısının arttığı durumlarda tüm algoritmalarda genel doğruluk oranında önemsiz düşmeler görülmüştür. F1 skoru da doğruluk oranlarına yakın sonuçlar vermiştir. RSIMCA, RQDA, tanjantboost ve gudermannianboost benzer, RLDA-OGK ve lojistik regresyon çok az bir farkla daha başarılı bulunmuştur. Üçüncü durumda sınıflandırıcıya tek taraflı uzaktan yanlış sınıflandırmalı veriler varlığında, F1-skoru için RQDA, RLDA artan gözlem değerlerine ve yanlış sınıflandırma oranlarına rağmen başarılarını korumuşlardır. Ancak diğer algoritmaların başarılarında ciddi düşüşler gözlenmiştir.

Wdbc veri kümesi, meme kanseri tanısı için yapılan analiz sonucunda, en başarısız algoritma, duyarlılık, doğruluk oranı ve F1-skoru için TS-DVM; belirlilik için KNN olmuştur. Lojistik regresyon ve RQDA bu algoritmalara yakın bir sonuç vermiştir. Ayrıca duyarlılık ve F1-skoru için incelendiğinde sağlamlaştırılmış algoritmalardan, RLDA ve RLDA-OGK birbirine benzer yüksek başarılar elde etmiştir.

Parkinson hastalığı teşhisi ile ilgili verilerin sınıflandırılmasında F1-Skoru incelendiğinde tanjantboost ve RQDA algoritmaları benzer başarıları ile birbirlerine alternatif olabilirler. En başarısız sonuç gerçek verilerde de TS-DVM algoritmasına aittir.

Sağlam lojistik regresyon, RSIMCA, RLDA, RLDA.OGK, RQDA algoritmaların çeşitli aykırı değer içeren veri setlerinde, çoğu zaman benzer başarıları elde ettiğini söyleyebiliriz. Aynı zamanda kayıp fonksiyonlardan geliştirilmiş tanjantboost ve gudermannianboost algoritmalarının da benzer doğrulama oranlarına sahip olduğu gözlenmiştir.

Bu çalışmada, sağlam yöntemlerin yanlış etiketleme, aykırı değere sahip eğitim kümesi gibi problemleri bulunan veri küemlerinde aykırı değerin etkisini azaltarak ortaya koymaya çalıştıkları öğrenme süreci, benzetim kümeleri ve gerçek hayat uygulamalarıyla incelenmiştir. Elde edilen sonuçlar, sınıflandırıcıya yakın ve yanlış etiketlemeye sahip gözlemlerin bulunduğu eğitim veri kümelerinde kayıp fonksiyonları sağlamlaştırılmış boosting algoritmaların başarılı olduğunu gösterirken, veri kümesinde aykırı değerler bulunması durumunda klasik yöntemlerin sağlamlaştırılmış olan alternatiflerinin daha başarılı olduğu gözlenmiştir. Çalışmada özellikle gerçek veri kümelerinde elde edilen sonuçlar tartışılırken gözlenen ve sonraki çalışmalar için ilham olabilecek durum ise dengesiz (imbalanced) veri sınıf bilgilerine ve aynı zamanda aykırı değerlere sahip olan veri kümelerindeki öğrenme süreçlerinin ve algoritmaların başarılarının incelenmesidir.

5. KAYNAKLAR

[1] P. Xanthopoulos, P. M. Pardalos ve T. B. Trafalis, Robust Data Mining, Springer New York Heidelberg Dordrecht, London, **2013**.

[2] J. Quinlan, Induction of Decision Trees, Machine Learning, Springer, 81-106, **1986**.

[3] R. Patton, J. Ogle ve T. Laird, SQL Server Analysis Services, Designing SQL Server 2000 Databases, Syngress, **2001**.

[4] M. Sugiyama, Statistical Reinforcement Learning: Modern Machine Learning, Tokyo, CRC Press Taylor and Francis Group, **2015**.

[5] D. Michie, D. J. Spiegelhalter, C. C. Taylor, Machine Learning, Neural and Statistical Classification, Metadata Version, 1 **(1994)** 10.

[6] X. Dong, Z. Yu, W. Cao, Y. Shi ve Q. Ma, A survey on ensemble learning, Front. Comput. Sci., cilt 14, no. 2, 241-258, **2020**.

[7] R. J. Henery, D. Michie, D. J. Spiegelhalter ve C. Taylor, Machine Learning, Neural and Statistical Classification, **1994**.

[8] O. Ibitoye, R. A. Khamis, A. Matrawy ve M. O. Shafiq, The Threat of Adversarial Attacks on Machine Learning in Network, Ottowa, **2019**.

[9] L. Kaelbling, M. Littman ve A. Moore, Reinforcement Learning: A Survey, Journal of Articial Intelligence Research 4, no. 4, 237-285, **1996**.

[10] A. Aldahiri, B. Alrashed ve W. Hussain, Trends in Using IoT with Machine Learning in Health Prediction System, Forecasting, cilt 3, 181-207, **2021**.

[11] M. Pelillo, Alhazen and the nearest neighbor rule, Pattern Recognition Letters, cilt 38, 34-37, **2014**.

[12] R. Nisbet, G. Miner ve K. Yale, Handbook of Statistical Analysis and Data Mining Applications, London: Academic Press, **2018**.

[13] M. Kowsher, A. Tahabilder ve S. Murad, Impact-Learning: A Robust Machine Learning Algorithm, ICCCM'20, Singapore, **2020**.

[14] G. Aksu ve N. Doğan, Veri Madenciliğinde Kullanılan Öğrenme Yöntemlerinin Farklı Koşullar Altında Karşılaştırılması, Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, cilt 51, no. 3, 71-100, **2018**.

[15] X. Wang, Y. Chen ve X. L. Wang, A Centroid-Based Outlier Detection Method, Proc. 2017 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, **2017**.

[16] M. A. L. Fernandez, Cross Validation, https://scholar.harvard.edu/files/malf /files/maluque-cross-validation_01.pdf, (Erişim Tarihi: **20 Nisan 2022**).

[17] M. Pavithra, P. P. Kumar, P. Divya, P. Manjubala ve S. Jayalakshmi, The Significance of Learning in Data Analytics: Supervised Learning Techniques, Global Journal of Internet Interventions and IT Fusion, cilt 4, no. 1, **2021**.

[18] C. Pelletier, S. Valero, J. Inglada, G. Dedieu ve N. Champion, New Iterative Learning Strategy To Improve Classificatıon Systems By Usıng Outlier Detection Techniques, IGARSS, Texas, **2017**.

[19] E. Acuña ve C. Rodriguez, On Detection Of Outliers And Their Effect In Supervised Classification, **IPCI**, Venice, **2004**.

[20] H. Sasaki, T. Takenouchi, R. P. Monti ve A. Hyvärinen, Robust contrastive learning and nonlinear ICA in the presence of outliers, Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), **2020**.

[21] J. Feng, H. Xu, S. Mannor ve S. Yan, Robust Logistic Regression and Classification, Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, **2014**.

[22] I. A. I. Ahmed and W. Cheng, The Performance of Robust Methods in Logistic, Open Journal of Statistics, no. 10, 127-138, **2020**.

[23] S. Mete, O. Çakır, O. Bayat, D. G. Duru ve A. D. Duru, Gözbebeği Hareketleri Temelli Duygu Durumu Sınıflandırılması, Bilişim Teknolojileri Dergisi, cilt 13, no. 2, **2020**.

[24] D. Bertsimas, J. Dunn, C. Pawlowski ve Y. D. Zhuo, Robust Classification, INFORMS Journal on Optimization, cilt 1, no. 1, **2019**.

[25] A. Ghosh, N. Manwani ve P. S. Sastry, On the Robustness of Decision Tree Learning under Label Noise, JMLR: Workshop and Conference Proceedings, **2016**.

[26] S. B. Kotsiantis, I. D. Zaharakis ve P. E. Pintelas, Machine learning: a review of classification and combining techniques, Artificial Intelligence Review, no. 26, 159-190, **2006**.

[27] S. Konsiantis, Supervised machine learning a review of classification techniques, Informatica, 249-268, **2007**.

[28] A. Jordan, Representation Schemes Used by Various Classification Techniques – A Comparative Assessment, cilt 12, no. 6, **2015**.

[29] H. Neeb ve C. Kurrus, Distributed K-Nearest Neighbors, https://stanford .edu/~rezab/classes/cme323/S16/projects_reports/neeb_kurrus.pdf. (Erişim tarihi: **9 Mayıs 2021**).

[30] T. M. Cover ve P. E. Hart, Nearest Neighbor Pattern Classification, IEEE Transactıons On Informatıon Theory, cilt 13, no. 1, **1967**.

[31] Y. Wu ve J. Feng, Development and Application of Artificial Neural Network, Wireless Pers Commun, no. 102, 1645-1656, **2019**.

[32] M. Nielsen, Neural Networks and Deep Learning, https://static.latexstudio .net/article/2018/0912/neuralnetworksanddeeplearning.pdf, (Erişim Tarihi: **2 Kasım 2022**).

[33] B. Chandra ve R. K. Sharma, Fast learning in Deep Neural Networks, Neurocomputing, cilt 171, 1205-1215, **2016**.

[34] W. S. Noble, What is a support vector machine?, Nature Biotechnology, no. 24, 1565-1567, **2006**.

[35] M. Rafiei, F. Rafiei, S. M. Tabatabaei, H. AlaviMajd, A. Rafieie ve S. Khodakarim, Valıdatıon Of Classıfıcatıon Models And Data Reductıon Methods Based On Gene Expressıon Data, JP Journal of Biostatistics, cilt 16, no. 2, 79-90, **2019**.

[36] M. Wedler, M. Stender, M. Klein, S. Ehlers ve N. Hoffmann, Surface similarity parameter: A new machine learning loss metric for oscillatory spatio-temporal data, Neural Networks, no. 156, 123-134, **2022**.

[37] O. Toka, Gudermannıan Kayıp Fonksiyonu Ve Gudermannıanboost İkili Sınıflandırma Yöntemi, Ankara, **2016**.

[38] S. Soaad, S. Yahaya, Y.-F. Hazlina ve Z. Omar, Robust Linear Discriminant Analysis, Journal of Mathematics and Statistics, cilt 12, no. 4, 312-316, **2016**.

[39] V. Todorov ve A. M. Pires, Comparative Performance of Several Robust Linear Discriminant Analysis Methods, Revstat - Statistical Journal, cilt 5, no. 1, 63-83, **2007**.

[40] A. Ghosh, R. SahaRay, S. Chakrabarty ve S. Bhadra, Robust generalised quadratic discriminant analysis, Pattern Recognition, no. 117, **2021**.

[41] A. L. Pomerantsev ve O. Y. Rodionova, Concept and role of extreme objects in PCA/SIMCA, Journal of Chemometrics, no. 28, 429-438, **2014**.

[42] K. V. Branden ve M. Hubert, Robust Classification in High Dimensions based on the SIMCA Method, Chemometrics and Intelligent Laboratory Systems, no. 79, 10-21, **2005**.

[43] D. Hendrycks, K. Lee ve M. Mazeika, Using Pre-Training Can Improve Model Robustness and Uncertainty, Proceedings of the 36th International Conference on Machine Learning, CA, **2019**.

[44] C. M. Salgado, C. Azevedo, H. Proença ve S. M. Vieira, Noise Versus Outliers, MIT Critical Data, Secondary Analysis of Electronic Health Records, Cambridge, Springer Open, 163-185, **2016**.

[45] B.Ville, Decision Trees, WIREs Comput Stat, no. 5, 448–455, **2013**.

[46] T. Hastie, R. Tibshirani ve J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer, **2017**.

[47] L. Jiang, Z. Zhou, T. Leung, L. Li ve L. Fei-Fei, MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels, Proceedings of the 35 th International Conference on Machine Learning, Stockholm, **2018**.

[48] B. Han, Q. Yao, X. Yu, G.Niu, M. Xu, W. Hu, I. Tsang ve M. Sugiyama, Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels, 32nd Conference on Neural Information Processing Systems (NIPS), Montreal, **2018**.

[49] J. Brownlee, Imbalanced Classification With Python, ebook Jason Brownlee, **2021**.

[50] R Core Team (2022): R: A Language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, **2022**

[51] Dua, D. and Graff, C., UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic). (Erişim Tarihi: **25 Aralık 2022**).

[52] Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM, Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection, BioMedical Engineering OnLine, https://archive.ics.uci.edu/ml/datasets/parkinsons. (Erişim Tarihi: **24 Aralık 2022**).

[53] X. Zhu, Semi-supervised learning literature survey, Computer Science, University of Wisconsin-Madison, cilt 2, no. 3, 4-4, **2006**.

[54] H. Shirazi ve N. Vasconcelos, On the Design of Loss Functions for Classification: theory, robustness to outliers, and SavageBoost, Advances in Neural Information Processing Systems 21 (NIPS), **2008**.

[55] G. Akgül, A. Çelik, Z. E. Aydın ve Z. K. Öztürk, Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı, Bilişim Teknolojileri Dergisi, cilt 13, no. 3, **2020**.

[56] X. Zhu, Semi-supervised Learning Literature Survey. http://pages.cs.wisc .edu/~jerryzhu/pub/ssl_survey_7_19_2008.pdf. (Erişim Tarihi: **25 Kasım 2021**).

[57] A.Uğur ve A.C.Kınacı, Yapay Zeka Teknikleri ve Yapay Sinir Ağları Kullanılarak Web Sayfalarının Sınıflandırılması, inet-tr'06 - XI. "Türkiye'de İnternet" Konferansı Bildirileri, Ankara, **2006**.

[58] J. Han, M. Kamber ve J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 585-630, **2012**.

[59] S. Venkatesh ve S. Gopal, Robust Heteroscedastic Probabilistic Neural Network for multiple source partial discharge pattern recognition - Significance of outliers on classification capability, Expert Systems with Applications, cilt 38, **2011**.

[60] P. J. Rousseeuw ve A. M. Leroy, Robust Regression and Outlier Detection, New York: John Wiley & Sons, **1987**.

[61] O. Alghushairy, R. Alsini, T. Soule ve X. Ma, A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams, Big Data and Cognitive Computing by MDPI, cilt 5, no. 1, **2021**.

[62] M. M. Breunig, H. P. Kriegel, R. T. Ng ve J. Sander, LOF: Identifying Density-Based Local Outliers, Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Dallas, **2000**.

[63] P.Chauhan ve M.Shukla, A Review on Outlier Detection Techniques on Data Stream by Using Different Approaches of K-Means Algorithm, International Conference on Advances in Computer Engineering and Applications (ICACEA), Ghaziabad, **2015**.

[64] R.Bala ve D.Kumar, Classification Using ANN: A Review, International Journal of Computational Intelligence Research, cilt 13, no. 7, **2017**.

[65] K. Kayaalp ve A. A. Süzen, Derin Öğrenme Ve Türkiye'deki Uygulamaları, Ankara, IKSAD Yayınevi, **2018**.

[66] D. Angluin ve P. Laird, Learning from noisy examples, Machine Learning *2*, cilt 2, 343-370, **1988**.

[67] H. Zhang, M. Cisse, Y. N. Dauphin ve D. Lopez-Paz, mixup: BEYOND EMPIRICAL RISK MINIMIZATION, Published as a conference paper at ICLR 2018, Vancouver, **2018**.

[68] C. M. Salgado, C. Azevedo ve H. Proença, Missing Data, MIT Critical Data, Secondary Analysis of Electronic Health Records, Massachusetts Institute of Technology Cambridge, SpringerOpen, 143-162, 2016.

[69] O. Ibitoye, R. Abou-Khamis, A. Matrawy ve M. O. Shafiq, The Threat of Adversarial Attacks on Machine Learning in Network, Researchgate, Ottowa, **2019**.