**T.C.**
**ONDOKUZ MAYIS ÜNİVERSİTESİ**
**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**
**İSTATİSTİK ANA BİLİM DALI**

# MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ İLE HEMATOLOJİ HASTALIKLARINDAN DEMİR EKSİKLİĞİ ANEMİSİNİN ERKEN TEŞHİS EDİLMESİ

Doktora Tezi

**Bünyamin SARIBACAK**

Danışman
**Doç. Dr. Erol TERZİ**

SAMSUN
2021

**T.C.**
**ONDOKUZ MAYIS ÜNİVERSİTESİ**
**LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**
**İSTATİSTİK ANA BİLİM DALI**

# MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ İLE HEMATOLOJİ HASTALIKLARINDAN DEMİR EKSİKLİĞİ ANEMİSİNİN ERKEN TEŞHİS EDİLMESİ

Doktora Tezi

**Bünyamin SARIBACAK**

Danışman
**Doç. Dr. Erol TERZİ**

SAMSUN
2021

## TEZ KABUL VE ONAYI

**Bünyamin SARIBACAK** tarafından, **Doç. Dr. Erol TERZİ** danışmanlığında hazırlanan **"Makine öğrenmesi sınıflandırma yöntemleri ile hematoloji hastalıklarından demir eksikliği anemisinin erken teşhis edilmesi"** başlıklı bu çalışma, jürimiz tarafından 09.02.2021 tarihinde yapılan sınav sonucunda oy birliği ile başarılı bulunarak Doktora Tezi olarak kabul edilmiştir.

| | **Unvanı Adı Soyadı**<br>**Üniversitesi**<br>**Ana Bilim Dalı** | **İmza** | **Sonuç** |
|---|---|---|---|
| **Başkan** | Prof. Dr. Muhammet BEKÇİ<br>Sivas Cumhuriyet Üniversitesi<br>Fen Fakültesi İstatistik ve Bilgisayar Bilimleri Anabilim Dalı | | ☐ Kabul<br>☐ Ret |
| **Üye (Danışman)** | Doç. Dr. Erol TERZİ<br>Ondokuz Mayıs Üniversitesi<br>Fen Edebiyat Fakültesi İstatistik Anabilim Dalı | | ☐ Kabul<br>☐ Ret |
| **Üye** | Doç. Dr. Pelin KASAP<br>Ondokuz Mayıs Üniversitesi<br>Fen Edebiyat Fakültesi İstatistik Anabilim Dalı | | ☐ Kabul<br>☐ Ret |
| **Üye** | Doç. Dr. Mustafa AKTAŞ<br>Ondokuz Mayıs Üniversitesi<br>Mühendislik Fakültesi Elektrik Elektronik Mühendisliği Anabilim Dalı | | ☐ Kabul<br>☐ Ret |
| **Üye** | Doç. Dr. Tolga ZAMAN<br>Çankırı Karatekin Üniversitesi<br>Fen Fakültesi İstatistik Anabilim Dalı | | ☐ Kabul<br>☐ Ret |

Bu tez, Enstitü Yönetim Kurulunca belirlenen ve yukarıda adları yazılı jüri üyeleri tarafından uygun görülmüştür.

ONAY

… / … / …

Prof. Dr. Ali BOLAT

Enstitü Müdürü

## BİLİMSEL ETİĞE UYGUNLUK BEYANI

Hazırladığım doktora tezinin bütün aşamalarında bilimsel etiğe ve akademik kurallara riayet ettiğimi, çalışmada doğrudan veya dolaylı olarak kullandığım her alıntıya kaynak gösterdiğimi ve yararlandığım eserlerin Kaynaklar' da gösterilenlerden oluştuğunu, her unsurun enstitü yazım kılavuzuna uygun yazıldığını ve TÜBİTAK Araştırma ve Yayın Etiği Kurulu Yönetmeliği'nin 3. bölüm 9. maddesinde belirtilen durumlara aykırı davranılmadığını taahhüt ve beyan ederim.

İmza

02 /03 / 2021

Bünyamin SARIBACAK

## TEZ ÇALIŞMASI ÖZGÜNLÜK RAPORU BEYANI

**Tez Başlığı:** Makine öğrenmesi sınıflandırma yöntemleri ile hematoloji hastalıklarından demir eksikliği anemisinin erken teşhis edilmesi

Yukarıda başlığı belirtilen tez çalışması için şahsım tarafından 06/01/2021 tarihinde intihal tespit programından alınmış olan özgünlük raporu sonucunda;

Benzerlik oranı : % 13

Tek kaynak oranı : % 2 çıkmıştır.

İmza

02 /03 / 2021

Doç. Dr. Erol TERZİ

# ÖZET

MAKİNE ÖĞRENMESİ SINIFLANDIRMA YÖNTEMLERİ İLE HEMATOLOJİ HASTALIKLARINDAN DEMİR EKSİKLİĞİ ANEMİSİNİN ERKEN TEŞHİS EDİLMESİ

Bünyamin SARIBACAK
Ondokuz Mayıs Üniversitesi
Lisansüstü Eğitim Enstitüsü
İstatistik Ana Bilim Dalı,
Doktora Şubat / 2021
Danışman: Doç. Dr. Erol TERZİ

Demir eksikliği anemisi (DEA) dünyada yaygın olarak görülen bir anemi türüdür. DEA tanısı, birçok tıbbi muayene bulguları, tahlil ve tetkikler sonucunda ortaya konulmaktadır. DEA tanısı ve dışında kalan hastaların ayırımı önemli bir konudur. Bunun için bilinen yöntemlerin dışında yeni yöntemlere de ihtiyaç duyulmaktadır. Veri bilimi, yapay zekâ ve makine öğrenimi yöntemleri hekimlerin daha hızlı, daha doğru ve daha güvenilir kararlar verebilmeleri için kullanılmaktadır. Bunlar genellikle bilgisayar desteği ve veri madenciliği teknikleri kullanılarak elde edilen verileri, çok daha hızlı ve güvenilir bir şekilde işleyerek sınıflandırabilmektedir. Günümüzde en çok tercih edilen veri madenciliği sınıflandırıcıları arasında yapay sinir ağları (YSA), destek vektör makinesi (DVM), *k*-en yakın komşuluk (*k*-EK) ve karar ağacı (KA) algoritmaları yer almaktadır. Bu dört algoritmanın sonuçlarının karşılaştırılması için Weka 3.8 yazılımı kullanılmıştır.

**Anahtar Kelimeler**: Sınıflandırma, Makine Öğrenmesi, Demir Eksikliği Anemisi

# ABSTRACT

IRON DEFICIENCY ANEMIA FROM HEMATOLOGICAL DISEASES WITH MACHINE LEARNING CLASSIFICATION METHODS EARLY DIAGNOSIS

Bünyamin SARIBACAK
Ondokuz Mayıs University
Institute of Graduate Studies
Department of Statistics
M.B.A, February/2021
Supervisor: Assoc. Prof. Dr. Erol TERZI

Iron deficiency anemia (IDA) is a common form of anemia in the world. The diagnosis of IDA is made as a result of many medical examination findings, tests and examinations. IDA diagnosis and distinction of patients outside of it is an important issue. For this, new methods are needed in addition to the known methods. In this study, data science, artificial intelligence and machine learning methods are used in order for physicians to make faster, more accurate and more reliable decisions. These are generally able to classify the data obtained using computer support and data mining techniques by processing it much faster and reliably. Among the most preferred data mining classifiers today are artificial neural network (ANN), support vector machine (SVM), *k*-nearest neighbor (*k*-NN) and decision tree (DT) algorithms. Weka 3.8 software was used to compare the results of these four algorithms.

**Keywords:** Classification, Machine learning, Iron deficiency anemia

# ÖNSÖZ VE TEŞEKKÜR

Bu çalışmamda, her türlü bilgi birikim ve tecrübesinden sınırsız yararlanma fırsatı bulduğum çok değerli danışmanım Doç. Dr. Erol TERZİ hocama çok teşekkür ediyorum. Uzun bir süre ara verdiğim akademik çalışmalarıma yeniden başlamaya beni teşvik eden ve devam eden süreçte her adımda yanımda olan Prof. Dr. Ahmet KÖROĞLU' na, araştırmanın uygulama aşamasındaki yardımlarından dolayı Samsun Eğitim ve Araştırma Hastanesi Hematoloji Hastalıkları uzmanı Dr. Sude Hatun AKTİMUR' a teşekkürü bir borç bilirim. Aynı hastanenin Hematoloji servisinde hemşire olarak görev yapan sevgili eşim Sibel SARIBACAK' a, hem manevi hem de akademik desteği için sonsuz teşekkürlerimi sunuyorum.

Çalışmalarımın en başından beri desteklerini her zaman hissettiren Prof. Dr. Mehmet Ali CENGİZ ve Prof. Dr. Yüksel TERZİ hocalarıma çok çok teşekkür ediyorum.

# İÇİNDEKİLER

# SİMGELER VE KISALTMALAR

**DEA** **:**Demir Eksikliği Anemisi

**DSÖ** **:**Dünya Sağlık Örgütü

**KA** **:**Karar Ağacı

**YSA** **:**Yapay Sinir Ağı

**DVM** **:**Destek Vektör Makinesi

***k*-EK** **:***k* En Yakın Komşuluk

**LR** **:**Lojistik Regresyon

**RO** **:**Rastgele Orman

**RBC** **:**Kırmızı Kan Hücresi

**CBC** **:**Tam Kan Sayımı

**TBA** **:**Temel Bileşenler Analizi

# ŞEKİLLER DİZİNİ

# TABLOLAR DİZİNİ

# 1. GİRİŞ

## 1.1. Genel Bilgiler

Veri Bilimi, diğer adıyla Veri Madenciliği veriden bilginin çıkarılması olarak tanımlanmaktadır (Özkan, 2008). Veri bilimi, istatistik, makine öğrenimi, yapay zekâ ve veri tabanı teknolojisini birleştiren çok disiplinli bir alandır (Sayad, 2015). Günümüzde, bilişim teknolojilerindeki gelişmeler sayesinde, çok büyük boyutlardaki veriler kayıt edilebilmektedir. Bu büyük miktarlardaki verilerin işlenebilmesi ve analizinin yapılabilmesi için klasik yöntemlerden farklı olarak makine öğrenmesi teknikleri geliştirilmiştir. Makine öğrenmesi, bilgisayarlara kendi başlarına karmaşık görevleri öğrenme, geliştirme ve yerine getirme becerisi kazandırılmasıdır (Mitchell, 1999). Veri madenciliğinde sıklıkla kullanılan metotların başında sınıflandırma gelmektedir. Sınıflandırma, basitçe bir veri kümesi üzerinde tanımlı olan çeşitli sınıflar arasında veriyi dağıtmaktır. Sınıflandırma algoritmaları, verilen eğitim kümesinden bu dağılım şeklini öğrenirler ve daha sonra sınıfının belirli olmadığı test verileri geldiğinde doğru şekilde sınıflandırmaya çalışırlar (Şeker, 2013).

Demir eksikliği, aneminin önemli bir belirleyicisidir ve Dünya Sağlık Örgütü' ne (DSÖ) göre, dünyadaki en yaygın anemi türleri arasındadır. Aneminin ana nedeni olarak her türlü yetersiz beslenme gösterilmektedir. Dünya genelinde yetersiz beslenmeye karşı gösterilen tüm çabalara rağmen ilerleme sınırlı kalmıştır. Halen bu hastalık ile mücadele eden dünya çapında 614 milyon kadın ve 280 milyon çocuk bulunmaktadır. Özellikle, hamile kadınların % 40’ ı, hamile olmayan kadınların % 33’ ü ve dünyadaki çocukların % 42' si bu hastalığın etkisi altındadır (WHO, 2020). Demir eksikliği iki yaşın altındaki çocuklarda beyin gelişimi üzerinde önemli ve geri döndürülemez etkilerinin yanı sıra, sonraki yaşamda öğrenme ve okul performansı üzerinde olumsuz sonuçlara sebep olmaktadır (Allali vd, 2017; Cusick vd, 2018). Yetişkinlerde demir eksikliğinin yorgunluk, fiziksel performans bozukluğu ve iş verimliliğinin azalması ve sosyal aktivitelerin etkilenmesi gibi olumsuz etkileri vardır (Andro vd, 2013; Haas vd, 2001). Dünya genelinde, sağlık hizmetlerindeki olumlu gelişmelere rağmen özellikle nüfus artışına bağlı olarak, anemiden etkilenen kişi sayılarında artış gözlenmektedir. Bu sonucun nedeni olarak, birincil ilgi alanı bu alanın dışında olan klinisyen tarafından verilen kan analizinin etkisiz değerlendirilmesi veya aneminin etkisiz tedavisi gösterilebilir.

Bu çalışmada, güvenilirliği yüksek, daha kolay yorumlanabilen bir hesaplama modeli oluşturulması amaçlanmaktadır. Bunun için istatistiksel analiz yöntemleri, yeni veri madenciliği ve makine öğrenmesi teknikleri ile uyum içerisinde kullanılmıştır. Bu çalışmada bilgisayar yardımıyla tıbbi ve sağlık alanlarındaki büyük miktarlardaki veri daha kullanılabilir hale getirilmektedir. Verileri analiz etmek için çok sayıda veri madenciliği algoritması ve araçları bulunmaktadır. Günümüzde en çok tercih edilen makine öğrenmesi sınıflandırma algoritmaları arasında karar ağacı (KA), destek vektör makinesi (DVM), *k* en yakın komşu (*k*-EK) ve yapay sinir ağları (YSA) gelmektedir.

KA, ağaç şeklinde oluşan hiyerarşik bir yapıdır. Ağacın en tepe noktasında kök düğüm yer almaktadır. Kök düğüm ayrı sınıfları temsil eden iki ya da daha fazla dala ayrılabilir (Jin vd, 2009). Hedef alan üzerindeki tüm öngörücülerin etkisi araştırılır ve en iyi ayrımı yapan öngörücü ile bölme işlemi gerçekleştirilir. Bu işlemler tekrarlanarak devam eder ve sonunda bir ağaç meydana getirir (Tsiptsis ve Chorianopoulos, 2011). KA, tıbbi karar vermede (sınıflandırma, teşhis vb.) yüksek sınıflandırma doğruluğu sağlayan güvenilir ve etkili bir karar verme tekniğidir ve tıbbi karar alma sürecinin farklı alanlarında kullanılmıştır (Podgorelec vd, 2002).

DVM iki grup sınıflandırma problemleri için oldukça etkili öğrenme yöntemlerinden birisidir (Cortes, 1995). DVM, örnekleri gruplara, aralarında bir karar sınırı oluşturarak ayırmaya çalışır. Bu sınırı iki grubun üyelerine en uzak yer olarak belirlemektedir (Ocak ve Seker, 2013). Eğitim setleri üzerinde bazı sapmalar olmasına rağmen parametreler iyi bir şekilde yapılandırılırsa küçük veri kümeleri üzerinde yüksek başarı oranları elde edilmektedir (Erfani vd, 2016). DVM algoritması fen ve mühendislik bilimlerde yaygın olarak kullanılmaktadır. DVM' nin büyük boyutlu tıbbi görüntüleme verilerini iyi derecede sınıflandırdığı görülmüştür (Gaonkar ve Davatzikos, 2013).

*k*-EK algoritması, istatistiksel örüntü tanıma uygulamalarında yüksek etkili parametrik olmayan bir sınıflandırma tekniğidir (Hwang ve Wen, 1998). Yeni üyenin sınıfı, kendisine en yakın k tane eğitim örneği arasından daha sık olanın türünden belirlenir. k parametresinin seçimi, verinin özelliğine göre belirlenir. Genellikle k değeri yükseldikçe sınıflandırma etkisi düşmektedir (Everitt vd, 2011). İkili sınıflandırma problemlerinde, birbirine bağlı ilişkilerden daha az etkilenmek için k' nın 1 seçilmesi önerilmektedir (Hall vd, 2008).

YSA insan beyninin çalışma şekli modellenerek oluşturulmuş bilgisayar yazılımlarıdır (Agatonovic ve Beresford, 2000). Son yıllarda birçok araştırmacı, kanser türlerinin sınıflandırılması, aktivite tanıma, görüntü işleme gibi konularda YSA ’nın etkili olduğunu savunmuştur (Yue vd, 2018; Mishra vd, 2019; Ting vd, 2019). Beyin hücreleri (nöronlar), sinirler (sinapsis) ile birbirine bağlanarak bir ağ meydana getirir. Sinir ağındaki tüm bağlantıların (sinapsisler) ağırlığı aynı değildir. Bir yapay sinir ağının hesaplama gücü düğümler arasındaki yoğunluğa göre belirlenmektedir (Bartosch-Härlid vd, 2008). YSA’ ları veriler arasındaki farklı ilişki biçimlerini inceleyerek öğrenme gerçekleştirirler. Elde edilen bu bilgiler, yeni verilerin sınıfı belirlenirken bir kural oluşturmak için kullanılır (Al-Nuaimy vd, 2000).

Sınıflandırma sürecinin öncesinde makine öğrenme algoritmaları, eğitim için bir dizi sınıflandırılmış örnek (eğitim seti) kullanılmaktadır. Sonrasında, eğitilmiş algoritmalar elde ettikleri kural ve kazanımları test verilerinin sınıflandırmasında kullanmaktadır. Sınıflandırma sonuçlarının doğruluğunu karşılaştırmak için temel bileşenler analizi, tahmin gücünü analiz etmek için çapraz doğrulama yöntemleri kullanılmıştır. Amacımız; yeni makine öğrenme tekniklerinin, demir eksikliği anemisi tanısının değerlendirilmesi ve doğrulanması, dijitalleştirilmiş bir eşik uyarısı geliştirilmesine yardımcı olmaktır.

### 1.2. Literatür Araştırması

Google Akademik ve çevrimiçi veri tabanları olan Science Direct, Science Citation Index, PubMed, IEEE gibi dijital platformlarda yapılan detaylı aramalar sonucunda, özellikle son yıllarda sağlık alanında makine öğrenme tekniklerinden yararlanılarak yapılmış bazı çalışmalar tespit edilmiştir. Laengsri ve arkadaşları yaptıkları çalışmada, Tayland’ da yetişkinlerde yaygın olarak bulunan demir eksikliği anemisi ve talasemi hastalığı arasındaki ayrımı yapmak için otomatikleştirilmiş bir tahmin modeli geliştirmeye çalışmışlardır. 146 talasemi ve 40 demir eksikliği anemili toplam 186 hastanın verilerini makine öğrenmesi algoritmalarından rastgele orman (RO), *k* en yakın komşu (*k*-EK), karar ağacı (KA), yapay sinir ağı (YSA) ve destek vektör makinesini (DVM) kullanılarak sınıflandırmışlardır. Bu çalışmanın sonucunda DVM ve RO modelleri, DEA' yi Talasemi’ den doğru bir şekilde ayırmak için sırasıyla en yüksek % 96,08 ve % 92,16 ikinci en yüksek doğru tahmin sonuçlarıyla iyi performans göstermiştir (Laengsri vd, 2019).

Chen ve ekibinin yaptığı çalışmada, kanın hemoglobin seviyesini incelemek ve anemi tanısını elde etmek için kan testi gerektirmeyen bir model ortaya koymaya çalışmışlardır. Renk algısının her zaman farklı insanlar arasında tutarlı olmamasından dolayı, anemiyi saptamak için standart bir anemi teşhisi prosedürü olan Palpebral konjonktiva renk dağılımının fiziksel muayenesini taklit etmeye çalışmışlardır. Bunun için destek vektör makinesi veya yapay sinir ağı kullanarak sınıflandırma modeli geliştirmişlerdir. Böylece bilgisayarlar, anemi hastalarını görsel bir tarama işlemi sonunda otomatik olarak tanımlayabileceklerdir (Chen vd, 2016).

Klinik teşhisler, çok sayıda laboratuvar test sonucunun uzmanlar tarafından yorumlanması ile yapılmaktadır. Sonuçlar bireysel değerler olarak rapor edilmektedir. Yuan ve arkadaşları çok değişkenli laboratuvar test sonuçlarının tanısal verimliliğini artırmak için çoklu veri elemanlarını birleştiren bir klinik karar destek algoritması geliştirdiler. Makine öğrenme sınıflandırma teknikleri ile 1538 hastaya ait kandaki demir durumunu doğru olarak sınıflandırmayı başarmışlardır (Yuan vd, 2016).

Demir eksikliği anemisi (DEA) dünya genelinde bir milyardan fazla insanı etkileyen bir beslenme bozukluğudur. DEA, karakteristik olarak küçük (mikrositik) ve hemoglobinde eksik olan kırmızı kan hücrelerinin (RBC) tespiti ile teşhis edilebilmektedir. Tipik olarak hematoloji analizörü tarafından yapılan tam kan sayımı sonuçları incelenmektedir. Bu aletlerin pahalı olması, taşınabilir olmaması ve eğitimli personel gerektirmesi, maliyetleri artırmaktadır. Hennek ve ekibi maliyetleri aşağı çekmek için makine öğrenmesi algoritmalarını kullanmışlardır (Hennek vd, 2016).

Demir eksikliği anemisi (DEA) ve β-talasemi ayırıcı tanısı zaman alıcı ve masraflı bir işlemdir ve tam kan sayımı (CBC) aneminin tanısında birincil test olarak kullanılan hızlı, ucuz ve kolay erişilebilir bir testtir. CBC, DEA ve β-talasemi arasında başarılı bir şekilde ayrım yapamadığından, ileri tekniklere ihtiyaç duyulmaktadır. Bugüne kadar, çok sayıda kırmızı kan hücresi (RBC) endeksi araştırılmış ve her bir indeks için çeşitli parametreler önerilmiştir. Ayyıldız ve arkadaşları yaptıkları bu çalışmada, destek vektör makinesi (DVM) ve *k*-en yakın komşu (*k*-EK) ile RBC indeksleri ve makine öğrenme teknikleri kullanılarak DEA ve β-talasemi birbirinden ayıran bir tanı ortaya koydular. Çalışmada sınıflandırıcı için giriş parametreleri olarak RBC endeksleri kullanılmıştır. Her iki tekniğin etkinliğini belirlemek için, DVM ve *k*-EK performansları ayrı ayrı değerlendirilmiştir. Makine öğrenme algoritmalarına girdi

olarak daha az parametre vererek ve daha yüksek performans elde etmeyi başarmışlardır (Ayyıldız vd, 2020).

Çocuklarda anemi, önleyici tedbirler konusundaki bilinçsizlik yüzünden dünya çapında bir sorun haline gelmektedir. Meena ve ekibi çocuklar arasındaki anemiyi tahmin etmek, çocukların beslenme faktörleri hakkında bilgi içeren bir veri tabanına, veri madenciliği tekniklerini kullanarak bir karar destek sistemi geliştirmişlerdir (Meena vd, 2019).

Demir Eksikliği Anemisi (DEA) ve Talasemi dünyada sık görülen bir hastalıktır. Hastane rutininde, DEA ve Talasemi tam kan sayımı (CBC) sonucundaki hemoglobin seviyesine göre tanınmaktadır. Daha sonra, görsel uzmanlar, insan hatasına maruz kalan ışık mikroskobu altında inceleme yapmaktadır. Bu araştırmada Ahmad ve arkadaşları DEA ve Talasemi ile ilişkili eritrositleri sınıflandırmak ve karakterize etmek için makine öğrenmesi sınıflandırıcılarını karşılaştırarak bir teknik önermişlerdir (Ahmad vd, 2018).

Teşhis doğruluğu tıbbi bakımın temel sorunlarından biri olmaya devam etmektedir. Sakhibgareeva ve arkadaşları yaptıkları bilimsel çalışmada bu sorunu çözmek için 200 farklı laboratuvar testinden elde edilen hasta verilerinin akıllı analizine dayanan tıbbi tahmin modelini yapay zekâ teknikleri ile ortaya koymaya çalışmışlardır. Çalışmada demir eksikliği anemisine sahip hastaların tespiti makine öğrenmesi ile yapılmaya çalışılmıştır (Sakhibgareeva vd, 2017).

İnsanda kırmızı kan hücrelerini (RBC) etkileyen en yaygın hastalıklardan biri anemidir. Elsalamony bu çalışmasında sağlıklı ve sağlıksız anemi hücrelerinin ayrımını, makine öğrenmesi yapay sinir ağlarını kullanarak yapmıştır ve yüksek sınıflandırma başarısı elde etmiştir (Elsalamony, 2016).

İlaslaner ve ekibi kan biyokimya parametreleri ile demir eksikliği anemisi arasındaki ilişkiyi değerlendiren bir karar destek sistemi oluşturmak için örüntü tanıma ve veri madenciliği tekniklerinden yararlanmışlardır. En yüksek performansı veri yapay sinir ağları (78.31) ile elde edilmiştir. Böylece; biyokimyasal parametrelerin demir eksikliği anemisinin saptanmasında etkili olduğunu gösterilmişlerdir. Bu durumun doktorun hastanın etkili tedavisini başlatmasına yardımcı olacağını varsaymaktadırlar (İlaslaner vd, 2019).

Kırmızı kan hücresi sayısı, hastanın genel sağlığını belirlemede hayati bir rol oynamaktadır. Kan hücrelerini saymak için kullanılan geleneksel yöntemler hatalı sonuçlar üretebilmektedir. Acharya ve arkadaşları araştırmalarında, kırmızı kan hücresini, kanın diğer bileşenlerinden ayırmak için bir görüntü işleme tekniği önermektedir. Makine öğrenmesi ile sınıflandırmaya yardımcı olan önemli özelliklerin belirlenmesi sonucunda, içinde demir eksikliği anemisinin de olduğu birçok hastalığı teşhis etmenin daha kolay olduğunu göstermişlerdir (Acharya vd, 2017).

Tukaram ve Vernekar yaptıkları çalışmada, kırmızı kan hücrelerini dört anormal tipte sınıflandırmak için yapay sinir ağı ve karar ağacı algoritmalarının sonuçlarını, sınıflandırma doğruluğu açısından karşılaştırmışlardır (Dalvi ve Vernekar, 2016).

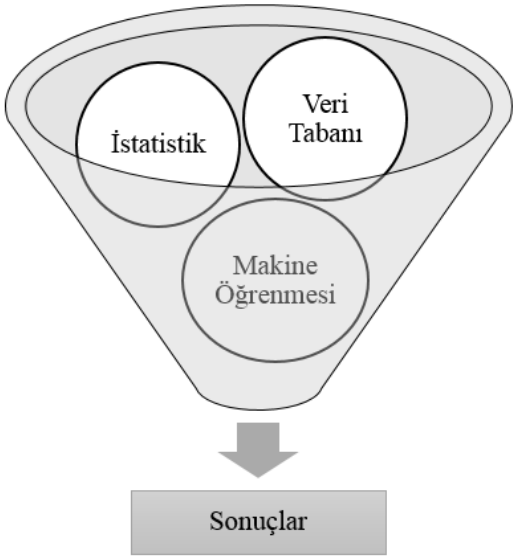
Hemoglobin üretimi için demir gereklidir ve eksikliği ile eritrositlerin boyutu küçülür, şeklini değiştirir ve normalden daha soluklaşır. Tyagi ve ekibi, normal RBC ve poikilosit hücreleri, çıkarılan özelliklere dayanarak yapay sinir ağı kullanılarak sınıflandırılmıştır (Tyagi vd, 2016).

Demir eksikliği anemisi, mikrositik aneminin önemli bir nedenidir. Doğrusal denklemlere dayanan mevcut yöntemler, iki anemi sınıfını ayırt etme konusunda yetkin olsa da, bu sorunların ortak oluşumunu tanımlayamamaktadır. Bellinger ve arkadaşları makine öğrenme algoritmalarının, karmaşık alanlarda doğru sınıflandırmayı mümkün kıldığını göstermişlerdir (Bellinger vd, 2015).

# 2. VERİ BİLİMİ

## 2.1. Veri madenciliği

Büyük miktardaki verinin içinden yararlı olan bilgileri keşfetme süreci olarak tanımlanmaktadır (Khatatneh ve Emary, 2009). Bu işlemler gerçekleştirilirken birçok disiplin uyum içinde birlikte çalışmaktadır. Bu disiplinler genellikle istatistik, makine öğrenmesi ve veri tabanlarıdır (Olson ve Delen, 2008). Sınıflandırma, örnekleme, tahmin, hipotez testi gibi istatistiksel yöntemler ile sinyal işleme, örüntü tanıma, modelleme, makine öğrenme gibi yapay zekâ teknikleri ve verilerin alınması, eklenmesi, güncellenmesi veya kaldırması için veri toplayan, depolayan ve yöneten veri tabanı işlemleri uyum içinde kullanılmaktadır.



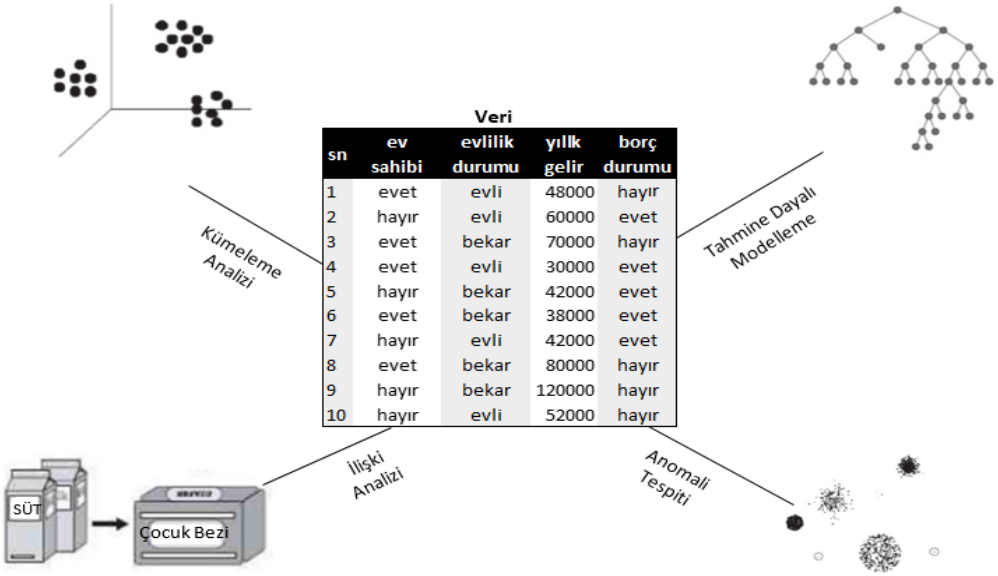

Şekil 2.1. Veri madenciliği disiplinler arası etkileşim

### 2.1.1. Veri madenciliği görevleri

Veri madenciliği görevleri genellikle tahmine dayalı görevler ve açıklayıcı görevler olmak üzere iki ana kategoriye ayrılmaktadır (Tan vd, 2016).

Tahmine dayalı görevlerin amacı açıklayıcı ve bağımsız değişkenlerden yararlanarak hedef veya bağımlı değişkeninin değerini tahmin etmektir.

Açıklayıcı görevler, sonuçları doğrulamak ve açıklamak için veriler arasındaki ilişkileri ortaya çıkarmaya çalışmaktadır. Bunun için kümeleme, sınıflandırma gibi tekniklerden yararlanırlar.




Şekil 2.2. Bazı veri madenciliği görevleri (Tan vd, 2016)

### 2.1.2. Veri madenciliği işlem basamakları

Veri madenciliği bilgi keşfinin başlangıcı olarak ifade edilmektedir ve bazı adımların ardışık tekrarı şeklindedir (Han ve Kamber, 2011). Bu adımlar Tablo 2.1' de gösterilmektedir.

Tablo 2. 1. Veri madenciliği işlem adımları (Han ve Kamber, 2011)

| İşlem Basamakları | Açıklaması |
|---|---|
| **Veri Temizleme** | Gürültülü ve tutarsız verinin, veri setinden çıkarılması |
| **Veri Bütünleştirme** | Çoklu veri kaynaklarının birbiri ile ilişkilendirilmesi |
| **Veri Seçimi** | Analizi yapılacak verinin veri tabanından alınması |
| **Veri Dönüştürme** | Verinin uygun formatlara dönüştürülmesi ya da birleştirilmesi |
| **Veri Madenciliği** | Veri örüntüsü çıkarabilmek için akıllı metotların uygulanması |
| **Örüntü Değerlendirme** | İlgi çekicilik ölçütlerine dayalı bilgiyi temsil eden desenleri tanımlamak |
| **Bilgi Gösterimi** | Elde edilen bilgiyi görsel teknikler kullanılarak sunmak |

## 2.2. Makine öğrenmesi

Yapay zekâ insan beyni model alınarak yapılan çalışmaların sonucunda ortaya çıkmıştır (Kocabaş, 2015). Yapay zekâ insan beyninin işlevlerini bilgi teknolojilerini kullanarak modelledikten sonra bilgisayarlara aktararak kullanılabilir sistemler oluşturmayı hedeflemektedir. Bu çalışmaların sonunda uzman sistemler, genetik algoritmalar, bulanık mantık, yapay sinir ağları, makine öğrenmesi gibi teknolojik sistemler ortaya çıkmıştır. Mitchell makine öğrenmesini, bir bilgisayar programının bir görevi yaparken edindiği tecrübe miktarı arttıkça, görevi gerçekleştirme performansın da artması olarak tanımlamaktadır (Mitchell, 1999). Makine başlangıçta öğrenme gerçekleştirmek için bir veri setine ihtiyaç duymaktadır. Öğrenme işlemini problemin tipine göre geliştirilmiş algoritmaları kullanarak gerçekleştirmektedir. Veri seti hazırlanırken, değişkenlerin seçimi, optimum kayıt sayısı, doğru öğrenme modelinin belirlenmesi, kullanılacak algoritmanın tespiti, öğrenme sürecini etkiyen faktörler arasında yer almaktadır. Bunun amacı en yüksek performansın elde edileceği modeli belirlemektir. Bu iş esnasında aynı öğrenme modeli değişik şekillerde test edilebileceği gibi farklı öğrenme modelleri de bir arada kullanılabilmektedir (Rai, 2011). Makine öğrenmesi yapay zekânın bir alt dalıdır. Bir bilgisayarın veya bilgisayar kontrolündeki bir robotun çeşitli faaliyetleri zeki canlılara benzer şekilde yerine getirme kabiliyeti olarak ifade edilmektedir (Copeland ve Proudfoot, 2007).

Alan Turing' in 1950 yılında yaptığı ve sonrasında "Turing Testi" olarak anılacak olan bir makinenin, bir insanla yaptığı konuşmayı ayırt edilemeden sürdürebiliyorsa, makinenin düşündüğünü iddia ettiği çalışması yapay zekâ alanındaki ilk bilimsel çalışmalardan birisi olarak gösterilmektedir (Turing, 1950).

1950' li yılların sonuna doğru Arthur Samuel' in geliştirdiği dama programının, sıradan bir amatörle mücadele edebilecek beceriye ulaştığı görülmektedir.

Dartmouth Workshop tarafından 1970' li yıllarda geliştirilen programlar sayesinde bilgisayarlar, matematik ve geometri problemlerini çözebilen ve İngilizce konuşabilen yeteneklere ulaşmıştır.

1980' erde ortaya çıkan " uzman sistemler " adı verilen bir yapay zekâ programı dünya çapındaki şirketler tarafından kullanılmıştır. Bu sistemlerde bilgi yapay zekâ araştırmalarının odak noktası haline gelmektedir (McCorduck ve Cfe, 2004).

1997' de IBM tarafından üretilen ve saniyede iki yüz milyon hareket yapabilme yeteneğine sahip süper bilgisayar, dünya satranç şampiyonu Garry Kasparov' u yenen ilk bilgisayar satranç oynama sistemi olmuştur.

21. yüzyılın ilk on yıllarında, büyük miktarda veriye erişim daha ucuz ve daha hızlı bilgisayarlar ve gelişmiş makine öğrenimi teknikleri, ekonomideki birçok soruna başarıyla uygulanmıştır (Manyika, 2011).

### 2.2.1. Makine öğrenmesi yöntemleri

Günümüzde bilgi teknolojileri araştırmacıları, canlıların öğrenme şekillerini makineler üzerine modellemeye çalışmaktadırlar. İlerleyen süreçte kendi kendine öğrenebilen ve öğrendiklerini yorumlama yeteneğine sahip makinelerin sayısında artış beklenmektedir. Makine öğrenmesi, makinelerin tanıma, kestirim yapma ve sınıflandırma gibi bazı davranışları gerçekleştirebilmesi üzerine yapılan çalışmaların bir sonucu olarak doğmuştur (Balaban ve Kartal, 2015). Makine öğrenme yöntemleri gözetimli öğrenme (supervised learning), gözetimsiz öğrenme (unsupervised learning) ve pekiştirmeli öğrenme (reinforcement learning) olmak üzere üç gruba ayrılmaktadır (Alpaydin, 2020).

#### 2.2.1.1. Gözetimli öğrenme

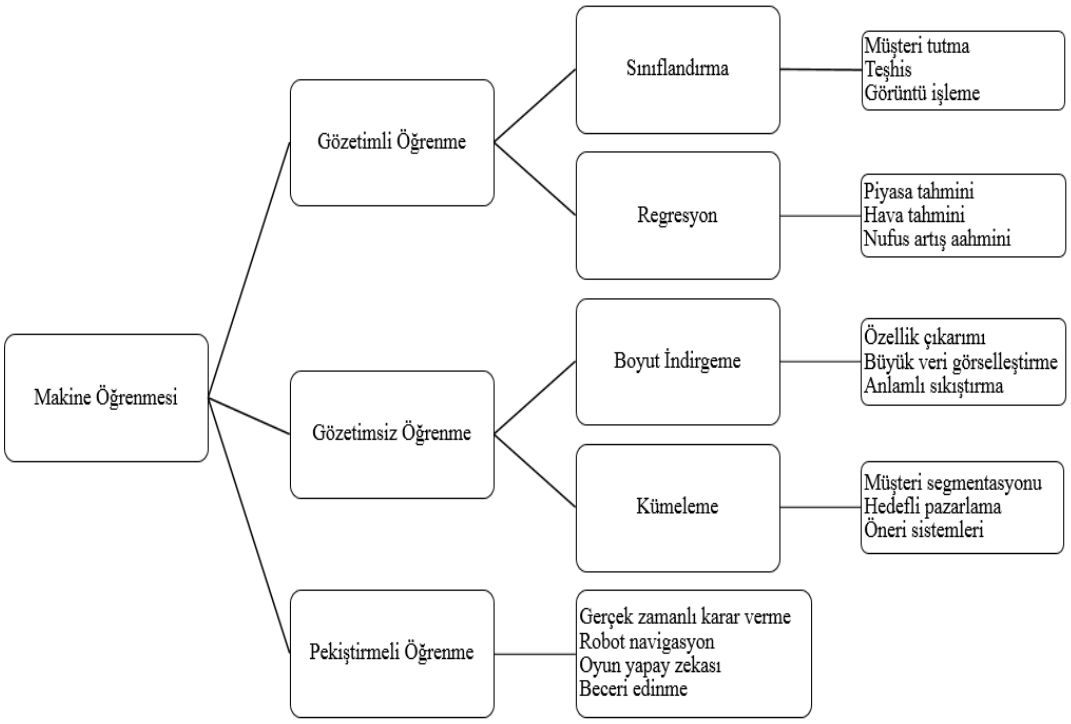
Gözetimli öğrenme, örnek girdi-çıktı çiftlerine dayalı olarak bir girdiyi bir çıktıya eşleyen işlevsel bir yöntemdir (Russell ve Norvig, 2010). Bu yöntem bir dizi eğitim örneğinden oluşan etiketli eğitim verilerini analiz ederek, gelecekteki girdi-çıktı gözlemlerini tahmin edebilen bir fonksiyon üretmektedir (Mohri vd, 2012). Burada amaç tahmin edilen çıktı değerleri ile bilinen çıktı değerleri arasındaki farkı en aza indirgemektir. Sıklıkla regresyon ve sınıflandırma problemlerinin çözümünde bu yöntem tercih edilmektedir. Çıktı değişkeninin sürekli yapıda olduğu durumlarda regresyon, karar ağaçları ve rastgele orman gibi algoritmalar, çıktı değişkeninin kesikli olduğu durumlarda da *k* en yakın komşuluk, karar ağaçları, lojistik regresyon, destek vektör makinesi algoritmaları kullanılmaktadır (Gürsakal, 2017). Bir bankanın kredi almak isteyen müşterilerini risk grubuna göre sınıflandırması, gözetimli öğrenme modeline örnek gösterilebilir (Balaban ve Kartal, 2015). Bu örnekte bankanın kredi verdiği müşterilerinden kazandığı deneyimler, yeni gelecek olan müşterilerin risk durumlarının belirlenmesinde kullanılmıştır.

#### 2.2.1.2. Gözetimsiz öğrenme

Gözetimsiz öğrenme yöntemi, önceden eğitilmemiş veriler arasında bağıntıların araştırılarak birbiriyle benzer özellikler taşıyan verilerin kümelenmesi olarak ifade edilmektedir (Hastie vd, 2009). Bu öğrenme yönteminde algoritmalar başlangıçta, önceden sınıfı bilinmeyen girdi verilerinin sınıfını tespit etmek için geriye doğru inceleyerek bir öğrenme gerçekleştirirler ve sonrasında elde ettikleri bu kazanımları yeni gelen verileri en uygun gruba atamakta kullanırlar. Algoritmalar bu işi genellikle verileri kümeleyerek, boyut indirgeyerek ya da birliktelik kurallarına bağlı şekilde gerçekleştirirler. Bir market müşterilerinin ürün satın alma davranışlarını inceledikten sonra birbiri ile yakın alışveriş alışkanlıkları olan müşterilerini aynı gruba atayarak ona uygun şekilde ürünlerinin stoklarını güncelleme veya raf düzenlemesi yapması gözetimsiz öğrenmeye örnek gösterilmektedir (Alpaydın, 2011).

#### 2.2.1.3. Pekiştirmeli öğrenme

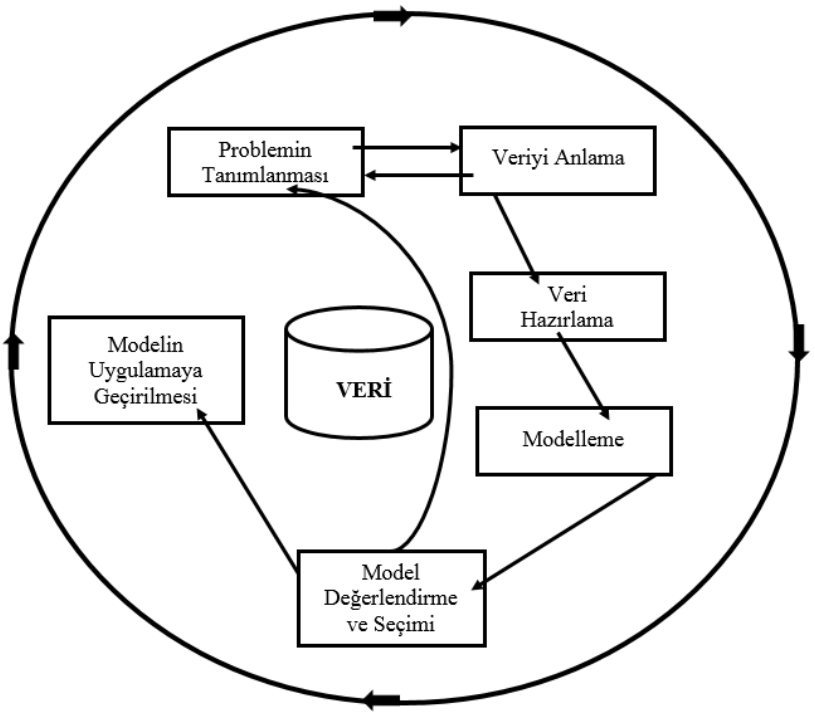
Makinenin sınırlı geribildirimin olduğu sıralı karar verme problemlerinde deneme-yanılma şeklinde etkileşime girerek öğrenmesi şeklinde tanımlanmaktadır (Van ve Wiering, 2012). Bu yöntemde olası durumların hedef olup olmadığı kontrol edilerek öğrenme süreci sürekli devam ettirilir. Amaç insan beyninin çalışma prensiplerine benzeyen algoritmalar üretmektir. Derin öğrenme algoritmaları bu öğrenme modeli üzerine kurgulanmaktadır.




Şekil 2.3. Makine öğrenmesi çalışma alanları

### 2.2.2. Makine öğrenmesi işlem adımları

Makine öğrenmesi ile veri madenciliği oldukça yakın ilişki içerisindedir. Veri madenciliği büyük miktardaki verinin içinden değerli olan bilgiyi çıkarmaya çalışırken makine öğrenmesi algoritmaları ile uyum içinde çalışmaktadır. Hem veri madenciliğinde hem de makine öğrenmesinde izlenen süreç birbirine benzemektedir. Bu konuda Shearer' ın ortaya koyduğu, veri madenciliği için çapraz endüstri standart süreç modeli (Cross-Industry Standart Process for Data Mining-CRISP) literatürde en çok kabul görenler arasındadır (Shearer, 2000).




Şekil 2.4. CRISP-DM modeli (Shearer, 2000)

Genellikle başarılı bir veri bilim projesi, iyi tanımlanmış bir soru veya ihtiyaçla başlamaktadır. Yani problemin neden kaynaklandığının, ulaşılmak istenen hedeflerin ve bu amaç doğrultusunda ihtiyaç duyulan gereksinimlerin başlangıçta iyi tasarlanmış bir planla ortaya konulmasıdır.

Problemin tanımlanmasından sonraki adım olan veri anlama aşaması probleme uygun veri toplanmaktadır. Veri daha yakından incelenerek, eksik, gürültülü ve kirli veriler tespit edilmektedir. Sonrasında ihtiyaca bağlı olarak ilave veriler, veri setine eklenmektedir. Yine bu süreçte veriyi problemlerinden ayıklamak ve veri kalitesini artırmak için doğruluk, tamlık, tutarlılık, güncellik, inanılırlık, katma değer, yorumlanabilirlik ve ulaşılabilirlik ölçütlerine uygunluğu kontrol edilmektedir (Balaban ve Kartal, 2015).

Veri hazırlama genellikle zaman alıcı bir süreçtir ve hatalara açıktır. Veri biliminde çöp sözcüğü, verilerin geçersiz, aralık dışı veya eksik değerlerle toplandığını ifade etmektedir. Bazı kısımları eksik olan verinin sisteme hiç dahil edilmemesi veya verinin eksik kısımlarının tamamlanması, bu tamamlama sırasında nasıl bir yöntem izleneceği gibi kararlar bu aşamada verilmektedir. Doğum tarihlerinin yaşa verilmesi gibi basit bir veri dönüşümü veya adres alanında kişinin il ilçe gibi bilgilerinin çıkarılması gibi veri zenginleştirme işlemleri buna örnek

gösterilmektedir. (Şeker ve Eşmekaya, 2017). Dikkatli bir şekilde taranmamış verileri analiz etmek, son derece yanıltıcı sonuçlara neden olabilmektedir.

Probleme ve öğrenme stratejisine uygun olarak bir makine öğrenmesi veya istatistiksel yöntem belirlenmesi, model kurma aşaması olarak tanımlanmaktadır (Flach, 2012). Bu aşamada problemin çözümünde farklı modeller kullanılarak elde edilen sonuçların karşılaştırılması, en uygun modelin belirlenmesi açısından doğru bir yaklaşım olacaktır. Nihai karar olarak tespit edilen model üzerinde iyileştirmeler yapmak, modele uygun olarak verinin yeniden düzenlenmesi sonuçlara olumlu katkı sağlayacaktır.

Model değerlendirme aşamasında, işin en başından itibaren oluşturulan stratejinin belirlenen hedefleri ne kadar sağladığı test edilmektedir. Bu işlem model performans ölçütleri kullanılarak yapılmaktadır. Doğruluk, duyarlılık, belirleyicilik, kesinlik, F-ölçütü literatürde yer alan ölçüt türlerinden bazılarıdır (Balaban ve Kartal, 2015).

Sistemin en önemli döngüsü Şekil 2.4'de görüldüğü gibi değerlendirme adımındaki sonuçlara göre yeniden sistemin en başına dönerek bütün adımların tekrarlanmasını sağlayan ana döngüdür. Son aşamada tercih edilen modelle ilgili tüm sorunlar giderildikten sonra modelin çalışan bir uygulaması geliştirilir.

### 2.3. Sınıflandırma

Sınıflandırma, bir veri kümesi (data set) üzerinde tanımlı olan çeşitli sınıflar arasında verinin dağıtılması olarak tanımlanmaktadır (Şeker, 2013). Makine öğrenimi terminolojisinde sınıflandırma, denetimli öğrenmenin bir örneği olarak kabul edilmektedir. Denetimli öğrenme, doğru tanımlanmış gözlemlerden oluşan bir eğitim setinin mevcut olduğu öğrenme şeklidir (Alpaydın, 2020).

#### 2.3.1. Veri setinin sınıflandırma için düzenlenmesi

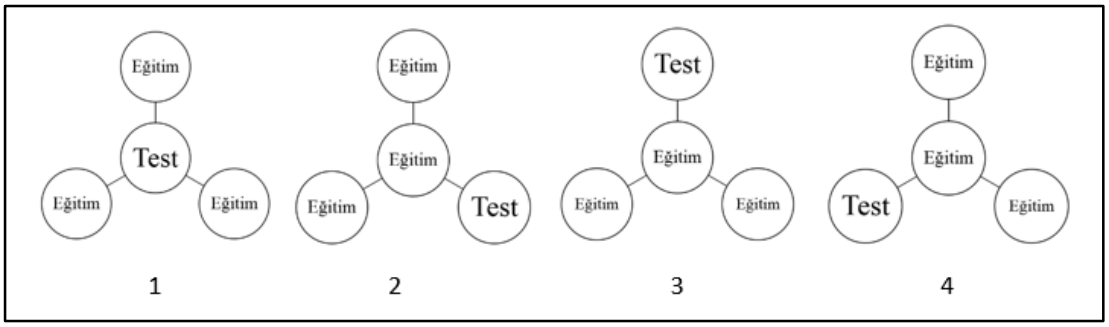
Sınıflandırma işlemi öncesinde, belirlenen modelin başarısını ölçmek için veri seti, eğitim seti (train set) ve test seti (test set ) olarak ikiye bölünmektedir. Eğitim setindeki veriler modelin eğitilmesinde, test setindeki veriler ise model performansının ölçülmesinde kullanılmaktadır. Ayrıca modelin tarafsız bir değerlendirmesini sağlamak için doğrulama seti (validation set) kullanılmaktadır. Test setindeki veriler sistem tarafından daha önce görülmeme özelliği ile doğrulama setindeki verilerden ayrılmaktadır (Uğuz, 2019). Birinci aşamada sınıflandırma algoritmaları eğitim seti

üzerinden verilerin dağılım şeklini öğrenirler ve ikinci aşamada elde ettikleri kazanımları sınıfının belirli olmadığı test verilerini doğru şekilde sınıflandırmak için kullanırlar. Kullanılan öğrenme algoritması, eğitim ve test setlerinin büyüklüğü, hatalı sınıflandırma, sınıf dağılımı gibi faktörler modelin performansını belirlemektedir.

### 2.3.2. Model performansının değerlendirilmesi

Model performansını değerlendirmek için dışarda tutma (holdout), tekrarlı holdout (repeated holdout), üçlü ayırma (three-way split), bootsstrap örnekleme (bootstrap sampling), temel bileşenler analizi (TBA), çapraz doğrulama (cross validation) gibi yöntemler kullanılmaktadır. Dışarda tutma yönteminde veri seti, eğitim seti ve test seti olarak ikiye bölünür. Eğitim setindeki veriler öğrenme için kullanılmaktadır. Test veri seti ile öğrenme düzeyi kontrol edilerek model performansı ölçülmektedir. Tekrarlı Holdout yöntemi, dışarda tutma yöntemindeki adımların rassal olarak birkaç kez tekrarlanmasıdır. Üçlü ayırma yönteminde veri seti, eğitim, test ve doğrulama veri seti olarak üçe bölünmektedir. Bu yöntemde model seçimi ve performans tahmini birlikte yapılmaktadır. Doğrulama veri seti ile tercih edilen algoritmaya ait parametreler hassas bir şekilde ele alınmaktadır. Modelin performans sonuçları test verileri kullanılarak elde edilmektedir. Bootstrap örnekleme yönteminde eğitim seti oluşturulurken veri setinden başlangıçta belirlenen sayı kadar örnek, rassal olarak alınmaktadır ve alınan örnekler daha sonra tekrar veri setine dahil edilmektedir. Bu işlem eğitim veri setinde tekrarlayan örnekler oluşmasına sebep olmaktadır. Eğitim veri seti dışında kalan örneklerin tamamı test veri setine dahil edilmektedir**.** TBA, değişken sayılarının fazlalığından kaynaklanan, eğitim süresinin uzaması, aşırı öğrenme (overfitting) ve çoklu doğrusal bağlantı (multicollinearity) problemlerini ortadan kaldırmak için değişken seçimi (feature selection) ve boyut indirgeme (dimensionality reduction) yöntemlerini kullanılmaktadır. Değişken seçiminde veri setindeki bazı değişkenler korunurken bazıları ise modelden çıkartılmaktadır. Boyut indirgemede mevcut değişkenlerin kombinasyonundan oluşan yeni değişkenler oluşturularak değişken sayısı azaltılmaktadır. Sonuçta veri setinin tüm özellikleri bir şekilde korunurken değişken sayısı azaltılmış olmaktadır (Muratlar, 2019). Çapraz doğrulama yönteminde veri seti k eşit parçaya bölünmektedir. Bu parçaların her biri bir defa test veri seti, kalan k-1 parçalar ise eğitim veri seti olarak kullanılmaktadır. İşlem k defa tekrarlanmaktadır. Tüm elde edilen performans değerlerinin ortalaması sonuç performans değeri olarak kabul edilmektedir. Bu çalışmada model

performansların ölçümünde k katlı çapraz doğrulama yöntemi kullanılmıştır. Bu durum Şekil 2,5'de k sayısı 4 alınarak gösterilmiştir. Genellikle uygulamalarda k değeri 10 seçilmektedir. Bilimsel çalışmalarda çoğunlukla veri seti, eğitim seti ve test seti olarak iki grubu ayrılırken bölünme oranları %70-%30, %40-%60 tercih edilmektedir.




Şekil 2. 5. 4 Kat çapraz doğrulama

### 2.3.3. Sınıflandırma algoritmaları için model performans değerlendirme ölçütleri

Makine öğrenmesinde ikili sınıflandırma modelinde algoritma performansının değerlendirilmesi için genellikle hata matrisi (confusion matrix ) kullanılmaktadır (Stehman, 1997). Bu matrisin her satırı, tahmin edilen bir sınıftaki örnekleri temsil ederken, her sütun gerçek bir sınıftaki örnekleri temsil etmektedir (Powers, 2011). Bu tablodan yararlanılarak hesaplanan performans ölçütleri Tablo 2.2'de gösterilmektedir.

Tablo 2. 2. Hata matrisi

| | | Gerçek sınıf | | |
|---|---|---|---|---|
| | | Pozitif | Negatif | Toplam |
| Tahmin edilen sınıf | Pozitif | Doğru Pozitif | Yanlış Pozitif | tpoz |
| | Negatif | Yanlış Negatif | Doğru negatif | tneg |
| | Toplam | poz | neg | m |

Bu tabloda doğru pozitif, aslında pozitif olan ve pozitif olarak sınıflandırılan örnekleri ifade etmektedir. Örneğin bir kişinin demir eksikliği anemisi hastası olup olmadığına bakılırken, doğru pozitif (dp) değeri, gerçekte demir eksiliği anemisi olan hastalardan modelin demir eksikliği anemisi hastası olarak tahmin ettiği hastaların sayısını vermektedir. Doğru negatif (dn) gerçekte demir eksikliği anemisi olmayan

hastalardan modelin demir eksikliği hastası değildir diye tahmin ettiği hastaların sayısıdır. Yanlış pozitif (yp), modelin gerçekte demir eksikliği anemisi hastası olan hastalardan modelin demir eksikliği anemisi hastası değildir şeklinde tahmin ettiği hastaların sayısıdır. Yanlış negatif, gerçek demir eksikliği anemisi hastası olmayan hastalardan modelin demir eksikliği anemisi hastasıdır şeklinde tahmin ettiği hastaların sayısıdır. Sınıflandırma performansını tespit etmek için kullanılan bazı ölçütler Şekil 2.6'da gösterilmiştir.

$$Doğruluk = \frac{dp + dn}{m}$$

$$Hata\ oranı = 1 - Doğruluk\ oranı$$

$$Duyarlılık\ oranı = \frac{dp}{poz} = \frac{dp}{dp + yn}$$

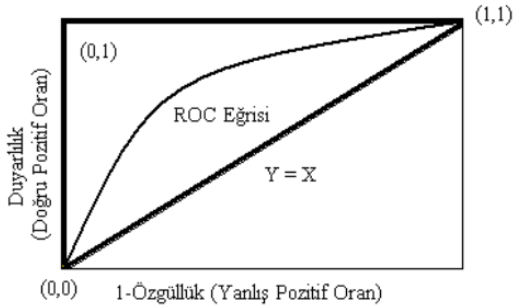
$$Belirleyicilik\ oranı = \frac{dn}{neg} = \frac{dn}{dn + yp}$$

$$Kesinlik = \frac{dp}{tpoz} = \frac{dp}{dp + yp}$$

$$F - ölçütü = \frac{2 * Duyarlılık\ oranı * Kesinlik}{Duyarlılık\ oranı + Kesinlik}$$

Şekil 2.6. Performans değerlendirme ölçütleri

Veri setinin dengesiz olduğu durumlarda yani farklı sınıflardaki gözlemlerin sayısı büyük ölçüde değiştiği zaman doğruluk yanıltıcı sonuçlar verebilmektedir. Hata matrisini değerlendirmek için en bilgilendirici ölçüt Matthews korelasyon katsayısıdır (Chicco ve Jurman, 2020). En uygun modeli tespit edebilmek için Şekil 2.6'da verilen ölçütlerin dışında ROC eğrisine de bakılmaktadır (Kılıç, 2013). Bu eğri ikinci dünya savaşı esnasında sinyallerin doğru tanımlanabilmesi için geliştirilmiş bir yöntemdir.




Şekil 2.7. ROC eğrisi (Tomak ve Yüksel, 2009)

ROC eğrisi; testin ayırt etme gücünün belirlenmesine özellikle tanı testi performanslarının değerlendirilmesi ve kıyaslanmasına yardımcı olmaktadır. En yararlı tanı testi, doğru pozitiflik oranı yüksek ve yanlış pozitiflik oranı düşük olan testtir. Mükemmele yakın bir tanı testi, hemen hemen dikey (0,0)'dan (0,1)'e ve sonra yatayda (1,1)'den geçen bir ROC eğrisine sahip olmalıdır (Şekil 2.7). Kısaca sol üst köşeye en yakın geçen ROC eğrisini veren test en kullanışlı testtir (Dirican, 2001).

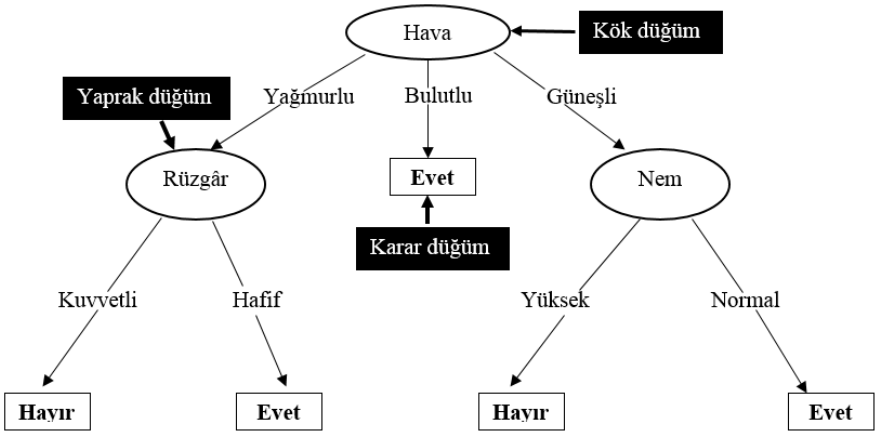
## 2.4. Makine öğrenmesi sınıflandırma algoritmaları

### 2.4.1. Karar ağaçları

Karar ağacı öğrenimi, istatistik, veri madenciliği ve makine öğreniminde kullanılan tahmine dayalı modelleme yaklaşımlarından birisidir. Bilinen karar ağaçları algoritmaları arasında AID, CHAID, CART, ID3, C4.5, C5.0, MARS, E-CHAID, SLIQ, SPRINT ve QUEST yer almaktadır. Bu algoritmalar kök, düğüm ve dallanma kriterlerinin seçiminde izlenen yollar ile birbirinden ayrılmaktadır. Karar ağacı tabanlı ilk algoritma 1970' li yılların başlarında Morgan ve Sonquist isimli uzmanlar tarafından geliştirilen AID algoritmasıdır. Bu algoritma, en iyi tahmini gerçekleştirmeye ve en kuvvetli ilişkiye sahip bağımsız değişkeni bulmaya dayanmaktadır. 1980 yılında G. V. Kass tarafından sınıflandırma ve regresyon işlemlerinde kullanılmak üzere istatistik tabanlı olan CHAID algoritması geliştirilmiştir. 1986 yılında J. Ross Quinlan tarafından ID3 adlı bir karar ağacı algoritması geliştirilmiştir. J. Ross Quinlan 1993 yılında ID3 algoritmasının eksik yönlerini gidermek için ID3' ün ileri bir sürümü olan C4.5' i ve C4.5' e göre daha hızlı, daha az bellek kullanan, daha kesin kurallar oluşturan C5.0 algoritmasını çıkarmıştır (Rokach ve Maimon, 2005).

Karar ağacı, bir ağaç yapısı biçiminde sınıflandırma veya regresyon modelleri oluşturmaktadır. Bu modelde veri kümesi gittikçe azalan alt kümelere ayrılırken, aynı

zamanda ilişkili bir karar ağacı aşamalı olarak ortaya çıkmaktadır (Sayad, 2020). Karar ağaçlarında, karar düğümleri ve yaprak düğümleri bulunmaktadır. Karar düğümleri veri setinde karar vermek, sınıflandırma yapmak ya da tahminde bulunmak için kullanılırken yaprak düğümler verilen kararları tutmaktadır (Balaban ve Kartal, 2015). Bir karar düğümünün iki veya daha fazla dalı olabilmektedir. Ağacın en tepe noktasında en iyi tahmin ediciye karşılık gelen kök düğüm yer almaktadır. Karar ağaçları hem sayısal hem de kategorik verileri işleyebilmektedir. Karar ağaçları, anlaşılırlıkları ve basitlikleri göz önüne alındığında en popüler makine öğrenimi algoritmaları arasındadır (Wu vd, 2008; Piryonesi ve El-Diraby, 2020).



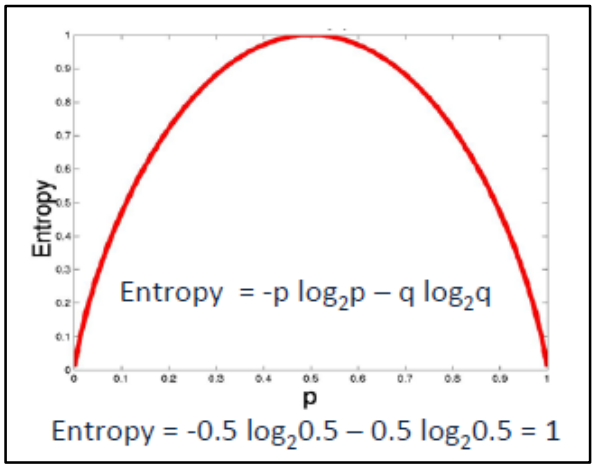

Şekil 2.8. Karar ağacı görünümü

Şekil 2.8'de verilen karar ağacı örneğinde, hava durumu bilgilerine (sıcaklık, nem, rüzgar) bakılarak tenis oynanıp oynanamayacağına karar verilmek istenmiştir. Havanın bulutlu olması durumunda tenis oynanabilmektedir. Eğer hava yağmurlu ise rüzgârın durumuna göre tenis oynanıp oynanmayacağına bakılmaktadır.

Karar ağaçları belirli bir veri setinin açıklaması, kategori ve genellemesine yardım etmek için matematiksel ve hesaplama teknikleri kombinasyonu olarak da tanımlanabilmektedir. Veri madenciliği işlemleri yapılırken veri;

(x,Y) = (x1, x2, x3, x4, ……,Y) şeklinde gelmektedir. Burada x değerleri sistemin girdilerini, Y değeri ise sistemin elde edilmek istenen çıktı değerini ifade etmektedir. Yani Şekil 2.8' deki örnekte dikdörtgen ile gösterilen ifade Y değerini, hava (yağmurlu, bulutlu, güneşli), rüzgar (kuvvetli, hafif) ve nem (yüksek, normal) durumları ise x değerleri olarak kabul edilebilir. Bir karar ağacı, bir kök düğümden yukarıdan aşağıya doğru oluşmaktadır ve verileri benzer değerlere sahip (homojen) örnekler içeren alt kümelere bölünmektedir. ID3 algoritması, bir örneğin

homojenliğini hesaplamak için entropi ölçüsünü kullanmaktadır. Entropi kısaca düzensizlik olarak tanımlanmaktadır ve ilk olarak Claude Shannon tarafından bir veri iletişim sisteminin üç unsurdan (veri kaynağı, iletişim kanalı, alıcı) oluştuğu şeklinde ifade edilmiştir (Shannon, 1948). Eğer örnek tamamen homojen ise entropi sıfır (0), örnek eşit olarak bölünmüşse entropi bir (1) değerini almaktadır.




Şekil 2.9 Entropi (Sayad, 2020)

Shannon'un entropi denklemi;

$$Entropi(Z) = -\sum_{i=1}^{n} p_i \, log_2 \, p_i \qquad (2.1)$$

Z: Sınıf etiketleri bulunan eğitim kümesini,

$D_i$: Sınıf sayısını (i=1, ……, n)

n: Entropisi hesaplanacak durum sayısı,

$P_i$: i durumunun olasılığını ifade etmektedir. $P_i = \frac{|D_{iz}|}{|Z|}$

Tablo 2.3'de bir kişinin gün içinde yaptığı aktiviteler (etiket) ve sayıları gösterilmiştir.

Tablo 2. 3. Günlük aktiviteler ve sayıları

| | Aktivite(etiket) | | | |
|---|---|---|---|---|
| Adet | Alışveriş | Tiyatro | Oyun | Ev |
| | 6 | 1 | 1 | 2 |

Tablo 2.3'de 4 farklı sınıf etiketi ve bunlara ait toplam 10 adet kayıt bulunmaktadır. Yani Denklem 2.1'e göre n değeri 4 olmaktadır. Bir sınıf etiketin gerçekleşme olasılığı bulunurken, etikete ait örnek sayısı, veri setindeki toplam örnek

sayısına bölünerek P(alışveriş) = 6/10, P(tiyatro) = 1/10, P(oyun) = 1/10 ve P(ev) = 2/10 şeklinde hesaplanmaktadır. Olasılık sonuçları Denklem 2.1'de yerine yazıldığında;

$$Entropi = -\left(\frac{6}{10}\right) * \log_2\left(\frac{6}{10}\right) - \left(\frac{1}{10}\right) * \log_2\left(\frac{1}{10}\right) - \left(\frac{1}{10}\right) * \log_2\left(\frac{1}{10}\right) - \left(\frac{2}{10}\right) * \log_2\left(\frac{2}{10}\right) = 1.571 \quad (2.2)$$

veri setinin entropisi 1.571 olarak bulunur. Sistemin genel entropisi bulunduktan sonra her bir özniteliğe ait entropi değerleri ayrı ayrı hesaplanarak elde edilen entropi değerleri, genel entropi değerinden çıkarılmaktadır. Bu işlemin sonucunda genel entropi değerini en çok azaltan öznitelik tespit edilmektedir. Entropide meydana gelen azalış miktarı kazanç olarak görülmektedir. Farklı bir ifade ile genel entropi değerini en çok düşüren öznitelik en fazla kazanç sağlayan olmaktadır. C4.5 algoritması kazanç ölçüsünü bölme kriteri olarak kullanmaktadır.

$$Bölme\ Bilgisi_A(Z) = -\sum_{j=1}^{k} \frac{|Z_j|}{|Z|} * log_2\left(\frac{|Z_j|}{|Z|}\right) \quad (2.3)$$

Nitelik A {a1, a2, a3, ,...,ak} şeklinde k farklı değer aldığı varsayıldığında ,

$Z_j\{Z_1, Z_2, Z_3\ldots, Z_k\}$ şeklinde k parçaya bölünebilmektedir. Bu işlem sayesinde sınıflandırma için en iyi çözümün bulunması ve en iyi karar ağacının oluşturulması amaçlanmaktadır. Ayırma işlemi sonrasında ihtiyaç duyulan bilginin nasıl hesaplandığı Denklem 2.4'de gösterilmiştir.

$$Entropi_A(Z) = -\sum_{j=1}^{k} \frac{|Z_j|}{|Z|} * Entropi(Z_j) \quad (2.4)$$

Bilgi kazancı, tüm veri setinin entropi değeri ile (2.1), oluşturulmuş alt grupların entropi değerinin (2.4) arasındaki fark olarak hesaplanmaktadır.

$$Bilgi\ Kazancı(A) = Entropi\ (Z) - Entropi_A(Z) \quad (2.5)$$

A niteliği kullanılarak yapılan bölme işlemi ile karar ağacı oluşturulurken ne kadar bilgi kazancı elde edileceği Denklem 2.5'de verilmiştir.

Kazanç oranı,

$$Kazanç\ Oranı(A) = \frac{Bilgi\ Kazancı(A)}{Bölme\ Bilgisi_A(Z)} \quad (2.6)$$

şeklinde hesaplanmaktadır. Ayrımın yapılacağı nitelik belirlenirken en yüksek kazanç oranını sağlayan niteliğe bakılmaktadır.

**Karar ağaçları algoritmasının avantajları ve dezavantajları**

Avantajları:

- Karar ağaçlarında, veri ön işleme aşamasında diğer algoritmalara göre veri hazırlamak için daha az emek harcanmaktadır.
- Eksik değerler içeren veri setleri için karar ağacı oluşturulabilmektedir.
- Karar ağaçları her türlü veriye (kategorik, sürekli) rahatça uygulanabilmektedir.
- Karar ağacı yapılarında verilerin normalleştirilmesi veya ölçeklendirilmesine ihtiyaç duyulmamaktadır
- Karar ağacı modeli son derece sezgisel olduğu için anlaşılması ve açıklanması da oldukça kolay olmaktadır (Zhao ve Zhang, 2008).

Dezavantajları:

- Kararsız yapıları sayesinde veri seti üzerinde yapılan küçük değişiklikler, optimal karar ağacı yapısında büyük bir değişikliğe sebep olabilmektedir.
- Genellikle karar ağacı algoritmalarında modeli eğitmek için diğerlerine göre daha fazla süre geçmektedir.
- Tek bir karar ağacı genellikle birçok yönüyle nispeten zayıftır ve daha iyi tahmin için rastgele orman olarak adlandırılan bir grup karar ağacına ihtiyaç duyulmaktadır fakat rastgele bir ormanın yorumlanması tek bir karar ağacına göre daha karmaşıktır.
- Sınırlı sayıdaki veri setlerine uygulandığında hesaplamalar diğer algoritmalara kıyasla çok daha karmaşık olabilmektedir (Zhao ve Zhang, 2008).

**Karar ağacı algoritması**

1. Adım: Veri setinden öğrenme kümesinin oluşturulması,
2. Adım: Ağacın kökünün tespit edilmesi için veri setindeki en ayırt edici özelliklerin belirlenmesi. Bunun için Entropy kullanılarak bilgi kazancı ölçülmektedir.

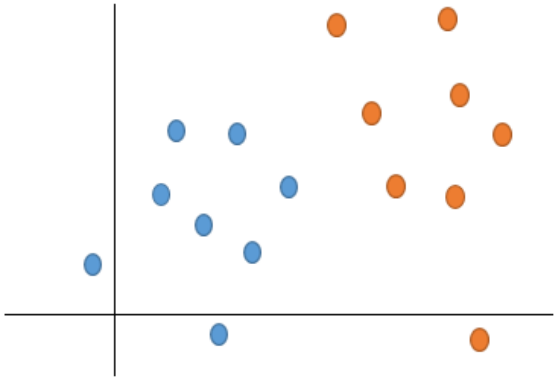
3. Adım: Ağacın alt düğümlerine ait veri kümelerinin belirlenmesi. Her alt küme için tekrar bilgi kazancı hesaplanarak en ayırt edici özellik belirlenir.
4. Adım: 3. adımda oluşturulan her alt veri kümesi için örneklerin hepsi aynı sınıfa ait ve bölecek nitelik kalmamışsa işlem sonlandırılır ve diğer durumlar için ikinci adımdan devam edilir.

### 2.4.2. K en yakın komşu algoritması

K en yakın komşu algoritması ilk olarak 1951 yılında Fix ve Hodges tarafından ortaya konulmuştur (Fix ve Hodges, 1951). Bu algoritma özellikle parametrik olmayan sınıflandırma ve regresyon problemlerinin çözümünde sıklıkla kullanılmıştır (Altman, 1992). Algoritma veri setine katılacak olan yeni verinin sınıfını belirlerken mevcut örneklem içindeki sınıfı belli olan *k* sayıdaki veri ile olan uzaklıklarını ölçerek en yakın komşunu bulmaya çalışmaktadır. Algoritmada sınıflandırma başarısını etkileyen faktörlerin en başında k değerinin seçimi gelmektedir. Ölçülendirme de $k$ sayısı belirlenirken geçmiş tecrübelere, örnekleme ve seçilen özelliklere bakılarak karar verilmektedir. Genellikle m sayıda örnek için $k$ değeri $k = \sqrt{m}$ olarak tespit edilmektedir. En iyi performansı gösteren başka bir deyişle en az hatayı veren $k$ değerini tespit etmek için birçok deneme yapılmaktadır. Algoritma uzaklık hesapları için genellikle Euclidean, Manhattan ve Minkowski gibi ölçülendirme yöntemlerini kullanmaktadır. Öklit mesafesi (Euclidean Distance), iki nokta arasındaki mesafenin Pisagor bağlantısını yardımıyla ölçülmesidir. İki boyutlu düzlemde, iki noktanın ayrı ayrı x ve y koordinatlarının hipotenüsüdür.

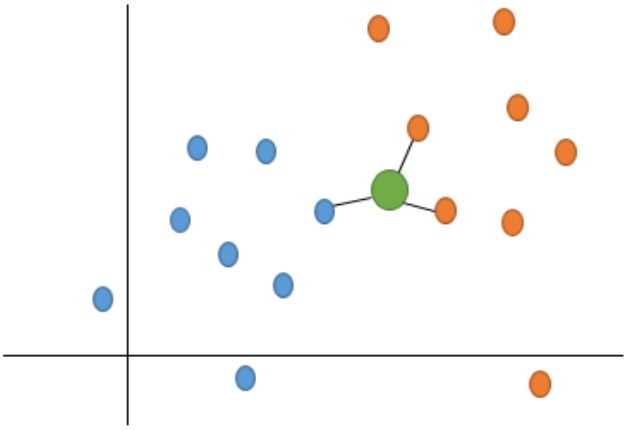
$$Fark = \sqrt{\left(X_i - X_{yeni}\right)^2 + \left(Y_i - Y_{yeni}\right)^2} \tag{2.7}$$

Şekil 2.10'da iki farklı sınıfa ait bir grup örneklem 2 boyutlu koordinat sistemi üzerine doğrusal ayrıştırma yöntemi kullanılarak yerleştirildiği varsayılırsa;



Şekil 2.10. İki boyutlu düzlemde 2 sınıf örneklem dağılımı

*k* 3 değeri için yeni gelen örneğin en yakın komşuları Şekil 2.11'de gösterilmiştir.



Şekil 2.11. En yakın 3 komşunun tespiti

Buna göre en yakın 3 örneğin ikisi turuncu diğeri mavi renkli olarak belirlenmiştir. Yeni gelen örneğin sınıfı *k* en yakın komşu algoritmasına göre turuncu olmaktadır.

**K en yakın komşu algoritmasının avantajları ve dezavantajları**

Avantajları:

- Öğrenme basit ve kolay olmasının yanı sıra eğitim süreci oldukça hızlıdır.
- Büyük ölçekli ve gürültülü eğitim verilerine karşı oldukça dirençlidir (Bhatia, 2010).

Dezavantajları:

- Yüksek miktarda bellek alanına gereksinim duymaktadır.
- Hesaplamaları karmaşık ve yavaş çalışmaktadır.

- Performansı k komşu sayısı, uzaklık ölçütü ve öznitelik sayısı gibi parametre değişmektedir (Liu ve Zhang, 2012).

***k* en yakın komşuluk algoritması**

1. Adım: Önce *k* parametresi belirlenir. Bu parametre verilen bir noktaya en yakın komşuların sayısıdır. Çalışmamızda *k* parametresi 2 olarak belirlenmiştir. Algoritma en yakın 2 komşuya göre sınıflandırma yapacaktır.
2. Adım: Veri setine dahil olacak yeni verinin, mevcut verilere göre uzaklığı tek tek hesaplanır. Çalışmamızda öklit mesafesi kullanılmıştır.
3. Adım: İlgili uzaklıklardan en yakın k komşu belirlenerek öznitelik değerlerine göre en yakın sınıfına atama yapılmaktadır.
4. Adım: Seçilen sınıf, tahmin edilmesi beklenen gözlem değerinin sınıfı olarak kabul edilmektedir.

### 2.4.3. Destek vektör makineleri

Destek vektör makineleri (DVM) istatistiksel öğrenme teorisine dayalı çoğunlukla sınıflandırma ve regresyon analizi problemlerinin çözümünde kullanılan denetimli öğrenme modellerinden biridir (Cortes ve Vapnik, 1995). DVM algoritması ilk olarak 1963 yılında istatistiksel öğrenme teorisinin ana geliştiricilerinden birisi olan Vladimir Vapnik ve ünlü matematikçi Alexey Yakovlevich Chervonenkis tarafından ortaya konulmuştur. Bu yöntem ilk olarak iki sınıflı doğrusal verilerin sınıflandırılması için kullanılmış olsa da sonrasında çok sınıflı ve doğrusal olmayan verilerin sınıflandırılmasında da kullanılmıştır. Özünde iki sınıfı birbirinden ayırabilen hiper düzlemin belirlenmesi prensibi bulunmaktadır (Vapnik, 1995). DVM son yıllarda görüntü sınıflandırma işlemleri (Vapnik, 2014), el ile yazılmış karakterlerin tanınması (Decoste ve Schölkopf, 2002), uydu verilerinin sınıflandırılması (Maity, 2016), biyolojik ve birçok farklı alandaki bilimsel çalışmalarda kullanılmıştır (Gaonkar ve Davatzikos, 2013; Cuingnet vd, 2011).

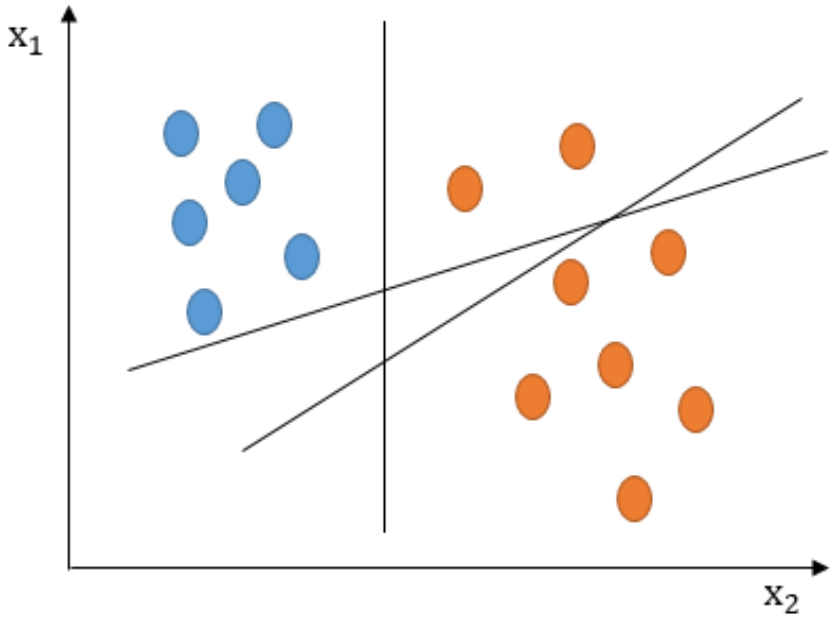
DVM' de D özelliğine ait $x_i$ ile temsil edilen her girdi $y_i = -1$ ya da $y_i = +1$ sınıflarından birisine ait olarak tanımlandığında girdilerin tamamı

$$\{x_{i,} y_i\} | i = 1 \dots n, y_i \in \{-1,1\}, x \in R^D \quad (2.8)$$

şeklinde gösterilmektedir ve bu durum farklı sınıfları birbirinden optimal olarak ayıracak doğrusal hiper düzlemin bulunmasına yardımcı olmaktadır.

$$Wx + b \tag{2.9}$$

(2.9)' da ağırlık vektörü w ile sabit değer b ile gösterilmektedir. İki grup arasındaki hiper düzlemin tek yönlü olmadığı Şekil 2.12 de gösterilmiştir.




Şekil 2.12. İki grup arasındaki çok yönlü hiper düzlem

**i)Doğrusal ayrılabilen veri setleri için DVM**

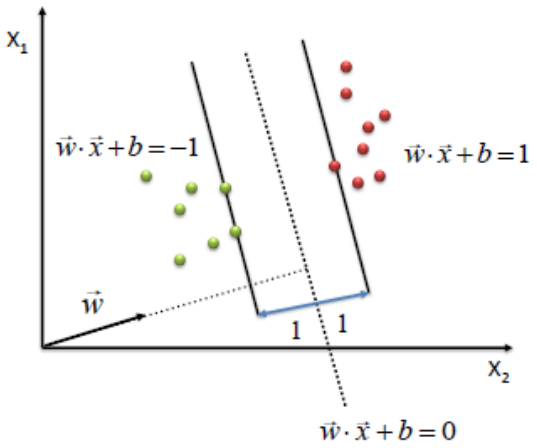
Hiper düzlemin tespit edilebilmesi için eğitim veri setinin doğrusal ayırımı gerekmektedir ve tüm örneklerin (2.10)' da eşitleri sağlaması gerekmektedir.

$$\begin{aligned} wx + b \geq +1 \quad & y_i = +1 \text{ için} \\ wx + b \leq -1 \quad & y_i = -1 \text{ için} \end{aligned} \tag{2.10}$$

Bu iki eşitliğin birleştirilmiş şekli (2.11) DVM' i meydana getirmektedir.

$$y_i(wx + b) \geq +1 \tag{2.11}$$

Optimal bir hiper düzlem tanımlamak için kenar boşluğunun (w) genişliğinin maksimize edilmesi gerekmektedir.




Şekil 2.13. Veriyi ikiye ayıran hiper düzlem ve marjinlerin uzaklığı (Sayad, 2020)

Eğitim örneğinin hiper düzleme olan uzaklığı (2.12)' de gösterilmiştir.

$$\begin{aligned} wx^{+} + b &= +1 \\ wx^{-} + b &= -1 \end{aligned} \tag{2.12}$$

Marjinler arasındaki uzaklık (2.13)' de verilmiştir.

$$d = \frac{(x^{+} - x^{-})w}{\|w\|} = \frac{\left(\left(\frac{1-b}{w}\right) - \left(\frac{-1-b}{w}\right)\right)w}{\|w\|} = \frac{2}{\|w\|} \tag{2.13}$$

Eğitim verilerinin DVM tarafından doğrusal olarak ayrılması bu mükemmel bir minimum değere ulaşıldığını göstermektedir. DVM analizinin başarısı, vektörleri örtüşmeyen iki sınıfı tamamen ayıran bir hiper düzlem üretmesine bağlıdır. Bununla birlikte her zaman mükemmel ayırma mümkün değildir. Bunun nedeni de modelin doğru sınıflandırılama yapabilmesini engelleyen farklı durumların mevcut olmasıdır. Böyle durumlarda DVM, marjı maksimize eden ve yanlış sınıflandırmaları en aza indiren hiper düzlemi bulmaya çalışır. Bu durum (2.14)' de gösterilmektedir.

$$\phi(w) = \frac{2}{\|w\|^2} \tag{2.14}$$

Bu tarz problemlerin çözümünde Lagrange çarpanları yöntemi kullanılmaktadır (Alpaydın, 2011).

$$L_P(w, b, a) = \frac{1}{2}\|w\| - \sum_{i=1}^{l} \alpha_i [\, y_i(wx + b) - 1] \tag{2.15}$$

(2.15)' de w ve b ye göre türevleri alınıp sıfıra eşitlendikten sonra Lagrange çarpanları $\alpha_i$ 'ye bağlı olarak yazılabilmektedir.

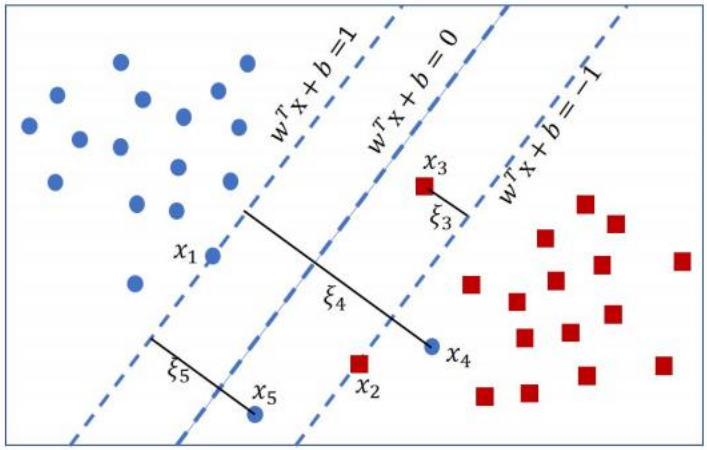
$$L_P(a) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j\, y_i y_j\, (x_i x_j) \quad \sum_{i=1}^{l} \alpha_i\, y_i = 0,\ \alpha_i > 0 \tag{2.16}$$

$$L_P(a) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j\, y_i y_j\, (x_i x_j) \quad \sum_{i=1}^{l} \alpha_i\, y_i = 0,\ \alpha_i > 0 \tag{2.17}$$

$$\mathrm{f(x)} = \mathrm{sgn}\big((\mathrm{w}x_i) + \mathrm{b}\big) = \mathrm{sign}\Big(\sum_{i=1}^{l} y_i\alpha_i(x_i x_j)\Big) \tag{2.18}$$

**ii) Belirli oranda hata ile doğrusal ayrılabilen veri kümeleri için DVM**

Çok boyutlu, gürültülü veri içeren, karmaşık veri setleri belirli bir hata ile doğrusal olarak ayrılabilmektedir (Li vd, 2009).Bu yöntemde iki sınıfı birbirinden ayırmak için gevşek sınır modelinden faydalanılmaktadır. Bu durum şekil 2.14'de gösterilmektedir.




Şekil 2.14. Belirli bir hata ile doğrusal ayrılabilme durumu (Le vd, 2018)

Şekilde 2.14' de görüldüğü gibi, gevşek sınır yaklaşımında modele, herhangi bir örneğin hatalı sınıflandırması durumunda ait olduğu karar sınırına olan uzaklığının ölçüsü olan ε aylak değişkeni eklenmektedir (Cortes ve Vapnik, 1995). Bu durumda ayırma hiper düzleminin bulunabilmesi için veri setindeki tüm örnekler (2.19) ve (2.20) deki eşitsizlikleri sağlaması gerekmektedir (Cortes ve Vapnik, 1995).

$$wx^{+} + b \geq +1 - \varepsilon \quad y_i{=}{+}1 \tag{2.19}$$

$$wx^{-} + b \leq -1 - \varepsilon \quad y_i{=}{-}1 \tag{2.20}$$

Problem, yanlış sınıflandırma olasılığını düşürmek için doğrusal ayrılma durumundaki bazı dönüşümler yapıldıktan sonra (2.21)' de gösterildiği gibi kareli optimizasyon problemine dönüştürülmektedir (Cortes ve Vapnik, 1995).

$$\phi(w) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\varepsilon_i \tag{2.21}$$

Kurulan modelde C katsayısı ceza parametresi olarak tanımlanmaktadır ve Lagrange çarpanının alabileceği üst sınır değerini gösteren ceza parametresini ifade etmektedir (Katagiri ve Abe, 2006).

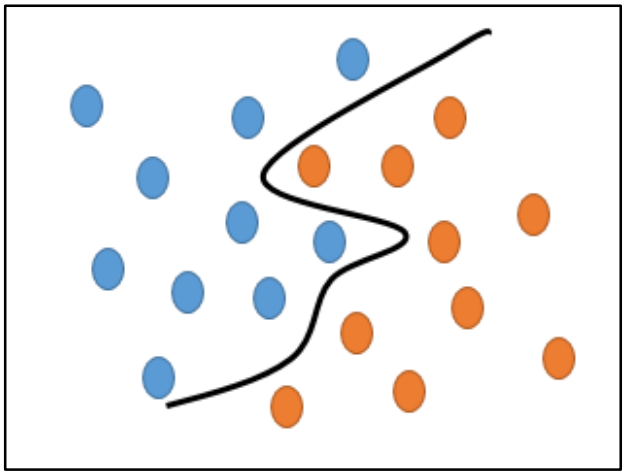
$$y_i(wx_i + b) - 1 + \varepsilon_i \geq 0 \tag{2.22}$$

Kareli optimizasyon probleminin çözümünde Lagrange fonksiyonu kullanılmaktadır (Cortes ve Vapnik, 1995). Problemle ilişkili modelinin çözümü sonucunda elde edilen hiper düzleme bağlı olarak elde edilen sınıflandırıcı (2.23)'de verilmiştir (Burges, 1998).

$$\mathrm{f(x)} = \mathrm{sgn}\big(\mathrm{(\,w}x_i) + \mathrm{b}\big) = \mathrm{sign}(\sum_{i=1}^{l} y_i\alpha_i(x_i x_j)) \tag{2.23}$$

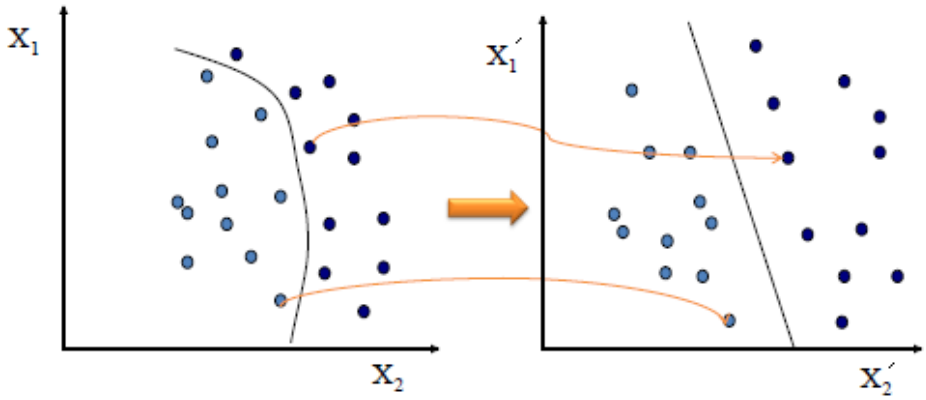
**iii)Doğrusal olmayan veri kümeleri için DVM**

Gerçek hayatta karşılaşılan bazı problemlerin analizi için oluşturulan veri setlerini her zaman doğrusal bir fonksiyonla veya belirli bir hata ile ayırmak mümkün olmamaktadır. Böyle durumlarda, doğrusal olamayan DVM algoritmaları kullanılmaktadır. Bu algoritmalar sınıfları ayırmak için bulunması oldukça zor olan sınıflandırma eğrisini tahmin etmeye çalışmaktadır. Bu durum Şekil 2.15' de gösterilmektedir. Bu gibi durumlarda p-boyutlu girdi vektörü x' in P boyutlu özellik vektörü Φ' ye dönüştürülmesi gerekmektedir (Cortes ve Vapnik, 1995).



Şekil 2.15. Doğrusal ayrılmama durumu

Bu işlem için doğrusal olmayan görüntüleme tekniği kullanılarak optimal ayırma düzlemi özellik uzayında tanımlanmaktadır (Busuttil, 2003). Doğrusal olmayan bir işlevin yüksek boyutlu bir özellik uzayında doğrusal bir öğrenme makinesi tarafından öğrenilmesi Şekil 2.15' de gösterilmektedir.




Şekil 2.16. Doğrusal olmayan görüntüleme tekniği (Sayad, 2020)

Bu işlem çekirdek numarası adı verilen, sistemin boyutluluğuna bağlı olmayan bir parametre tarafından kontrol edilmektedir. Bu çekirdek fonksiyonu, doğrusal ayırmanın gerçekleştirilmesini yapabilmek için verileri daha yüksek boyutlu bir özellik uzayına dönüştürmektedir.

Verinin iç çarpımının görüntüsü, verilerin görüntülerinin iç çarpımıdır. Bu durum doğrusal DVM için $X_i.X_j$, doğrusal olmayan DVM için $\phi(X_i).\phi(X_j)$ ve çekirdek fonksiyonu için $k.(X_i.X_j)$ şeklinde ifade edilmektedir.

Tablo 2. 4. Bazı çekirdek fonksiyonları

| | |
|---|---|
| Polinom çekirdek fonksiyonu | $k.(X_i.X_j) = (X_i.X_j)^d$ |
| Radyal tabanlı çekirdek fonksiyonu | $k.(X_i.X_j) = \exp(-\frac{\|X_i, X_j\|^2}{2\sigma^2})$ |

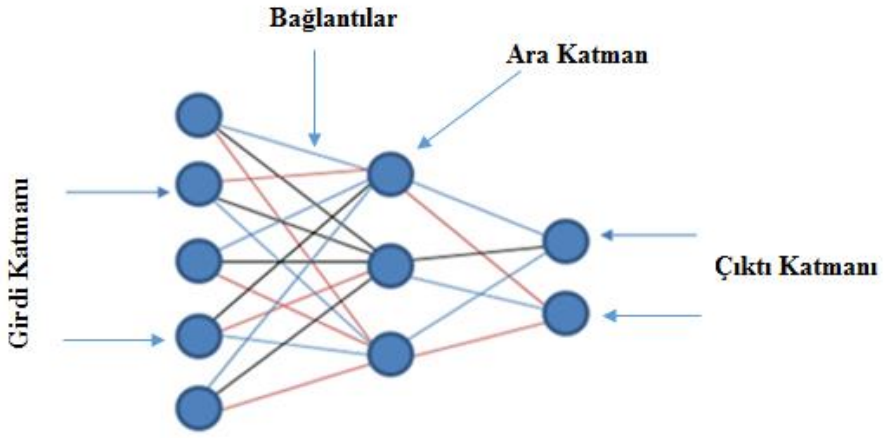
**Destek vektör makinesi algoritması**

1. Adım: İki sınıfı birbirinden ayırabilen hiper düzlemin belirlenmesi
2. Adım: Hiper düzlemin tespit edilebilmesi için eğitim veri setinin doğrusal ayrılıp ayrılmadığının test edilmesi ve eğitim örneğinin hiper düzleme olan uzaklığının hesaplanması
3. Adım: İki grup arasındaki hiper düzlemin tek yönlü olmadığı durumlarda optimal bir hiper düzlem tanımlamak ve yanlış sınıflandırma olasılığını düşürmek için kenar boşluğu genişliğinin maksimize edilmesi
4. Adım: Eğer veri seti doğrusal bir fonksiyonla veya belirli bir hata ile ayırmak mümkün değilse bu işlem için doğrusal olmayan görüntüleme tekniği kullanılması

### 2.4.4. Yapay sinir ağları

Yapay sinir ağları(YSA), insan beyninin ve sinir sisteminin davranışlarını taklit eden geliştirilmiş bir bilgi işleme tekniğidir (Chen vd, 2019). YSA bilgiyi öğrenerek elde etmektedir. Öğrendiği bilgilerden yeni bilgiler üretebilmektedir. Keşfedebilme yeteneği sayesinde dışardan gelen etkilere insan davranışlarına benzer tepkiler verebilmektedir. İlişkilendirme, sınıflandırma ve optimizasyon problemlerinin çözümünde sıklıkla kullanılmaktadır (Öztemel, 2003). YSA hakkında ilk bilgiler 1890 yıllarda William James tarafından ortaya atılmıştır. 1940 yıllara kadar bazı bilim adamları tarafından çok fazla mühendislik değeri olmayan çalışmalar yapılmıştır. İlk yapay sinir ağı modeli, 1943 yılında Warren McCulloch ve Walter Pitts tarafından geliştirilmiştir ve bir sinir hücresinin matematiksel modelini ifade etmektedir. YSA 1950 ve 1960' lı yıllarda, yapay zekâ kavramı ile birlikte anılmaya başlamıştır. YSA üzerine yapılan araştırmalar bir ara duraksasa da 1980' li yıllarda tekrar hız kazanmaya başlamıştır. Aynı yıllarda Hopfield isimli bilim adamı yapay sinir ağlarının bilgisayar programlama teknikleri kullanarak birçok alanda çözülmesi zor olan problemlerin

çözümünde kullanılabileceğini göstermiştir. Danışmansız ve geriye yayılmalı öğrenme modellerindeki başarısı, gezgin satıcı probleminin çözülmesine olan katkısı, çok katmanlı algılayıcıların ortaya çıkmasını sağlamıştır. Bilgisayar teknolojilerindeki gelişmelerle birlikte 1990' lı yıllardan günümüze kadar yapay sinir ağları, günlük hayatta kullanılan birçok sisteme entegre edilerek insanlar için daha faydalı hale gelmeye başlamıştır. YSA üç katmandan meydana gelmektedir ve bu katmanların bir araya gelmesinden sinir ağı oluşmaktadır (Öztemel, 2012).




Şekil 2. 17. Yapay sinir ağı görünümü

Girdi katmanı, dışardan girdileri alan nöronlardan (sinirler) oluşmaktadır. Bu katmandaki nöronların görevi, girdi değerlerini bir sonraki katmana iletmektir. Girdiler, yapay sinir hücresine dışardan alınan verileri ifade etmektedir. Bu veriler ağın öğrenmesinde kullanılmaktadır. Girdiler, (x),n elemanlı sütun vektörü şeklinde gösterilmektedir.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{2.24}$$

Girdi katmanı tarafından iletilen sinyaller ara katmanda işlenmektedir. Ağırlık(w) olarak ifade edilirler ve gelen verilerin hücre üzerindeki etkisini belirleyen değerlerdir. Ağırlıklar, n elemanlı satır vektörü olarak gösterilmektedir.

$$W = [w_1 \cdots w_n] \tag{2.25}$$

Hücreye giren net girdi, girişlerin ilişkin olduğu ağırlıklarla çarpımlarının toplamı şeklinde hesaplanmaktadır.

$$Net\ Girdi = \sum_{i=1}^{n} w_i\, x_i \tag{2.26}$$

Elde edilen toplam ağırlık, transfer fonksiyonu yardımı ile bir sonraki ara katmana veya çıktı katmanına iletmektedir. Transfer fonksiyonu,

$$y = F(x) \tag{2.27}$$

şeklinde gösterilmektedir. Ağ tasarımcısı denemelerden elde ettiği tecrübelere göre, probleme uygun transfer fonksiyonun seçimine karar vermektedir. Yapay sinir hücre modellerinde genellikle doğrusal, adım, eşik, sigmoid, hiperbolik tanjant gibi transfer fonksiyonları tercih edilmektedir. Bu fonksiyonlar,

Doğrusal fonksiyon: $F(x) = x$

Adım Fonksiyonu: $y = F(x) = \begin{cases} 1\ x > 0 \\ -1\ x \leq 0 \end{cases}$

Eşik Fonksiyonu: $y = F(x) = \begin{cases} 1\ x > 0 \\ 0\ x \leq 0 \end{cases}$ (2.28)

Sigmoid fonksiyonu: $y = F(x) = \dfrac{1}{1 + e^{-x}}$

Hiperbolik Tanjant fonksiyonu: $y = F(x) = \dfrac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$

şeklinde gösterilmektedir.

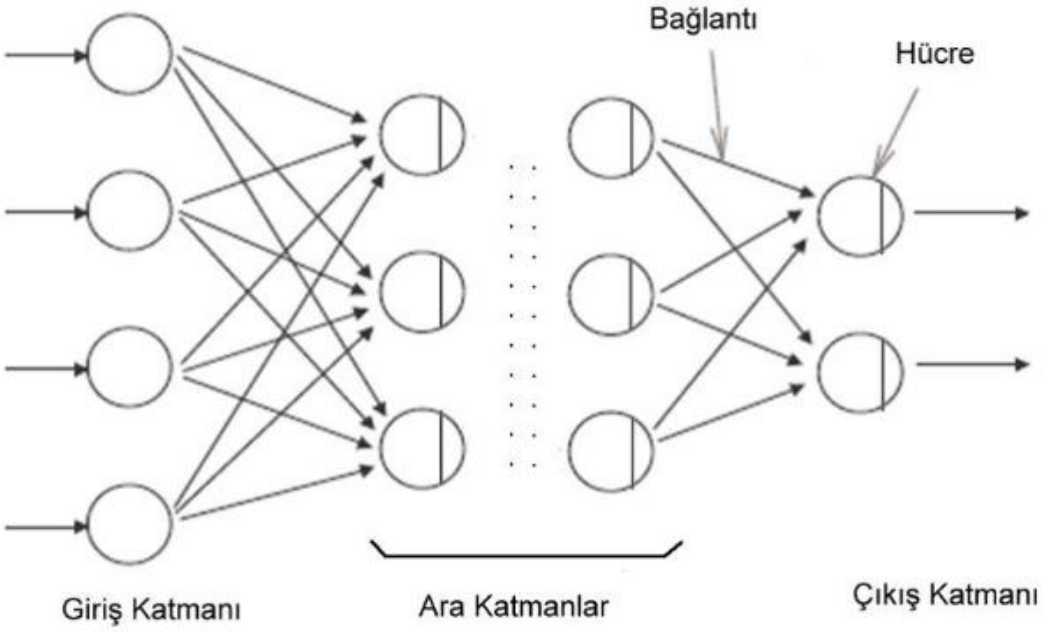
Problemin çeşidine göre ara katmanların ve katmanlardaki nöronların sayısına ağın tasarımcısı karar vermektedir. Bu sayıların azlığı bazen ağın çıktıları için beklenilen tahmin doğruluğunu, fazla olması durumunda ise yeni girdi değerleri için tahmin doğruluğunu azaltabilmektedir. Ağın en ucunda çıktı katmanı yer almaktadır. Ara katmandan aldığı verileri dış ortamlara ya da diğer ağlara ileten nöronlardan oluşmaktadır. Bu katman, yapay sinir ağının problemin sonucuna ait değerlerini tutmaktadır (Kargı, 2013).

#### 2.4.4.1. Mimari yapılarına göre YSA sınıflandırılması

Yapay sinir ağları mimari yapılarına göre ileri beslemeli (feedforward) ve geri beslemeli (feedback) olarak iki gruba ayrılmaktadır. Bir katmandaki nöronlar sadece kendinden sonraki katmana veri iletmektedir.

### i) İleri beslemeli yapay sinir ağları

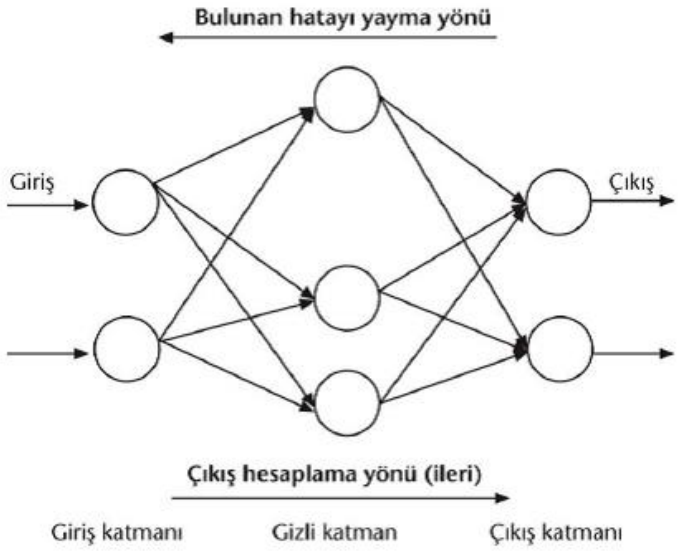
İleri beslemeli YSA nöronlar arasında hiyaraşik bir yapı bulunmaktadır. Bu ağda sinyaller, girişten çıkışa doğru tek yönlü olarak hareket etmektedir. Giriş katmanı, dış ortamdan aldığı bilgileri üzerinde hiçbir değişiklik yapmadan ara katmandaki (gizli katman) hücrelere iletmektedir. Nöronlar katmanlar arasında iletişim kurarken aynı katmandaki nöronlar arasında bir bağlantı olmamaktadır. Bu yüzden nöronlar arasında döngüsel bir oluşum meydana gelmemektedir. Bu durum ağın hızını artırmaktadır (Graupe, 2013). Üç katmanlı ileri beslemeli sinir ağının bir örneği Şekil 2.18'de gösterilmiştir.




Şekil 2. 18 İleri beslemeli yapay sinir ağı görünümü ( Öztemel, 2012)

### ii) Geri beslemeli yapay sinir ağları

Geri beslemeli yapay sinir ağlarında en az bir hücrenin çıkışı kendisine ya da diğer hücrelere giriş olarak verilmektedir. Genellikle geri besleme bir geciktirme elemanı üzerinden yapılmaktadır. Geri besleme, bir katmandaki hücreler arasında olduğu gibi katmanlar arasındaki hücreler arasında da olabilmektedir. Bu yapısı ile geri beslemeli YSA, doğrusal olmayan dinamik bir davranış göstermektedir. Bu yüzden, geri beslemenin yapılış şekline göre farklı yapıda ve davranışta geri beslemeli YSA yapıları elde edilebilmektedir. Şekil 2.19'da iki katmanlı ve çıkışlarından giriş katmanına geri beslemeli bir YSA yapısı görülmektedir (Kabalcı, 2015).




Şekil 2.19 Geri beslemeli yapay sinir ağı görünümü (Kabalcı, 2017)

### 2.4.4.2. Yapay sinir ağı tasarımı

Bir sinir ağı modeli meydana getirmek için nöronların bağlantı şekilleri, oluşturdukları katman sayısı ve katmanlar arasındaki veri iletim şekli, kullanılacak öğrenme yöntemi ve algoritmasının belirlenmesi gerekmektedir. Oluşturulan ağın yapı ve mimarisi, ağın fonksiyonelliğini ve performansını önemli ölçüde etkilemektedir. Bunun için YSA tasarımcısının başlangıçta, çözülmek istenen probleme göre uygun ağ modelini belirlemesi gerekmektedir. Bunlar arasında genellikle ağda kullanılan katman sayısı (tek katmanlı, çok katmanlı), öğrenme algoritması (danışmanlı, danışmansız), öğrenme kuralı, iletişim yönü (ileri beslemeli, geri beslemeli) önemli yer tutmaktadır. Doğru verilen kararlar yapay sinir ağının daha hızlı ve başarılı sonuçlar üretmesini sağlayacaktır. Bunun için aşağıdaki işlem adımları takip edilmektedir.

i) Öncelikli olarak probleme uygun ağın şekli (topolojisi) belirlenmektedir. Bu da büyük ölçüde ağda kullanılması düşünülen öğrenme algoritması ile alakalıdır. Girdilerin hangi sınıfa ait olduğunun belirlenmek istendiği tahmin-öngörü, sınıflandırma gibi problemlerin çözümünde genellikle tek veya çok katmanlı algılayıcıların yer aldığı ağ mimarisi (topolojisi) tercih edilmektedir.

ii) Katmanlar aynı doğrultudaki işlemci elemanlardan meydana gelmektedir. Doğru katman sayısını belirlemek için birkaç deneme yaparak ağ performansı ölçülmektedir.

iii) Ağ için önemli unsurlardan biri de işlemci elaman (nöron) sayısının belirlenmesidir. Bu da tıpkı katman sayılarının belirlenmesinde olduğu gibi birkaç deneme ile ağ performansı ölçülerek yapılmaktadır. Bunun için atılması gereken adım başlangıçtaki nöron sayısını istenilen performansa ulaşıncaya kadar arttırmak veya tam tersi arzu edilen performansa ulaşılıncaya kadar azaltmaktır. İşlemci elamanın sayısının yetersiz seçilmesi ağın öğrenme yeteneğinin, fazla seçilmesi ise genelleme yeteneğinin azalmasına sebep olmaktadır. Birçok problem için iki veya üç katmanlı bir ağ iyi sonuçlar üretebilmektedir.

iv) YSA tasarımında işlemci elemanlarının karakteristik özelliklerinin belirlenmesi önem arz etmektedir. Verilerin özelliklerine göre toplama ve transfer fonksiyonun seçilmesi gerekmektedir. Bu iş için genellikle sigmoid, hiperbolik tanjant ve doğrusal fonksiyonlar kullanılmaktadır (Bayır, 2006).

v) YSA genellikle verilerin doğrusal olmaması durumlarında tercih edilmektedir. Bu yüzden veriler ağ ile iletişime geçmeden önce normalizasyon işlemi yapılmaktadır. Sistem performansını artırmak için veriler [0,1] ya da [-1,1] aralıklarından herhangi birine ölçeklendirilmektedir (Saraç, 2004).

vi) Öğrenme katsayısı ve momentumun seçimi de diğer önemli faktörlerden birisidir ve öğrenme sürecinde ağırlıklardaki değişim, Öğrenme Oranı/Katsayısı (λ) şeklinde ifade edilmektedir. Genellikle bu oran salınıma (oscillation) sebebiyet vermeyecek kadar büyük alınmaya çalışılmaktadır. Öğrenme katsayısı genellikle [0,1] aralığında tercih edilmektedir. Öğrenme oranı büyük seçilmesi durumunda salınıma sebep olur ve ağın minimum değere ulaşması güçleşir tersi durumda ise ağın öğrenme süresini uzatmaktadır.

vii) Performans fonksiyonun, YSA öğrenme performansı ölçmek için kullanılmaktadır. Bunun için ileri beslemeli ağlarda genellikle hata kareler ortalaması (Mean Square Error-MSE) kullanılmaktadır. Denklem 2.29' da gösterilen bu fonksiyon istenen çıktı değerleri ile ağ tarafından üretilen değerler arasındaki farkın kareleri toplamının ortalamasını hesaplamaktadır. Bu denklemde $y_i$ istenen (gerçek) çıktı değerlerini, $\hat{y}_i$ ağ tarafından üretilen çıktı değerlerini ve n veri sayısını göstermektedir (Kargı, 2013).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2.29)$$

Hata değerinin sıfıra yaklaşması, arzu edilen çıktı değerlerine yaklaşıldığı anlamına gelmektedir. Bu amaçla kullanılan diğer performans fonksiyonları arasında hata kareler ortalamasının karekökü (Root Mean Square Error-RMSE) ve ortalama mutlak hata (Mean Absolute Error- MAE) yer almaktadır.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2.30)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2.31)$$

**YSA algoritmasının avantajları ve dezavantajları**

Avantajları:

- YSA eğitim aşamasında kendisine verilen örneklerden veriler arasındaki ilişkileri sezgisel olarak çok hızlı öğrenir ve gerçekleştirdiği öğrenmeler sayesinde genellemeler yapabilmektedir.
- YSA paralel işlem yapabilme yetenekleri sayesinde aynı anda birçok görevi yerine getirebilirler.
- YSA eksik ve hatalı bilgi içeren verileri işleyerek çıktı üretebilirler.
- YSA doğrusal olmayan problemlerin çözümünde yüksek başarı oranına sahiptirler.
- YSA bir veya birkaç nöronun zarar görmesine rağmen sonuç üretebilirler ve bu gibi durumlardan geleneksel bilgi işleme teknikleri kadar olumsuz etkilenmezler.

Dezavantajları:

- YSA probleme uygun ağ yapısının belirlenmesinde belirli bir kural yoktur. Uygun ağ yapısını elde etmek için deneyim gerekmektedir.
- YSA çalışmalarında sayısal veriler kullandığı için problemler ağa tanıtılmadan önce sayısal verilere dönüştürülür. Kullanıcının yeteneğine bağlı olarak tercih edilecek gösterimin ağın performansı üzerindeki etkisi yüksek olacaktır.

- Tercih edilen ağın eğitim süresinin önceden kestirimi mümkün olmamaktadır. Ağın hata değerlerinin azaltılması sayesinde tamamlanan eğitimler, optimum sonuçlar üretmezler.
- YSA' nın problemlerin çözümüne dair açıklayıcı bilgi paylaşmaması ağa olan güveni azaltmaktadır.

**Yapay sinir ağı algoritması**

1. Adım: Yapay sinir ağının girdi değerlerinin ve girdi değerlerine karşılık gelen çıktı değerlerinin gösterilmesi
2. Adım: Modele dahil edilen net girdi,

$$Net\ Girdi = \sum_{i=1}^{n} w_i\, x_i$$ formülü ile hesaplanır

3. Adım: Modelin çıktısı hesaplanır. Modelin çıktı değeri 1 veya 0' dır. Eğer net girdi eşik değerinden büyükse çıktı 1, küçükse 0 değerini almaktadır. Gerçekleşen çıktı ile beklenen çıktı değerleri karşılaştırılır.
4. Adım: Eğer gerçekleşen çıktı ile beklenen çıktı değerleri aynı ise ağrılıklarda değişiklik yapılmaz. Beklenen çıktı değeri 0, gerçekleşen çıktı değerinin 1 olması durumunda ağrılık değerleri azaltılmaktadır. Bu,

$$W_n = W_0 - \lambda X$$ vektörü ile hesaplanır.

Burada λ öğrenme katsayısını, $W_n$ yeni ağırlık vektörünü, $W_0$ ise eski ağırlık değerini ifade etmektedir. Eğer beklenen çıktı değeri 1, gerçekleşen çıktı değerinin 0 ise bu defa ağrılık değerleri artırılmaktadır.

$$W_n = W_0 - \lambda X$$

5. Adım: Tüm girdi kümesindeki örnekler için doğru sınıflandırmalar gerçekleşinceye kadar bu adımlar tekrarlanmaktadır.

### 2.4.5. Lojistik Regresyon

Lojistik regresyon (LR) en az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ve bağımsız değişkenler arasındaki ilişkiyi tanımlayabilmek için bir model oluşturmak olarak tanımlanmaktadır (Çokluk, 2010). LR alanında dikkat çeken ilk çalışmalar Berkson tarafından 1944 yılında yapılmıştır. Cox ve Andersson 1970 ve 1980' li yıllarda birçok bilimsel çalışmada LR kullanmıştır ve verilerin lojistik modele uyumu araştırılmıştır (Bircan, 2004). Lojistik regresyon modellerinin yaygın bir şekilde kullanılır hale gelmesinde en büyük etken doğrusal regresyon için geçerli

varsayımların hiçbirinin LR' da aranmamasıdır. LR ile doğrusal regresyon arasındaki en önemli farklardan birisi de LR' da bağımlı değişken kategoriktir (Aktaş, 2009). LR analizi sınıflandırma çalışmalarında başarılı sonuçlar vermektedir (Girginer ve Cankuş, 2008). LR ile doğrusal regresyon analizi yöntemleri arasında üç önemli fark bulunmaktadır.

i) Doğrusal regresyon analizinde kestirilecek olan bağımlı değişken sürekli değerler alırken, LR analizinde bağımlı değişken kesikli değer almaktadır.
ii) Doğrusal regresyon analizinde bağımlı değişkenin değeri tahmin edilmeye çalışılırken, LR analizinde bağımlı değişkenin alabileceği değerlerden birinin gerçekleşme olasılığı tahmin edilmeye çalışılmaktadır.
iii) Doğrusal regresyon analizinde bağımsız değişkenin çoklu normal dağılım göstermesi şartı aranırken, LR analizinde değişkenlerin dağılımı ile ilgili hiçbir ön şart bulunmamaktadır (Elhan, 1997).

LR analizi üç farklı şekilde gerçekleştirilebilmektedir. Bağımlı değişkenin iki kategorili olması durumunda, İkili Lojistik Regresyon (Binary Logistic Regression), bağımlı değişkenin ikiden fazla kategorili ve sıralanabilir (ordinal) olması durumunda, Sıralı Lojistik Regresyon ( Ordinal Logistic Regression), bağımlı değişkenin ikiden fazla kategorili ve sırasız (multinominal) olması durumunda Çok Kategorili Lojistik Regresyon ( Multinominal Logistic Regression) kullanılmaktadır (Karabulut ve Alpar, 2011).

#### 2.4.5.1. İkili lojistik regresyon

Bağımlı değişkenin iki kategorili olduğu durumları ifade eden regresyon modelidir (Alpar, 2013). Lojistik regresyon modellerinde temel kavram lojittir. Bu modelde üstünlük oranı (odds ratio) hesaplanmaktadır ve Lojit Odds oranının doğal logaritması olarak tanımlanmaktadır Üstünlük oranı araştırılan iki farklı durumun odds katsayılarının birbirine oranı şeklinde gösterilmektedir (Gujarati, 1999).

$$O = \frac{P}{1-P} \tag{2.32}$$

Burada P ile bir olayın olasılığını, O ise olayın üstünlüğünü ifade etmektedir. Bu oran bağımlı değişkenin bağımsız değişkenler üzerine etkisini açıklamakta kullanılmaktadır (Mertler ve Vannatta, 2005). Üstünlüğü 1' in altında olan olayların olasılığı 0.5' in altında olmaktadır. Denklem 2.33 'de Log-olasılık dönüşümüyle modellenen lojistik regresyon fonksiyonu gösterilmiştir.

$$Lojit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n \quad (2.33)$$

Olasılık ($p_i$) 0,5'in altında değerler aldığında Lojit($p_i$) negatif, olasılık($p_i$) 0,5'in üzerinde değerler aldığında Lojit($p_i$) pozitif değerler almaktadır. Olasılık değeri yükseldikçe Lojit($p_i$) değeri artmaktadır. Burada $x_i$'ler sürekli veya ikilik açıklayıcı değişkenler olduğu durumlarda, $\beta_i$' ler maksimum olabilirlik yöntemleri kullanılarak tahmin edilen regresyon katsayılarıdır (Hosner ve Lemeshow, 1989).

Üstünlük oranı, lojistik regresyon kullanılan çalışmalarda bağımsız değişkenlerin etkilerinin araştırılmasına, ikili değişkenler arasındaki ilişkiler için güven aralıkları yardımıyla tahminde bulunulmasına ve vaka kontrol çalışmalarında özel ve uygun bir değerlendirme yapılmasına olanak sağlamaktadır (Bland ve Altman, 2000). Her bir değişken katsayısının modelde istatistiksel olarak anlamlı olup olmadığına (2.34, 2.35) denklemleri kullanılarak Wald istatistiği hesaplanarak bakılmaktadır (Bircan, 2004).

$$t = \frac{\hat{\beta}}{SH_{\hat{\beta}}} \quad (2.34)$$

$$t^2 = Wald \quad (2.35)$$

#### 2.4.5.2. Sıralı lojistik regresyon

Sıralı lojistik regresyon analizi sıralı (ordinal) bir bağımlı değişken ile bir veya birden çok bağımsız değişken arasındaki ilişkiyi belirlemek için kullanılmaktadır (Özdamar, 2013). Bu yöntemde bağımlı değişken en az üç kategoride olmak şartıyla sıralama küçükten büyüğe doğru olmaktadır (Şerbetçi, 2013). Sıralı lojistik regresyon modeli gözlemlenebilir kategorik Y bağımlı değişkeninin altında gözlemlenemeyen bir $Y^*$gizli değişkenin olduğu varsayımına dayanmaktadır (McCullagh, 1980).

Sıralı lojistik regresyon modeli genel olarak,

$$Y^* = \beta_i * x_i + u_i \quad (2.36)$$

şeklinde ifade edilmektedir. Kesme parametresi bağımlı değişken ile gizli değişken arasındaki ilişkiyi belirtmektedir.

$$\begin{aligned} Y = 1 \; iken \; Y^* &\leq \propto_1 \\ Y = 2 \; iken \; \propto_1 &< Y^* \\ &\leq \propto_2 \end{aligned} \quad (2.37)$$

$$Y = J\ iken\ \propto_{j-1} < Y^*$$

Denklem 2.37' de $\propto$ kesme noktasını göstermektedir. Kesme noktaları $\beta_i$'lerle tahmin edilmektedir ve $0<\propto 1<\propto 2<\ldots<\propto(j-1)$ şeklinde pozitiflik sınırlandırması bulunmaktadır (Akın ve Şentürk, 2012). J sayıda ordinal kategoride Odds oranı referans alınarak model Denklem 2.38' de genel olarak gösterilmiştir.

$$\ln(Y_j) = \propto_j - (\beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_i * x_i) \quad (2.38)$$

Sonucun güvenilirliği, katsayıların anlamlılığı açısından sıralı lojistik modelinde paralel eğriler varsayımının sağlanması gerekmektedir (Ayhan, 2006). Bu varsayım Wald $\chi^2$ gibi olabilirlik oran testi kullanılarak test edilebilmektedir. Bu modelde ilişkili regresyon katsayılarının sonuç değişkeni üzerindeki tüm kategorilerde aynı olması beklenmektedir.

#### 2.4.5.3. Çok kategorili lojistik regresyon

Çok kategorili lojistik regresyon analizinde ikiden fazla kategorik bağımlı değişken ile birçok bağımsız değişken arasındaki ilişki incelenmektedir. Bu analizde modeldeki katsayıların kestirimi ve anlamlılıkları test edilmektedir. Modeli oluşturmak için x' i p açıklayıcı değişken ve sabit terimi içeren p+1 uzunluğunda vektör olduğu varsayıldığında çoklu lojistik regresyon modelinin logit fonksiyonları denklem 2.39' ve denklem 2.40' daki gibi gösterilmektedir (Alpar, 2013).

Referans olarak kategori 0 alındığında, kategori 1' in kategori 0' a göre,

$$g_1(x) = ln\left[\frac{(p(y=1)|x)}{(p(y=0)|x)}\right] = \beta_{10} + \beta_{11} * x_1 + \beta_{12} * x_2 + \cdots + \beta_{1p} * x_p \quad (2.39)$$

kategori 2' nin kategori 0' a göre,

$$g_2(x) = ln\left[\frac{(p(y=2)|x)}{(p(y=0)|x)}\right] = \beta_{20} + \beta_{21} * x_1 + \beta_{22} * x_2 + \cdots + \beta_{2p} * x_p \quad (2.40)$$

Ortak değişken vektöründe bağımlı değişkenin her bir kategorisinin koşullu olasılığının hesaplanması denklem 2.41' de gösterilmiştir.

$$p(Y=j|x) = \frac{e^{g_j(x)}}{\sum_{k=0}^{2} e^{g_k(x)}} \quad (2.41)$$

x=a, x=b değerleri için bağımlı değişkenin Y=j kategorisinin referans kategori Y=0' a göre Odds oranının bulunması denklem 2.42' de gösterilmiştir.

$$OR_j(a,b) = \frac{p(Y=j|x=a)/p(Y=0|x=a)}{p(Y=j|x=b)/p(Y=0|x=b)} \tag{2.42}$$

# 3. MATERYAL VE YÖNTEM

Bu çalışma DEA etki eden risk faktörlerini kabul edilebilir bir hassasiyetle tahmin etmeyi amaçlamaktadır. Günümüzde bilgisayar ve özellikle yapay zekâ alanındaki yaşanan teknolojik gelişmeler sağlık alanındaki problemlerin çözümünde yoğun olarak kullanılmaktadır (Khan vd, 2019; De ve Chakraborty, 2020). Yapay zekânın bir alt dalı olan makine öğrenmesi, önceki gözlemlerden yararlanılarak doğru tahminler yapabilmek amacıyla geliştirilmiş sistematik tekniklerden oluşmaktadır (Schapire, 2003). Bu çalışmada istatistik, veri tabanı ve makine öğrenmesi gibi disiplinler arasında uyumlu bir ilişki kurularak aynı veri setine ait iki uygulama yapılmıştır.

Uygulama I' de DEA' ne etki eden risk faktörleri geriye dönük (retrospektif) olarak incelenmiştir. Birinci adımda seçilen parametrelerin ayrı ayrı hastalık gruplarına etkilerinin araştırıldığı tek değişkenli istatiksel analizler yapılmıştır. İkinci adımda sadece iki değere sahip olabilecek bir sonucun olasılığını öngörmek için anlamlı bulunan tüm parametrelerin modele katıldığı lojistik regresyon çok değişkenli istatiksel analizi yapılmıştır.

Uygulama II' de, güvenilirliği yüksek, daha kolay yorumlanabilen bir hesaplama modeli oluşturulması amaçlanmaktadır. Bu nedenle istatistiksel analiz yöntemleri, yeni veri madenciliği ve makine öğrenmesi teknikleri ile uyum içerisinde kullanılmıştır. Bu çalışmada bilgisayar yardımıyla tıbbi ve sağlık alanlarındaki büyük miktarlardaki verileri daha kullanılabilir hale getirilmektedir. Verileri analiz etmek için çok sayıda veri madenciliği algoritması ve araçları bulunmaktadır. Veri madenciliği için Weka 3.8 yazılımı kullanılmıştır. Verileri analize hazır gale getirmek için bir dizi işlem yapıldıktan sonra, Karar Ağacı (J.48), Destek Vektör Makinesi, Yapay Sinir Ağları ve K en Yakın Komşu algoritmaları kullanılarak ve performans ölçütleri karşılaştırılmıştır. Sınıflandırma sonuçlarının doğruluğunu karşılaştırmak için temel bileşenler analizi, tahmin gücünü analiz etmek için çapraz doğrulama yöntemleri kullanılmıştır. Amacımız; yeni makine öğrenme tekniklerinin, demir eksikliği anemisi tanısının değerlendirilmesi ve doğrulanması, dijitalleştirilmiş bir eşik uyarısı geliştirilmesine yardımcı olmaktır.

### 3.1. Uygulama I

Ekim 2017-Mart 2020 tarihleri arasında, Sağlık Bilimleri Üniversitesi Samsun Eğitim ve Araştırma Hastanesi Hematoloji Polikliniğine başvuran malaise and fatigue (ICD-10 kodu: R53) tanısı almış 516 vaka retrospektif olarak incelendi. Hastalardan 359 kişiye laboratuvar sonuçları ile DEA tanısı konulmuştur. Kalan 157 vaka da aynı tanı ile değerlendirilip, laboratuvar değerleri DEA ile uyumlu çıkmamıştır. Bu 516 vakanın yaş, cinsiyet, tam kan sayımı (CBC) değerleri, hemoglobin (Hb), hematokrit (Hct), ortalama hücre hacmi (MCV), ortalama hücre Hb konsantrasyonu (MCHC), kırmızı kan hücresi dağılım genişliği (RDW), kırmızı kan hücresi (RBC) ve DEA parametreleri (ferritin, serum demir, serum demir bağlama kapasitesi, transferrin saturasyonu) kaydedildi. Yukarıdaki kısaltmalar Tablo 3.1' de gösterilmiştir.

Tablo 3. 1. Bu çalışmada kullanılan laboratuvar testi kısaltmalarının listesi

| **Laboratuvar Testi** | **Kısaltması** |
|---|---|
| Hemoglobin | Hb |
| Hematocrit | Hct |
| Mean Corpuscular Volume | MCV |
| Mean Corpuscular Hemoglobin Concentration | MCHC |
| Red Blood Cell Count | RBC |
| Red Blood Cell Distribution Width | RDW |
| İron | FE |
| Unsaturated Iron Binding Capacity | UIBC |
| Ferritin | FERR |
| Hastalık Tanısı | HT |

### 3.2. İstatistiksel analizler

Bu çalışmanın istatiksel analizlerinde, SPSS 21 yazılımı tercih edilmiştir. Bütün istatistiksel analizlerde anlamlılık düzeyi $\alpha$=0,05 olarak kabul edilmiştir. Bulguların yorumlanmasında frekans tabloları ve tanımlayıcı istatistikler kullanılmıştır. Normal dağılıma uygun ölçüm değerleri için parametrik yöntemler kullanılmıştır. Parametrik yöntemlere uygun şekilde, iki bağımsız grubun ölçüm değerleriyle karşılaştırılmasında "Independent Sample-t" test yöntemi kullanılmıştır. Normal dağılıma uygun olmayan ölçüm değerleri için parametrik olmayan yöntemler kullanılmıştır. Parametrik olmayan yöntemlere uygun şekilde, iki bağımsız grubun ölçüm değerleriyle karşılaştırılmasında "Mann-Whitney U" test yöntemi kullanılmıştır. Hastalık riskini etkileyen faktörlerin tespit edilmesinde "İkili (Binary) Lojistik Regresyon(LR)" modeli kullanılmıştır.

Demir eksikliği anemisine etki eden risk faktörlerinin geriye dönük (retrospektif) incelendiği bu çalışmamızda, araştırmaya konu olan 516 hastaya ait bilgiler Tablo 3.2'de verilmiştir.

Tablo 3. 2. Gruplara göre bazı parametrelerin karşılaştırılmaları.

| **Değişken (N=516)** | **Sağlam (n=157)** | | **Hasta (n=359)** | | **İstatistiksel analiz* Olasılık** |
|---|---|---|---|---|---|
| | $\bar{X} \pm S.S.$ | **Median [Min-Max]** | $\bar{X} \pm S.S.$ | **Median [Min-Max]** | |
| Yaş (yıl) | 54.37±14.40 | 54.0 [21.0-89.0] | 43.29±14.11 | 41.0 [17.0-87.0] | Z=-7.969 **p=0.000** |
| Hb | 13.38±1.28 | 13.2 [11.0-17.0] | 10.32±1.79 | 10.3 [6.2-15.8] | t=21.906 **p=0.000** |
| Hct | 38.84±3.62 | 38.5 [32.0-48.8] | 31.20±4.76 | 31.1 [18.3-48.2] | t=19.955 **p=0.000** |
| Mcv | 84.90±6.38 | 85.2 [34.9-99.7] | 71.76±9.34 | 71.9 [51.9-111.6] | Z=-14.272 **p=0.000** |
| Mch | 29.45±2.27 | 29.4 [21.5-38.3] | 23.75±3.86 | 23.9 [14.6-40.1] | Z=-14.500 **p=0.000** |
| Mchc | 34.32±1.07 | 34.3 [25.1-36.2] | 32.99±1.34 | 33.1 [28.2-36.2] | Z=-11.158 **p=0.000** |
| Eritrosit RBC | 4.50±0.53 | 4.5 [3.1-5.9] | 4.37±0.58 | 4.4 [2.0-6.6] | Z=-2.650 **p=0.008** |
| Rdw | 15.21±3.48 | 14.1 [12.0-31.5] | 18.39±4.43 | 17.3 [12.1-46.9] | Z=-11.167 **p=0.000** |
| Demir | 86.49±26.81 | 80.0 [40.0-217.0] | 32.36±38.64 | 23.0 [10.0-464.0] | Z=-16.174 **p=0.000** |
| Total dbk | 252.68±50.93 | 251.0 [135.0-404.0] | 390.35±96.81 | 405.0 [68.0-717.0] | Z=-13.987 **p=0.000** |
| Satürasyon (%) | 0.36±0.16 | 0.3 [0.1-0.9] | 0.12±0.31 | 0.1 [0.0-4.6] | Z=-16,150 **p=0.0000** |
| Ferritin | 108.68±79.95 | 83.0 [28.0-640.0] | 52.59±170.94 | 7.0 [10.0-650.0] | Z=-14.382 **p=0.000** |

*Normal dağılıma sahip olan iki bağımsız grubun ölçüm değerleriyle karşılaştırılmasında "Independent Sample-t" test (t-tablo değeri); normal dağılıma sahip olmayan iki bağımsız grubun ölçüm değerleriyle karşılaştırılmasında "Mann-Whitney U" test (Z-tablo değeri) istatistikleri kullanılmıştır.

Gruplara göre yaş (yıl), Hb, Hct, Mcv, Mch, Mchc, Eritorsit RBC, Rdw, Demir, Total dbk, satürasyon (%) ve ferritin değerleri açısından istatistiksel olarak anlamlı farklılık tespit edilmiştir ($p<0,05$).

Tablo 3. 3. Gruplar ile cinsiyet arasındaki ilişkilerin incelenmesi

| **Değişken (N=516)** | **Sağlam (n=157)** | | **Hasta (n=359)** | | **İstatistiksel analiz* Olasılık** |
|---|---|---|---|---|---|
| | **n** | **%** | **n** | **%** | |
| **Cinsiyet** | | | | | |
| Kadın | 107 | 68.2 | 332 | 92.5 | $\chi^2$=49.016 |
| Erkek | 50 | 31.8 | 27 | 7.5 | **p=0.000** |

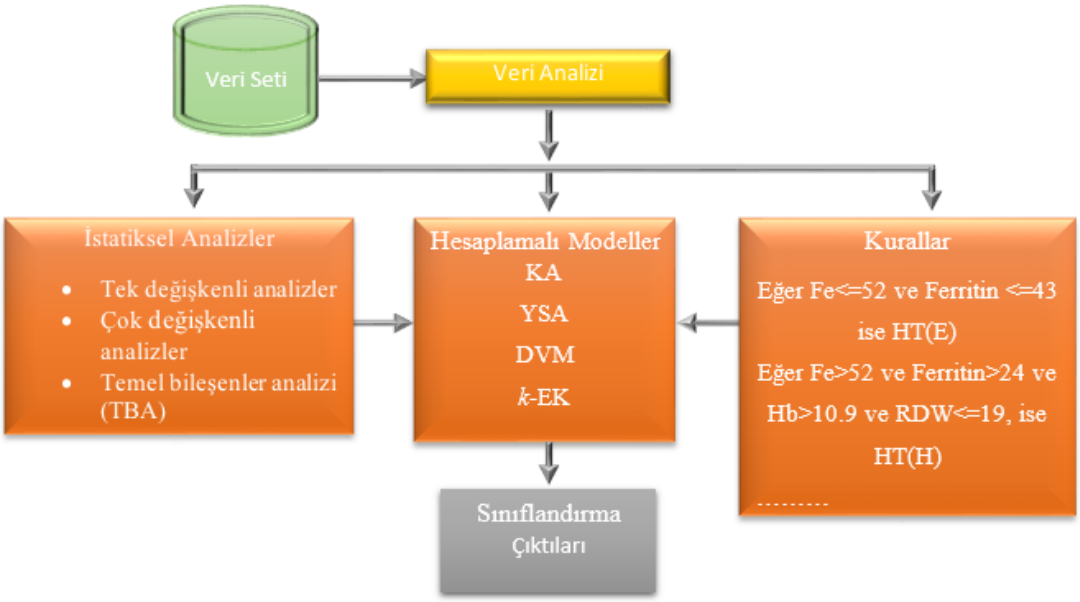
*İki nitel değişkenin ilişkilerinin incelenmesinde $\chi^2$-çapraz tablosu kullanılmıştır.

Gruplar ile cinsiyet arasında istatistiksel olarak anlamlı ilişki tespit edilmiştir ($\chi^2$=49,016; p=0,000). Sağlam grupta 50 kişinin (%31,8) erkek, hasta grupta 332 kişinin (%92,5) kadın olduğu tespit edilmiştir.

Tablo 3. 4. Hastalık durumunu etkileyen faktörlerin LR modeliyle incelenmesi

| | | Standart | | | | | OR, 95% GA | |
|---|---|---|---|---|---|---|---|---|
| **Değişken** | **β** | **Hata** | **Wald** | **sd** | **p** | **OR** | **Alt** | **Üst** |
| Yaş (yıl) | -0.057 | 0.015 | 14.670 | 1 | **0.000** | 0.944 | 0.917 | 0.972 |
| Hb | -2.710 | 0.587 | 21.305 | 1 | **0.000** | 0.067 | 0.021 | 0.210 |
| Mcv | 0.218 | 0.099 | 4.832 | 1 | **0.028** | 1.243 | 1.024 | 1.509 |
| Eritrosit RBC | 4.614 | 1.522 | 9.196 | 1 | **0.002** | 1.091 | 1.011 | 1.991 |
| Rdw | 0.123 | 0.059 | 4.360 | 1 | **0.037** | 1.131 | 1.008 | 1.269 |
| Demir | -0.044 | 0.016 | 7.129 | 1 | **0.008** | 0.957 | 0.927 | 0.988 |
| Total dbk | 0.022 | 0.005 | 20.919 | 1 | **0.000** | 1.022 | 1.013 | 1.032 |
| Satürasyon (%) | 5.726 | 3.217 | 3.167 | 1 | 0.075 | 3.068 | 0.560 | 16.808 |
| Sabit | -8.316 | 8.537 | 0.949 | 1 | 0.330 | 0.000 | | |

*Hosmer&Lemeshow test $\chi^2$=2,236; p=0,973; CCR=93,8%



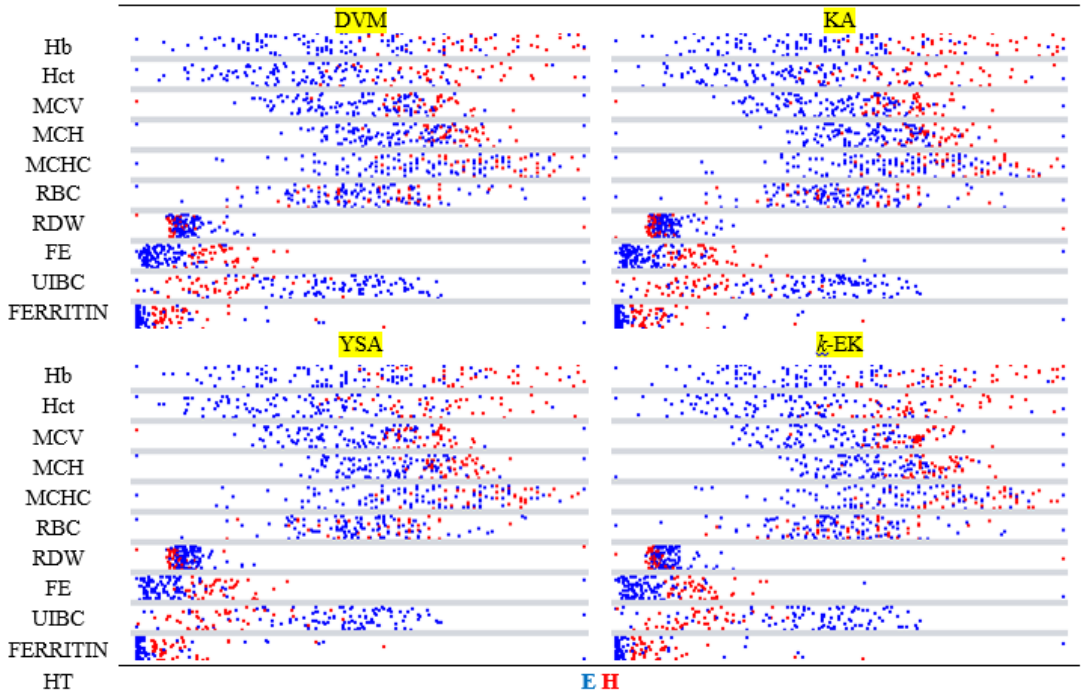

Şekil 3.1 Hesaplamalı iş akışı

Sınıflandırma işlemi öncesinde veri seti, %66 eğitim seti, %34 test seti şeklinde 2' ye bölünmüştür. Karşılaştırma sonuçları arasındaki uyumsuzluğu önlemek için tüm algoritmalara veri seti aynı şekilde uygulanmıştır.

Tablo 3. 5. Demir eksikliği anemisi tanısı konan ve dışındakilerin ayırt edilebilmesi için DVM, KA, YSA, *k*-EK modellerinin performans karşılaştırması

| Sınıflandırıcı | Toplam Örnek Sayısı | Doğru Sınıflandırılmış Örnek Sayısı | % | Karışıklık Matrisi | | |
|---|---|---|---|---|---|---|
| | | | | a | b | ← Sınıflandırma |
| **DVM** | 175 | 157 | 89.72 | 114<br>5 | 13<br>43 | |
| **KA** | 175 | 171 | 97.72 | 123<br>0 | 4<br>48 | a=HT(E)<br>b=HT(H) |
| **YSA** | 175 | 161 | 92 | 118<br>5 | 9<br>43 | |
| ***k*-EK** | 175 | 160 | 91.43 | 117<br>5 | 10<br>43 | |

Tablo 3.5'de görüldüğü gibi HT(E) ve HT(H) grupları arasında KA algoritması, % 97.72 en yüksek ayırma oranına sahiptir. YSA ve *k*-EK algoritmaları ise sırasıyla en iyi ikinci ve üçüncü tahmin sonuçlarına ulaşarak çok iyi performans göstermiştir.




Şekil 3.2. Sınıflandırıcıların hata tahmini

Mavi ve Kırmızı sırasıyla HT(E) ve HT (H) vakalarını simgelemektedir. Algoritmalara göre çok değişkenli analiz için belirlenen özelliklerin her birinin hastaların ayırımında etkisi görülmektedir. Tüm algoritmalardaki nokta yoğunlukları yakından incelendiğinde Hb, RDW, FE ve FERRITIN' in grupları ayırma güçleri yüksektir.

### 3.3. Sınıflandırmanın temel bileşenler analizi (TBA) ile doğrulanması

TBA sınıflandırma gücü yüksek bir tekniktir. Büyük boyutlu verilerde, verinin güçlü yönlerini ön plana çıkarmak için boyut indirgeyerek veriyi sıkıştırmaktadır (Abdi ve Williams, 2010). Bu çalışmamızda veri setinin tamamı eğitim seti olarak kullanılmıştır.

Tablo 3. 6. TBA sonrası demir eksikliği anemisi tanısı konan ve dışındakilerin ayırt edilebilmesi için DVM, KA, YSA, *k*-EK modellerinin performans karşılaştırması

| **Sınıflandırıcı** | **Toplam Örnek Sayısı** | **Doğru Sınıflandırılmış Örnek Sayısı** | **%** | **Karışıklık Matrisi** | | |
|---|---|---|---|---|---|---|
| | | | | **a** | **b** | **← Sınıflandırma** |
| **DVM** | 516 | 465 | 90.12 | 325<br>117 | 34<br>140 | |
| **KA** | 516 | 489 | 94.77 | 336<br>4 | 23<br>153 | a=HT(E) |
| **YSA** | 516 | 498 | 96.52 | 345<br>4 | 14<br>153 | b=HT(H) |
| ***k*-EK** | 516 | 516 | 100 | 359<br>0 | 0<br>157 | |

Tablo 3.6’ da *k*-EK algoritması, grupları %100 ile hatasız tahmin etmiştir. Diğerleri çok küçük yanılma payı ile yüksek tahmin gücüne sahiptirler.

### 3.4. Sınıflandırma başarılarının çapraz doğrulama ile testi

Test sonuçlarının değerlendirmesi için aynı veri seti 10 kat çapraz doğrulama kullanılarak yeniden test edildi. Bu işlemin amacı doğruluk yüzdesini elde etmek için oluşturulan veri madenciliği uygulamasını test etmektir. Öncesinde veri seti, 10 kat çapraz doğrulama işlemi için on parçaya ayrılmaktadır. Verinin dokuz parçası eğitim verisi olarak bir parçası ise test için kullanılmaktadır. 10 eğitim ve test süreci işleminin sonucunda ortalama doğruluk elde edilmektedir. 10 kat çapraz doğrulama işleminden sonra elde edilen değerlendirme sonuçları Tablo 3.7' de gösterilmektedir.

Tablo 3. 7. Çapraz doğrulama test sonuçları

| Sınıflandırma Çıktıları | Algoritmalar | | | |
|---|---|---|---|---|
| | **DVM** | **KA** | **YSA** | ***k*-EK** |
| Doğru Sınıflandırılmış Örnekler | 89.9225 % | 96.124 % | 93.8984 % | 91.8605 % |
| Hatalı Sınıflandırılmış Örnekler | 10.0775 % | 3.876 % | 6.2016 % | 8.1395 % |
| Kappa statistic | 0.7694 | 0.9081 | 0.8556 | 0.8118 |
| Mean absolute error (MAE) | 0.1008 | 0.044 | 0.0818 | 0.0832 |
| Root mean squared error(RMSE) | 0.3175 | 0.1897 | 0.2308 | 0.2847 |
| Relative absolute error (RAE) | 23.7839 % | 10.3818 % | 19.3052 % | 19.6337 % |
| Root relative squared error(RRSE) | 68.995 % | 41.2401 % | 50.1581 % | 61.8757 % |

Tablo 3.7' de 10 kat çapraz doğrulama işleminden sonra KA algoritması %96.124 sınıflandırma doğruluğu vermektedir.

Tablo 3. 8 Sınıflandırma algoritmalarının HT sonuçları

| **Hasta No** | **HB** | **HCT** | **MCV** | **MCH** | **MCHC** | **RBC** | **RDW** | **DEMIR** | **TDBK** | **FERRITIN** | **HT**(bilinmiyor) | **Tahmin Sonuçları** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | **HT** | **DVM** | **KA** | **YSA** | ***k*-EK** |
| **1** | 8.2 | 26.3 | 62.2 | 19.5 | 31.3 | 4.22 | 18.6 | 14 | 463 | 3 | ? | E | 1 | 1 | 1 | 0.998 |
| **2** | 8.1 | 26.7 | 62.1 | 19 | 30.7 | 4.31 | 19 | 11 | 449 | 3 | ? | E | 1 | 1 | 1 | 0.998 |
| **3** | 9.8 | 31.8 | 57.3 | 17.7 | 30.8 | 5.54 | 22.9 | 11 | 405 | 12 | ? | E | 1 | 1 | 1 | 0.998 |
| **4** | 9.6 | 29.8 | 65.4 | 21 | 32.1 | 4.55 | 18 | 34 | 432 | 4 | ? | E | 1 | 1 | 1 | 0.998 |
| **5** | 7.8 | 24.9 | 62.9 | 19.7 | 31.3 | 3.95 | 17.8 | 10 | 472 | 1 | ? | E | 1 | 1 | 1 | 0.998 |
| **6** | 14.9 | 40.5 | 86.1 | 31.1 | 36.1 | 4.7 | 13.5 | 124 | 283 | 205 | ? | H | 1 | 0.986 | 0.979 | 0.998 |
| **7** | 15.3 | 43.2 | 82 | 29.1 | 35.5 | 5.2 | 3.5 | 88 | 288 | 240 | ? | H | 1 | 0.987 | 0.977 | 0.998 |
| **8** | 15 | 43.7 | 89.6 | 30.7 | 34.2 | 4.8 | 13.7 | 92 | 262 | 49 | ? | H | 1 | 0.988 | 0.974 | 0.998 |
| **9** | 13.7 | 40.7 | 88.4 | 30.8 | 34.1 | 4.2 | 13 | 157 | 240 | 124 | ? | H | 1 | 0.989 | 0.977 | 0.998 |
| **10** | 12.5 | 36.2 | 99.7 | 34.4 | 35.5 | 3.6 | 15.8 | 107 | 211 | 110 | ? | H | 1 | 0.990 | 0.972 | 0.998 |

?HT(E) veya HT(H) durumlarının önceden bilinmediğini ifade etmektedir.

Tablo 3.8' de HT bilinmeyen hastalara ait kan değerleri makineye girilerek grupları tahmin edilmek istenmiştir. Sonuçlar DVM algoritmasının hastalık gruplarının her ikisini de HT(E) ve HT(H) %100 doğrulukla tahmin ettiğini göstermiştir.

# 4. BULGULAR

## 4.1. Uygulama I test sonuçları

Bu aşamada hematoloji servisine başvuran kişilerin yaş, cinsiyet ve tam kan sayımı (CBC) analizinden elde edilen sonuçlarının DEA tanısını değerlendirmedeki güçlerini araştırıyoruz. RBC, Hb, Hct, MCV, MCH, MCHC, FE, UIBC, FERRITIN ve RDW' den oluşan RBC indeksleri, otomatik cihazlar kullanılarak CBC analizinden üretilmektedir. CBC sonuçlarından ilk edindiğimiz bilgilere göre DEA tanısı konulan hastaların Yaş (yıl), Hb, Hct, MCV, MCH, MCHC, Eritrosit RBC, Demir, satürasyon (%) ve ferritin değerleri, sağlam gruba göre daha düşük olduğu tespit edilmiştir. Aynı şekilde, hasta grubun Rdw ve UIBC değerleri, sağlam gruba göre istatistiksel olarak anlamlı düzeyde daha yüksek çıkmıştır ve bu beklenen bir durumdur (Tablo 3.2). Araştırmaya konu olan kadınların ağırlıklı olarak hasta olduğu, erkeklerin ise ağırlıklı olarak sağlam olduğu belirlenmiştir (Tablo 3.3). Hastalık risk durumunu tespit etmek için tek değişkenli analizler sonucunda (Tablo 3.2) gruplara göre anlamlı çıkan tüm parametrelerin modele dahil edilerek yapılan Lojistik regresyon sonuçlarına göre, Yaş (yıl) 1 birim arttığında, hasta olma riski (OR=1-0,944=0,056) %5,6 azalacaktır. Hb değeri 1 birim arttığında, hasta olma riski (OR=1-0,067=0,933) %93,3 azalacaktır. Mcv değeri 1 birim arttığında, hasta olma riski %24,3 artacaktır. Eritrosit RBC değeri 0,01 birim arttığında, hasta olma riski %9,1 artacaktır. Rdw değeri 1 birim arttığında, hasta olma riski %13,1 artacaktır. Demir değeri 1 birim arttığında, hasta olma riski (OR=1-0,957=0,043) %4,3 azalacaktır. UIBC değeri 1 birim arttığında, hasta olma riski %2,2 artacaktır( Tablo 3.4). Hem tek değişkenli (Tablo 3.2 ) hem de çok değişkenli (Tablo 3.4 ) analizlerin sonuçları, sadece bir parametrenin iki koşul arasında ayrım yapmak için yeterli olmadığını aksine tüm parametrelerin iki grubu ayırmada etkili olduğunu göstermiştir. Çoklu karşılaştırma testlerinde Hb, MCV, RBC, RDW ve FE' in daha güçlü ayırıcılar olduğu görülmüştür (Tablo 3.4).

## 4.2. Uygulama II test sonuçları

Bu aşamada DEA tanısı konulan ve dışındakilerin ayırt edilebilmesi için DVM, KA, YSA, *k*-EK modellerinin performans karşılaştırması yapılmıştır. Sınıflandırma işlemi öncesinde veri seti, %66 eğitim seti, %34 test seti şeklinde ikiye bölünmüştür. Bölünme oranlarının belirlenmesinde tüm sınıflandırıcılar için optimum doğruluk değerlerinin elde edilmesi amaçlanmıştır. Karşılaştırma sonuçları arasındaki

uyumsuzluğu önlemek için tüm algoritmalara veri seti aynı şekilde uygulanmıştır (Tablo 3.5). Bu sonuçların doğruluğu, veri setine TBA uyguladıktan sonra veriler tekrar sınıflandırılarak test edilmiştir. Bu sınıflandırma işlemi esnasında veri setinin tamamı eğitim seti olarak kullanılmıştır. Sonucunda, sıkıştırılmış ve özellik sayısı indirgenmiş verinin algoritmalara göre doğru sınıflandırma oranlarının pozitif yönde değiştiği gözlemlenmiştir (Tablo 3.6). Hatta bu analizin sonucunda *k*-EK algoritmasının şaşırtıcı bir şekilde verileri % 100 doğru sınıflandırdığı görülmüştür. Her iki bağımsız geçerlilik testinin sonuçları çapraz doğrulama testinin sonuçları ile karşılaştırılmıştır (Tablo 3.7). Sonuçlar öngörülerimizi yüksek oranda doğrulamaktadır. DVM algoritması her üç testin sonucunda, tüm algoritmalar içinde 89.92 doğru sınıflandırma aritmetik ortalaması ve ± 0.2 standart sapma ile en düşük oranı elde etmiştir. Onu sırasıyla 94.14 aritmetik ortalama ve ± 2.28 standart sapma ile YSA, 94.43 aritmetik ortalama ve ± 4.83 standart sapma ile *k*-EK sınıflandırıcıları izlemektedir. Özellikle KA algoritması 96.21 aritmetik ortalama ve ± 1.48 standart sapma ile diğerlerinden daha iyi sonuçlar vermektedir. Weka 3.8, KA oluştururken J48 algoritması kullanmaktadır. Ağacın en tepesinde kök düğüm olarak demir yer almaktadır. Demir'i sırasıyla Ferritin, Hb, RDW izlemektedir. Ağaç oluşurken kişi, demir değerinin elli ikiden küçük eşit olması durumunda eğer Ferritin değeri de kırk üçten küçükse HT(E) gurubuna, eğer Ferritin değeri kırk üçten büyükse Hb değerinin on ikiden küçük olması halinde HT(E) gurubuna atanmaktadır.

# 5. TARTIŞMA VE SONUÇ

Tıbbi alandaki gelişmelere rağmen demir eksikliği anemili hastaları diğerlerinden ayırma güçlüğünün devam ettiği görülmektedir. Özellikle bu alanın dışındaki klinisyen tarafından verilen kan analizinin etkisiz değerlendirilmesi veya etkisiz tedavisi bu düşünceyi desteklemektedir. Yeni tip hesaplamalı modellerin (data mining, machine learning) bu tarz sorunların çözümüne katkı sağlayacaktır. Bunun için disiplinler arası ortak yürütülecek çalışmalara ihtiyaç duyulmaktadır. Bu tarz çalışmalarda araştırmacıların (tıp doktorları, veri bilim uzmanları vs.) veri toplama sürecinde hastalıklar konusunda detaylı bilgiye sahip olmaları veri analiz aşamaları açısından olumlu bir yaklaşım olacaktır.

Tezde istatistiksel analizler Bağımsız Örneklemler T Testi, Mann-Whitney U Testi, Lojistik Regresyon testlerinin yanı sıra, makine öğrenme ve sınıflandırma yöntemlerinden Karar Ağacı, K en Yakın Komşuluk, Yapay Sinir Ağları ve Destek Vektör Makineleri tanıtılmış ve makine öğrenme sınıflama algoritmalarının performansları karşılaştırılmıştır

Bu çalışmada iki uygulama yapılmıştır. Uygulama I' de DEA' ne etki eden risk faktörleri geriye dönük (retrospektif) olarak incelenmiştir. Hastalığa etki eden faktörlerin, etki oranlarının önceden bilinmesi hekimler ve hastalar açısından önemlidir.

Uygulama II' de makine öğrenme sınıflandırma algoritmaları *k*-EK, KA, YSA ve DVM' in performansları karşılaştırılmıştır. Karşılaştırma işlemlerinin birinci aşamasında veri seti %66 eğitim seti, %34 test seti olarak iki parçaya bölünerek tüm sınıflandırıcılara eşit şekilde uygulanmıştır. İkinci aşamada veri seti temel bileşenler analizi uygulanarak özellik indirgeme yöntemi ile sıkıştırıldıktan sonra tekrar sınıflandırma işlemine tabi tutulmuştur. Bu işlem esnasında veri setinin tamamı eğitim seti olarak tüm algoritmalara uygulanmıştır. TBA sonrası yapılan sınıflandırma sonuçlarına göre tüm algoritmaların doğru sınıflandırma oranlarının arttığı görülmüştür. Üçüncü aşamada veri seti 10 kat çapraz doğrulama tekniği kullanılarak algoritmalar tarafından tekrar sınıflandırılmıştır. Son aşamada makine öğrenmesi sınıflandırma algoritmalarının veri seti içinden rastgele seçilen on hastaya ait kan sonuçlarını inceleyerek HT' i tahmin etmesi istenmiştir. Tüm algoritmaların sonuçları yüzde yüze yakın doğruluk değerleriyle tahmin ettiği gözlemlenmiştir (Tablo 3.8).

Sonuç olarak makine öğrenmesi sınıflandırma algoritmaları tüm aşamalarda oldukça yüksek doğru sınıflandırma oranları vermiştir. Bu sonuç literatür çalışmalarıyla da uyumlu çıkmıştır (Laengsri vd, 2019). KA algoritmasının diğerlerine göre daha iyi sonuçlar verdiği gözlemlenmiştir. Bu algoritma tarafından oluşturulan karar ağacı yardımı ile başlangıçta tanımlanan değişkenlerin sayısı azaltılarak karar verme süreçleri desteklenebilir.

Sonuç olarak Uygulama I ve Uygulama II test sonuçlarının birbiriyle uyumlu olduğu gözlemlenmiştir. FE, Ferritin, Hb, RDW gibi hastalığa etki eden faktörlerin her iki uygulamada da ön plana çıkan özellikler olduğu görülmüştür. Yöresel farklılıklardan arındırılmış büyük boyutlu güncel DEA hastalık verilerini içeren veri setlerine uygulanacak yeni makine öğrenmesi tekniklerinin, anemiye etki eden risk faktörlerinin belirlenmesinde daha iyi sonuçlar vereceği düşünülmektedir. Başlangıçta tanımlanan değişkenlerin sayısının azaltılabilmesi ve hekimlerin karar verme süreçlerini büyük oranda kolaylaştıracaktır. Ayrıca birçok aşamadan geçerek elde edilen sonuçların, kullanımı kolay dijital uygulamalara dönüştürülmesi zaman ve emek kaybını önemli ölçüde azaltacaktır. Daha fazla araştırma için, araştırmacılar bu araştırmanın sonuçlarını, hekimlerin DEA hastalarının diğer hastalardan ayırımını kolaylaştıran mobil bir uygulama haline getirebilirler.

# KAYNAKLAR


Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews:* computational statistics, 2(4), 433-459.

Acharya, V., & Kumar, P. (2017, September). Identification and red blood cell classification using computer aided system to diagnose blood disorders. *International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 2098-2104). IEEE.

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis, 22*(5), 717-727.

Ahmad, I., Norul, S., & Zahratul, R. (2018). Morphological features analysis for erythrocyte classification in IDA and thalassemia. *Int. J. Adv. Comput. Sci. Appl., 9*(12), 274-280.

Akin, H. B., & Şentürk, E. (2012). Bireylerin mutluluk düzeylerinin ordinal lojistik regresyon analizi ile incelenmesi-analysing levels of happiness of individuals with ordinal logistic analysis. *Öneri Dergisi, 10*(37), 183-193.

Aktaş, C. (2009). Lojistik Regresyon Analizi: Öğrencilerin Sigara İçme Alışkanlığı Üzerine Bir Uygulama. *Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 1*(26), 107-122.

Allali, S., Brousse, V., Sacri, A. S., Chalumeau, M., & de Montalembert, M. (2017). Anemia in children: prevalence, causes, diagnostic work-up, and long-term consequences. *Expert review of hematology, 10*(11), 1023-1028.

Al-Nuaimy, W., Huang, Y., Nakhkash, M., Fang, M. T. C., Nguyen, V. T., & Eriksen, A. (2000). Automatic detection of buried utilities and solid objects with GPR using neural networks and pattern recognition. *Journal of applied Geophysics, 43*(2-4), 157-165.

Alpar, R. (2013). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, Ankara: 4. *Baskı, Detay Yayıncılık.*

Alpaydın, E. (2011). Yapay öğrenme. *Boğaziçi Üniversitesi Yayınevi.*

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175-185.

Andro, M., Le Squere, P., Estivin, S., & Gentric, A. (2013). Anaemia and cognitive performances in the elderly: a systematic review. *European Journal of Neurology, 20*(9), 1234-1240.

Ayhan, S. (2006). Sıralı lojistik regresyon analiziyle Türkiye'deki hemşirelerin iş bırakma niyetini etkileyen faktörlerin belirlenmesi. *Yayımlanmamış Yüksek Lisans Tezi. Eskişehir: Osmangazi Üniversitesi Fen Bilimleri Enstitüsü.*

Ayyıldız, H., & Tuncer, S. A. (2020). Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anemia and beta thalassemia via Neighborhood Component Analysis Feature Selection-Based machine learning. *Chemometrics and Intelligent Laboratory Systems, 196*, 103886.

Balaban, M. E., & Kartal, E. (2015). Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları. *Çağlayan Kitabevi, İstanbul*.

Bartosch-Härlid, A., Andersson, B., Aho, U., Nilsson, J., & Andersson, R. (2008). Artificial neural networks in pancreatic disease. *British journal of surgery, 95*(7), 817-826.

Bayır, F. (2006). Yapay sinir ağları ve tahmin modellemesi üzerine bir uygulama. *Yayımlanmamış Yüksek Lisans Tezi, İstanbul, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü.*

Bellinger, C., Amid, A., Japkowicz, N., & Victor, H. (2015, December). Multi-label classification of anemia patients. IEEE 14th *International Conference on Machine Learning and Applications (ICMLA)* (pp. 825-830). IEEE.

Bhatia, N. (2010). Survey of nearest neighbor techniques. *arXiv preprint arXiv:1007.0085*.

Bircan, H. (2004). Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama. *Kocaeli Üniversitesi Sosyal Bilimler Dergisi,* (8), 185-208.

Bland J M & Altman D G (2000). The odds ratio. *Bmj*, *320*(7247), 1468

Busuttil, S. (2003). Support vector machines. In Proceedings of the Computer Science Annual Research Workshop, *Villa Bighi, Kalkara, University of Malta*

Chen, Y. M., Miaou, S. G., & Bian, H. (2016). Examining palpebral conjunctiva for anemia assessment with image processing methods. *Computer methods and programs in biomedicine, 137,* 125-135.

Chen, Y. Y., Lin, Y. H., Kung, C. C., Chung, M. H., & Yen, I. (2019). Design and implementation of cloud analytics-assisted smart power meters considering advanced artificial intelligence as edge analytics in demand-side management for smart homes. *Sensors, 19*(9), 2047.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics, 21*(1), 6.

Copeland, B. J., & Proudfoot, D. (2007). Artificial intelligence: History, foundations, and philosophical issues. *In Philosophy of Psychology and Cognitive Science* (pp. 429-482). North-Holland.

Cortes, C. (1995). WSupport-vector network. *Machine learning, 20*, 1-25.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20(*3), 273-297.

Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., & Colliot, O. (2011). Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. *Medical image analysis, 15*(5), 729-737.

Cusick, S. E., Georgieff, M. K., & Rao, R. (2018). Approaches for reducing the risk of early-life iron deficiency-induced brain dysfunction in children. *Nutrients, 10*(2), 227.

Çokluk, Ö. (2010). Lojistik regresyon analizi: Kavram ve uygulama. *Kuram ve Uygulamada Eğitim Bilimleri, 10*(3), 1357-1407.

Dalvi, P. T., & Vernekar, N. (2016, May). Computer aided detection of abnormal red blood cells. *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 1741-1746). IEEE.

De, S., & Chakraborty, B. (2020). Case-Based Reasoning (CBR)-Based Anemia Severity Detection System (ASDS) Using Machine Learning Algorithm. *In International Conference on Advanced Machine Learning Technologies and Applications* (pp. 621-632). Springer, Singapore.

Decoste, D., & Schölkopf, B. (2002). Training invariant support vector machines. *Machine learning, 46*(1-3), 161-190.

Dirican, A. (2001). Tanı testi performanslarının değerlendirilmesi ve kıyaslanması. *Cerrahpaşa Tıp Dergisi, 32*(1), 25-30.

Elhan, A. H. (1997). Lojistik regresyon analizinin incelenmesi ve tıpta bir uygulaması. *Biyoistatistik Yüksek Lisans Tezi. AÜ*, 4-29.

Elsalamony, H. A. (2016). Healthy and unhealthy red blood cell detection in human blood smears using neural networks. *Micron, 83*, 32-41.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning. *Pattern Recognition, 58*, 121-134.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Miscellaneous clustering methods. *Cluster analysis*, 215-255.

Fix, E., & Hodges, J. L. (1951). *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties USAF School of Aviation Medicine, Randolph Field* (pp. 1-21). Texas, Tech. Report 4.

Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press.

Gaonkar, B., & Davatzikos, C. (2013). Analytic estimation of statistical significance maps for support vector machine based multivariate image analysis and classification. *Neuroimage, 78*, 270-283.

Girginer, N., & Cankuş, B. (2008). Tramvay yolcu memnuniyetinin lojistik regresyon analiziyle ölçülmesi: Estram örneği. Y*önetim ve Ekonomi: Celal Bayar Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 15*(1), 181-193.

Graupe, D. (2013). *Principles of artificial neural networks* (Vol. 7). World Scientific.

Gujarati D N (1999). Temel Ekonometri (Çev. Ü. Şenesen & G. Şenesen), *Literatür Yayıncılık, İstanbul*

Gürsakal, N. (2017). Makine Öğrenmesi ve Derin Öğrenme. *Dora Basım Yayın Dağıtım, Bursa.*

Haas, J. D., & Brownlie IV, T. (2001). Iron deficiency and reduced work capacity: a critical review of the research to determine a causal relationship. *The Journal of nutrition, 131*(2), 676S-690S.

Hall, P., Park, B. U., & Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics, 36*(5), 2135-2152.

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems, 5*(4), 83-124.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

Hennek, J. W., Kumar, A. A., Wiltschko, A. B., Patton, M. R., Lee, S. Y. R., Brugnara, C., & Whitesides, G. M. (2016). Diagnosis of iron deficiency anemia using density-based fractionation of red blood cells. *Lab on a Chip, 16*(20), 3929-3939.

Hosner, D. W., & Lemeshow, S. (1989). Applied logistic regression. *New York: Jhon Wiley & Son.*

Hwang, W. J., & Wen, K. W. (1998). Fast kNN classification algorithm based on partial distance search. *Electronics letters, 34*(21), 2062-2063.

İlaslaner, T., & Güven, A. (2019, October). Investigation of the Effects Biochemistry on Iron Deficiency Anemia. *Medical Technologies Congress (TIPTEKNO)* (pp. 1-4). IEEE.

Jin, C., De-Lin, L., & Fen-Xiang, M. (2009, July). An improved ID3 decision tree algorithm. *4th International Conference on Computer Science & Education* (pp. 127-130). IEEE.

Kabalci, E. (2017). Yapay Sinir Ağları. (Artificial Neural Networks).

Karabulut, E., & Alpar, R. (2011). Lojistik Regresyon. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, *Detay Yayıncılık Ankara*.

Kargı, V. S. A. (2013). Yapay sinir ağ modelleri ve bir tekstil firmasında uygulama.

Katagiri, S., & Abe, S. (2006). Incremental training of support vector machines using hyperspheres. *Pattern Recognition Letters, 27*(13), 1495-1507.

Khan, J. R., Chowdhury, S., Islam, H., & Raheem, E. (2019). Machine learning algorithms to predict the childhood anemia in Bangladesh. *Journal of Data Science, 17*(1), 195-218.

Khatatneh, K., & El Emary, I. M. (2009). Enhancing the Performance of Entropy Algorithm using Minimum Tree in Decision Tree Classifier. *In Advances in Computational Algorithms and Data Analysis* (pp. 333-347). Springer, Dordrecht.

Kılıç, S. (2013). Klinik Karar Vermede ROC Analizi. *Journal of Mood Disorders, 3*(3).

Kocabaş, Ş. (2015). Yapay zekâ araştırma ve uygulama alanları Erişim: 11 Eylül 2020, http://sakirkocabas.blogspot.com/search/label/YAPAY%20ZEKA

Laengsri, V., Shoombuatong, W., Adirojananon, W., Nantasenamart, C., Prachayasittikul, V., & Nuchnoi, P. (2019). ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC medical informatics and decision making, 19*(1), 212.

Le, H. M., Tran, T. D., & Van Tran, L. A. N. G. (2018). Automatic heart disease prediction using feature selection and data mining technique. *Journal of Computer Science and Cybernetics, 34*(1), 33-48.

Li, S., Li, H., Li, M., Shyr, Y., Xie, L., & Li, Y. (2009). Improved prediction of lysine acetylation by support vector machines. *Protein and peptide letters, 16*(8), 977-983.

Liu, H., & Zhang, S. (2012). Noisy data elimination using mutual k-nearest neighbor for classification mining. *Journal of Systems and Software, 85*(5), 1067-1074.

Luo, Y., Szolovits, P., Dighe, A. S., & Baron, J. M. (2016). Using machine learning to predict laboratory test results. *American journal of clinical pathology, 145*(6), 778-788.

Maity, A. (2016). Supervised Classification of RADARSAT-2 Polarimetric Data for Different Land Features. *arXiv preprint arXiv:1608.00501*.

Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

McCorduck, P., & Cfe, C. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence.* CRC Press.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological), 42*(2), 109-127.

Meena, K., Tayal, D. K., Gupta, V., & Fatima, A. (2019). Using classification techniques for statistical analysis of Anemia. *Artificial intelligence in medicine, 94*, 138-152.

Mertler C A & Vannatta R A (2005). Advanced and multivariate statistical methods: Practical application and interpretation. *Pyrczak Publishing. Glendale*.

Mishra, P., Dey, S., Ghosh, S. S., Seal, D. B., & Goswami, S. (2019, September). Human Activity Recognition using Deep Neural Network. *International Conference on Data Science and Engineering (ICDSE)* (pp. 77-83). IEEE.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM, 42*(11), 30-36.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning. Adaptive computation and machine learning.* MIT Press, 31, 32.

Muratlar, E. R. (2019). Makine öğrenmesine çok değişkenli istatistiksel yaklaşımlar: Temel bileşenler analizi. Erişim: 15. Şubat 2021. https://www.veribilimiokulu.com/blog/makine-ogrenmesine-cok-degiskenli-istatistiksel-yaklasimlar-temel-bilesenler-analizi/

Ocak, I., & Seker, S. E. (2013). Calculation of surface settlements caused by EPBM tunneling using artificial neural network, SVM, and Gaussian processes. *Environmental earth sciences, 70*(3), 1263-1276.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

Özdamar, K. (2013). Paket programlar ile istatistiksel veri analizi (Cilt 1). *Ankara: Nisan Kitapevi,* 27-36

Özkan, Y. (2008). Veri madenciliği yöntemleri. *Papatya Yayıncılık Eğitim, İstanbul.*

Öztemel, E. (2012). Yapay sinir ağları. *Papatya Yayıncılık Eğitim, İstanbul.*

Piryonesi, S. M., & El-Diraby, T. E. (2020). Data analytics in asset management: Cost-effective prediction of the pavement condition index. *Journal of Infrastructure Systems, 26*(1), 04019036.

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: an overview and their use in medicine. *Journal of medical systems, 26*(5), 445-463.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Rai, P. (2011). Model Selection and Feature Selection.

Rokach, L., & Maimon, O. (2005). Clustering methods. *In Data mining and knowledge discovery handbook* (pp. 321-352). Springer, Boston, MA.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Sakhibgareeva, M. V., & Zaozersky, A. Y. (2017). Developing an artificial intelligence-based system for medical prediction. *Bulletin of Russian State Medical University,* (6).

Saraç, T. (2004). Yapay Sinir Ağları, Seminer Projesi. Gazi Üniversitesi Endüstri Mühendisliği Ana Bilim Dalı.

Sayad, S. (2020). Data Science. Erişim: 17 Ağustos 2020. https://www.saedsayad.com/data_mining.htm

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. *Nonlinear estimation and classification* (pp. 149-171). Springer, New York, NY.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal, 27*(3), 379-423.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing, 5*(4), 13-22.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment, 62*(1), 77-89.

Şeker, Ş. E., & Eşmekaya, E. (2017). Eksik Verilerin Tamamlanması (Imputation). *YBS Ansiklopedi, 4,* 10-17.

Şeker, Ş.E. (2013). İş zekâsı ve veri madenciliği. *Cinius Yayınları Eğitim, İstanbul.*

Şerbetçi, A. (2013). Sıralı lojistik regresyon analizi ile istatistik ve ekonometri, derslerinde başarıyı etkileyen faktörlerin belirlenmesi: Atatürk Üniversitesi İktisadi ve İdari Bilimler Fakültesi öğrencileri üzerine bir uygulama. *Kahramanmaraş Sütçü İmam Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 3*(1), 89-110.

Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data minin*g. Pearson Education India.

Ting, F. F., Tan, Y. J., & Sim, K. S. (2019). Convolutional neural network improvement for breast cancer classification. *Expert Systems with Applications, 120*, 103-115.

Tomak, L., & Yüksel, B. (2009). İşlem karakteristik eğrisi analizi ve eğri altında kalan alanların karşılaştırılması. *Journal of Experimental and Clinical Medicine, 27*(2).

Tsiptsis, K. K., & Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation.* John Wiley & Sons.

Turing, A. (1950). Computing machinery and intelligence. Mind, 59:236, 433.

Tyagi, M., Saini, L. M., & Dahyia, N. (2016). Detection of poikilocyte cells in iron deficiency anaemia using artificial neural network. *International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC)* (pp. 108-112). IEEE.

Uğuz, S. 2019. Makine öğrenmesi teorik yönleri ve Python uygulamaları ile bir yapay zekâ ekolü. *Nobel Yayıncılık, Ankara.*

Van Otterlo, M., & Wiering, M. (2012). Reinforcement learning and markov decision processes. *In Reinforcement Learning* (pp. 3-42). Springer, Berlin, Heidelberg.

Vapnik, V. N. (1995). T he Nature of Statistical Learning Theory. New York: Springer2V erlag.

Vapnik, V. N. (2014). Invited Speaker. *IPMU Information Processing and Management*

WHO, (2020). WHO guidance helps detect iron deficiency and protect brain development. Erişim: 17 Ağustos 2020. https://www.who.int/news-room/detail/20-04-2020-who-guidance-helps-detect-iron-deficiency-and-protect-brain-development

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems, 14*(1), 1-37.

Yue, Z., Gao, F., Xiong, Q., Wang, J., Huang, T., Yang, E., & Zhou, H. (2019). A novel semi-supervised convolutional neural network method for synthetic aperture radar image recognition. *Cognitive Computation*, 1-12.

Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research, 41*(12), 1955-1959.

# EKLER

## EK1 Etik Kurul Kararı



T.C.
SAĞLIK BAKANLIĞI
Samsun Eğitim ve Araştırma Hastanesi

Sayı : 33646832-799
Konu : Etik Kurul Kararı (Uzm.Dr.Sude Hatun AKTİMUR)

**GİRİŞİMSEL OLMAYAN KLİNİK ARAŞTIRMALAR ETİK KURULU**

| Girişimsel Olmayan Klinik Araştırmlar Etik Kurulu Kararları | Oturum Tarihi | Oturum sayısı |
|---|---|---|
| | 27.05.2020 | 2020/ 7 |

**Protokol No: GOKA/2020/7/20**

Sorumlu araştırmacısı Uzm.Dr. Sude AKTİMUR olan **"Demir eksikliği anemisi tanısı alan hastalarda; tanının makine öğrenmesi yöntemleri kullanılarak teyidi ve değerlendirilmesi"** isimli çalışma incelenmiş olup ,çalışmanın yürütülmesi Girişimsel Olmayan Etik Kurul tarafından uygun görülmüştür.

e-imzalıdır.

Doç.Dr.Mahcube ÇUBUKCU

Girişimsel Olmayan Etik Kurul Başkanı

Kadıköy Mh. Barış Bulvarı no:199
Telefon: Faks No:
e-Posta: nesrin.yazici@saglik.gov.tr İnternet Adresi: Nesrin YAZICI TUEK + ARGE

Bilgi için: Nesrin YAZICI
SAĞLIK TEKNİKERİ
Telefon No: (0 362) 311 15 00

## EK2 Uygulama Algoritması

1. Adım: Verilerin kayıt edilmesi
2. Adım: Veri setinin, istatistiksel ve makine öğrenmesi analizlerine uygun formata dönüştürülmesi
3. Adım: Veri setindeki tüm değişkenlerin normal dağılım gösterip göstermediğinin tespit edilmesi
4. Adım: Veri setindeki değişkenlere ait standart sapma, aritmetik ortalama, medyan gibi önemli istatistiksel özelliklerin belirlenmesi
5. Adım: Normal dağılım gösteren değişkenlerin hastalık durumlarına etkilerinin Bağımsız Örneklemler T Test' i ile normal dağılım göstermeyen değişkenlerin Mann Witney-U Test' i kullanılarak araştırılması
6. Adım: Hastalık grupları ile cinsiyet arasında bir ilişki olup olmadığının araştırılması
7. Adım: Lojistik Regresyon analizi kullanılarak hastalık durumuna etki eden faktörlerin güçlerinin belirlenmesi
8. Adım: KA, DVM, *k*-EK, YSA sınıflandırıcılarının hastalık durumlarını ayırmadaki doğruluk oranlarının belirlenmesi (Bu analizde veri seti % 64 eğitim seti, % 36 test seti olarak iki parçaya bölünmüştür)
9. Adım: KA, DVM, *k*-EK, YSA sınıflandırıcılarının hastalık durumlarını ayırmadaki güçlerinin grafiksel gösterimi
10. Adım: TBA kullanılarak yeniden belirlenen sınıflandırma algoritmalarının doğruluk oranlarının Adım 8' de elde edilen sonuçlarla karşılaştırılması (Bu analizde veri setinin tamamı eğitim seti olarak kullanılmıştır)
11. Adım: Çapraz Doğrulama yöntemi ile sınıflandırıcılara ait ortalama doğruluk değerlerinin tespit edilmesi (Bu analizde 10 kat çapraz doğrulama tercih edilmiştir)
12. Adım: Test setinin oluşturulması (Veri setinden rassal olarak seçilen on adet kayıt kullanılmıştır ve bu kayıtlarda HT değişkenine ait tüm değerler bilinmiyor anlamında ? simgesi ile gösterilmiştir)
13. Adım: Analizlerden elde edilen tüm bulguların anlaşılır bir dilde ifade edilmesi
14. Adım: Analiz sonuçlarının tartışılarak yorumlanması

# ÖZ GEÇMİŞ

Bünyamin SARIBACAK 14.10.1973 tarihinde Samsun'da doğdu. Samsun Havza Lisesi'ni bitirdikten sonra Gazi Üniversitesi Fakültesi'nden 1995 yılında mezun oldu. 2004 yılında OMÜ LEE İstatistik Yüksek Lisans Programını bitirdi. Mezuniyetinden bu yana OMÜ Eğitim Fakültesi Bilgisayar ve Öğretim Teknolojileri Eğitimi Bölümü'nde Öğretim Görevlisi olarak görev yapan Bünyamin SARIBACAK, orta derecede İngilizce bilmektedir. Temel ilgi alanları, Veri Tabanları, Robotik ve Kodlamadır.

**İletişim Bilgileri**

E mail: bunyamin@omu.edu.tr

Telefon: 0 533 711 48 09

Orcid No: 0000-0003-2775-776X

**Yayınlanmış Çalışmalar:**

Sarıbacak, B., Terzi, E., 2018. Classification of data by using machine learning methods. 11. International Statistics Days Conference 3 - 7 October 2018 Muğla Sıtkı Koçman UniversityDepartment of Statistics, TURKEY

Sarıbacak, B., Terzi, E., 2018. Performance comparisons of some classification algorithms used in machine learning. 11. International Statistics Days Conference 3 - 7 October 2018 Muğla Sıtkı Koçman UniversityDepartment of Statistics, TURKEY