T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ALS VAKA TESPİTİNDE BAYES AĞLARI VE MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Hasan Aykut KARABOĞA

DOKTORA TEZİ

İstatistik Anabilim Dalı

İstatistik Programı

Danışman

Doç. Dr. İbrahim DEMİR

Temmuz, 2021

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ALS VAKA TESPİTİNDE BAYES AĞLARI VE MAKİNE ÖĞRENMESİ YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Hasan Aykut KARABOĞA tarafından hazırlanan tez çalışması 13.07.2021 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı, İstatistik Programı **DOKTORA TEZİ** olarak kabul edilmiştir.

Doç. Dr. İbrahim DEMİR
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Doç. Dr. İbrahim DEMİR, Danışman
Yıldız Teknik Üniversitesi _____

Prof. Dr. Ersoy ÖZ, Üye
Yıldız Teknik Üniversitesi _____

Doç. Dr. Reşit ÇELİK, Üye
Yıldız Teknik Üniversitesi _____

Prof. Dr. Mehmet Hakan SATMAN, Üye
İstanbul Üniversitesi _____

Dr. Öğr. Üyesi Serkan AKOĞUL, Üye
Pamukkale Üniversitesi _____

Danışmanım Doç. Dr. İbrahim DEMİR sorumluluğunda tarafımca hazırlanan ALS Vaka Tespitinde Bayes Ağları ve Makine Öğrenmesi Yöntemlerinin Karşılaştırılması başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

Hasan Aykut KARABOĞA

İmza

Eşime

ve

Biricik oğluma

TEŞEKKÜR

Yıldız Teknik Üniversitesi'nde yüksek lisansa eğitimime başladığımdan beri benden akademik birikimini ve manevi desteğini esirgemeyen, tüm çalışmalarımda beni doğru kaynağa yönlendiren, çalışmalarımın her adımında zamanını, bilgisini ve tecrübesini esirgemeden beni destekleyen çok kıymetli hocam sayın Doç. Dr. İbrahim DEMİR'e en içten teşekkürlerimi sunarım.

Tez izleme jürimde yer alan, derin akademik bilgisi ve deneyimiyle bana yol gösteren kıymetli hocam Prof. Dr. Mehmet Hakan SATMAN'a teşekkürlerimi sunarım. Tezimin tüm süreçlerinde rehberlik ederek beni desteklemiş, ufuk açıcı fikirleriyle ve bilimsel sohbetleri ile tezimin bilimsel bir eser olarak hazırlanmasından keyif almamı sağlamıştır. Her zaman doğru yönlendirmeleri ve yerinde müdahaleleri ile tezin ilerleme aşamalarında beni yönlendiren değerli hocam Prof. Dr. Ersoy ÖZ'e katkılarından dolayı teşekkür ederim.

Bilimsel ilerlemenin ısrarla devam etmek olduğunu öğütleyen, istatistik bilimine yaklaşımı ile doğru soruları sorarak araştırmalarıma başlamamı sağlayan, bana akademik araştırmayı sevdiren kıymetli hocam Doç. Dr. Reşit ÇELİK'e de teşekkür ederim.

Bu zorlu süreçte bana hep destek olan, sabırla çözüm arayan, sıkıntılı zamanlarımda gözlerinde bana olan inancını hissettiğim sevgili eşim ve meslektaşım Arş. Gör. Dr. Tuğba KARABOĞA'ya teşekkür ederim. Ayrıca dualarını benden esirgemeyen annem ve babama saygı ve şükranlarımı sunarım.

Hasan Aykut KARABOĞA

İÇİNDEKİLER

ŞEKİL LİSTESİ ix

TABLO LİSTESİ xi

ÖZET xii

ABSTRACT xiv

1 GİRİŞ 1

1.1 Literatür Özeti ... 1

1.2 Tezin Amacı ... 6

1.3 Hipotez ... 6

1.4 Tezin Yapısı ... 7

2 BAYES AĞLARI 8

2.1 Grafiksel Modeller, Yollar ve Döngü ... 10

2.2 Grafiksel Ayrım ve Koşullu Bağımsızlık ... 13

2.3 Bayes Ağı ... 14

2.4 Yönsel Ayrılma Kuralı (D-Ayırma Kuralı) ... 17

2.5 Markov Özelliği ... 18

2.6 Bayes Ağlarında Öğrenme ... 19

2.7 Kalite Ölçüsü ... 20

2.8 Yapı Öğrenme ... 21

2.9 Kısıt Bazlı Algoritmalar ... 22

2.10 Skor Bazlı Algoritmalar ... 24

2.11 Karma Algoritmalar ... 25

2.12 Parametre Öğrenme ... 26

3 MAKİNE ÖĞRENMESİ YÖNTEMLERİ 28

3.1 Yapay Sinir Ağları (YSA) ... 29

3.2 Lojistik Regresyon (LR) ... 33

3.3 Naïve Bayes Algoritması (NB) ... 36

3.4 J48 Algoritması ... 38

3.5 Destek Vektör Makineleri (DVM) ... 41

3.6 KStar Algoritması ... 49

3.7 k-En Yakın Komşu Algoritması (k-NN) ... 52

3.8 Performans Karşılaştırma Kriterleri ... 55

3.9 K-kat Çapraz Doğrulama 58
4 UYGULAMA **60**
4.1 Amyotrofik Lateral Skleroz (ALS) Hastalığı 61
4.2 Katılımcılar 63
4.3 Analiz Sonuçları 64
5 SONUÇ VE ÖNERİLER **76**
KAYNAKÇA **79**
A EK BİLGİLER **91**
TEZDEN ÜRETİLMİŞ YAYINLAR **92**

SİMGE LİSTESİ

Desc(A)	A düğümünün çocuğu
Eb(A)	A düğümünün ebeveyni
info(.)	Bilgi gereksinimi fonksiyonu
freq (.)	Bir kümeye ait frekans değerleri
$\prod$	Çarpım fonksiyonu
V	Değişken Kümesi
ξ	Destek vektör makinelerinde soft hata
μi	Destek vektör makinelerinde güncellenmiş lagrange parametreleri
αi	Destek vektör makinelerinde i. lagrange çarpanı
$\perp\!\!\!\perp_G$	Grafiksel bağımsızlık
κ	Kappa istatistiği
∂	Kısmi türev
$p(.)$	Olasılık
$\perp\!\!\!\perp_p$	Olasılıksal bağımsızlık
$\bar{\gamma}$	Öznitelik alt kümesinin korelasyonu
sign (.)	Signum fonksiyonu
$\sum$	Toplam fonksiyonu
A_{ij}	V_i düğümü V_j düğümü ile bağlantı

KISALTMA LİSTESİ

ALS	Amyotrofik Lateral Skleroz
ACC	Doğruluk
BA	Bayes Ağı
ÇKA	Çok Katmanlı Algılayıcı
DVM	Destek Vektör Makinesi
ERR	Hata Oranı
FM	F-Ölçütü
FN	Yanlış Negatif
FP	Yanlış Pozitif
FPR	Yanlış Pozitif Oran
GM	Geometrik Ortalama
kStar	kStar Algoritması
kNN	kNN Algoritması
LR	Lojistik Regresyon
ML	Makine Öğrenme Yöntemleri
MCC	Matthews Korelasyon Katsayısı
NB	Naive Bayes
PREC	Kesinlik
ROC	Alıcı İşlem Karakteristiği Eğrisi
SENS	Duyarlılık
SPEC	Özgüllük
TN	Doğru Negatif
TP	Doğru Pozitif
YI	Youden indeksi
YSA	Yapay Sinir Ağı

ŞEKİL LİSTESİ

Şekil 2.1 Bir ağ örneği 11
Şekil 2.2 Döngüsel bir Bayes ağı 12
Şekil 2.3 Dört düğümden oluşan yön verilmiş ve yön verilmemiş grafik örnekleri [54] 13
Şekil 2.4 Yakınsak bağlantı, seri bağlantı ve ayrık bağlantı için grafiksel ayrım ve koşullu bağımsızlık [49] 14
Şekil 2.5 Aile ebeveyn, çocuk ve torun dışı düğümler için verilen Bayes ağı örneği 16
Şekil 2.6 Basit bir Bayes ağı örneği ve olasılıkları 17
Şekil 2.7 Bir Bayes ağında **(a)** Serisel, **(b)** Iraksayan ve **(c)** Yakınsayan bağlantılar 18
Şekil 3.1 Bir nöronun çalışma prensibi 30
Şekil 3.2 Yapay sinir ağlarında kullanılan çeşitli aktivasyon fonksiyonları [76] 31
Şekil 3.3 Çok katmanlı algılayıcı ağ yapısı 32
Şekil 3.4 Lojistik regresyon yapısı [80] 35
Şekil 3.5 Naïve Bayes yapısı 37
Şekil 3.6 Karar ağacı yapısı 39
Şekil 3.7 Lineer DVM için optimal ayırıcı hiper düzlem 43
Şekil 3.8 Lineer olmayan DVM için çekirdek fonksiyonu ile yüksek boyutlu örnek uzaya haritalama [98] 47
Şekil 3.9 Radyal temel fonksiyon çekirdeği ile oluşturulan bir DVM sınıflandırıcı [103] 49
Şekil 3.10 k-NN ile sınıf belirleme örneği 54
Şekil 3.11 5-kat çapraz doğrulama 59
Şekil 4.1 ALS hastalığının BA ve diğer makine öğrenmesi yöntemleri ile modelleme süreci 60
Şekil 4.2 Elde edilen Bayes ağı modeli 64
Şekil 4.3 Yöntemlerin genel karşılaştırma sonuçlarının grafiksel gösterimi 66
Şekil 4.4 Yöntemlerin ROC analizi sonuçları; **(a)** Bayes Ağları **(b)** Yapay Sinir Ağları **(c)** Lojistik Regresyon **(d)** Naive Bayes **(e)** J48 Algoritması **(f)** DVM **(g)** Kstar **(h)** kNN 70

Şekil 4.5 Cinsiyete karşı parkin düzeyinin (ng/mL) hastalık tipine göre gösterimi.....73

TABLO LİSTESİ

Tablo 1.1 Bayes Ağları İlgili Sağlık Alanında Yapılan Çalışmaların Özeti 4

Tablo 2.1 Aile Ebeveyn, Çocuk ve Torun Dışı Çizelgesi.....16

Tablo 3.1 İki Sınıf İçin Karşılaştırma Matrisi 55

Tablo 4.1 Hastaların Karakteristik Özellikleri 63

Tablo 4.2 Yöntemlerin Genel Karşılaştırma Sonuçları 65

Tablo 4.3 Yöntemlerin ALS, Kontrol, Nörolojik Kontrol ve Parkinson Hastalığı Alt Grupları Bazında Karşılaştırılması.....67

Tablo 4.4 Hasta Tipi için Bayes Ağları Modelinin Sorguları 71

Tablo 4.5 Bayes Ağları ile Tutulum Tiplerinin Etkileşimleri Bazında Elde Edilen Sorgular 74

ÖZET

ALS Vaka Tespitinde Bayes Ağları ve Makine Öğrenmesi Yöntemlerinin Karşılaştırılması

Hasan Aykut KARABOĞA

İstatistik Anabilim Dalı

Doktora Tezi

Danışman: Doç. Dr. İbrahim DEMİR

Bayes Ağları (BA), son on yılda birçok alanda kullanımı hızla yayılan grafiksel istatistiksel modellerdendir. Ekonomi, finans, eğitim, mühendislik, tıp ve biyomedikal alanlarda veri madenciliği amacıyla kullanılabilen bu yöntem karmaşık ilişkili veri setlerinin modellenmesinde önemli rol oynamaktadır.

Çalışmamızda amaçlanan BA yaklaşımı kullanarak Amyotrofik Lateral Skleroz (ALS) hastalığını modellemek ve tahminde bunmaktır. ALS klinik tanısı erken dönemde zor olan bir hastalıktır. Ancak ALS tanısı koymada kan testleri, diğer tanı yöntemlerine göre daha az zaman alan ve düşük maliyetli yöntemlerdir. ALS araştırmalarında, hastalığın genetik mimarisini tahmin etmek amacıyla makine öğrenmesi yöntemleri kullanılmıştır. Bu çalışmada ise Parkin protein düzeyi ve hastalık önsel bilgileri birlikte kullanılarak bireye ALS tanısı koymada yardımcı olacak bir model geliştirmektir. Elde edilen sonuçlar makine öğrenmesi yöntemleri ile karşılaştırılmıştır.

Karşılaştırma sonuçlarına göre, Bayes Ağlarının doğruluk değeri 0,887 ve ROC eğrisi altındaki alan (AUC) 0,970 olarak bulunmuştur. Bu sonuçlara göre cinsiyet ve yaşın ALS vaka türünün tahmin edilmesinde etkili değişkenler olduğu doğrulanmıştır. Ayrıca ALS hastalarında tutulum yaşama olasılığının çok yüksek olduğu tespit edilmiştir. Son olarak, Parkin protein düzeyinin Parkinson hastalığı üzerinde etkili olduğu gibi ALS hastalığı üzerinde de etkisi olabileceği ortaya çıkarılmıştır. Bu proteinin düzeyi Parkinson hastalarında çok düşük seviyelerde iken ALS hastalarında tüm kontrol gruplarına göre daha yüksek bulunmuştur.

Anahtar Kelimeler: Bayes ağları, tahmin modeli, motor nöron hastalığı, Amyotrofik lateral skleroz, Parkinson hastalığı, makine öğrenmesi.

ABSTRACT

Comparison of Bayesian Networks and Machine Learning Methods in ALS Desease Detection

Hasan Aykut KARABOĞA

Department of Statistics

Doctor of Philosophy Thesis

Advisor: Assoc. Prof. Dr. Ibrahim DEMIR

Bayesian Networks (BN) are one of the graphical statistical models that have been widely used in many fields in the last decade. This method, which can be used for data mining in economics, finance, education, engineering, medicine and biomedical fields, plays an important role in modeling complex data sets.

The aim of this study is to model and predict the disease of amyotrophic lateral sclerosis (ALS) using the BN approach. The clinical diagnosis of ALS is difficult in the early stages. However, blood tests are less time-consuming and low-cost methods compared to other diagnostic methods. In ALS research, machine learning methods have been used to predict the genetic architecture of the disease. In this study, BN were used to model ALS disease by using Parkin protein level and other individual informations. The results were compared with other popular machine learning methods.

According to the comparison results, the accuracy value of BN was 0.887 and the area under the ROC curve (AUC) was 0.970. According to these results, it has

been confirmed that gender and age are effective variables on ALS. In addition, it has been determined that the probability of experiencing involvement is very high in ALS patients. Finally, it has been revealed that Parkin protein level may have an effect on Parkinson's disease as well as on ALS disease. The level of this protein was very low in Parkinson's patients. It was found to be higher in ALS patients than in all control groups.

Keywords: Bayesian networks, prediction model, motor neuron disease, Amyotrophic lateral sclerosis, Parkinson's disease, machine learning.

1
GİRİŞ

1.1 Literatür Özeti

Bayes ağları bilimsel araştırmalarda karar vericilerin işini oldukça kolaylaştıran önemli bir tekniktir. Karmaşık verileri modelleme becerisi, tüm değişkenler arası ilişkileri dikkate alan yapısı ile modelleme ve tahmin için tercih edilen tekniklerden olmasını sağlamaktadır. Bunun yanında her yeni bilgi ile ağın tahmin değerlerinin kesinliğinin artması – yani sonsal olasılıklarının güncellenmesi – Bayes ağlarını modellemede kullanılan diğer makine öğrenmesi yöntemlerine göre üstün kılmaktadır. Bayes ağlarının belirtilen bu özelliklerini dikkate alan birçok çalışmaya literatürde rastlanmaktadır.

Oniśko vd. karaciğer teşhisinde kullanılan ve karaciğer biyopsisi yapmadan belirli medikal değerleri kullanarak hastanın tedavisi için gerekli bilgileri sağlayan HEPAR modelini kullanmıştır. Modelde hastalıkların birbirinden bağımsız olduğu ve bir kişinin yalnızca bir hastalığı olduğu varsayılmaktadır. Modelde eksik veya az sayıda veri olduğunda koşullu olasılıklar hesaplanmasını sağlamak için Noisy-OR geçidi kullanılmıştır. Karşılaştırma yapılan model sonucunda karaciğer hastalıklarının değerlendirilmesinde küçük veri setlerinde için Noisy-OR geçidi ile birleştirilen HEPAR modelinin daha başarılı sonuçları ortaya çıkmıştır [1].

Galapero ve ark., tarihsel verileri çevresel etmenlerle birleştirilerek kuzuların solunum yolu hastalığının incelenmesinde ilk kez BA kullanmışlardır. Çalışma sonucunda elde edilen çıktılar çiftliklerin verimliliğini arttırmak için belirlenen risk faktörleri gözetilerek kuzuların solunum yolu hastalıklarının kontrol altında tutulmasında kullanılmıştır [2].

Li ve Wang ürün yorgunluğunun ortadan kaldırılmasında ve müşteri tercihlerinin analizinde bir BA önermişlerdir. Bu amaçla akıllı telefonların 8 farklı özelliği

alınarak ürün yorgunluğu üzerindeki farklı etkileri analiz edilmiştir. Bunun için 100 müşteri seçilip 2 farklı ürün için ürünlerin özelliği ve bu özelliğin kullanım kolaylığını yüksek-orta-düşük olarak değerlendirmeleri istenmiştir (capacity-complexity). Ürün yorgunluğunun analizinde özelliklerin tek tek etkileri ile birlikte bunların kombinasyonları da analiz edilmiştir. Bunun için BA'nın müşteri tercihlerindeki modele yansıtma gücünden yararlanılmıştır [3].

Jongsawat ve Premchaiswadi, kullanıcıların Bayes Ağlarını herhangi bir program yüklemesi yapmadan kolaylıkla kullanacağı, istediği verileri modele alıp, istediği veriyi almayacağı analiz algoritması bulunan SMILE(Structural Modelling, Interface and Learning Engine) tabanlı bir web uygulaması geliştirmiştir [4].

Liu ve ark., dört farklı BA skorlama fonksiyonunun bir veri kümesinin dağılımını ortaya çıkarma başarısını araştırmışlardır. Yaklaşık algoritmaların öğrenilen ağları etkilemediğinden emin olmak için optimal bir yapı öğrenme algoritması kullanmışlardır. Analizin odak noktası yüzlerce veya binlerce değişken içeren büyük verilerde ağın öğrenme başarısının en iyi hangi algoritmada olduğunu ortaya çıkarmaktır [5].

Finans çalışmalarında özellikle risk değerlendirme ve risk yöntemi ile ilgili konularda BA'nın kullanımı ve popülerliği artmaktadır [6]. Bu güne kadar yönetim ve finans alanında BA iflas veya başarısızlık tahmini [7–9], kredi riski modellemesi ve kredi skorları [10–13], portföy riski analizi [14], firma performansı değerlendirmesi [15], şirket değerlendirmesi [16] ve tedarik zinciri analizi [17] gibi konularda kullanılmıştır.

Bu konuda risk değerlendirmesi ve analizi dışında diğer finans konularında da çalışmalar mevcuttur. Örneğin Zuo ve Kita, BA ile hisse senedi fiyat tahmini yapmışlardır. Çalışmada ağları sırasıyla karar verme, doğrulama ve tahmin için kurmuşlardır. Çalışma sonucunda BA'nın fiyat tahmininde AR, MA, ARMA ve ARCH modellerinden daha iyi sonuçlar verdiğini ortaya koymuşlardır [18].

Baesens ve ark. ise üç farklı gerçek kredi derecelendirme veri setinde Bayes Ağları sınıflandırıcısının performansını değerlendirmişlerdir. Çalışma sonucunda

ise Örtülü Markov Yaklaşımı destekli MCMC Bayes Ağının çok iyi performans gösterdiğini tespit etmişlerdir [19].

Gemela 1993 ile 1997 arasında toplanan 4400 finansal raporu Çek ekonomisinin temel sektörlerinden olan mühendislik sektörü ve alt sektörleri için temel finansal analiz yapabilen bir Bayes Ağında kullanmıştır. Model ise uzman bilgisi yardımıyla 10 temel finansal oran kullanılarak oluşturulmuştur [20]. İkinci çalışmasında ise Gemela, finansal derecelendirme analizinde yapay sinir ağlarını ve Bayes Ağlarını karşılaştırmıştır [21].

Wu ve ark., kırsal mülk ipoteklerini etkileyen faktörleri bulmak için iki farklı Karınca Kolonisi ile optimize edilmiş BA modelini kullanmışlardır. Verilerin modellenmesi için algoritmalar ChainACO ve K2ACO olarak isimlendirilen algoritmalar kullanılmıştır. Bu çalışma, ChainACO'nun K2ACO'dan daha ucuz olduğunu, ancak bu avantaja karşı K2ACO'nun daha iyi BA yapısına sahip olduğunu göstermiştir [22].

Artış azalış analizinde Zuo ve Kita Nikkei, Dow30 ve FSTE100 endekslerinin tahmininde BA kullanmışlardır. Sonuçlar Psikolojik sınır ve Trend tahmini analizleri ile karşılaştırılmıştır. BAnın geleneksel ağlara göre daha fazla getiri sağladığı belirtilmiştir [18].

Weber ve ark., koruma uygulamalarına, değişkenlerin güvenilirliğine ve risk analizlerine ilişkin Bayes Ağı çalışmalarının genel bir incelemesini yapmışlardır. Bu bakımdan Arıza Ağaçları (FT), Markov Zincirleri (MC) ve Petri Ağları (PN) gibi klasik bağımlılık yöntemlerini de tartışmışlardır [23].

Cattel ve Love ise müteahhitlerin risk profillerini BA ile ortaya çıkarmaya çalışmışlardır [6]. Olbryś, portföy getirilerini Polonya finansal piyasalarındaki 10 ekonomik değişken ile oluşturduğu BA yardımıyla incelemiştir [24]. Finansal endeksler ile ilgili sektörlerin farklılıklarını ortaya çıkarmak için yapılan en önemli çalışmalardan birini Chernyak & Chernyak yapmıştır [25]. Tarantola ve ark. ise çalışmalarında Bayes Ağlarını müşteri memnuniyeti araştırmasında kullanmayı önermiştir [26]. Shen finansal riskleri Bayesçi yöntemlerle

modellemenin e-lojistik yatırımlarına uygulanabilirliğini göstermek için BA modelleri önermiştir [27].

Görüldüğü üzere birçok farklı alanda BA ile ilgili çalışmalar yapılmaktadır. Sağlık konuları ile ilgili yapılan bazı örnek çalışmaların literatür özeti Tablo 1.1'de verilmiştir.

Tablo 1.1. Bayes ağları ilgili sağlık alanında yapılan çalışmaların özeti

Yazar/ Yıl	Çalışma Konusu	Araştırma Bulguları
Oniśko vd. (1998)	Karaciğer Bozukluklarının Modellenmesi	Çalışmada karaciğer teşhisinde kullanılan ve karaciğer biyopsisi yapmadan belirli medikal değerleri kullanarak hastanın tedavisi için gerekli bilgileri sağlayan HEPAR modeli kullanılmıştır [28].
Seixas vd. (2014)	Bunama, Alzheimer, Hafif Bilişsel Bozukluk Tahmini	Çalışmada bunama benzeri hastalıklarda BA kullanımını önerilmiş ve BN diğer yöntemlere göre daha başarılı sonuçlar üretmiştir [29].
Rodrigues vd. (2018)	İlaç Yan Etkilerinin Tahmini	Çalışmada ilaçların yan etkilerinin incelenmesi için BA kullanılmış ve %80 doğrulukla yan etkileri tahmin etmiştir. Elde edilen modelin başarılı bir değerlendirme sağladığını eklemişlerdir [30].
Arora vd. (2019)	Risk Modelleme ve Tahmini	Kanser ve kalp krizi riskinin tahmininde BA kullanılmıştır. Yöntemin regresyon analizinden daha kullanışlı olduğunu ve kişisel risk değerlendirmesinde ucuz bir yöntem olacağını vurgulamıştır [31].
Zhao ve Weng (2011)	Pankreas Kanseri Tahmini	Pankreas kanseri verilerinin analizinde BA tercih edilmiş ve kNN ile DVM yöntemleri ile karşılaştırılmıştır. Sonuç kısmında BA'nın kNN ve DVM yöntemine göre daha başarılı bir model olduğunu tespit etmişlerdir [32].
Lee vd. (2015)	Akciğer Kanseri Risk Faktörlerinin Modellenmesi	Radyasyon pnömonisinin risk faktörleri arasındaki olası etkileşimleri tespit etmek için BA kullanılmıştır. BA'nın AUC değerinin (0,81) olduğu ve bu değerin çok değişkenli lojistik regresyondan (0,77) önemli ölçüde iyi olduğu belirtilmiştir [33].
Orak (2020)	Arsenik Maruziyeti ve Doğum Ağırlığı İlişkisi	Çalışmada arsenik maruziyeti ve sağlık riski değerlendirmesi BA ile analiz edilmiştir. Farklı bebek ağırlıklarının incelendiği çalışmada annenin arsenik maruziyetinin bebeğin doğum ağırlığını azalttığı belirtilmiştir. Ortaya atılan BA modelinin, %82 duyarlılık ve %72 özgüllük ile doğum ağırlığının tahminini sağladığı vurgulanmıştır [34].
Ong (2019)	Pulmoner Hipertansiyon Hastalığı,	Çalışmada çocuklukta başlayan Pulmoner hipertansiyon hastalığının hastalık alt tiplerinin ayırt edilmesinde eldeki klinik verilere BA uygulanmasını

	Hastalık Alt Tiplerinin Tahmini	önermektedir. BN'ün hastalardan alınan büyük veri kümeleri kullanılarak hastalıklarının tanımlanması ve tedavisinde kullanılabilecek bir itici güç sağlayacağı tahmin edilmiştir [35].
Marvin vd. (2017)	İnsan Sağlığı Risk Değerlendirmesi	Çalışmada BA, metal ve metal oksit nanomalzemelerin insan sağlığı üzerindeki biyolojik etki riskinin değerlendirmesi için kullanılmıştır. Elde edilen bulgulara göre literatür ile benzer olarak metal ve metal oksit nanomalzemelerin tehlike potansiyelinin tahmin doğruluğu %72 ve biyolojik etki için tahmin doğruluğu %71 olarak elde edilmiştir [36].
Antal vd. (2004)	Yumurtalık tümörlerinin modellenmesi	Çalışmada literatürde yumurtalık tümörleri konusunda verilen bilgiler ile eldeki verileri kullanılarak klinik modeller için BA modeli oluşturulmuştur. Uzman bilgisi ile geliştirilen bu modelin yumurtalık tümörlerinin sınıflandırılmasında bilgilendirici bir model olarak kullanılabileceğini vurgulamışlardır [37].
Sciarretta vd. (2011)	Hipertansiyonu Olan Hastalarda Kalp Yetmezliği Riskinin Modellenmesi	Çalışmada BA metaanaliz yöntemi ile hipertansiyon ve yüksek kardiyovasküler riski olan hastalar konusunda yapılan çalışmaların bir modeli oluşturulmuştur. Makalede toplam 223.313 hastanın değerlendirildiği vurgulanmıştır. Hipertansiyon hastalarının kalp yetmezliği riskinin önlenmesinde en önemli ilaçların Diüretikler ve renin-anjiyotensin sistem inhibitörleri olduğu belirtilmiştir [38].
Mestizo-Gutiérrez vd. (2019)	Parkinson Hastalığının Modellenmesi	Çalışmada hastaların gen ifade düzeylerinin Parkinson Hastalığı için BA Modeli oluşturulmuştur. Oluşturulan modelin doğruluk düzeyi %76,1 olarak raporlanmıştır. Genlerin ilişkisinin Parkinson hastalığı için BA ile modellenmesinin önemli bir yöntem olacağı belirtilmiştir [39].
Sa-ngamuang vd. (2018)	Dang Humması Teşhisi	Tropikal bölgelerde görülen Dang humması hastalığının diğer akut ateşli hastalıklardan ayırt etmek için BA kullanılmıştır. Model yaklaşık %80 özgüllük ve %70 duyarlılık ile doğru tahminde bulunulmasını sağlamıştır [40].
Sato ve Sato (2015)	Alzheimer Hastalığı, Uyku Apnesi ve Kalp Hastalıklarının Teşhisi	Çalışmada Alzheimer hastalığı, uyku apnesi ve kalp hastalıklarının teşhisinde BA kullanılmıştır. Veri kalitesinin ağın kalitesini değiştireceğini ve bu nedenle karar verme süreçlerinde yardımcı bir araç olarak kullanılması gerektiğini belirtmişlerdir [41].

1.2 Tezin Amacı

Literatür incelendiğinde BA kullanımının son yıllarda yaygınlaştığı görülmüştür. Aynı zamanda inanç ağları olarak da ifade edilen BA karmaşık veri yapılarının daha kolay anlaşılmasını sağlaması bakımından özellikle davranış, eğitim ve finans, tıp gibi insan faktörünü içeren alanlarda daha kullanışlı bir yöntem olabileceği düşünülmektedir. BA'nda kullanılan veri setlerinin daha iyi şekilde kesikli hale getirildiği durumlarda daha güvenilir BA'nın elde edilebileceği belirtilmektedir [42].

BA'nın birçok alanda kullanımının hızla artması, yeni algoritmaların ortaya çıkması, veri madenciliği gibi alanlarda kullanımının yaygınlaşmasını sağlamaktadır. Bu çalışma ile ülkemizde konuya ilgi duyan araştırmacılara yeni bir bakış açısı kazandırarak birçok alanda geniş bir araştırmacı kitlesine hitap etmesini beklemekteyiz.

Bu tezin temel amacı BA ile ALS tanısı koymaya yardımcı bir model oluşturmaktır. BA'nın diğer makine öğrenmesi yöntemlerine göre kullanışlı yapısını ve yorumun daha kolay ve anlaşılır olması yönüyle ayrışmaktadır. Bu amaçla BA sonuçlarını irdelemek için literatürdeki popüler makine öğrenmesi yöntemlerinden Yapay Sinir Ağları, Lojistik Regresyon, Naïve Bayes Algoritması, J48 Algoritması, Destek Vektör Makineleri, KStar Algoritması ve k-En Yakın Komşu Algoritması ile karşılaştırılmıştır.

1.3 Hipotez

Güncel literatürde yaygın olarak kullanılmakta olan BA'nın ve belirtilen diğer makine öğrenmesi yöntemlerinin ALS vaka tespitinde yardımcı araç olarak kullanılabileceği gösterilmek istenmektedir.

Birçok kıstas bakımından değerlendirmenin yapıldığı bu çalışmada kullanılan BA, diğer makine öğrenmesi yöntemlerinden farklı olarak eksik verilerle de çalıştığından hastaların bilinen özellikleri ile mümkün olan hastalığının tahmin edilebileceğini savunmaktadır.

1.4 Tezin Yapısı

Bölüm 2'de BA'nın genel tanıtımına yer verilmiştir. Bu kısımda BA'nın en temel özelliklerine, ağın grafiksel gösterimine, ağın olasılıksal yapısına ve ağda düğümlerin ilişkilerine değinilmiştir. BA'da öğrenme süreçleri de bu bölümde açıklanmıştır.

Bölüm 3'te literatürde en sık kullanılan makine öğrenmesi yöntemlerinden çalışmada kullanılanlarına değinilmiştir. Bu yöntemler açıklanarak yöntemlerin özellikleri vurgulanmıştır. Ayrıca çalışmada ağ performanslarının karşılaştırılmasında kullanılan karşılaştırma ölçütlerine yer verilmiştir.

Bölüm 4'te ALS hastalığının özelliklerinden bahsedilmiştir. Ardından uygulamada kullanılan veri seti ile ilgili temel bilgiler açıklanmıştır. Bir sonraki adımda veriler BA ve diğer makine öğrenmesi algoritmaları ile modellenmiş ve elde edilen sonuçlar karşılaştırılmıştır. Karşılaştırma sonuçları son adımda değerlendirilerek karşılaştırmalar tamamlanmıştır. Bu bölümde ayrıca ALS ve diğer kontrol gruplarına göre çeşitli senaryolar oluşturularak değişken değişimleri irdelenmiştir.

Bölüm 5'te ALS hastalığı literatürü ile sonuçlar birlikte değerlendirilmiştir. Elde edilen temel çıktılar ile modeller değerlendirilmiş ve ilerleyen aşamalarda yapılabilecek araştırmalar tartışılmıştır.

2
BAYES AĞLARI

Bayes Ağları, olasılığı belirsizliğin bir ölçüsü olarak kullanan olasılıksal uzman sistemlerin (probabilistic expert systems) bir türüdür [43]. Bayes Ağları aynı zamanda Bayesçi İnanç Ağları (Bayesian Belief Networks), Nedensel Olasılık Ağları (Causal Probabilistic Networks), Olasılıksal Sebep - Sonuç Modelleri (Probablilistic Cause - Effect Models), Olasılıksal Etki Diyagramları (Probabilistic Influence Diagrams) olarak da isimlendirilmektedir.

Yapay zekâ ve makine öğrenmesi konularındaki birçok çalışmada Bayes Ağları sıklıkla kullanılmaktadır. Özellikle karmaşık koşullu ilişki yapılarının ortaya çıkarılmasında Bayes Ağları tercih edilmektedir. Bu teknikler değişkenlerin koşullu olasılıkları ile temsil edilen anlaşılması zor ilişkilerin matematiksel modelini yönlendirilmiş çevirimsiz oklu grafikler (DAG =**D**irected **A**cyclic **G**raph) ile görselleştirir. Bayes Ağlarının yapısı "***Grafiksel Kısım***" ile gösterilirken ilişkili değişkenler arasındaki koşullu olasılıklar koşullu olasılık dağılımları ile hesaplanarak "***Koşullu Olasılıklar Tablosu***" nda gösterilmektedir. Diğer bir ifade ile Bayes Ağları tablodaki rastgele değişkenler arasındaki bağımlılık ilişkilerinin anlaşılmasını ve görselleştirilmesini sağlayan grafiksel olasılık yapısı ortaya çıkaran istatistiksel bir yöntemdir [44]. Aynı zamanda bu ağ koşullu olasılıklar için çıkarsama yapılmasını sağlar. Bayes Teoremi yardımıyla araştırmacıların eldeki verilerin yanında uzman bilgisini de ağa katmasını sağlar [2].

Bayes Ağları, ele alınan değişkenler arasındaki olasılıksal ilişkileri ortaya koyan grafiksel bir modeldir. Bayes Ağlarının oluşturulmasında grafik teorisinden yararlanılır. Değişkenler hakkındaki olasılık bilgilerinin grafiksel olarak gösterimini sağlamak, değişkenlerin ortak olasılık fonksiyonunu görsel olarak incelemek, gözlemlenen verilerden en etkili sonuçları çıkarmak grafiksel modellerin en önemli özelliklerindendir.

Grafiksel modeller ile verilerin analiz edilmesinin bazı avantajları vardır. Bunların ilki modelin tüm değişkenlerden yararlanması dolayısıyla bazı verilerin eksik olduğu durumlarda da kolaylıkla kullanılmasıdır. İkincisi, nedensel ilişkileri ortaya çıkarmayı, bu ilişkilere dayalı olarak problemi daha iyi anlamayı ve değişkenlerin etkileşiminin sonuçlarını tahmin etmeyi sağlar. Üçüncüsü, veriden elde edilen bilgi ile değişkenler hakkındaki önsel bilginin Bayesçi yaklaşım ile birleştirilerek gösterilmesini sağladığından güçlü bir yöntemdir. Yani gerçek bilgi ile sübjektif bilginin birlikte kullanımına olanak sağlar. Dördüncüsü, Bayesçi istatistiksel yöntemler Bayes Ağları ile birlikte kullanıldığında modellemeye etkili ve prensipli bir yaklaşım sunarak verilerin modele aşırı uyumunun önüne geçer. Yani modeli düzgünleştirmek için uygun olan bazı verileri ayırıp onları kullanmak yerine tüm uygun verileri kullanır [45].

Bayes Ağlarının farklı makine öğrenmesi yöntemlerine göre diğer avantajları şöyledir [3]:

1. İnsan davranışlarının belirsizliğini diğer yöntemlerle modellemek neredeyse imkânsızdır. Oysa Bayes Ağları ile X özelliği müşterinin bu özelliğin kullanışlılık ve yararlılığına olan *inancının derecesi* olarak alınıp bu derecenin olasılığı modelde gösterilebilir.
2. Bayes Ağlarında bir değişkenin olasılık fonksiyonu düğümün ebeveynlerine bağlıdır. Bu nedenle daha az sayıda parametre tahmin edilmelidir. Bu ise çok sayıda değişkenin bulunduğu ağlarda önemli bir avantaj sağlar.
3. İleri beslemeli yapay sinir ağları veya bulanık mantık gibi akıllı sistemler çoğunlukla tek yönlüdür. Verilen girdilerle gerçek çıktılara yakın sonuçlar tahmin etse de “Hangi ürün kapasitesi yüksek olduğunda ürün sevilecektir?” sorusuna cevap vermemektedir. Fakat bu şekilde çift yönlü çıkarsamaya imkân sağlayan Bayes Ağlarıdır [46].
4. Bayes Ağlarının nedensel semantik yapısı nedensel önsellerin daha basit anlaşılmasına imkân sağlamaktadır. Bu özellik veriler karmaşık ve pahalı

olduğunda kullanıcının ön bilgisini kullanarak BN oluşturmasına yardımcı olmaktadır.

Bayes Ağları kesin olmayan karmaşık bilgiler içeren problemlerin modellenmesinde oldukça kullanışlı bir yöntemdir [47]. Modelde verilen düğümlerin koşullu olasılıklarının ortaya çıkarılmasında oldukça birçok farklı deneysel ve istatistiksel teknik mevcuttur [14]. Bununla birlikte, incelenen problem için model tutarlılığını garanti eden Bayes Ağı modeli oluşturmak ve geliştirmek için özel bir yöntem veya kılavuz yoktur. Bunun için araştırmacılar, geliştirilen modellerin veriler üzerinde sağlamlıklarını tespit etmek ve duyarlılık analizini doğrulamak için bazı yöntemler geliştirmişlerdir. Model geçerliliği, belirtilen yöntemlerin bir parçasıdır [48]. Bayes Ağlarının ilham verici özelliklerinden en iyisi kanıt yayılımıdır. Bahsedilen bu karakteristik özellik ile her bir düğümün olasılıklarını bir düğümden iki yönlü olarak yeni bir bilgi ile bütün yapı boyunca güncellemeye izin verir [45].

Grafiksel ağlar yön verilmiş ve yön verilmemiş grafiksel modeller olmak üzere ikiye ayrılmaktadır. Yön verilmiş modellerde iki değişken arasındaki doğrudan etkiyi göstermek için yönlü oklar kullanılır. Yön verilmiş grafiklerde tutarsızlıkların ortaya çıkmaması için başlangıç düğümüne dönüş yapılmasını sağlayan yönlü oklar art arda gelmemelidir. Böylece döngüsel olmayan grafikler elde edilecektir [49]. Bayes Ağları yön verilmiş döngüsel olmayan grafiksel modellerdendir. Markov Ağları ise yön verilmemiş grafiksel modellerdendir [36, 37].

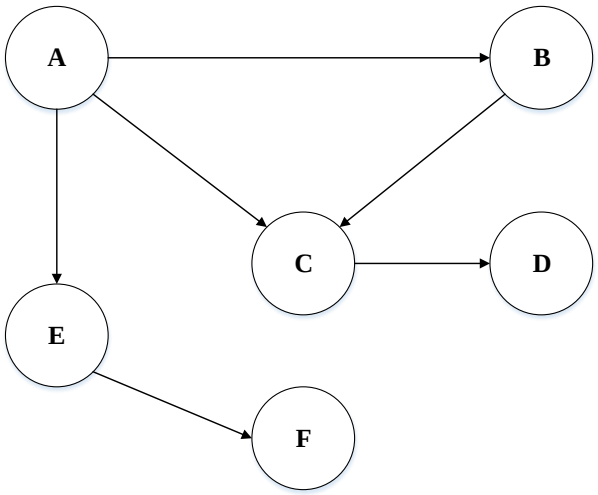
Bayes ağlarda, Markov ağlarda olduğu gibi bir yön verilmemiş grafik kullanılarak koşullu bağımsızlık varsayımları sağlanır. Grafiksel modellerin en önemli özelliklerinden birisi ***"Koşullu Bağımsızlık"***tır. Grafikte yer alan düğümlerin birbirine bağlayan kenarların oluşturulması ve değişkenler arası ilişkilerin ortaya çıkarılması konularında koşullu bağımsızlık ilişkileri oldukça önemlidir.

2.1 Grafiksel Modeller, Yollar ve Döngü

Bir grafiksel modelde $V=\{V_1, V_2, \ldots, V_n\}$ değişkenler kümesine sahip olunduğu varsayılsın. V kümesi, değişkenlerin bir kümesidir her bir değişken bir düğüme

karşılık gelir. Ayrıca bu düğümler yay veya ok şeklindeki bağlantılar ile birbirlerine bağlanırlar. V_i ve V_j düğümleri arasında bir bağlantı varsa bu bağlantı A_{ij} ile gösterilecektir. Ağda verilen tüm bağlantıların kümesi $A=\{A_{ij}$: V_i düğümü V_j düğümü ile bağlantılıdır} şeklinde gösterilir. Yani, V ve A kümeleri, tam bir grafiği tanımlar.

Bu grafik G=(V,A) ile verilmektedir. G=(V,A) grafiği V sonlu düğümler kümesi ve bu düğümlerin arasındaki bağlantıların sonlu kümesi A'dan oluşan bir yapı olarak tanımlanır. A kümesi, V'den alınan rastgele değişken çiftlerinden oluşur. Yani A kümesi (V_i, V_j) düzenlenmiş çiftler kümesidir. Burada verilen her (V_i, V_j), V'nin elemanlarıdır. Bir grafikte, düğümler rastgele değişkenleri ve kenarlar rastgele değişken çiftlerinde var olan ilişkiyi temsil etmektedirler [30, 38, 39].

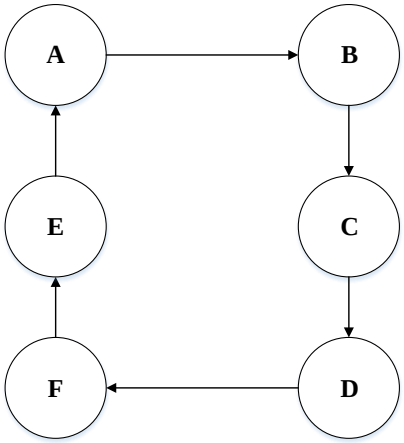


Şekil 2.1 Bir ağ örneği

Bir Bayes Ağı modelinde $(V_1, V_2) \in V$ şeklindeki iki düğüm arasında bulunan düğümlerin farklı sıralamaları ile oluşturulan bağlantılı yapıya ***yol*** denir. Şekil 2.1'de verilen ağ örneğinde A→B→C→D, A→C→D veya A→E→F yolları görülmektedir. Ağda verilen bir yolun uzunluğu yolda bulunan düğüm sayısının bir eksiği olacaktır. A→B→C→D yolunda 4 düğüm vardır ve yolun uzunluğu 3'tür.

Aynı düğüm ile başlayan ve biten yola ***kapalı yol*** denir. Aynı grafikte A ile C düğümleri arasındaki bağlantının yönü ters çevrilecek olursa uzunluğu 3 olan kapalı bir yol elde edilmiş olacaktır.

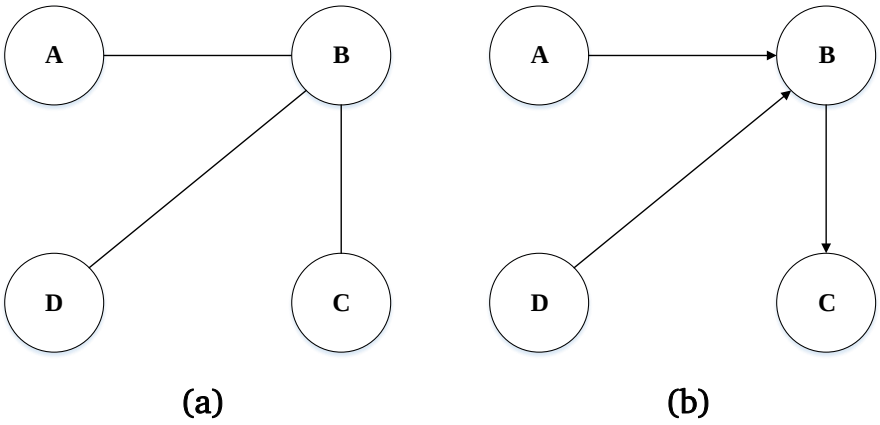
Bir grafikte aynı düğüm ile başlayıp aynı düğüm ile biten yola döngü denir. Şekil 2.2'de verilen örnek ağda A→B→C→D→ F→ E→A yolu döngüsel bir yolu göstermektedir.



Şekil 2.2 Döngüsel bir Bayes ağı

G=(V,A) grafiğinde A'daki kenarlar arasında sadece yön verilmemiş kenarlar varsa, G grafiğine ***yön verilmemiş grafik*** denir. A'da sadece yön verilmiş kenarlar varsa G grafiğine ***yön verilmiş grafik*** denir.

Verilen bir G=(V,A) grafiğinde, (V_i, V_j) ve (V_j, V_i)' nin her ikisi de A'nın elemanı ise, V_i'den V_j'ye olan bir kenara ***yön verilmemiş kenar*** veya ***yön verilmemiş bağlantı*** denir. Bu bağlantı V_i — V_j ile gösterilir. Verilen grafikte $(V_i, V_j) \in V$ ancak $(V_j, V_i) \notin V$ ise, V_i 'den V_j' ye doğru olan bir kenara ***yön verilmiş kenar*** veya ***yön verilmiş bağlantı*** denir ve $V_i \rightarrow V_j$ şeklinde gösterilir. Aşağıdaki şekilde yön verilmiş ve yön verilmemiş grafik örnekleri verilmiştir [54].



Şekil 2.3 Dört düğümden oluşan yön verilmiş ve yön verilmemiş grafik örnekleri [54]

Yukarıda verilen Şekil 2.3(a) ve Şekil 2.3(b) grafiklerinin aynı düğümler arası kenarlara sahip olduğu görülmektedir. Verilen grafiklerden Şekil 2.3(a) grafiği yön verilmemiş kenarlara sahiptir ve bu grafik V={A,B,C,D} düğümler kümesine ve A={(A,B),(B,A),(B,C),(C,B),(B,D),(D,B)} kenarlar kümesine sahiptir. Diğer grafik olan Şekil 2.3(b) grafiği ise V={A,B,C,D} düğümler kümesine sahiptir ancak A={(A,B),(B,C),(D,B)} yön verilmiş kenarlar kümesine sahiptir.

Bir grafik eğer hem döngü olmama hem de yön verilmiş olma özelliklerini aynı anda sağlarsa bu grafiğe ***yön verilmiş döngüsel olmayan grafik*** denir.

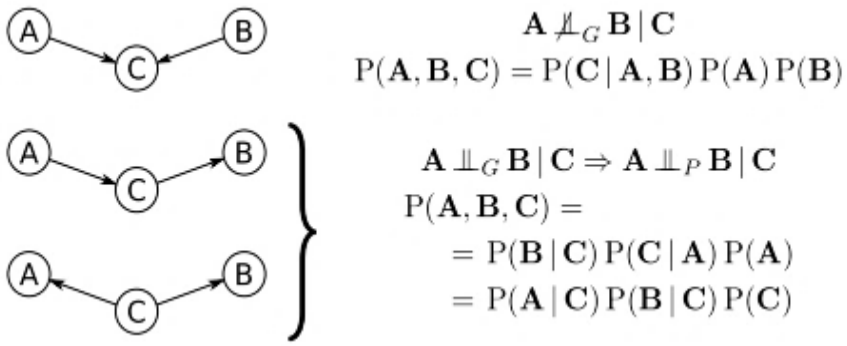
2.2 Grafiksel Ayrım ve Koşullu Bağımsızlık

Bayes Ağları, çok değişkenli olasılık tablolarının grafiksel olarak gösterilmesini sağlayan grafiksel modellerin bir çeşididir. Bu ağlar iki temel kısımdan oluşmaktadır. Bunlardan ilki $\mathbf{X}=\{X_1, X_2, \ldots X_p\}$ olan ve değişkenlerin özelliklerini tanımlayan olasılık tablolarıdır. **X** değişkeninin çok değişkenli olasılık dağılımı global dağılım, her bir X_i **X** değişkenlerinin dağılımları ise yerel dağılımlar olarak adlandırılmaktadır. İkincisi ise G = (**V**, A) olan ve DAG (***Directed Acyclic Graph***) olarak adlandırılan grafiksel yapıdır. Her bir $v \in$ V düğümü bir X_i değişkeni ile ilişkilendirilir. Grafikte verilen her bir $a \in$ A ise düğümleri birbirine bağlayan yönlendirilmiş oklardır. Bu oklar değişkenler arasındaki direkt olasılıksal ilişkiyi göstermektedir. Eğer değişkenler arasında ok yoksa bu

değişkenlerin olasılıksal bağımsızlığını veya koşullu olasılıksal bağımsızlığını ortaya koymaktadır [55].

Grafiksel bağımsızlık $\perp\!\!\!\perp_G$ şeklinde gösterilmektedir ve olasılıksal bağımsızlık ($\perp\!\!\!\perp_p$) ile birlikte değişkenler arasındaki olasılıksal ilişkinin kolayca ifade edilebilmesine olanak sağlamaktadır.

A, B ve C gibi üç farklı olayın olduğu düşünülsün. C olayının olasılığı P(C)≠0 olmak üzere, C olayı biliniyorken A ve B olayları birbirinden bağımsız ise, bu duruma C olayı verildiğinde A ve B olayları koşullu bağımsızdır denir. $(A \perp B \backslash C)_p$ veya I(A,B\C) şeklinde gösterilir. Olasılıksal olarak koşullu bağımsızlık P(A,B\C)=P(A\C).P(B\C) şeklinde ifade edilir. İstatistiksel açıdan, I(A,B\C) koşullu bağımsızlığı mevcutken C olayı bilindiğinde A olayında meydana gelecek bir değişimin B olayını etkilemeyeceği şeklinde yorumlanacaktır. B olayında meydana gelecek bir değişim de A olayını etkilemeyecektir.



Şekil 2.4 Yakınsak bağlantı, seri bağlantı ve ayrık bağlantı için grafiksel ayrım ve koşullu bağımsızlık [49]

2.3 Bayes Ağı

Çok değişkenli bir veri kümesinde nedensellik ilişkilerini grafiksel yöntemlerle göstermeye yarayan yönteme Bayes Ağı denir. Bir Bayes Ağını oluşturan üç temel bileşen vardır. Bunlar:

1. X={X_1 , X_2 ,…, X_n }Rastlantı değişkenleri kümesi
2. Değişkenlerin tümünü içeren ağın ortak olasılık dağılımı $P(U) = \prod_x P(X/Eb(X))$.

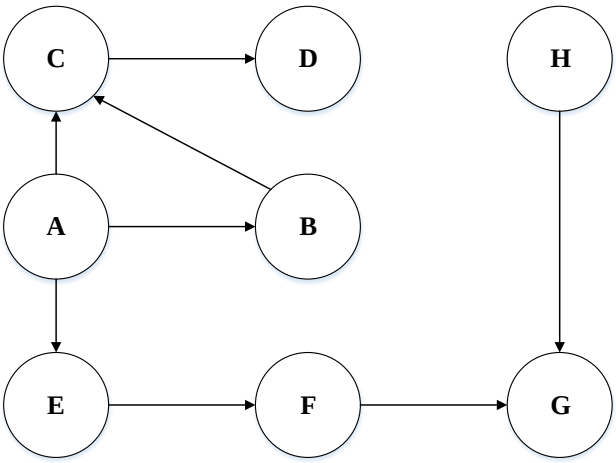
3. G=(V,A) biçiminde ifade edilen yönlü döngüsel olmayan bir grafik.

Ağda verilen yönlü bir kenar, nedensellik ilişkilerini ifade eder. Bayes Ağını oluşturan düğümler ebeveyn düğüm, torun düğüm(ler) ve torun dışı düğüm(ler) şeklinde üç temel kavram bulunmaktadır.

G=(V,A) grafiği ve herhangi bir V_i düğümü verildiğinde, bu düğümün *bitişiklik kümesi* (adjacency set), V_i 'den doğrudan ulaşılabilen düğümlerin kümesidir ve Bitş.(V_i) ile gösterilir. Yani, Bitş.(V_i) = { $V_j \in V$: $A_{ij} \in A$}'dir [54].

V_i' den V_j' ye yön verilmiş bir bağlantı $V_i \rightarrow V_j$ ise, tüm i ve j'ler için, V_i' ye V_j' nin *ebeveyni*, V_j 'ye V_i 'nin *çocuğu* denir. Mesela bir Bayes ağının içerisinde bulunan A ve B şeklindeki iki düğüm için A'dan B' ye doğru yönlü bir kenar çizilmişse A, B'nin ebeveynidir ve Eb(B)={A} veya Pa(B)={A} biçiminde gösterilir. A değişkeninin, B değişkeninin bir nedeni olduğu sonucuna ulaşılır. A değişkeninde meydana gelecek bir değişimin B değişkeninde doğrudan bir değişime neden olacaktır. B düğümüne A'nın çocuğu denir ve Desc(A)={B} şeklinde ifade edilir. Bir G grafiğinde, V_i düğümünün tüm ebeveynlerinin kümesi, tüm i'ler için Eb(G) ile gösterilir. Bir yön verilmiş grafikte, bir düğümün çocuğu ile o düğüme ait bitişik küme aynıdır. Eğer C düğümü A düğümünün ebeveyni veya çocuğu değilse Nondesc(A)={C}şeklinde gösterilir ve *torun dışı* düğüm olarak adlandırılır [56].

Bir düğümün *ailesi*, düğümün kendisini ve düğümün tüm ebeveynlerini içerir. Bir G grafiğinde, V_i 'den V_j 'ye bir yol ortaya çıkarsa tüm i, j'ler için, V_i, V_j 'nin *atası* ve V_j, V_i 'nin *torunudur* denir. Çocuğu olmayan düğüme *kısır düğüm* denir.



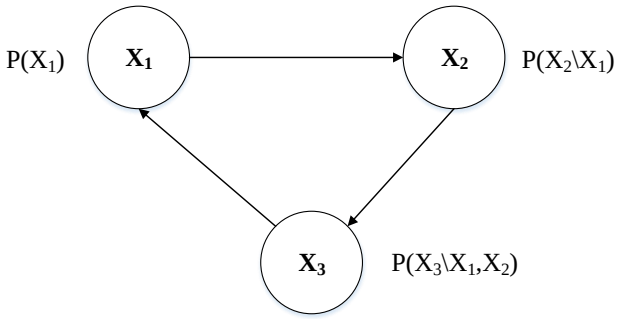
Şekil 2.5 Aile ebeveyn, çocuk ve torun dışı düğümler için verilen Bayes ağı örneği

Şekil 2.5' te verilen ağı örnek alalım. Bu ağ için her bir düğüm ile ilgili Tablo 2.1 ortaya çıkarılmıştır.

Tablo 2.1 Aile ebeveyn, çocuk ve torun dışı çizelgesi

Düğüm	Ebeveyn	Çocuk	Torun dışı	Aile
A	Ø	B, C, E	D, F, G, H	A, B, C, E
B	A	C	D, E, F, G, H	A, B, C
C	A, B	D	E, F, G, H	A, B, C, D
D	C	Ø	A, B, E, F, G, H	C, D
E	A	F	B, C, D, G, H	A, E, F
F	E	G	A, B, C, D, H	E, F, G
G	F, H	Ø	A, B, C, D, E	F, H, G
H	Ø	G	A, B, C, D, E, F	G, H

Bir diğer örnek olarak Şekil 2.6 verilmiştir. Şekil 2.6' da gösterilen ağda X_1 değişkeni X_2 ve X_3 değişkenlerinin ebeveyn düğümüdür. X_1 değişkeni ile X_2 değişkenleri de X_3 değişkeninin ebeveyn düğümleridir. Görüldüğü gibi X_2 değişkeni hem ebeveyn hem de çocuk düğümü olarak ağda yer almıştır.



Şekil 2.6 Basit bir Bayes ağı örneği ve olasılıkları

Eğer ağda herhangi bir düğüme okla bağlanmamış düğümler varsa bu düğüm ile diğer düğümlerin olasılıksal olarak bağımsız olduğunu ve ağda marjinal olasılıkları ile bulunduğunu gösterir. Bu ağın birleşik olasılıkları ise ağda yer alan düğümlerin koşullu olasılıklarının çarpımıdır. X_i çocuk ve X_j de ebeveyn düğümü olmak üzere ağın olasılığı verilen formül ile hesaplanır.

$$P(X_1, X_2, \dots, X_N) = \prod_{i=1}^{N} P(X_i \backslash X_j) \qquad (2.1)$$

Burada N değişken yani düğüm sayısıdır. Bayes Ağları, DAG yapısı ve Koşullu olasılıklar tablosu olmak üzere iki kısımdan oluşmaktadır. Bu tabloda ebeveyn düğümler ile her bir çocuk düğümün alınan değerlere göre koşullu olasılıkları bulunmaktadır.

2.4 Yönsel Ayrılma Kuralı (D-Ayırma Kuralı)

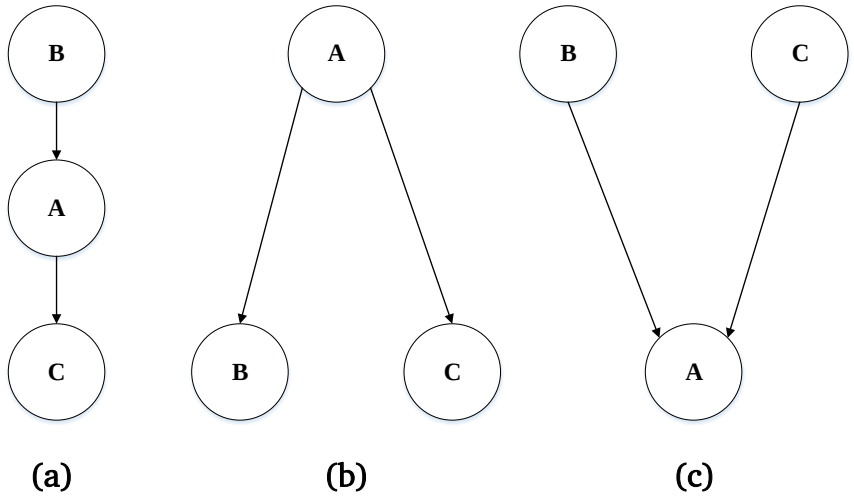
Bayes Ağlarda temel olarak Yakınsak, Iraksak ve Serisel bağlantı olmak üzere 3 çeşit bağlantı vardır. Serisel ve Iraksayan bağlantılarda bir olay ile ilgili kuvvetli bilgi var ise koşullu bağımsızlık koşulu sağlanacaktır. Ancak yakınsayan bağlantılarda koşullu bağımsızlığın sağlanması için kuvvetli ya da zayıf bir olay ile koşullu bağımsızlık sağlanacaktır.

Bayes ağın içerisinde A, B düğümleri ile bu düğümlerin arasındaki tüm yollar için X orta değişkeni bulunduğunu varsayalım. Eğer X;

1. X düğümünün durumu biliniyor ve serisel ya da ıraksayan kenarları varsa,
2. Yakınsayan kenarlar var ve ne X ne de X' in torunları olayda bulunmuyor

ise X, A' dan B' yi yönsel ayırmaktadır ya da d-ayırmaktadır denir.

A ve B düğümlerinin D-ayrılması $(A \perp B)_G$ biçiminde gösterilir. Eğer bir düğümün en az iki ebeveyni bulunuyorsa, o düğümün yakınsayan kenarları bulunmaktadır. Olasılıksal olarak X' in A ile B'yi d-ayırması, X bilindiğinde A ve B düğümlerini koşullu bağımsız olması anlamına gelmektedir.



Şekil 2.7 Bir Bayes ağında **(a)** Serisel, **(b)** Iraksayan ve **(c)** Yakınsayan bağlantılar

Şekil 2.7'de bağlantı çeşitleri verilmiştir. Buna göre birinci grafikte A düğümü bilindiğinde B ve C düğümleri koşullu bağımsızdır. İki numaralı grafikte A verildiğinde B ve C düğümleri koşullu bağımsızdır. Üçüncü grafikte ise A düğümü bilindiğinde B ve C düğümleri koşullu bağımlı olacaktır.

Birinci ve ikinci grafiklerde B ve C düğümleri arasındaki kenarlarda A kenarı yakınsamadığı için B⊥C\A koşullu bağımsızlık ilişkisi vardır. Ancak üçüncü grafikte A kenarına doğru yakınsayan iki kenar olduğu için koşullu bağımlılık ilişkisi bulunmaktadır.

Yönsel ayrım için $(A \perp B)_G$ notasyonu tercih edilmiştir. Bu notasyon olasılıksal bağımsızlık notasyonu $(A \perp B)_P$ ile karıştırılmamalıdır.

2.5 Markov Özelliği

G=(V,A) grafiği ve P ortak olasılık dağılımı verilen bir Bayes Ağı verildiğinde grafikte yer alan herhangi bir X düğümünün ebeveynleri verildiğinde X düğümü

ebeveyni ya da torunu olmayan herhangi bir düğümden yönsel ayrılmıştır. Yani ağda bulunan tüm v ∈ V düğümleri için v⊥Nondesc(v)\Pa(v) ise, Bayesci ağ Markov koşulunu sağlamaktadır. Markov koşulu dikkate alındığında yönsel ayrılma ve koşullu bağımsızlık arasında güçlü bir ilişki vardır [47].

Teorem 3.1:

G=(V,A) yönlü döngüsel olmayan grafikte herhangi (X,Y,Z) düğümleri ve P olasılık fonksiyonları için:

i. G ve P her zaman tutarlı ise $(X \perp Y \backslash Z)_G \Rightarrow (X \perp Y \backslash Z)_P$

ii. $(X \perp Y \backslash Z)_P$ G ile tutarlı tüm dağılımları sağlıyorsa $(X \perp Y \backslash Z)_G$ olacaktır [47].

2.6 Bayes Ağlarında Öğrenme

Yapay zeka ve makine öğrenmesi yöntemlerinden devşirilen "***öğrenme"*** kelimesi BA' nın model seçimi ile tahmini ifade etmektedir. BA' da öğrenme genellikle iki aşamada gerçekleştirilir [35, 38, 41]:

1. ***Yapı öğrenmesi***: Yani modelin DAG yapısının öğrenilmesidir.

2. ***Parametre öğrenme***: Önceki adımda öğrenilen DAG yapısının gösterdiği Lokal (yerel) dağılımlarının öğrenilmesidir.

Her iki adımda da öğrenme işlemi veri seti tarafından sağlanan bilgilerin kullanılması yoluyla verinin kendisinden öğrenme yani ***Öğretmensiz öğrenme*** veya modellenecek alanda uzman kişilerle yapılan görüşmeler baz alınarak öğrenme yani ***Öğretmenli öğrenme*** yoluyla yapılabilir [55].

Bu iki yöntemin birleştirilmesi en yaygın olan yöntemdir. Modelleme aşamasında önsel bilgi bir uzmanın modeli direkt olarak kurması için yeterli değildir. Hatta gen analizi gibi konularda DAG modelinde yer alacak oldukça büyük sayıda değişken olduğunda yapının oluşturulması imkânsızdır [55].

Bu işin doğası gereği tüm bilgiler Bayes teoremi ile işlenmektedir. Burada veri seti $\mathcal{D}$ ve BA da $\mathcal{B}$=(G, **X**) olsun. Eğer **X**' in global dağılım parametreleri Θ olarak belirtirsek, $\mathcal{D}$ verisinin modellemesi için Θ parametresi seçilen parametrik

dağılımlar ailesinde $\mathbf{X}'$ i benzersiz olarak tanımladığı varsayılabilir ve $\mathcal{B}=$(G, $\mathbf{X}$) şeklinde gösterilir. BA' nın öğrenimi şu şekilde formüle edilebilir:

$$\underbrace{\mathrm{P}(\mathcal{B} \mid \mathcal{D}) = \mathrm{P}(G, \Theta \mid \mathcal{D})}_{\text{Öğrenme}} \quad \underbrace{= \quad \mathrm{P}(G \mid \mathcal{D})}_{\text{Yapı Öğrenme}} . \underbrace{\mathrm{P}(\Theta \mid G, \mathcal{D})}_{\text{Parametre}} \qquad (2.2)$$

Formülden de anlaşıldığı üzere $\Pr(G, \Theta \mid \mathcal{D})$ ifadesi ayrıştırıldığında öğrenme süreci ortaya çıkmaktadır. Yapı öğrenmesi öncelikle DAG yapısının bulunması ve aşağıda verilen formülün maksimize edilmesi ile bulunacaktır.

$$\Pr(G \mid \mathcal{D}) \propto \Pr(G)\Pr(\mathcal{D} \mid G) = \Pr(G) = \int \Pr(\mathcal{D} \mid G, \Theta)\Pr(\Theta \mid G)\mathrm{d}\,\Theta \qquad (2.3)$$

Açık olarak belirtilmelidir ki ikinci kısmı G'nin Θ parametrelerini tahmin etmeden hesaplamak mümkün değildir. Bunun için verilen eşitlikten $\Pr(G \mid D)$ olasılığını herhangi bir özel seçilen Θ parametresinden bağımsız hale getirmek için denklemden integrali alınarak ayrıştırılmalıdır.

Pr (G) önsel dağılımı, $\mathbf{X}'$de yer alan değişkenler arasındaki koşullu bağımsızlık ilişkileri hakkındaki bir önsel bilginin belirtilmesi için ideal bir yol ortaya koyar. Önceki çalışmalarda elde edilen önsezileri hesaba katmak için DAG'da mevcut olan ya da olmayan bir veya daha fazla ok gerekebilir. Ayrıca modellenen örnekte akla en uygun ilişki yapısının ifade edilmesinde bazı okların belirli bir doğrultuda yönlendirilmesi gerektiği belirtilebilmektedir [35, 41].

Öğrenme yöntemleri, kalite ölçüsü ve arama algoritması olmak üzere iki elemandan oluşmaktadır [54].

2.7 Kalite Ölçüsü

Bu ölçü adından da anlaşılacağı gibi en iyi ağın seçilmesini sağlar. Yani kalite ölçüsü, mümkün durumdaki Bayes ağlarının bir kümesinden hangisinin en iyi olduğuna karar vermek için kullanılan bir ölçüdür. Parametre tahminlerinin niteliğini ve grafiksel ağın yapısını ölçen kalite ölçüsü oldukça geniş çaplıdır. Literatürde kullanılan kalite ölçüleri şunlardır [54]:

1. Bayesçi Kalite Ölçüleri (Bayesian quality mesaures)

a. Geiger ve Heckerman ölçüsü
b. Cooper- Herskovits ölçüsü
c. Standart Bayes ölçüsü

2. En küçük tanımlı uzaklık ölçüleri (Minimum description lenght measures)
3. Bilgi ölçüleri (Information measures)

2.8 Yapı Öğrenme

Bir Bayes Ağda yapı nedensel ilişkilerin grafiksel gösterimini sağlayan grafiksel yapıdır. Yapı öğrenme algoritmaları yalnızca veriden yararlanarak ağın öğrenilmesini sağlayan yöntemlerdir. Bu algoritmalarda herhangi bir ön bilgi ya da uzman görüşü kullanılmaz. Bu yöntem genellikle karmaşık ilişkilerin olduğu ve fazla sayıda düğümün olduğu ağlarda kullanılmaktadır. Yöntemlerin en temel amacı en uygun parametre değerlerini bularak veriyi en iyi şekilde temsil edecek grafiksel yapıya ulaşmaktır [44]. Fakat bu da kolay değildir çünkü elde edilecek tüm mümkün ağlar arasından en iyisini seçmek oldukça zor olacaktır. Bu durum 1977 Robinson tarafından gösterilmiştir. Düğüm sayısının artması hesaplamada çok büyük yükler ortaya çıkaracaktır. Özetle kaliteli Bayes ağların en iyi olanını seçmek için arama algoritmaları kullanılmalıdır. Bu algoritmalar 1980'li yıllardan itibaren bilgisayar teknolojilerindeki hızlı gelişim ile ortaya çıkmış ve Bayes ağların yaygınlaşmasını sağlamıştır [56]. Değişkenlerin çok küçük sayıları için bile tüm mümkün ağların sayısını değerlendirmek ve denemek çok zordur. Bu nedenle, çalışmalarda arama algoritmaları en çok üzerinde durulan algoritmalardır. Bu algoritmaların en yaygın olanları ise K2 - algoritması ve B algoritmasıdır [42].

Yapı öğrenmesi için literatürde olasılık, bilgi ve optimizasyon teorisinden elde edilen çeşitli sonuçların kullanılması sayesinde çeşitli algoritmalar ortaya atılmıştır. Yöntemlerin teorik arka planına ve terminolojinin kafa karıştırıcı boyuttaki çeşitliliğine rağmen, bunlar kısıt tabanlı, skor tabanlı ve karma algoritmalar olmak üzere üç temel başlık altında toplanmaktadır. Yapı öğrenme algoritmaları genel olarak bir dizi ortak varsayım altında çalışmaktadır [55]:

- DAG içindeki düğümler ve **X**'deki rastgele değişkenler arasında bire bir uyum olması gerekmektedir. Bu madde özellikle bazı düğümlerin tek bir değişkenin deterministik işlevleri olmamalıdır demektir.
- **X**'deki değişkenler arasındaki tüm ilişkiler koşullu bağımsız olmalıdır, çünkü bunlar sadece bir BN tarafından ifade edilebilecek ilişki türlerini tanımlamaktadır.
- **X**'deki değişkenlerin her bir olası kombinasyonu, imkânsız olay olsa bile geçerli ve gözlemlenebilir bir olayı temsil etmelidir. Bu varsayım, kesin olarak belirlenmiş Markov örtüleri ve dolayısıyla benzersiz bir şekilde tanımlanabilir bir modele sahip olması gereken pozitif bir küresel dağılımı ifade etmektedir. Kısıt tabanlı algoritmalar, bu durum doğru olmadığı zaman bile çalışmaktadır. Çünkü mükemmel bir haritanın varlığı, Markov örtülerinin tekliği için yeterli bir koşuldur [33, 41].

Gözlemler, düğüm kümesinin bağımsız gerçekleşmesi olarak kabul edilir. Eğer zamansal veya mekânsal bağımlılığın bir biçimi varsa, "***Dinamik Bayes Ağlar"*** da olduğu gibi, ağın tanımlanmasında özel olarak hesaba katılmalıdır.

2.9 Kısıt Bazlı Algoritmalar

Kısıtlamaya dayalı algoritmalar, Pearl'ün haritalar ve bunların nedensel grafik modellerine uygulanması ile ilgili çığır açıcı çalışmalarına dayanmaktadır. Geliştirdiği ***Tümevarımsal Nedensellik (****Inductive Causation = IC****) Algoritması*** [57] koşullu bağımsızlık testleri kullanarak BA'ların DAG yapısını öğrenmek için bir çerçeve sağlamaktadır.

Algoritmanın ilk adımı, hangi değişken çiftinin yönünden bağımsız olarak bir ok tarafından bağlandığının tanımlanmasıdır. Bu değişkenler, herhangi bir değişken altkümesi göz önüne alındığında bağımsız olamazlar, çünkü bunlar d-ayrıştırılamazlar. Bu adım aynı zamanda doymuş modelden başlayarak tam bir grafik ile geriye doğru seçim prosedürü olarak da görülebilir. İlişkiler koşullu bağımsızlık için istatistiksel testlere dayanarak budanır.

İkinci adımda, ortak bir komşu olan C düğümü ile bitişik olmayan A ve B düğümlerinin tüm çiftleri arasındaki v-yapılarının tanımlanır.

Tanım olarak v-yapıları, üçüncü bir düğüme bağlı olan ve bitişik olmayan iki düğümün bağlantısı koşullu bağımsız olmadığı tek temel bağlantıdır. Bu nedenle, C'yi içeren ve A ile B'yi d-ayıran düğümlerin bir alt kümesi varsa bu üç düğüm C düğümünün merkezde olduğu bir v-yapısının parçasıdır. Bu durum, A ve B için C'yi içeren ortak komşularının olası her alt kümesine karşı bir koşullu bağımsızlık testi gerçekleştirerek doğrulanabilir. İkinci adımın sonunda hem iskelet hem de ağın v-yapıları bilinir, dolayısıyla BA'nın ait olduğu eşdeğerlik sınıfı benzersiz olarak tanımlanır. IC algoritmasının üçüncü ve son adımı, önceki adımların tanımladığı denklik sınıfını tanımlayan kısmi yönlendirilmiş DAG'ı elde etmek için zorunlu okları tanımlar ve bunları tekrar tekrar yönlendirir.

IC algoritmasının en önemli problemi, olası koşullu bağımsızlık ilişkilerinin üstel olarak artması nedeniyle gerçek dünya problemlerine uygulanamamasıdır. Değişkenin Markov örtüsü elde edildikten sonra düğüm için çözüm kümeleri elde edilir. Daha sonra ise kenar ekleme çalışması yapılarak ağ tamamlanır. Burada bazı Kısıtlama tabanlı algoritmalar verilmiştir [58].

Bu kısımda Büyüme-daralma (GS), Artımsal ilişki (IAMB), Aralıklı artımsal ilişki (Inter-IAMB), Hızlı artımsal ilişki (Fast-IAMB), En az en çok ebeveynler ve çocuklar (MMPC) algoritmaları açıklanmıştır [35, 40 – 42].

Büyüme-daralma algoritması (Grow-Shrink Markov blanket): Bu algoritmada ilk olarak tüm Markov örtüleri elde edilmektedir. Ardından Markov örtüleri iki aşamalı büyüme daralma algoritması yardımıyla elde edilebilir. Markov örtüleri bulunduktan sonra yönsüz kenarlar yönlendirilir ve döngüsellik ortadan kaldırılır. Bu yöntemde koşullu bağımsızlık testleri kullanılarak yerel ağ yapıları ortaya çıkarılmaktadır [59].

Artımsal ilişki algoritması: Bu yöntem artımsal ilişkili Markov örtüsü yöntemi yardımıyla düğümlerin Markov örtülerinin hesaplanmasına dayanır. Yöntemde ilk aşamada düğümlerin ilişkili düğümler kümesi aday Markov örtülerine eklenir. Bunun için kullanılan fonksiyonun değeri maksimum olana dek algoritma işlemlere devam eder. İkinci adımda ise elde edilen Markov örtüsünde yer

almayan düğümler modelden çıkarılır. Ardından elde edilen Markov örtüleri kullanılarak kenar yönlendirme işlemi yapılıp Bayes ağı elde edilir.

Aralıklı artımsal ilişki algoritması: Artımsal İlişki algortiması ile benzer olan aralıklı artımsal ilişki algoritması ondan farklı olarak Markov örtüleri elde edilirken ileriye doğru aşamalı seçim işlemini kullanılmasıdır. Bu yöntemin avantajı daha az sayıda koşul kümesinin kullanılarak yapı öğrenme işleminin tamamlanmasıdır.

Hızlı artımsal ilişki algoritması: Diğer yöntemlere benzer şekilde bu yöntemde de büyüme ve daralma aşamaları uygulanmaktadır. Ancak Fast-IAMB algoritmasında IAMB ve Inter-IAMB algoritmalarından farklı olarak G^2 koşullu istatistiksel testi kullanılmaktadır. Burada Fast-IAMB algoritmasında en yüksek G^2 değerine sahip düğümler Markov örtüsüne eklenmektedir [56].

En az en çok ebeveynler ve çocuklar algoritması: Bu algoritma da diğer algoritmalar gibi iki aşamalıdır. Algoritmada ebeveynler ve çocuklardan oluşan kümeyi oluşturularak çalışılır. En temel fark ilk aşamadır. İlk aşamada tüm ebeveyn ve çocuklar için aday ebevenler ve çocuklar kümesi oluşturulur. İkinci aşamada da ilişkili olanlar kümede kalırken ilişkili olmayanlar kümeden çıkarılarak ağ yapısı elde edilir [56].

2.10 Skor Bazlı Algoritmalar

Skor bazlı öğrenme algoritmaları, keşifsel (heuristic) optimizasyon tekniklerinin bir Bayes ağının yapısının öğrenilmesinde uygulamasıdır. Her bir ağın veri ile uyumunu temsil eden bir ağ skoru (network score) vardır. Bu değer algoritmayı maksimize etmeyi sağlar. Konu ile ilgili çeşitli algoritmalar aşağıda verilmiştir[55]:

Rastgele yeniden başlangıçlar (*Random restarts*) veya Tabu aramalar (*Tabu search*) yardımıyla yapılan Tepe Tırmanışı (*hill-climbing*) tipi ***Açgözlü Arama*** (*greedy search*) algoritmalarıdır [60]. Bu algoritmalarda genellikle boş bir ağ yapısı alınarak ve ok eklenerek, çıkarılarak veya okun yönünü ters çevrilerek ağ skoru artık geliştirilemeyene kadar tüm seçenekler denenir.

"En uygun" modellerin tekrarlayıcı bir yolla seçimi ve özelliklerinin melezlenmesi yoluyla doğal evrimi taklit eden ***genetik algoritmalar*** (*genetic algorithms*)[61] kullanılabilmektedir. Bu durumda arama uzayı, iki ağın yapısını birleştiren *çaprazlama* operatörü ve rastgele değişiklikler getiren *mutasyon* operatörü gibi stokastik operatörler aracılığıyla araştırılmaktadır.

Benzetimli tavlama algoritmasında, ağ puanını arttıran değişiklikleri kabul ederek stokastik bir yerel arama gerçekleştirilmektedir. Ayrıca puan azalmasıyla ters orantılı olarak belirlenen bir olasılıkla onu azaltan değişikliklere izin verilmektedir.

K2-Algoritması: Bu algoritmada ilk olarak düğümler sıralanır ve bağlantılar olmadan en basit ağ tanımlanır. Her bir düğüm için ağ modeli tam olarak elde edilenceye kadar sırasıyla X_i için, X_i 'den daha küçük numaralı düğümün ebeveyn kümesi olan $Eb(m)_i$'ye eklenir. K2 algoritması kullanıldığı durumda en önemli şey ilk olarak düğümlerin bir sıralanmasını yapmaktır [62].

B-Algoritması: B-algoritması K2-algoritmasından farklı olarak düğümlerin herhangi bir sıralanmasını gerektirmez. K2-algoritmasına benzer şekilde bu algoritma da boş bir ebeveyn kümesiyle işlemlere başlar. Algoritmanın her bir adımında bir döngü elde edilmeden yeni bir bağlantı eklenir. Mümkün olan en iyi ağa ulaşana kadar bağlantı ekleme işlemine devam edilir [54].

2.11 Karma Algoritmalar

Karma öğrenme algoritmaları, kısıt bazlı ve skor bazlı algoritmaların zayıflıklarını ortadan kaldırmak ve çeşitli durumlarda daha güvenilir ağ yapıları oluşturmak için birleştirilmesi ile oluşturulmaktadır. Bunun için ilk olarak kısıt tabanlı yöntem uygulanır ve daha sonra skor tabanlı yöntemlerle bu ağ en iyi ağ haline getirilir. Yani ilk olarak düğümlerle ilişkili olan bir düğüm kümesi meydana oluşturulur. Daha sonra ise kenarları yönlendirilmek için yalnızca düğümle ilişkisi olan çözüm kümesi üzerinden skor tabanlı yöntemler uygulanarak ağ meydana getirilir [55].

Bu ailenin en tanınmış iki üyesi, ***Seyrek Aday Algoritması*** (*Sparse Candidate = SC*) ve ***Maksimum-Minimum Tepe Tırmanışı*** *(Max-Min Hill-Climbing = MMHC*) algoritmasıdır [58].

Her iki algoritma da kısıtlama ve maksimize etme (*restrict and maximise*) aşamalarına dayanır. İlk algoritmada X_i düğümünün ebeveyn adaylar kümesi tüm düğümler kümesi olan **V**' den yalnızca X_i düğümü ile ilişkili olan $\boldsymbol{C}_i \subset \mathbf{V}$ olacak şekilde alt kümeye düşürülür. Bu küme temel kümeye göre hem daha küçük hem de daha düzenli bir arama uzayı elde edilmesini sağlar. İkinci aşamada ise $\boldsymbol{C}_i$ ebeveyn kümeleri ile çeşitli kısıtların zorlamasıyla verilen ağ skorunu maksimum yapmak hedeflenmektedir [55].

Seyrek Aday Algoritmasında bu iki adım ağa eklenecek ok kalmayana ya da ağ skoru geliştirilemeyene kadar devam eder. Sezgisel yöntemin seçimi uygulama aşamasına bırakılmıştır. Diğer yandan ***Maksimum-Minimum Tepe Tırmanışı*** algoritmasında kısıtlama ve maksimizasyon işlemleri tek bir kez uygulanır. Maksimum-minimum ebeveyn ve çocuklar yöntemi (***Max-Min Parents and Children = MMPC***) ise $\boldsymbol{C}_i$ ebeveyn adayları kümesini ve optimum ağı bulmak için tepe tırmanışı açgözlü aramayı (*hill-climbing greedy search*) kullanır [55].

2.12 Parametre Öğrenme

Parametre öğrenme aynı zamanda parametre tahmini olarak da belirtilmektedir. Tahmin yöntemleri en çok olabilirlik tahmini, Bayesçi tahmin ve düzgünleştirilmiş tahmin olmak üzere üç başlık altında toplanmaktadır. Bu yöntemlerden en çok kullanılanı en çok olabilirlik tahminidir.

BA'nın ağ yapısı verilerden öğrenildikten sonra, küresel dağılımın parametrelerini tahmin etme ve güncelleme görevi, dağılımın yerel dağılımlara ayrıştırılması ile büyük ölçüde basitleştirilmiştir. Literatürde ***Bayes Tahmini*** ve ***Maksimum Olasılık Tahmini*** yaklaşımları yaygındır. BN'lerin ağ yapısını öğrenmek için kullanılan yaklaşımın parametre öğrenmede hangi yaklaşımların kullanılabileceğini belirlemesi gerekmemektedir. Örneğin, hem parametre öğrenme hem de yapı öğrenmenin her ikisinde de sonsal yoğunlukların kullanılması BN'nin yorumlanmasını ve ağdan çıkarsama yapmayı daha açık hale

getirerek kolaylaştırır. Bununla birlikte, yapısal öğrenme için Monte Carlo permütasyon testi kullanılması ve parametre öğrenme için sonsal tahminlerin kullanılması, daralma yaklaşımları ve maksimum olabilirlik parametre tahminlerinin kullanılması kadar yaygındır [35, 41].

Uygulamada yerel dağılımlar az sayıda değişken içermesine ve boyutları genellikle BN'nin büyüklüğü ile ölçeklenmemesine rağmen bazı durumlarda parametre tahmini hala sorunludur. Örneğin, modelde yer alan örneklem büyüklüğünün değişkenlerin sayısından çok daha küçük olması giderek yaygınlaşmaktadır. Bu, birkaç on veya yüz gözlem ve binlerce gen içeren mikro-işlemler gibi yüksek verimli biyolojik veri kümeleridir. “Küçük n, büyük p” olarak adlandırılan bu ortamda, hem yapı hem de parametre öğrenmede özel bir dikkat gösterilmedikçe, tahminler yüksek bir değişkenliğe sahip olacaktır [38, 49–52].

3
MAKİNE ÖĞRENMESİ YÖNTEMLERİ

Makine Öğrenmesi (ML) yöntemleri Yapay Zekânın (AI) alt alanlarından biridir ve klinik araştırmalarda yaygın halde kullanılmaktadır [53, 54]. Makine öğrenmesi yöntemleri, gerçek durum ile tahmin edilen durum arasındaki farkı en küçük hale getirmek için kullanılan çeşitli algoritmalar bütünüdür. Bu algoritmalar ile eldeki verilerden çeşitli modeller oluşturulur ve oluşturulan modeller ile gerçekleşecek durumlar için tahminler yürütülür. Eldeki veri seti büyüdükçe değişkenler arasındaki ilişkinin ortaya çıkarılması daha da zorlaşmaktadır. Veri madenciliği ise ortaya çıkarılması zorlu olan ilişkileri ortaya çıkarmayı ve büyük veri yığınlarını anlamlı bilgilere dönüştürmeyi hedeflemektedir. Hedeflerin gerçekleştirilebilmesi ise bilgisayar programlarına ve güçlü algoritmalara ihtiyaç vardır [55, 56]. Makine öğrenmesi algoritmaları bu aşamada devreye girmektedir. Veri madenciliği makine öğrenmesi algoritmalarının büyük veri setlerine uygulanmasıdır [71].

Ancak bu durumda da makine öğrenmesi yöntemlerinin yani üretilen algoritmaların hesaplama karmaşıklığı ortaya çıkmaktadır. Karmaşıklık ise eğitim karmaşıklığı ve eğitilmiş algoritmayı uygulamanın karmaşıklığı olmak üzere iki kısma ayrılır. Algoritmaların eğitimi çok sık gerçekleşmez. Zaman açısından kritik değildir ve eğitim bu nedenle daha uzun sürebilir. Bununla birlikte, model test aşamasında hızlı bir şekilde karar vermek gereklidir ve bu nedenle testin daha az maliyetli ve daha hızlı olması gerekir [70].

Makine öğrenmesi yöntemleri, veri madenciliği için kullanılan algoritmaları içermektedir. Bu yöntemler ile sınıflandırma, tahmin, kümeleme ve birliktelik kurallarının ortaya çıkarılması şeklinde 4 temel görev gerçekleştirilir. Makine öğrenmesi yöntemleri denetimli, denetimsiz ve destekleyici öğrenme yöntemleri olmak üzere 3 başlık altında toplanmaktadır [72]. Veriler ile çıktı olarak

etiketlenmiş bağımlı değişkenlerin olduğu durumlarda denetimli öğrenme, bağımlı değişken bilinmiyorsa denetimsiz öğrenme kullanılmaktadır.

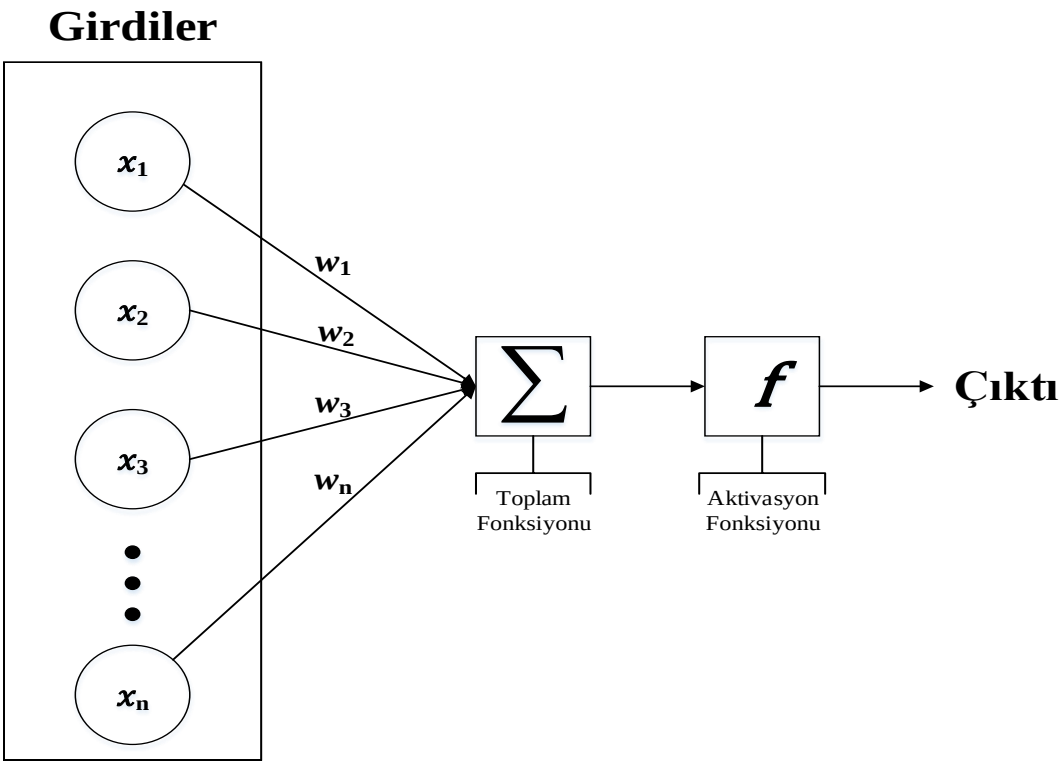
Denetimli öğrenme yöntemleri, yaş, cinsiyet, başlangıç türü gibi bilinen durumlara dayalı olarak bilinmeyen durumlar (örneğin hastalık türü) hakkında tahminlerde bulunmayı amaçlamaktadır [53, 58]. Sınıflandırma, benzerlik tespiti ve regresyon, denetimli makine öğrenimi yöntemlerinin en yaygın görevleri arasındadır [73]. Özetle, yapılacak çalışmanın amacına göre farklı makine öğrenmesi yaklaşımları tercih edilmelidir.

Çalışmamızda, Bayes Ağları ile birlikte Yapay Sinir Ağları, Lojistik Regresyon, Naif Bayes Algoritması, J48 Algoritması, Destek Vektör Makineleri, KStar Algoritması ve K-En Yakın Komşu Algoritması olmak üzere yedi popüler denetimli makine öğrenmesi yöntemi kullanılmıştır.

3.1 Yapay Sinir Ağları (YSA)

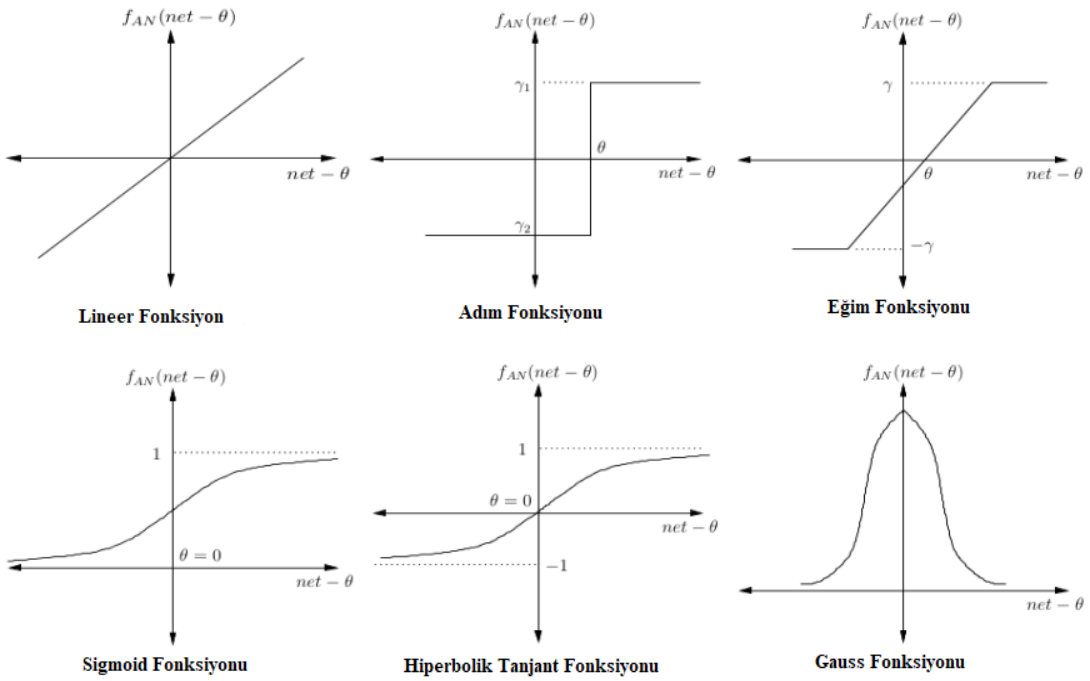
YSA, öğrenme ve genelleme yetenekleri ile insan beynini taklit eden önemli makine öğrenme yöntemlerindendir. İlk örneği 1957 yılında F. Rosenblatt tarafından geliştirilen yapay sinir ağları başlangıçta temel matematiksel mantık problemlerinden olan AND ve OR problemlerinin çözümü için kullanılmıştır. Ancak XOR gibi daha zor problemlerin çözümünde geliştirilen bu ağlar yetersiz kalmıştır [74]. Bu problemlerin çözümü için daha karmaşık yapıda olan Çok Katmanlı Algılayıcılar (ÇKA) geliştirilmiştir. ÇKA günümüzde de en popüler YSA algoritmalarındandır [75].

Basit bir nöron modelinde 3 ana unsur bulunur. Bunlar girdi katmanı, birleştirme (net) fonksiyonu ve aktivasyon fonksiyonudur.



Şekil 3.1 Bir nöronun çalışma prensibi

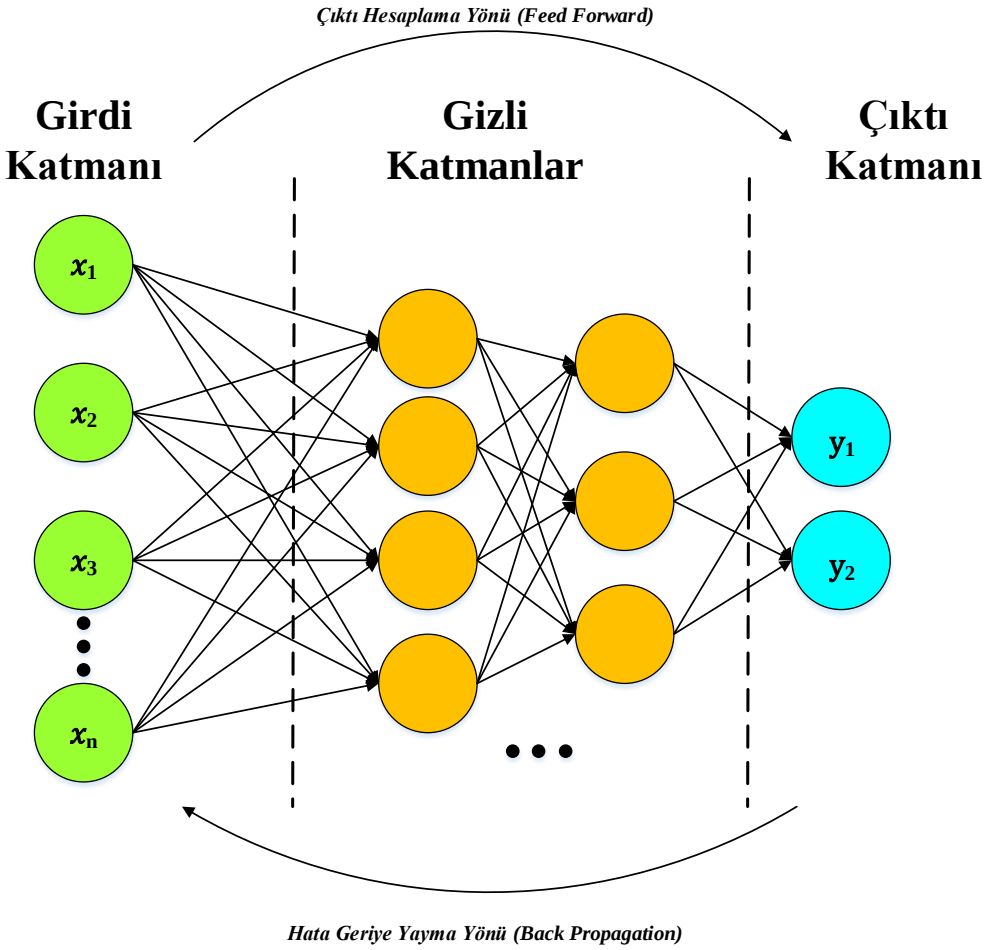
Şekil 3.1' de bir yapay nöronun çalışma prensibi gösterilmiştir. Buna göre veriler Girdi katmanı ile modele alınır. Girdi olarak ele alınan $(x_1, x_2, x_3 \ldots, x_n)$ verileri ile $(w_1, w_2, w_3 \ldots, w_n)$ ağırlıkları çarpılarak toplamı alınır. Elde edilen değer ise aktivasyon fonksiyonuna aktarılır. Bu değer lineer fonksiyon, hiperbolik tanjant fonksiyonu, sigmoid fonksiyonu gibi aktivasyon fonksiyonlarıyla işlenerek (bir sonraki katmana iletilmek üzere) çıktı olarak üretilir. Çeşitli aktivasyon fonksiyonları Şekil 3.2 ' de verilmiştir.



Şekil 3.2 Yapay sinir ağlarında kullanılan çeşitli aktivasyon fonksiyonları [76]

Ayrıca birleştirme fonksiyonu yalnızca toplam fonksiyonu olmak zorunda değildir. Birleştirme fonksiyonu olarak birçok farklı matematiksel fonksiyon kullanılmaktadır. Bu fonksiyonlar problemin yapısına göre değişmektedir. Toplam fonksiyonu dışında yaygın olarak kullanılan birleştirme fonksiyonları çarpım fonksiyonu, maksimum fonksiyonu, minimum fonksiyonu, signum fonksiyonudur [76].

Bahsedilen fonksiyonlar istenen çıktı ile uyumlu olarak tahminler yapılmasını sağlamalıdır. ÇKA hem lineer hem de lineer olmayan problemlerin çözümünde etkili sonuçlar üretmektedir. Bu nedenle en çok tercih edilen denetimli öğrenme algoritmalarındandır. Bir ÇKA'nın ağ yapısı Şekil 3.3' te verilmiştir. Verilen ağ girdi katmanı, gizli katman ve çıktı katmanı olmak üzere 3 temel katmandan oluşmaktadır. Ağın tasarımında gizli katman sayısı, ağda bilginin yayılma durumu, ağın besleme yapısı, öğrenme algoritması, hata algoritması ve her katmandaki düğüm sayısı ağın mimarisini oluşturmaktadır.



Şekil 3.3 Çok katmanlı algılayıcı ağ yapısı

Şekil incelendiğinde ağın hata geriye yayılımlı ve ileri beslemeli bir ağ olduğu görülmektedir. Doğrusal olmayan, karmaşık modeller ile eksik verilerde çalışması en önemli avantajlarındandır. Ağın gerçekleştirmesi istenen görevin karmaşıklığına göre ağın yapısı da karmaşıklaşmaktadır. Gizli katman sayısı ve katmanlarda yer alan düğüm sayıları görev yapısına göre değişmektedir. Bu nedenle ağın gücü sadece ağın büyüklüğü ile değil ağdaki birimlerin tasarımı ile de ilgilidir [71]. Modellerin eğitimi esnasında hatalar geriye yayılım algoritmaları ile optimize edilir. Diğer taraftan istatistiksel yöntemlerde bulunan katı hipotezlerin gerekli olmaması ÇKA'yı modelleme konusunda avantajlı kılmaktadır [61, 63].

3.2 Lojistik Regresyon (LR)

Bireylerin var olan iki gruptan hangisine üye olduğu lojistik regresyon analizi ile tahmin edilmektedir. Bunun için klasik regresyonda olduğu gibi bir regresyon denklemi oluşturulur. LR, bağımlı değişkenin yapısı nedeni ile standart regresyon modellerinden farklılaşmaktadır. Ancak LR da lineer regresyon modellerinde olduğu gibi bağımlı ve bağımsız değişkenlerin ilişkileri araştırılmaktadır. Burada en önemli fark LR da bağımlı değişkenin dikotom olmasıdır. Bağımsız değişkenler ise dikotom, kesikli, sürekli ya da bunların bir karışımı olabilmektedir. Uygulama açısından LR, standart lineer regresyonla benzerdir [78].

LR, hipotez olarak ayırma analizi ve regresyon analizi ile benzer soruları yanıtlamaktadır. LR, bağımlı değişkenin ve bağımsız değişkenlerin dağılımı, bağımlı değişkenin ve bağımsız değişkenlerin doğrusal ilişkili olması, grup içi eş-varyanslılık, varyans-kovaryans matrislerinin eşitliği gibi varsayımları olmadığından bu tekniklerden daha esnek yapıdadır [79]. Az sayıda veri ile çalışabilmesi de avantajlarından biridir. Bir diğer önemli nokta ise parametre tahmini ve çıkarsama yaklaşımıdır. Klasik doğrusal regresyon analizinde bağımlı değişkenin normallik varsayımı sağlandığında En Çok Olabilirlik tahmini ile En Küçük Kareler tahmini eşit sonuçlar üretmektedir. Bu yöntemlerden En Küçük Kareler tahmini yalnızca bağımlı değişkenin normal dağılım varsayımı sağlandığında kullanılabilmektedir. En Çok Olabilirlik yöntemi hem doğrusal hem de doğrusal olmayan karmaşık modellerin tahmininde uygulanabilir. Lojistik regresyon doğrusal olmayan bir model olduğundan En Küçük Kareler tahmini kullanılamaz ve iteratif bir yaklaşım olan En Çok Olabilirlik tahmini tercih edilir [80].

Lojistik regresyonda temel olarak lojistik fonksiyonu kullanılmaktadır. Fonksiyon denklem (3.1)'de verilmiştir.

$$\hat{Y} = f(z) = \frac{e^{z}}{1+e^{z}} \tag{3.1}$$

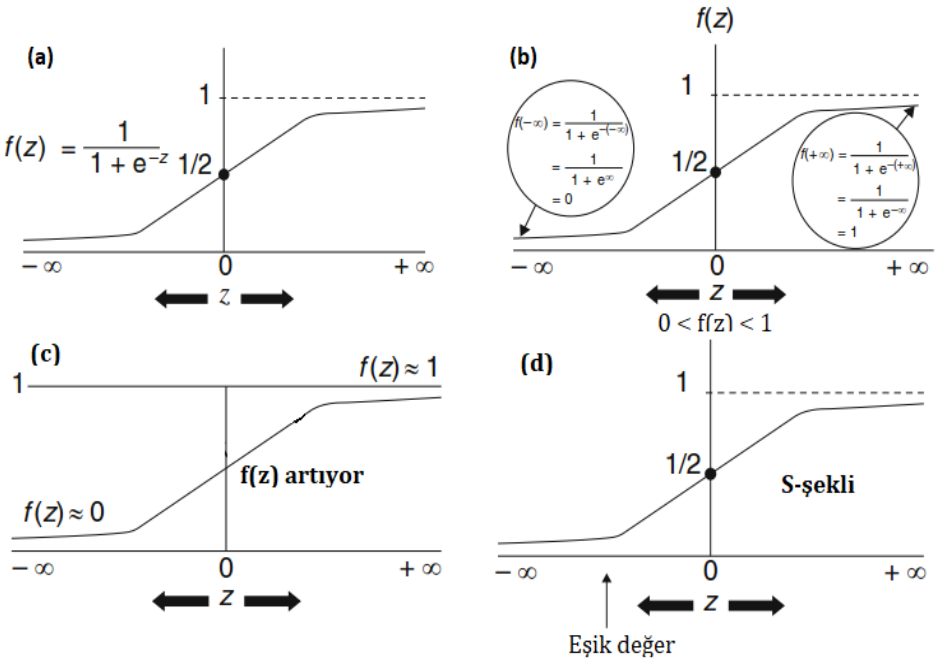
Bu eşitlikte z değeri klasik regresyon denklemidir. Buna göre z ifadesi aşağıda verilmiştir.

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{3.2}$$

Verilen bu denklemlerin olasılıksal değere dönüştürülmesi için bir orana ihtiyaç duyulmaktadır. Bu oran odds oranıdır. Odss oranı istenen durumun istenmeyen durumun kaç katı olasılıkla gerçekleşebileceğini ifade eden değerdir. Katsayıların hesaplanmasında lojit fonksiyonu kullanılır. Lojit fonksiyonu odds oranının doğal logaritmasıdır. Denklemde i numaralı deneğin tahmin edilen sınıf olasılığı $\hat{Y}_i = f(z)$ olacaktır. Lojit fonksiyonu denklem (3.3)'te verilmiştir.

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{3.3}$$

Olasılık değerleri 0 ile 1 aralığında değerler alıyorken lojit fonksiyonu -∞ ile +∞ aralığında değerler alır. İkiden çok durumun olduğu durumlarda da bağımlı değişkeni tahmin etmek için LR uygulanabilmektedir [80]. Lojistik regresyon biyoloji ve sağlık bilimi uygulamalarında en yaygın kullanılan yöntemlerden biridir [81]. Diğer modelleme ve makine öğrenmesi yöntemleri de yaygın olarak kullanılsa da lojistik regresyon, hastalık durumu ile ilgili kesikli epidemiyolojik verilerin analizinde kullanılan en popüler modelleme yaklaşımıdır [65, 66]. Bunun nedeni Şekil 3.4'te gösterilmiştir.



Şekil 3.4 Lojistik regresyon yapısı [80]

Lojistik regresyonun matematiksel formunu tanımlayan lojistik fonksiyonu (a)'da gösterilmiştir. Fonksiyonda verilen ve daha önce de ifade edilen z değeri -∞ olduğunda f(z)=0 ve z değeri +∞ olduğunda f(z)=1 olacaktır. Modelin her zaman 0 ile 1 arasında bir olasılık değerini temsil etmesi için tasarlanmıştır. Epidemiyolojik olarak ifade edilecek olursa böyle bir olasılık, bir bireyin hastalığa yakalanma riskini göstermektedir. Bu değer her durumda 0 ile 1 arasında olacaktır. Bu sayede hiçbir zaman 0'ın altında ya da 1'in üstünde olasılık değeri risk tahmini yapılmayacaktır. Şekil 3.4 (c) de gösterildiği gibi olasılık 0'dan 1'e doğru artmaktadır. Fonksiyonun S şekli epidemiyolojik olarak z hastalığa yakalanma risk faktörlerinin bir birleşimini oluşturuyorsa f(z) verilen z değeri için hastalık riskini göstermektedir. Eşik değerin altında hastalık riski oldukça düşük iken eşik değerin üstünde oldukça yüksek riski temsil etmektedir. Özetle S-şekli bir modelin epidemiyolojik bir araştırma sorusunun çok değişkenli doğasını göstermesi için geniş çapta uygulanabilir yapısını ortaya koymaktadır [80].

3.3 Naïve Bayes Algoritması (NB)

Naïve Bayes Algoritması, Bayes Teoremine dayanan ve veri madenciliği alanında araştırmacıların en çok tercih ettiği önemli makine öğrenmesi yöntemlerinden biridir. Bu yöntem değişkenlerin bağımsızlığı koşuluna dayanan klasik bir Bayes Ağıdır. Denetimli öğrenme yöntemi olan NB, makine öğrenmesi yöntemi olmasının yanında belirsizlikleri kontrol etmeye yarayan istatistiksel bir yaklaşımdır. NB yönteminde tahmin edilecek sınıfların birbirinden bağımsız olması gereklidir [82]. Basit olmasına rağmen tıbbi uygulamalarda NB teşhis ve tahmine dayalı sorunların çözümünde oldukça başarılı sonuçlar üretmektedir [69, 70]. Algoritmanın formülü aşağıda verilmiştir:

$$p = (C = c_j \mid X_i = x_i) = \frac{p(C = c_j)\, p(X_i = x_i \mid C = c_j)}{p(X_i = x_i)}; i = 1,\ldots,n;\ j = 1,\ldots,k \tag{3.4}$$

Denklemde verilen C sınıf değişkeni ve c_j sınıfın değerlerini; X ise gözlenen değişkenleri (sınıf değişkenine ait gözlenen özellikleri) ifade eder. Verilen formüle göre sınıf değişkeni tahmin edilir. Naïve Bayes algoritmasında öznitelik değerleri koşullu olarak bağımsız olduğu varsayıldığından Bayes Teoremi yardımıyla denklem (3.5)'te verildiği gibi yazılabilir.

$$p = (C = c_j \mid X_i = x_i) \propto p(C = c_j) \prod_i^n p(X_i = x_i \mid C = c_j) \tag{3.5}$$

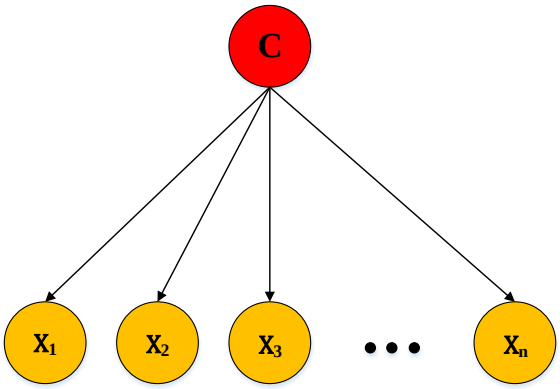
$$p = (X \mid C = c_j) = \prod_i^n p(X_i = x_i \mid C = c_j) \tag{3.6}$$

Denklem (3.7)'de verildiği gibi hesaplanan bu sonsal olasılıklar içerisindeki en büyük olasılık değerine göre sınıf belirlenir. Sonsal olasılıkların kullanıldığı bu ifade en büyük sonsal olasılık (Maximum A Posteriori: MAP) olarak da bilinir [85]. Denklem (3.8)' de C_{MAP} bir örnek için tahmin edilen sınıf değerini göstermektedir.

$$\arg\max_{c_j \in C} \left(p(c_j) \prod_i^n p(x_i \mid c_j) \right) \tag{3.7}$$

$$C_{MAP} = \arg\max_{C} \prod_{i}^{n} p(x_i \mid c_j) \tag{3.8}$$

Gözlenen değişkenler kesikli ya da sürekli formda olabilmektedir. Kesikli olan değişkenler için belirli bir aralık verilmektedir. Sürekli değişkenler var ise bu değişkenlerin bir dağılımdan geldiği varsayılır. Genel olarak bu dağılım normal dağılımdır [86]. Bir Naïve Bayes sınıflandırıcısının grafiksel olarak gösterimi Şekil 3.5' te gösterilen forma sahiptir. Şekilden de anlaşılacağı gibi tüm yaylar sınıf değişkeninden bağımsız (tahminci) özniteliklere yönlendirilir. Ağda varsayım gereği diğer değişkenler arasında bir yay yoktur.



Şekil 3.5 Naïve Bayes yapısı

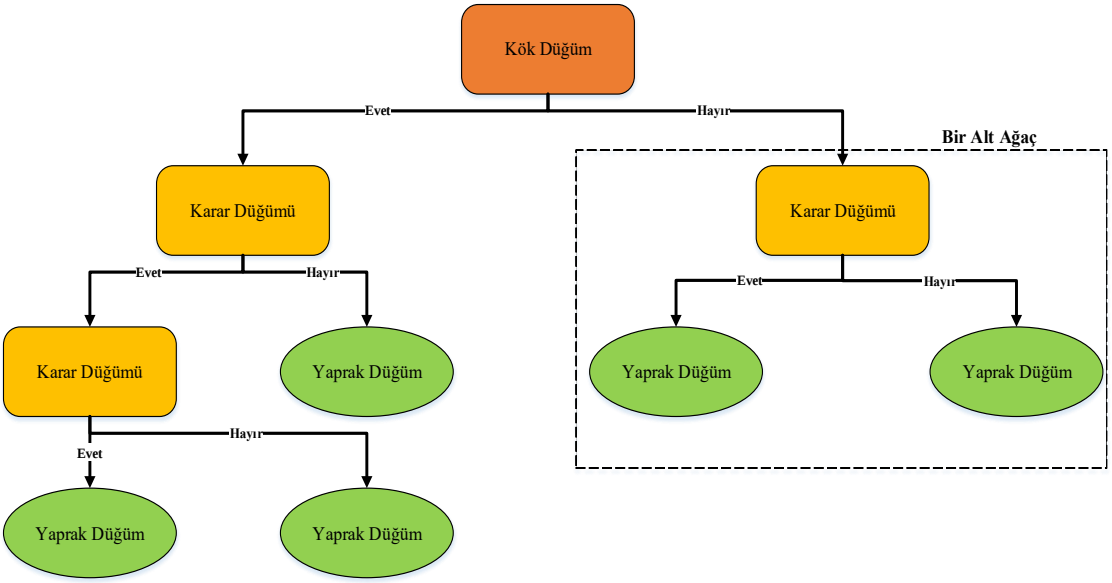
NB algoritmasının ortaya koyduğu rekabetçi performans oldukça şaşırtıcıdır. Çünkü NB algoritmasının dayandığı en temel varsayım olan '***koşullu bağımsızlık varsayımı***' gerçek dünya uygulamalarında nadiren doğrudur [87]. Ayrıca hiçbir saklı değişkenin ya da latent özniteliğin sınıf tahmini sürecini etkilememesi gerekmektedir [88]. Bu başarısını yanlış sınıflandırma oranı olarak da bilinen sıfır-bir kayıp fonksiyonuna borçludur. Değişkenlerin bağımsız olmadığı durumlarda dahi bu varsayım geniş bir marjla ihlal edildiğinde bile sonuçların uygun sınıfta olabileceğini göstermektedir [89]. Yani NB'in yeni bilgi geldikçe her bir sınıfın tahmin edilen sonsal olasılıklarını değiştirebileceği anlamına gelir. Ancak olasılıklar değişse de maksimum sonsal olasılığa sahip sınıf genellikle değişmez. Dolayısıyla, tahmin edilen olasılık zayıflamasına rağmen sınıflandırma hala doğrudur.

3.4 J48 Algoritması

J48 Algoritması karar ağaçları, popüler makine öğrenmesi algoritmalarını içermektedir [90]. J48 algoritması en önemli karar ağacı algoritmalarından biridir. Özellikle küçük boyutlu verilerin sınıflandırması için ortaya atılan bir yöntemdir. Bu algoritma ID3 [91] ve C4.5 algoritmalarının [78, 79] geliştirilmiş halidir. Bu algoritma karar ağacının oluşturulmasında C4.5, C5.0 ve ID3 algoritmalarını kullanıyorken tahmin için gini indeksi, entropi azaltma veya bilgi arttırma gibi ölçütler kullanılmaktadır [76, 79]. Önemli özelliklerinden bir diğeri ise diğer karar ağaçlarına göre daha küçük bir ağaç oluşturarak tahmin yapabilmesidir. Bu ise J48 algoritmasının benzerlerine göre daha başarılı sonuçlar üretmesini sağlamaktadır [94].

J48 Algoritması, C4.5 algoritmasının WEKA programı için geliştirilen sürümüdür. Sınıflandırma aşamasında modelleme için yukarıdan aşağıya bölerek bir karar ağacı oluşturulur [90]. Ağacın oluşturulmasında eksik veriler göz ardı edilmektedir. Daha sonra bu veriler gözlemlenen özelliklerden tahmin edilebilmektedir. Karar ağaçlarında olduğu gibi kurallar tanımlanır ve bu kurallar kullanılarak analiz gerçekleştirilir. Büyük ağaçlar için bilginin daha anlaşılır olması ve verilerin daha iyi kategorize edilmesi için budama işlemi de gerçekleştirilir. Budama işlemi ileriye doğru ya da geriye doğru yapılabilmektedir [60, 71, 76, 81].

Ağaç diyagramında değişkenler düğümleri oluşturur. Değişkenler kesikli değişkenler olmakla beraber, sürekli değişkenler ile de ağaç diyagramı oluşturulabilmektedir. Diyagram bir kök düğüm ile başlar. Karar düğümleri ve yaprak düğümler ile devam eder. Şekil 3.6’ da bir karar ağacının temel yapısı verilmiştir. Alt ağaçlara geçtikçe burada yer alan düğümler bir değişkeni ifade etmektedir. Her bir yaprak düğüm ise buradaki sınıfta yer alacak değerleri göstermektedir. Diyagram verileri olabildiğince benzer özellikler gösteren küçük alt sınıflara ayırmaktadır [71, 81].



Şekil 3.6 Karar ağacı yapısı

Hangi düğümün kök düğüm olacağına, alt ağaçlarda hangi niteliğin test edileceğini seçmek için ID3 algoritması kullanılmaktadır. Her aşamada örnekleri sınıflandırmak için en yararlı olan özelliği belirlemek gerekmektedir. ID3, ağacı büyütürken her adımda aday özellikler arasından seçim yapmak için bilgi kazancı ölçüsünü kullanır. Bu ölçü belirli bir özelliğin eğitim örneklerini hedef sınıflandırmalarına göre ne kadar iyi ayırdığını ölçen istatistiksel bir özelliktir [71, 76]. En yüksek bilgi kazancını veren öznitelik ayırıcı öznitelik olarak alınır. Bu öznitelik, ortaya çıkan bölümlendirmede alt örneklemleri sınıflandırmak için gereken bilgileri en aza indirir ve en az rastlantısallığı sağlar. Böyle bir yaklaşım, belirli bir alt örneklemi sınıflandırmak için gereken beklenen test sayısını en aza indirir. Bunun için en basit olmasa da basit bir ağacın bulunmasını garanti eder [95].

Bilgi kazancı kriterine göre bir mesaj ile iletilen bilgi, bu bilginin olasılığına bağlıdır. S ile ifade edilen bir dizi vakadan rastgele bir örnek seçilip bunun C_j sınıfına ait olduğu bildirildiğinde bu bilginin bir olasılığı $p_i = \frac{freq(C_j, S)}{|S|}$, dır. İletilen bilginin büyüklüğü denklem (3.9) ile bit cinsinden ölçülebilir.

$$-\log_2\left(\frac{freq(C_j,S)}{|S|}\right) \tag{3.9}$$

Denklemde $freq(C_j,S)$ S örneklemindeki $\{C_1,C_2...,C_k\}$ sınıflarından C_j sınıfına ait elemanların frekans değeridir. $|S|$, S örnekleminin büyüklüğünü göstermektedir. Sınıf üyeliğiyle ilgili böyle bir mesajdan beklenen bilgiyi bulmak için denklem (3.10) kullanılır.

$$Info(S)=\sum_{j=1}^{k}-\left(\frac{freq(C_j,S)}{|S|}\right)\times\log_2\left(\frac{freq(C_j,S)}{|S|}\right) \tag{3.10}$$

T kümesini, $\{T_1,T_2...,T_n\}$ şeklinde birbirini dışlayan n sonuçlu (outcome) alt kümelerine ayıran (training cases) olası bir testimiz olduğunu varsayalım. $Info(T)$, T'deki bir vakanın sınıfını belirlemek için gereken ortalama bilgi miktarını ölçer. Bu miktar aynı zamanda S kümesinin entropisi olarak da bilinir [96]. Entropi, beklenmeyen bir durumun gerçekleşme olasılığıdır ve bir eğitim veri setindeki karmaşıklığın ölçüsüdür. Bilgi kazancı ölçüsü, basitçe, bu tanımdan yola çıkılarak örnekleri bu özelliğe (değişkene) göre bölümlere ayırmanın neden olduğu entropide beklenen azalmadır [71, 78, 81].

T veri kümesini bir A testinin n farklı sonucuna göre bölümlendirildikten sonra yukarıda verilen ölçüme benzer bir ölçüm ele alalınsın. Beklenen bilgi gereksinimi, alt kümeler üzerinde ağırlıklı toplam olarak bulunabilir.

$$Info_A(T)=\sum_{i=1}^{n}\frac{|T_i|}{|T|}\times Info(T_i) \tag{5.11}$$

A testine göre T'nin bölümlendirilmesinden kazanılan bilgi denklem (3.12) ile ölçülür. Kazanç kriteri, seçilen test ile sınıf arasındaki karşılıklı bilgi olarak da bilinen bu bilgi kazancını maksimize etmek için bir test seçer.

$$Gain(A)=Info(T)-Info_A(T) \tag{3.12}$$

Kazanç kriterinin çok sayıda çıktısı olan durumlarda sonuçları sapmalı olmaktadır. Yalnızca bir örneği tanımlayan ID gibi bir öznitelik seçildiğinde bilgi kazancı sıfır olacaktır çünkü tüm örneklem bir sınıf oluşturarak bölümlenecektir. Tahmin açısından bakıldığında, böyle bir ayrım oldukça yararsızdır. Kazanç kriterinin doğasında bulunan sapma, birçok sonucu olan testlere atfedilebilen görünür kazancın ayarlandığı bir tür normalleştirme ile düzeltilebilir.

Normalleştirme kazanç oranı ölçüsü kullanılarak yapılabilir. Kazanç oranı ölçüsü, özelliğin verileri ne kadar geniş ve tekdüze bir şekilde böldüğüne duyarlı olan bölünmüş bilgi adı verilen bir terimi dâhil ederek öznitelikleri cezalandırır [96].

$$Split\ Info(A) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2\left(\frac{|T_i|}{|T|}\right) \tag{3.13}$$

Bu değer T'yi n alt gruba bölerek üretilen potansiyel bilgileri temsil ederken, bilgi kazancı aynı bölümden ortaya çıkan sınıflandırmayla ilgili bilgileri ölçer.

$$Gain\ Ratio(A) = \frac{Gain(A)}{Split\ Info(A)} \tag{3.14}$$

Denklem (3.14)' te verilen kazanç oranı, sınıflandırma için yararlı görünen, bölünme tarafından üretilen bilgi oranını ifade eder. Bu değer küçük olduğunda, bölünmenin önemsiz miktarda katkısı olacaktır. Bundan kaçınmak için, yukarıdaki oranı maksimize edecek bir test seçilecektir. Böylece bilgi kazanımı en azından incelenen tüm testlerden ortalama kazanç kadar büyük olacaktır [71, 78, 82].

3.5 Destek Vektör Makineleri (DVM)

DVM, eldeki verileri kullanarak tutarlı bir tahminci üretmek amacıyla istatistiksel öğrenme teorisinden ve yapısal riskin en küçüklenmesinden yararlanan istatistiksel algoritmalardır. İstatistiksel öğrenme bir dizi eğitim örneğinden bir fonksiyonun tahmin edilmesi anlamına gelir. Bunu yapmak için, bir öğrenen makinenin belirli bir fonksiyon kümesinden bir fonksiyon seçmesi gerekir. Bu durum tahmin edilen fonksiyonun henüz bilinmeyen gerçek fonksiyondan farklı

olduğuna dair deneysel riski en aza indirir. Deneysel risk, seçilen fonksiyonların karmaşıklığına ve eğitim örneklemine bağlıdır. Bu nedenle öğrenen bir makine karmaşıklığına göre belirlenen en iyi fonksiyon dizisini ve bu kümedeki en iyi fonksiyonu bulmalıdır. Ancak pratikte risk sınırlaması kolayca hesaplanamaz ve de çözümün kalitesini analiz etmek için çok yararlı değildir [83, 84].

Parametrik olmayan denetimli öğrenme yöntemlerinden olan DVM, Cortes ve Vapnik tarafından 1995 yılında geliştirilmiştir. Yöntemin en önemli özelliği matematiksel temellerinin güçlü olmasıdır. Parametrik yöntemlerdeki kısıtlamalar da bu yöntemde yoktur. Bunun için birçok makine öğrenmesi çalışmasında DVM kullanılmaktadır.

DVM algoritmalarının temel amacı ele alınan veri setini iki sınıfa ayıran optimum ayırıcı düzlemin bulunmasıdır. Yani sınıflar arası uzaklığı maksimize edecek bir düzlemi tespit etmektedir. Ancak burada karar fonksiyonunun genelleme hatasını düzeltmeyecekleri beklendiğinden veya örnekleri değerlendirmenin hesaplama maliyetini düşürmek istediğimizden karar sınırından çok uzak olan örneklerin etkisi kaldırılmak istenebilir. Bunun için eğitim örneklemlerinde sınıfların uç noktalarında yer alan örnekleri dikkate alarak oluşturulan destek vektörlerinden yararlanılmaktadır [63, 84, 85].

DVM verileri iki temel kategoriye ayırmaya çalışmaktadır. Bunun için n boyutlu hiper düzlem üretilmektedir [71]. Temel olarak verilerin lineer ayrımı mümkün ise sistem optimizasyonu Lineer DVM yapılır. Eğer mümkün değil ise lineer olmayan DVM ile quadratik optimizasyon sağlanır [57, 83, 86]. Bunun için modeller çekirdek (kernel) fonksiyonlarını kullanmaktadır. Farklı çekirdek fonksiyonları ile farklı sonuçlar elde edilebilmektedir. Seçilen çekirdek fonksiyonu sistemin performansını etkilemektedir [101].

Bir destek vektör makinesi sınıflandırma, regresyon veya öz nitelik seçimi gibi diğer görevler için kullanılabilir. Sınıflandırma görevinde yöntemin amacı, dış gözlemlere dayalı olarak bir nesneyi birkaç sınıftan birine atayan bir kural bulmaktır [99]. Belirli bir eğitim örneklemi kümesi için:

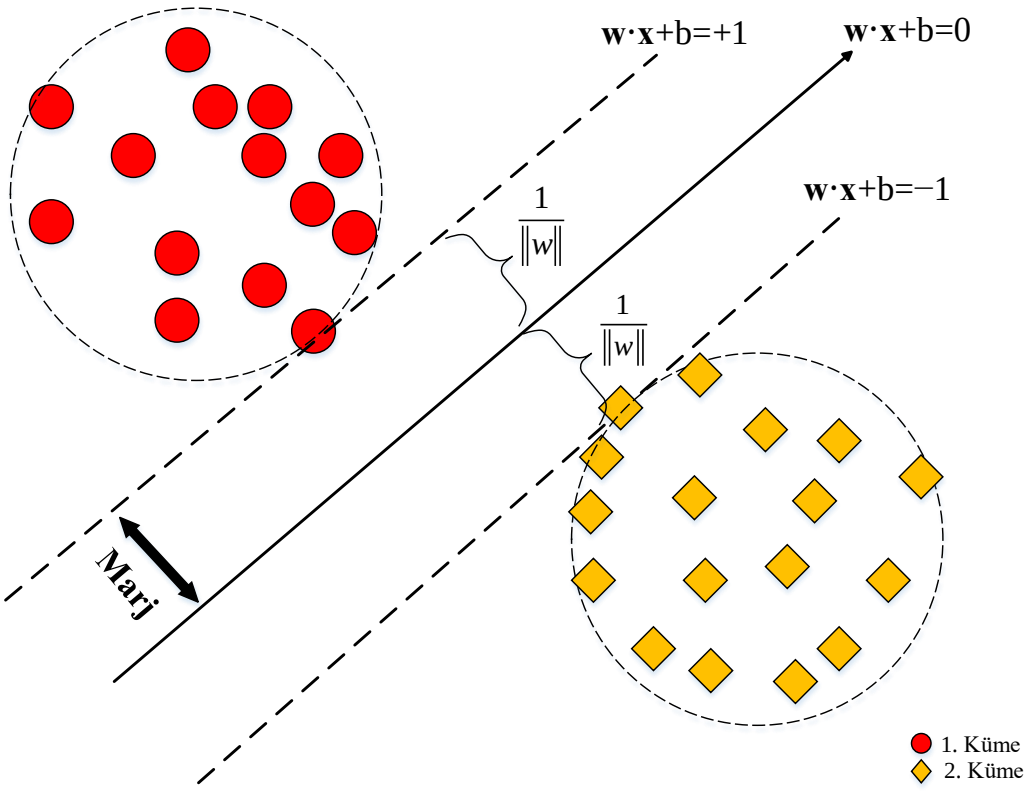
$$D = \left\{ \left(x_i, y_i \right) \mid x_i \in R^p, y_i \in \left\{ -1,1 \right\} \right\}_{i=1}^{n} \tag{3.15}$$

Verilen kümede x_i girdi p boyutlu gerçek değerli girdi vektörlerini ve y_i de girdi vektörlerinin ait olduğu sınıfa karşılık gelen sınıf değerini göstermektedir. DVM, $y_i = 1$ olan noktaları $y_i = -1$ olanlardan ayıran maksimum sınır hiper düzlemini bulmayı amaçlamaktadır. Eğitim örneğinin bir hiper düzlem ile ayrılabilir olduğunu varsayalım ve düzlemin fonksiyonunu belirleyelim. Bu fonksiyon denklem (3.16)'te verilmiştir.

$$(\mathbf{w}.\mathbf{x}) + b = 0 \; \mathbf{w} \in \mathbf{R}^N, b \in \mathbf{R} \tag{3.16}$$

Lineer DVM'nin grafiksel gösterimi Şekil 3.12' de verilmiştir. Bu fonksiyonun karar fonksiyonu olarak gösterimi denklem (3.17)'de tanımlanmıştır:

$$f(\mathbf{x}) = sign((\mathbf{w}.\mathbf{x}) + b) \tag{3.17}$$



Şekil 3.7 Lineer DVM için optimal ayırıcı hiper düzlem

Hiper düzlemlerin sınıfı için, fonksiyonun kapasitesinin başka bir miktar olan marj ile sınırlandırılabileceği gösterilmiştir. Marj, bir örneğin karar yüzeyine olan minimum mesafesi olarak tanımlanır. Marj, denklemde verilen ağırlık vektörü olan **w**'nin uzunluğuna bağlıdır. Eğitim örneğinin ayrılabilir olduğunu varsaydığımız için, **w** ve b hiper düzleme en yakın noktaların $\left|(\mathbf{w}.\mathbf{x}_i)+b\right|=1$ şeklinde düzlemin kanonik temsili sağlanır. Öyleyse farklı sınıflardan $\mathbf{x}_1$ ve $\mathbf{x}_2$ şeklinde iki örnek düşünürsek bu örneklerde uzaklıklar $\left|(\mathbf{w}.\mathbf{x}_1)+b\right|=1$ ve = 1 ve $\left|(\mathbf{w}.\mathbf{x}_2)+b\right|=1$ şeklindedir. Bu iki örnek arasındaki marj $\frac{\mathbf{w}}{\|\mathbf{w}\|}(\mathbf{x}_1.\mathbf{x}_2)=\frac{2}{\|\mathbf{w}\|}$ şeklinde hesaplanacaktır [98]. Verileri ayıran tüm hiper düzlemler arasında, sınıflar arasında maksimum ayırma marjını sağlayan benzersiz bir tane vardır:

$$\underset{\{\mathbf{w},b\}}{\text{Max}}\ \text{Min}\left\{\|\mathbf{x}-\mathbf{x}_i\| : \mathbf{x}\in\mathbf{R}^N((\mathbf{w}.\mathbf{x})+b=0), i=1,...n\right\} \tag{3.18}$$

Optimum hiper düzlemi oluşturmak için aşağıda verilen optimizasyon probleminin çözülmesi gerekmektedir:

$$\underset{\{\mathbf{w},b\}}{\text{Min}}\ \tfrac{1}{2}\|\mathbf{x}\|^2 \tag{3.19}$$

Denklem y_i ile kısıtlandığında:

$$y_i((\mathbf{w}.\mathbf{x}_i)+b=0)\geq 1, \quad i=1,...n \tag{3.20}$$

Optimum bir hiper düzlem böylece ortaya koyulmuştur. Bu düzlem ile negatif ve pozitif örnekler hiper düzlem ile iki tarafa ayrılmıştır. Denklem (3.20)'de verilen bu kısıtlama optimizasyonu problemi, Lagranj fonksiyonu ve $a_i\geq 0$ şeklinde verilen Lagranj çarpanlarının tanıtılmasıyla çözülebilir.

$$L(\mathbf{w},b,\alpha)=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{n}\alpha_i(y_i((\mathbf{w}.\mathbf{x}_i)+b)-1) \tag{3.21}$$

Lagranj L {**w**, b} ilk değişkenlerine göre minimize edilmeli ve ikili değişkenler a_i 'ye göre maksimize edilmelidir. En uygun nokta bir eyer noktasıdır ve birincil değişkenler için aşağıdaki denklemler elde edilmelidir:

$$\frac{\partial L}{\partial b} = 0; \quad \frac{\partial L}{\partial \mathbf{w}} = 0 \tag{3.22}$$

Hesaplanmalıdır. Bunlar şu şekle dönüştürülür:

$$\sum_{i=1}^{n} \alpha_i y_i = 0; \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \tag{3.23}$$

Destek Vektörleri, sıfır olmayan α_i 'ye karşılık gelen modellerdir ve sıfır olmayan α_i "*Destek Değerleri"* olarak adlandırılır. Karush-Kuhn-Tucker (KKT) tamamlayıcı optimizasyon koşullarına göre, eşitlik olarak karşılanmayan Denklem 12.5'teki tüm kısıtlamalar için α_i sıfır olmalıdır. Böylece:

$$\alpha_i (y_i ((\mathbf{w}.\mathbf{x}_i) + b) - 1) = 0 \quad i = 1, ..., n \tag{3.24}$$

Elde edilir ve tüm destek vektörleri sınırda bulunurken kalan tüm eğitim örnekleri çözümle ilgisizdir. Hiper düzlem, kendisine en yakın olan desenler tarafından tam olarak sağlanır. Lineer olmayan ve denklem (3.19)'da verilen bir probleme birincil problem denir ve belirli koşullar altında, birincil ve ikincil problemler aynı objektif değerlere sahip olacaktır. Bu nedenle birincil problemden daha kolay olabilecek ikincil problemi çözülebilmektedir. Birincil dereceden verilen denklemdeki değişkenler, α_i için ikincil dereceden bir dual problem haline getirilerek çözülecek bir optimizasyon problemine dönüştürülür [84 – 86]:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i . \mathbf{x}_j) \tag{3.25}$$

$$\alpha_i \geq 0; \; i = 1, ...n; \; \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.26}$$

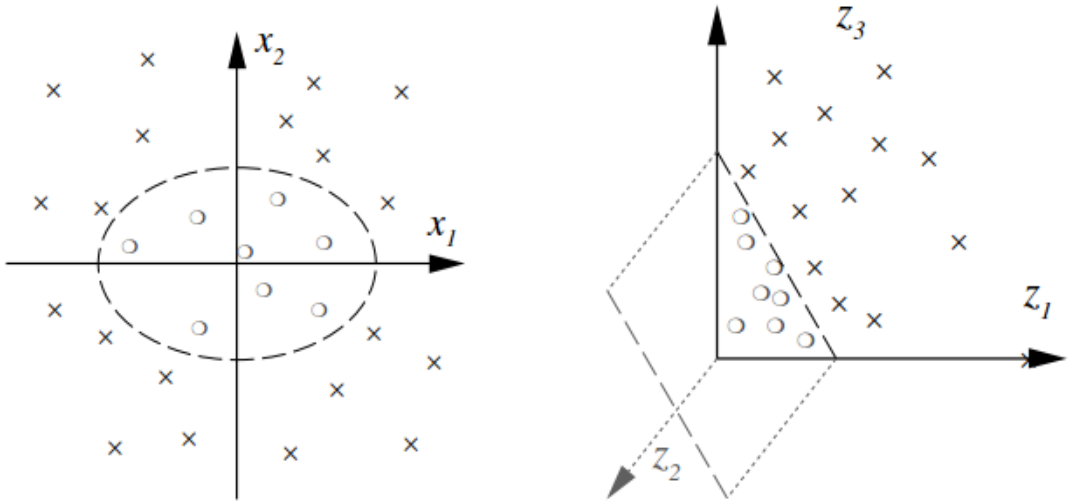
Denklemde sunulan hiper düzlem karar fonksiyonu açıkça denklem (3.27)'deki şekilde yazılabilir:

$$f(\mathbf{x}) = sign(\sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}.\mathbf{x}_i) + b) \tag{3.27}$$

Verilen b denklem (3.24)'ten ve $\mathbf{x}_i, i \in I \equiv \{i : \alpha_i \neq 0\}$ destek vektör kümesinden yararlanılarak hesaplanır:

$$b = \frac{1}{|I|} \sum_{i \in I} \left(y_i - \sum_{j=1}^{n} \alpha_j y_j (\mathbf{x}_i.\mathbf{x}_j) \right) \tag{3.28}$$

Doğrusal sınıflandırıcı fonksiyonların seçimi çok sınırlıdır. Ancak Lineer olmayan çekirdek seçimi yöntemi ile maksimum marjlı hiper düzlemler kullanılarak hem doğrusal modellere hem de doğrusal olmayan karar fonksiyonu elde etmek mümkündür. DVM'de çekirdek numarasını kullanmak, maksimum marj hiper düzleminin bir F özellik uzayına uyumlu olmasını sağlar. F özellik uzayı, $\Phi : R^N \rightarrow F$ şeklinde verilen ve genellikle orijinal girdi uzayından çok daha yüksek boyuta sahip doğrusal olmayan bir haritadır. Lineer olmayan girdilerin dönüşümü için non-lineer çekirdek belirlenir ve sonrasında lineer DVM yöntemi kullanılır. Eğitim örnekleminden yararlanarak ayrılabilir doğrusal olmayan veriler daha yüksek boyutlu doğrusal yüksek boyutlu bir özellik uzayı halinde haritalandırılmaktadır. Yani $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$ şeklinde ifade edilebilen bir K çekirdek fonksiyonu varsa eğitim algoritmasında K'yı bilmemiz yeterli olacaktır. Ayrıca Φ'nin bilinmesine gerek yoktur. Şekil 3.8' de bir çekirdek fonksiyonu için ele alınan verinin yüksek boyutlu şekilde haritalandırılması gösterilmiştir.



Şekil 3.8 Lineer olmayan DVM için çekirdek fonksiyonu ile yüksek boyutlu örnek uzaya haritalama [98]

Şekildeki dönüşüm $\Phi : \mathbf{R}^2 \to \mathbf{R}^3,\ (x_1, x_2) \to (z_1, z_2, z_3) \equiv (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ olarak ifade edilebilir. Yapılan çalışmalarda birçok farklı çekirdek fonksiyonu kullanılmaktadır. Polinom çekirdeği, Radyal temel fonksiyon çekirdeği, lineer çekirdek, sigmoid çekirdeği en yaygın kullanılan çekirdeklerdendir. Bu çekirdek fonksiyonları sırasıyla denklem (3.29)-(3.32)'de verilmiştir.

$$\left((\mathbf{x}^T.\mathbf{x}_i) + \eta\right)^d \tag{3.29}$$

$$\exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}_i\right\|^2\right),\quad \gamma > 0 \tag{3.30}$$

$$(\mathbf{x}^T.\mathbf{x}_i) \tag{3.31}$$

$$\tanh\left(\gamma(\mathbf{x}^T.\mathbf{x}_i) + \eta\right),\quad \gamma > 0 \tag{3.32}$$

Doğrusal olarak ayrılamayan eğitim örneklemlerinin olduğu durumlarda ise Esnek Marjlı Destek Vektör makineleri kullanılmaktadır. Bunun için ξ_i şeklinde sert sınır kısıtlamalarını gevşeten gevşek değişkenler tanımlanmaktadır [102].

$$y_i((\mathbf{w}.\Phi(\mathbf{x}_i)) + b) \geq 1 - \xi_i,\quad \xi_i \geq 0,\ i = 1, \ldots, n \tag{3.33}$$

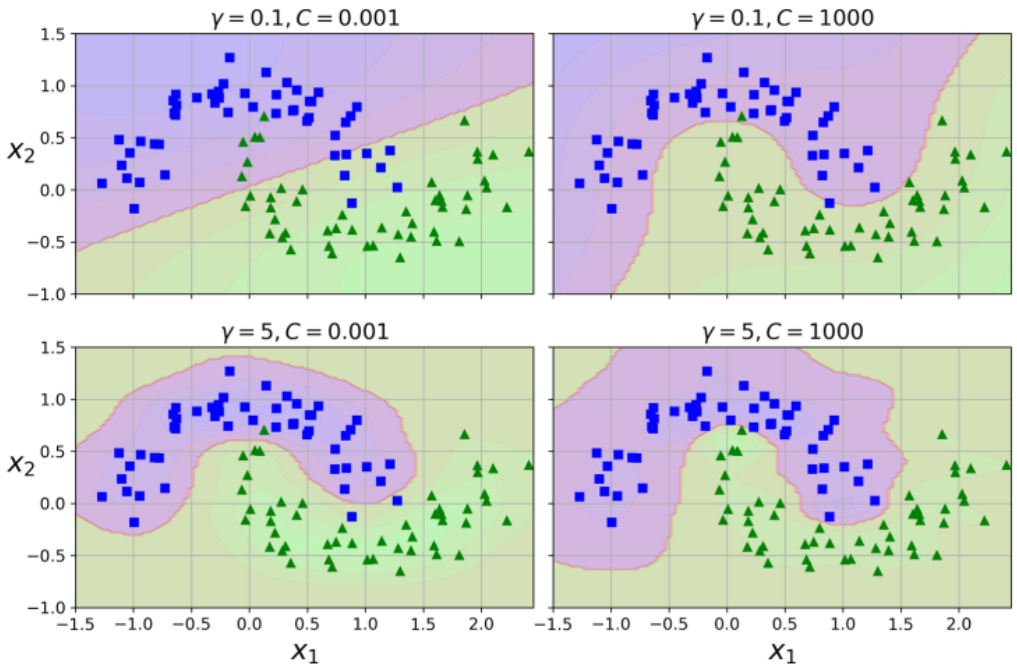
Denklemde hem ||**w**|| marj değişkeni yoluyla hem de gevşek değişkenlerin toplamı $\sum_{i=1}^{n} \xi_i$ yoluyla kontrol edilerek genelleştiren bir sınıflayıcı bulunur. C-DVM olarak adlandırılan olası bir yumuşak marj sınıflandırıcısının gerçekleştirilmesi için aşağıdaki amaç fonksiyonun en küçüklemek gerekmektedir [98].

$$\underset{\{\mathbf{w},b,\xi\}}{\text{Min}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{3.34}$$

C > 0 olarak verilen düzenleme sabiti, deneysel hata ile karmaşıklık terimi arasındaki dengelemeyi belirler. Lagrange çarpanlarının probleme dâhil edilmesi ile denklem (3.35)'te verilen ikili problem ortaya çıkmaktadır.

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i.\mathbf{x}_j) \tag{3.35}$$

Öyle ki $0 \le \alpha_i \le C;\ i = 1,...n;\ \sum_{i=1}^{n}\alpha_i y_i = 0$ olarak verilmektedir. Denklemin daha önce verilen durumdan tek farkı α_i için üst sınır C'nin belirlenmiş olmasıdır. Çözüm önceki denklemle aynıdır.



Şekil 3.9 Radyal temel fonksiyon çekirdeği ile oluşturulan bir DVM sınıflandırıcı [103]

Şekil 3.9'da Radyal temel fonksiyon çekirdeği ile oluşturulmuş örnek bir sınıflandırma probleminde γ ve C hiper parametrelerinin değişimine göre sınıflandırma performansı görülmektedir. Gama değerinin artması ile karar sınırının daha düzensiz hale geldiği, azalan gama değerinde ise daha düzgün karar sınırının oluştuğu anlaşılmaktadır. C hiper parametresine benzer şekilde gama parametresi de bir düzenlilik hiper parametresi gibi davranmaktadır.

Analizlerde öncelikli olarak her zaman doğrusal çekirdekler tercih edilmelidir. Bu seçim çalışmada veri setinin büyük olmasına ve değişken sayısının fazla olmasına göre ayarlanmaktadır. Eğer büyük veri setleri ile ve fazla sayıda değişken ile çalışılıyorsa doğrusal çekirdekler daha kullanışlı ve daha hızlı sonuçlar üretmektedir. Ancak küçük veri kümelerinde çalışılıyorsa Gauss Radyal Temel Fonksiyon Çekirdeği kullanımı daha iyi sonuçlar üretmektedir [103].

3.6 KStar Algoritması

K-Star algoritması WEKA programında yer alan örnek tabanlı öğrenme algoritmalarından biridir [74]. K-Star algoritması küme sayısı önceden bilinmediğinde otomatik olarak küme sayısını ortaya çıkaran bir yöntemidir

[104]. Öğrenme aşamasında örneklerin benzerliklerini belirlemek için bir uzaklık fonksiyonu ve benzerlikleri kullanarak yeni gelen örneğin sınıfını belirleyen bir sınıflandırma fonksiyonu kullanılmaktadır [105]. Yani bir test örneği benzerlik fonksiyonları ile bu örneğe benzeyen eğitim örneklerinin sınıfına atanmaktadır.

Hem sözel (sembolik) hem de sayısal (gerçek) değerli verilerle çalışabilen K-Star algoritması iki örneği birbirine dönüştürmenin karmaşıklığını ortaya çıkarmak için aralarındaki uzaklığı hesaplamaktadır. K-Star algoritması uzaklık ölçüsü olarak entropi kullanmaktadır [91, 92]. Algoritma bu yönüyle verilerin uzaklıklarının ölçüsü olarak entropi kullanan k-NN algoritmasına benzemektedir [107]. Ancak diğer örnek tabanlı öğrenme algoritmalardan farklıdır.

Ölçü olarak entropi kullanımı sembolik verilerde, gerçek verilerde ve eksik verilerde yaklaşımı tutarlı hale getirir ve bu yöntemin en büyük faydasıdır. Çünkü gerçek değerli özniteliklerle yapılan bir sınıflandırma sembolik değerli öznitelikler kullanıldığında başarısız olacaktır. Sembolik verilerle yapılan bir sınıflandırma ise gerçek değerli öznitelikler kullanıldığında başarılı sonuçlar üretemeyecektir. Eksik değerlerin sınıflandırıcılar tarafından ele alınması benzer problemler ortaya çıkarır. Genellikle eksik değerler ayrı bir değer olarak ele alınır, maksimum derecede farklı olarak düşünülür, ortalama değerle ikame edilir ve aksi takdirde göz ardı edilir. K-Star algoritması tüm bu problemleri entropi kullanımı ile tek başına çözmektedir [92, 94].

K-Star algoritması işlemleri iki ana adımda gerçekleştirmektedir. İlk adımda bir a örneğinin b örneğine dönüştürülmesi için (a) → (b) şeklinde tüm sınırlı dönüşümler kümesi ortaya çıkarılır. Yöntem, Kolmogorov mesafesini iki örneği birbirine bağlayan en kısa yol olarak kullanmaktadır. Bu nedenle K-Star mesafesi ele alınan iki örneğin arasındaki tüm olası mesafelerin toplam değerini ifade etmektedir [109].

I bir sonsuz örnekler kümesi ve T, I üzerindeki sonlu dönüşümler kümesi olsun. Ele alınan örnek kümesinde her bir $t \in T$ ($t: I \rightarrow I$) örnekten örneğe dönüşüm

haritasını göstersin. Ayrıca T'nin (σ(a) = a) şeklinde bir a örneğinin kendisine dönüşümünde bütünlüğü sağlayan bir ayırıcı σ özelliği düşünülsün. Verilen σ özelliği T* dönüştürülmüş elemanlar kümesinden örneği kendine dönüşüm kodlarını sonlandırır. T*, I üzerinde bire bir dönüşümü tanımlayan üyelerden oluşur. P, T* kümesi üzerinde tanımlanan bir olasılık fonksiyonudur [109]. Burada $\bar{t} = t_1, ..., t_n$ olmak üzere;

$$\bar{t}(a) = t_n(t_{n-1}(...t_1(a)...)) \tag{3.36}$$

T* ve P olasılığı, I üzerinde bir dönüşümü birebir ve benzersiz şekilde tanımlar. P* olasılığı, t dönüşümleri kullanılarak bir a örneğinden b örneğine giden tüm yolların olasılığı olmak üzere;

$$P^*(b \mid a) = \sum_{\bar{t} \in P : \bar{t}(a) = b} p(\bar{t}) \tag{3.37}$$

Bu değer olasılık fonksiyonu özelliklerini sağlar. K* fonksiyonu ise bu olasılıklar dikkate alınarak denklem (3.38)'de verilmiştir.

$$K^*(b \mid a) = -\log_2 P^*(b \mid a) \tag{3.38}$$

Fonksiyon simetrik olmadığından $K^*(a \mid a)$ sıfırdan farklıdır. Ayrıca $K^*(b \mid a)$ ile $K^*(a \mid b)$ birbirine eşit olmak zorunda değildir. Ek olarak $K^*(b \mid a) \geq 0$ olmak üzere $K^*(c \mid b) + K^*(b \mid a) \geq K^*(c \mid a)$ özelliğinin sağlanmasıdır. Bu özellikleri ile K* fonksiyonu tam olarak bir uzaklık fonksiyonu değildir [105].

Fonksiyonun hesaplanmasında gerçek değerli örnekler için olasılık yoğunluk fonksiyonu, sembolik değerli örnekler için frekanslara dayanan bir olasılık fonksiyonu kullanılır. Ele alınan kümede herhangi bir P* fonksiyonu için etkili örnek sayısı,

$$n_0 \le \frac{\left(\sum_b P^*(b \mid a)\right)^2}{\sum_b P^*(b \mid a)^2} \le N \tag{3.39}$$

Formülü ile ifade edilir. Burada verilen formülde N toplam örnek sayısını ve n_0 da bu öznitelik için *a* örneğine en yakın mesafedeki örneklerin sayısıdır. Bir a örneği olarak n_0 ile N değerleri arasında bir değerin seçilir. Eğer n_0 seçilirse en yakın komşu algoritmasını veriyorken N seçilirse eşit ağırlıklı örneklerin sayısını verir. Kolaylık sağlamak için bu değer b "harmanlama parametresi" kullanılarak belirlenir. Bu parametre, % 0 ile % 100 arasında değişir.

3.7 k-En Yakın Komşu Algoritması (k-NN)

k-NN algoritması, verilerin en yakın komşularına göre sınıfını tespit etmektedir. k-NN veri madenciliği çalışmalarında kullanılan popüler algoritmalardandır [74]. Basit ve anlaşılır olması nedeniyle tercih edilmektedir [110]. Eğitim aşamasında eldeki tüm örneklemi kullanarak örnekler arasında en yakın komşuların bulunması sağlanır. Böylece yeni eklenen bir örneğin özelliklerine bakılarak hangi kümede olması gerektiğine karar verilir.

Algoritmada k parametre değeri ile benzerlik fonksiyonu performansı etkilemektedir. Ele alınan bir verinin k en yakın komşusunun durumuna göre komşularının bulunduğu sınıfa dâhil olma olasılığını hesaplamaktadır. Bu yönüyle de tamamıyla kara kutu olan YSA'dan üstün özelliktedir [77]. Ancak komşular arası uzaklık ölçülerinin belirlenmesi zordur. Manhattan, Öklid, Minkowski, Chebyshev ve Mahalanobis uzaklık ölçüleri yaygın olarak kullanılmaktadır [72, 81,97]. Denklem (3.40)' da Minkowski uzaklığı verilmiştir.

$$L^p\left(x_a, x_b\right) = \left(\sum_{i=1}^{m} \left|x_{ai} - x_{bi}\right|^p\right)^{\frac{1}{p}} \tag{3.40}$$

Denklemde x_a ve x_b ele alınan iki örneği ve $X = (x_1,x_2,...,x_m)$ ise değişkenlerin özelliklerini ifade etmektedir. Bu denklemde p değeri yerine 2 yazıldığında ölçüler arasında en yaygın kullanılan ölçü Öklid uzaklığı elde edilir.

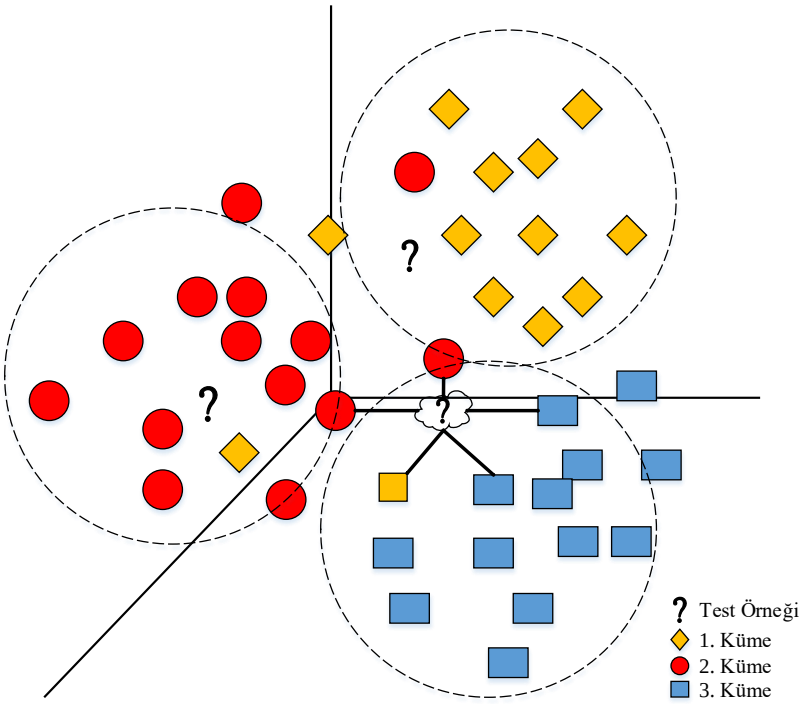
$$L^2\left(x_a, x_b\right) = \left(\sum_{i=1}^{m} \left|x_{ai} - x_{bi}\right|^2\right)^{\frac{1}{2}} \tag{3.41}$$

Öklid uzaklığı denklem (3.41)'de verilmiştir. Denklem (3.40)' ta p = 1 olduğunda, Minkowski uzaklığı Manhattan uzaklığına eşit olacaktır ve bu uzaklık ikili (binary) öngörüleri olan örnekler için yaygın olarak kullanılan bir ölçüdür [86]. Değişkenlerin ölçeklerinin aynı olması durumunda Öklid uzaklığının kullanılması uygun iken farklı olması durumunda Manhattan uzaklığı kullanılmaktadır. Her boyuttaki ham değerlerin kullanılması durumunda toplam uzaklık herhangi bir boyuttaki ölçek değişikliğinden etkilenecektir. Büyük boyutlu değişkenler yanlış sonuçlar üretilmesine neden olacaktır. Bunun için veri normalleştirme yapılması gerekmektedir [111]. Normalleştirme ile veriler [0,1] aralığında bir değere dönüştürülmektedir. Normalleştirme için minimum-maksimum normalleştirme formülü ya da z-skor normalleştirme formüllerinde yararlanılabilir. Bu formüller denklem (3.42) ve (3.43)' te verilmiştir.

$$x_{ij}' = \frac{x_{ij} - \min(x_i)}{\mathrm{m}\,aks(x_i) - \min(x_i)} \tag{3.42}$$

$$x_{ij}' = \frac{x_{ij} - \mu_{x_i}}{\sigma_{x_i}} \tag{3.43}$$

Denklemlerde x_{ij}, $X = (x_1,x_2,...,x_m)$ de yer alan x_i değişkeninin j numaralı elemanıdır. x_{ij}' ise normalleştirilmiş değerini değişkenlerin özelliklerini ifade etmektedir.



Şekil 3.10 k-NN ile sınıf belirleme örneği

Şekil 3.10' da 3 sınıftan oluşan bir veri kümesi örneği verilmiştir. Örnekte yer alan 3 test verisi için uzaklık ölçüleri kullanılarak sınıflar belirlenecektir. Belirlenecek sınıflar için en önemli kıstas en yakın komşu sayısının belirlenmesidir. Komşu sayısının belirlenebilmesi için etkili bir yöntem yoktur. K değeri arttırılıp azaltılarak modelin esnekliği arttırılıp azaltılabilmektedir [77].

Oldukça basit olması, dağılım varsayımı olmaması bu algoritmanın hızlı öğrenen ve etkili bir yöntem olmasını sağlamaktadır. Buna karşın fazla sayıda değişkenin olması durumunda sınıflandırma aşaması yavaş olmaktadır. Süreçte tüm uzaklıklar kaydedileceğinden fazla miktarda işlemci gücüne de ihtiyaç duyulmaktadır. İlgisiz ve gürültülü tahminci değişkenlerin olması da modelin başarısını etkilemektedir. Çünkü bu değişkenler benzer örneklerin tahmin esnasında birbirinden uzaklaşmasına neden olabilmektedir. Bu nedenle alakasız ve gürültü tahminci değişkenlerin analizden çıkarılması k-NN için önemli bir ön işleme adımıdır [86].

Model için bir diğer problem ise eksik ve kayıp gözlemlerin olmasıdır. Eksik veri olması durumunda örnekler arasındaki mesafeyi hesaplamak mümkün

olmadığından, bir örnek için bir veya daha fazla tahminci değişkenin değeri eksikse örnekler arasındaki mesafelerin hesaplanması sorunlu olabilir. Kayıp değerlerin yerini doldurmak ya da bu verileri analizden çıkarmak iki seçenektir. Özellikle değişkenlerin az sayıda kayıp değer içerdiği durumlarda eksik verileri çeşitli istatistiksel yaklaşımlarla tamamlamak, çalışmalarda bilgi kaybını azaltacaktır [72, 81, 96].

3.8 Performans Karşılaştırma Kriterleri

Makine öğrenmesi modellerinin performansını karşılaştırmak için verilerin yapısına ve görevin niteliğine bağlı olarak kullanılabilecek çeşitli kriterler vardır [53, 57, 60]. Uygulama kısmında verilen örneklemimizde her sınıftaki örneklem sayıları birbirinden farklıdır. Ayrıca makine öğrenmesi çalışmalarında genel olarak iki sınıf varken, bu çalışmada dört farklı sınıf vardır. Sınıf sayısının artırılması sonuçları etkileyebilir [112]. Sonuçları değerlendirmek için kullanılan bazı yöntemler dengesiz verilere duyarlı olduğundan değerlendirmede Geometrik Ortalama ve Youden indeksi gibi karşılaştırma kıstasları da kullanılmıştır [113].

Tablo 3.1 İki sınıf için karşılaştırma matrisi

	Tahmin Edilen Sınıf		
Gerçek Sınıf	**Pozitif**	**Negatif**	***Toplam***
Pozitif	Doğru Pozitif	Yanlış Negatif	p
Negatif	Yanlış Pozitif	Doğru Negatif	n
Toplam	p'	n'	

Bu bölümde etkili olan algoritmaların belirlenmesinde kullanılan kriterler verilmiştir. Bu kriterler Doğruluk (ACC), Geometrik Ortalama (GM), Hata Oranı (ERR), Kesinlik (PREC), Duyarlılık (SENS), Özgüllük (SPEC), F-Ölçütü (FM), Matthew korelasyon katsayısı (MCC), Youden indeksi (YI), Kappa (κ), Yanlış Pozitif Oran (FPR) ve Alıcı İşletim Karakteristik Alanı (ROC) olarak sıralanabilir. Bu formülleri Doğru pozitif (TP), Doğru negatif (TN), Yanlış pozitif (FP) ve

Yanlış negatif (FN) değerleri kullanılarak hesaplamak mümkündür. Bu değerler sınıfların tahmin sonuçlarının karşılaştırılmasında kullanılan karşılaştırma matrislerinden elde edilmektedir. İki sınıf için karşılaştırma matrisi Tablo 3.1' de verilmiştir. Karşılaştırma matrisinde belirtilen TP; doğru pozitif tahmin sayısını, FP; yanlış pozitif tahmin sayısını, TN; doğru negatif tahmin sayısını ve FN; yanlış negatif tahmin sayısını ifade etmektedir.

Doğruluk, toplam model tahminleri içindeki gerçek pozitif ve gerçek negatif tahminlerin oranını yansıtır. Tanıma oranı olarak da bilinen doğruluk, bir sınıflandırıcının öngörü yeteneğini göstermektedir [95]. Doğruluk en yaygın kullanılan performans ölçülerindendir [114].

$$ACC = \frac{TP + TN}{p + n} \tag{3.44}$$

Geometrik ortalama, sınıflandırmadaki hem çoğunluk hem de azınlık alt gruplarının sonuçları arasındaki dengeyi belirleyen bir ölçüdür [115]. Doğruluk, sınıf dağılımındaki değişikliklerden etkilenir, ancak geometrik ortalama etkilenmez. Bu nedenle geometrik ortalama, dengesiz veri seti için daha uygundur [113].

$$GM = \sqrt{\left(\frac{TP}{TP + FN}\right) \cdot \left(\frac{TN}{TN + FP}\right)} \tag{3.45}$$

Hata oranı, doğruluğu tamamlayıcı niteliktedir. Doğruluk ölçümünden farklı olarak, bu metrik hem pozitif hem de negatif sınıflar için yanlış sınıflandırılan örneklerin sayısını gösterir.

$$ERR = \frac{FP + FN}{p + n} = 1 - ACC \tag{3.46}$$

Kesinlik, model için kaç tane pozitif tahminin gerçekten pozitif olduğunu gösterir. Gerçek pozitif ve gerçek negatif oranları temsil eden duyarlılık ve özgüllük birbirini tamamlayıcı niteliktedir. Gerçek pozitif oran olarak da bilinen

duyarlılık, doğru pozitif örnek sayısının pozitif olarak sınıflandırılan sayıya oranıdır, özgüllük ise negatif örnekler için aynı şekilde hesaplanan orandır [70].

$$PREC = \frac{TP}{TP + FP} \tag{3.47}$$

$$SENS = \frac{TP}{TP + FN} \tag{3.48}$$

$$SPEC = \frac{TN}{FP + TN} \tag{3.49}$$

Kesinlik ve duyarlılık arasındaki denge, F-ölçüsü ile temsil edilir. Daha yüksek F-Ölçüsü, daha iyi sınıflandırıcı performansını gösterir. Bu değer aynı zamanda, duyarlılık ve kesinliğin harmonik ortalamasına da eşittir [116].

$$F-Ölçüsü = \frac{2*TP}{2*TP + FP + FN} \tag{3.50}$$

Matthew korelasyon katsayısı, dengesiz verilerden en az etkilenen ve gözlemlenen ve tahmin edilen sınıflandırmalar arasındaki korelasyonu hesaplayan karşılaştırma katsayısıdır. Youden indeksi, bir sınıflandırıcının yanlış sınıflandırma potansiyelini değerlendirir.

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \tag{3.51}$$

$$FPR = \frac{FP}{FP + TN} \tag{3.52}$$

$$YI = TPR + TNR - 1 \tag{3.53}$$

Youden indeksi, doğru pozitif oranı ile yanlış pozitif oranı kullanılarak hesaplanmaktadır. Doğru pozitif oranı (TPR), duyarlılık değerine eşittir. Yanlış pozitif oranı (FPR) ise denklem (3.52)'de verilmiştir. Tamamen tesadüfen elde edilebilecek doğruluk Kappa ölçüsü tarafından hesaplanmaktadır [117]. Kappa istatistiği, bir algoritmanın performans başarısı ile ilgilenir. Kappa istatistiği 1'e

yaklaştıkça, performans yükselir. İki kategorik değişken arasındaki gözlemlenen ve beklenen değerler $p_0$ve p_e olmak üzere Kappa istatistiği denklem (3.54)'de verilmiştir [118].

$$\kappa = \frac{(p_0 - p_e)}{1 - p_e} \tag{3.54}$$

$$p_0 = \frac{TP + TN}{p + n} \tag{3.55}$$

$$p_e = \frac{(TP + FP) * (TP + FN) + (FN + TN) * (TP + TN)}{(TP + FN + FP + TN)} \tag{3.56}$$

Alıcı İşlem Karakteristikleri Eğrisi (ROC eğrisi), doğru ve yanlış pozitif oranları arasında uygun bir denge belirlemek için 1- özgüllüğe karşı hassasiyeti grafiksel olarak gösterir. ROC eğrisi, klinik çalışmalardaki önemli karşılaştırma ölçütlerinden biridir. Bu yöntem, alt sınıfları karşılaştırırken çizilen eğrinin altındaki alanı kullanır. Eğri altında kalan alan (AUC) toplamının daha büyük olması daha iyi sınıflandırma sonuçlarını göstermektedir [119].

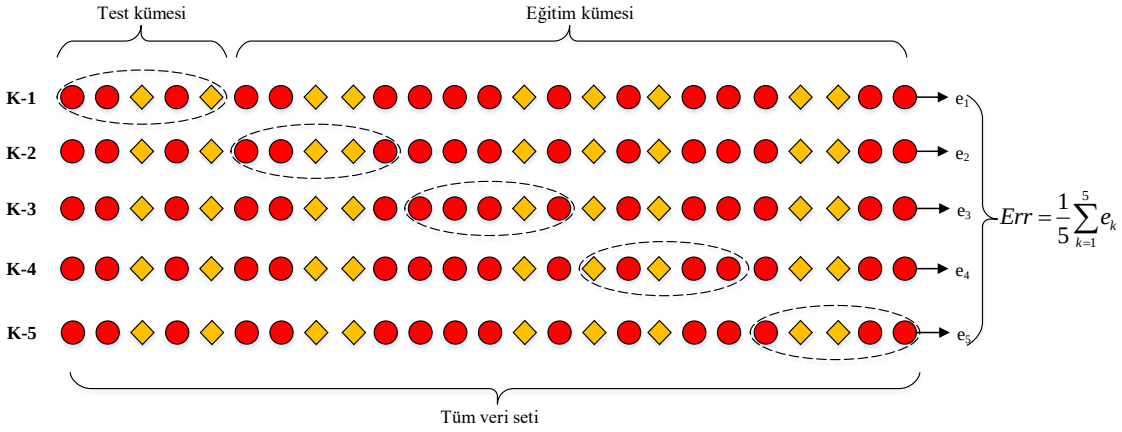
ROC alanı, modelin sınıf tahmini esnasında hatalardan kaçınma özelliğini ölçer. AUC değeri pratikte duyarlılık ve özgüllük değerleri ile yakından ilgilidir [120]. Bu değer mükemmel sınıflandırma yapılıp yapılmadığını göstermek için ROC eğrisi ile birlikte kullanılan bir ölçüdür [70].

3.9 K-kat Çapraz Doğrulama

Makine öğrenmesi yöntemlerine başarının kritik noktalarından birisi de eğitim ve test verilerinin doğru belirlenmesidir. Verilerin tüm veri setini temsil edecek şekilde ayrılabilmesi gerekmektedir. Model başarısının ve geçerliliğinin test edilmesinde farklı uygulamalar mevcuttur. Bunun için veri setleri doğrudan eğitim ve test olarak iki gruba ayrılabilmektedir. Verilerin ayrımında % 50 - % 50, % 75 - % 25, % 80 - % 20 gibi oranlar belirlenerek rastgele bölümlendirme yapılmaktadır. Ancak bu yaklaşımda veri setinin bölünme durumuna göre model performansı değişkenlik göstermektedir. Belirli sınıfa ait örnekler orantısız

şekilde dağıtılabilmektedir. Modelde ortaya çıkan hataların en aza indirilmesi için k-kat çapraz doğrulama seçilmektedir [95].

K-kat çapraz doğrulamada veri seti k alt kümeye bölünmektedir. Her adımda, k alt kümeden biri test kümesi, diğer k-1 alt küme ise eğitim kümesi olarak kullanılır. Modelin hata hesaplaması k denemenin ortalama alınarak hesaplanır. Böylece tüm veriler hem eğitimde hem de testte kullanılmış olacağından hatalar en aza indirilmiş olacaktır [57, 81, 107]. 5-kat çapraz doğrulama için sürecin bir örneği Şekil 3.11'de gösterilmiştir.



Şekil 3.11 5-kat çapraz doğrulama

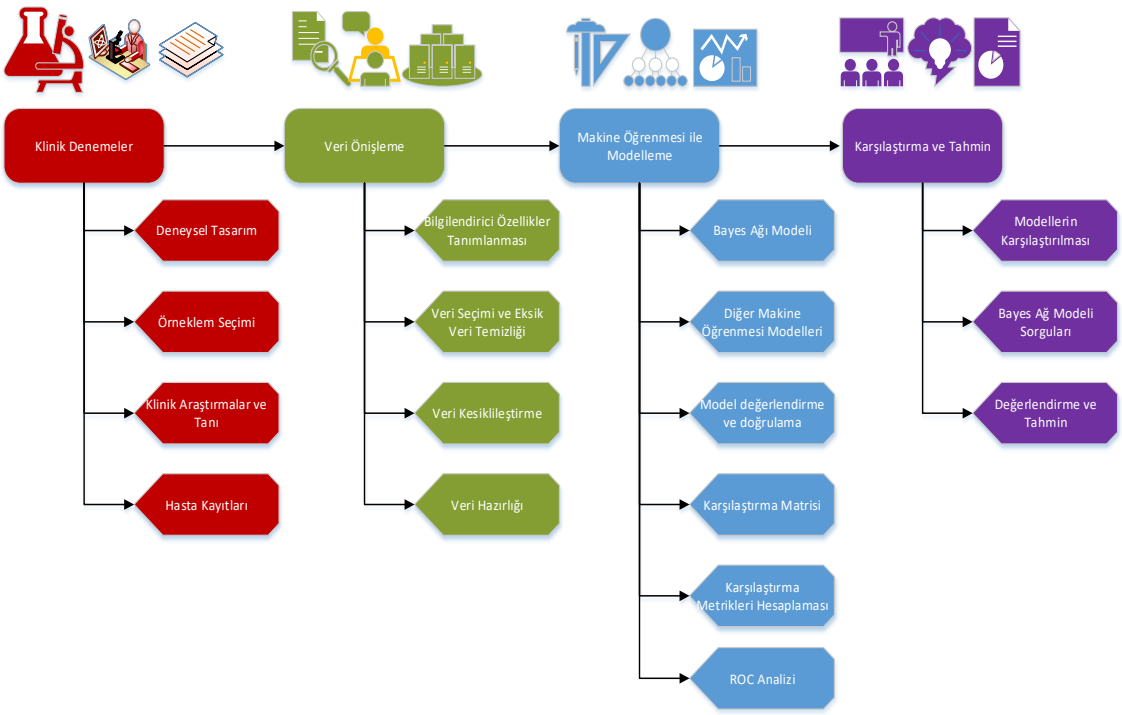
Ayrıca analizlerde tahmin sonuçları oluşturmak için 5 kat çapraz doğrulama tercih edilmiştir. Mevcut veriler beş parçaya bölünmüş, ilk dört parça eğitim amaçlı kullanılmış ve son parça test için kullanılmıştır [116]. 5 kat çapraz doğrulama, model sağlamlığını artırmak için yaygın olarak kullanılan doğrulama yöntemlerinden biridir [68].

Modelde k sayısı beş kattan fazla seçildiğinde hata oranları birbiri ile bağlantılı olacaktır. En ideal olan beş kat çapraz doğrulama yapılarak hataların ilişkisini en aza indirmek ve test için yeterince büyük alt gruplar oluşturmaktır [71].

4
UYGULAMA

Bu kısımda Bayes Ağlarının ve diğer makine öğrenme algoritmalarının ALS vaka tespitinde kullanılabilecek modelleme sonuçları birbiri ile karşılaştırılarak gösterilecektir.

Bu bölümde, çalışmada kullanılan veriler özetlenmiş ve ALS ile ilgili ele alınan değişkenlerin etkileşimlerini modellemek için kullanılan makine öğrenmesi süreçlerinin grafiksel olarak gösterilmiştir. Çalışmanın uygulama süreci Şekil 4.1' de verilmiştir.



Şekil 4.1 ALS hastalığının BA ve diğer makine öğrenmesi yöntemleri ile modelleme süreci

Şekil 4.1' de ilk adım, deneysel verilerin elde edilmesi için gerekli klinik çalışmalardır. ALS ile ilgili özellikler belirlendikten sonra ikinci adım, veri ön işleme aşamasıdır. Bir sonraki adımda veriler Bayes ağları ve diğer makine öğrenmesi algoritmaları ile modellenmiş ve elde edilen sonuçlar

karşılaştırılmıştır. Karşılaştırma sonuçları son adımda değerlendirilerek çalışma tamamlanmıştır.

4.1 Amyotrofik Lateral Skleroz (ALS) Hastalığı

ALS, temel olarak üst ve alt motor nöronların ilerleyici bozulmasından kaynaklanan nadir bir nörolojik bozukluktur. Günümüzde bu hastalığın bir tedavisi bulunmamaktadır ve hastalığın ilerlemesini durdurmak da mümkün değildir [122]. ALS başlangıçta sadece bir eli veya sadece bir bacağı etkileyerek düz bir çizgide yürümeyi zorlaştırabilir. Hastalık ilerledikçe hastalarda şiddetli kas güçsüzlüğü, kas kütlesinde azalma, konuşma bozukluğu, yutkunma, ince ve kaba motor becerilerinde azalma ve solunum güçlüğü ortaya çıkar. Bu değişim ise genellikle tanıdan sonraki 2-5 yıl içinde hastada felce ve ölüme yol açar [123].

ALS çok bileşenli bir hastalıktır. Bu hastalığın birbirini etkileyen birçok etkeni vardır. ALS vakalarının yaklaşık %10'u aileseldir (fALS) ve vakaların %90'ı sporadiktir (sALS)[124]. Yani ALS vakaları belirli bir sosyal çevrede ya da coğrafyada kümelenmeyen, tesadüfi ve genelde de nadir olarak görülen yapıdadır.

Etyolojisi büyük ölçüde bilinmemekle birlikte, çeşitli genlerdeki mutasyonlar ALS ile ilişkilendirilmiştir [111, 112]. Ayrıca protein katlanması (agregasyon), endoplazmik retikulum stresi, oksidatif stres, mitokondriyal bozukluk, nöro-enflamasyon, apoptotik hücre ölümü, glutamat eksitotoksisitesi, RNA mekanizmalarındaki anormallikler ve ubikuitin-proteazom sisteminin anormal işlevi (UPS) gibi bazı temel biyokimyasal mekanizmalarda gerçekleşen bozulaların ALS hastalığına neden olduğu düşünülmektedir [127].

ALS tipik olarak yetişkinlerde görülen bir hastalıktır, ancak genç bireylerde de görülebilmektedir. Erkeklerde bir miktar daha yaygın görülen bu hastalıkta hastalık gelişiminde cinsiyete bağlı farklılıklar bulunmaktadır [114, 115]. ALS, dünyanın her yerinden, her kesimden insanda ortaya çıkabilir. Farklı popülasyonlara dayalı ve farklı coğrafi bölgelerde yapılan çalışmalarda yılda

100.000'de 0,6 ila 11 vaka arasında değişen ALS tespit oranı bildirilmiştir. ALS prevalansı 100.000 kişide 4,1 ile 8,4 arasındadır ([130]'de incelenmiştir).

ALS'nin klinik tanısı erken dönemde zordur çünkü hastalar herhangi bir üst veya alt motor nöron bulgusu göstermeyebilir [131]. Ayrıca ALS semptomları oldukça heterojen olabilir ve birçok nörolojik hastalığa benzerlik gösterebilir. Şu anda tanı, Dünya Nöroloji Federasyonu'nun El Escorial Kriteri'ne göre ve tam nörolojik muayene, radyolojik ve elektrofizyolojik incelemelere dayalı olarak yapılmaktadır [132]. Tüm bu testler 3 ile 6 ay arasında sürebilir ve erken semptomların ortaya çıkması ile tanı arasında gecikmeye neden olabilir. ALS'nin daha etkin ve erken teşhisi ile hastanın yaşam süresini uzatmak ve yaşam kalitesini yükseltmek mümkün olacaktır.

Hastalığın daha kısa zamanda ve etkin şekilde belirlenmesi için kan testleri kullanılabilmektedir. Kan testleri diğer tanı yöntemlerine göre daha az zaman alan ve düşük maliyetli yöntemlerdir. Ayrıca kan testleri ile elde edilen değerler ile hastayı tanımlayıcı diğer değişkenler arasındaki ilişki oldukça önemlidir. Bu tez çalışmasında, kan plazmasındaki Parkin protein konsantrasyonu ve hasta ile ilgili diğer tanı değişkenlerini kullanarak ALS riskinin tahmini için Bayes Ağlarına dayalı istatistiksel bir makine öğrenme modeli geliştirmeyi amaçlamaktadır. Literatürde ALS hastalığının genetik mimarisini incelemek için makine öğrenmesi yöntemleri kullanılmıştır [67].

Bu amaçla, ALS tanısı için biyobelirteç olarak Parkin proteininin potansiyel kullanımınının araştırıldığı bir deneysel çalışma ile veriler elde edilmiştir. Hastaların yaş, cinsiyet, hastalık başlangıcı, kronik hastalık bilgilerini içeren kayıtlar da alınarak veri seti oluşturulmuştur.

Özetle bu çalışmada:

(1) Bayes Ağlarını kullanarak tahmine dayalı bir model geliştirilmiş,

(2) Model performansını diğer makine öğrenme yöntemleriyle karşılaştırarak incelenmiş ve

(3) Yukarıda belirtilen değişkenlerin değerlendirilmesi için hasta tipine ve tutulum tiplerine dayalı sorgular oluşturulmuştur.

4.2 Katılımcılar

Bu veri seti, ALS hastalarından alınan kan plazması ile multipl skleroz, frontal demans ve Parkinson hastalığı gibi diğer nörolojik vakalar arasındaki Parkin proteini düzeyindeki farklılıkları araştıran deneysel bir çalışmadan elde edilmiştir. Hastaların amnezisi detaylı olarak alındığı için veri setinde eksik veri bulunmamaktadır.

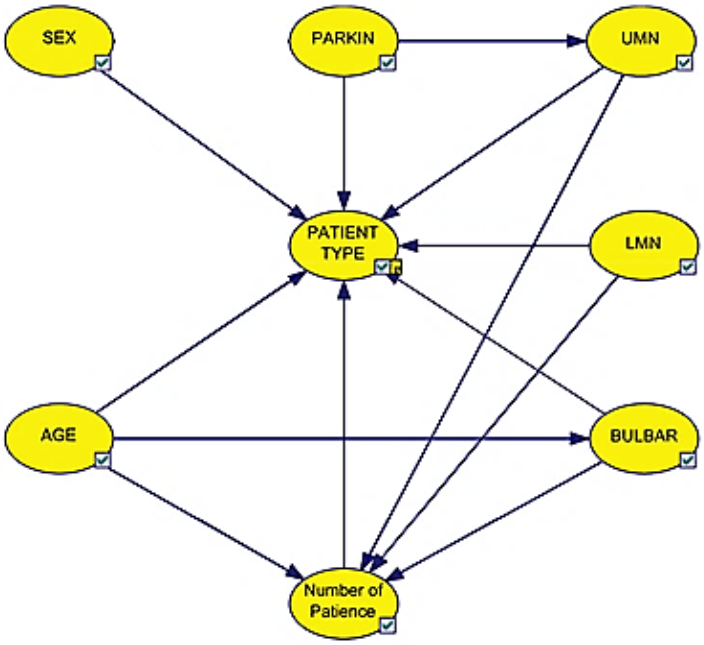
Tablo 4.1 Hastaların karakteristik özellikleri

Değişken	Değer	Frekans	% Değeri
CİNSİYET	Kadın	79	38,7
	Erkek	125	61,3
YAŞ	36'dan az	29	14,2
	36 – 52 arası	70	34,3
	52 – 67 arası	79	38,7
	67'den büyük	26	12,7
UMN	Hayır	129	63,2
	Evet	75	36,8
LMN	Hayır	178	87,3
	Evet	26	12,7
BULBAR	Hayır	182	89,2
	Evet	22	10,8
Toplam Kronik Hastalık Sayısı	Beş	1	0,5
	Dört	1	0,5
	Üç	12	5,9
	İki	20	9,8
	Bir	112	54,9
	Hiç	58	28,4
PARKIN Seviyesi (ng/mL)	3.74'ten yüksek	31	15,2
	2.79 – 3.74 arası	17	8,3
	2.06 – 2.79 arası	36	17,6
	1.36 – 2.06 arası	52	25,5
	1.36'den az	68	33,3
Hastalık Tipi	ALS	103	50,5
	Control	42	20,6
	N-Control	40	19,6
	Parkinson	19	9,3

Çalışmada kullanılan deneklerin özellikleri Tabloda verilmiştir. Cinsiyet, yaş, üst motor nöron (UMN), alt motor nöron (LMN) ve Bulbar tutulum tipleri, toplam kronik hastalık sayısı ve Parkin proteini düzeyi (ng/ml) değişkenleri hastalık tipi ile ilgilidir. Buna göre çalışmadaki verilerin %50,5'i ALS hastalarından ve %9,3'ü Parkinson hastalarından oluşmaktadır. Nörolojik Kontrol (N-Kontrol) grubu, bu hastalıklar dışında farklı nörolojik hastalıkları olan kişileri içermektedir. Kontrol grubu ise tamamen sağlıklı bireylerden oluşmaktadır. Çalışmada toplam 204 kişi yer almaktadır. Tüm hastalara El Escorial kriterlerine göre İstanbul Üniversitesi Tıp Fakültesi'ndeki nörologlar tarafından tanı ve tedavi uygulanmıştır [132].

4.3 Analiz Sonuçları

Kullanılan verilerden elde edilen Bayes Ağı modeli Şekil 4.2' de verilmiştir. Oklar ağdaki değişkenler arasındaki ilişkiyi göstermektedir. Okun yönü aynı zamanda etkinin yönünü de göstermektedir. Elde edilen Ağ, Bayes ağlarına dayalı bir makine öğrenme programı olan GeNIe 2.1 programının academik sürümü kullanılarak oluşturulmuştur [133].



Şekil 4.2 Elde edilen Bayes ağı modeli

Ağ modeline göre tutulum tipleri, yaş, cinsiyet, Parkin protein yoğunluğu ve hastalık sayısı hastalık türünü tipini doğrudan etkilemektedir. Ayrıca tutulum

tiplerinin de hastalık sayısını etkilediği görülmektedir. Kontrol grubu dışındaki kişilerde en az bir hastalık olduğu için tutulumun hastalık sayısını etkilemesi beklenmektedir. ALS hastalığında en önemli semptomlardan birinin UMN tutulumu olduğu bilinmektedir. Analiz sonucunda elde ettiğimiz modelde Parkin protein yoğunluğunun UMN tutulumunu etkilediği görülmüştür.

Karşılaştırma için kullanılan diğer makine öğrenme yöntemleri için WEKA programı kullanılmıştır. Bu program, Waikato Üniversitesi tarafından makine öğrenmesi algoritmalarının kullanılmasını kolaylaştırmak için oluşturulmuş Java tabanlı açık kaynak kodlu bir yazılımdır [74].

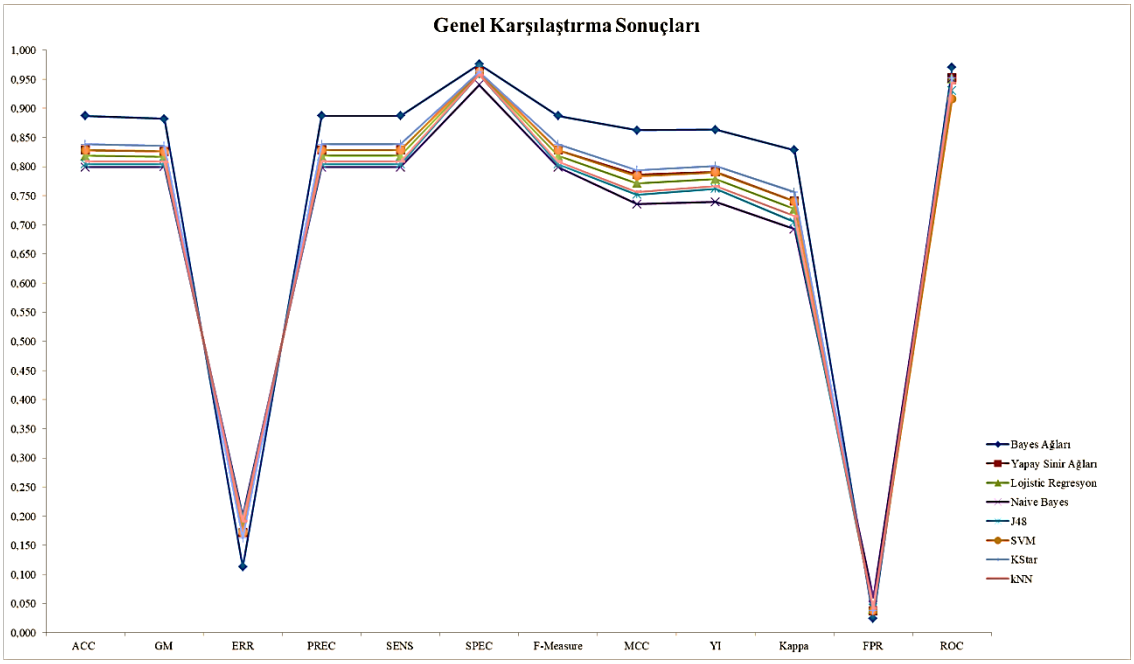
Algoritmaların daha önce belirtilen performans kriterlerine göre elde edilen performansları aşağıda verilmiştir. Genelleştirilmiş sonuçlar Tablo 4.2' de ve her sınıf için elde edilen sonuçlar Tablo 4.3' te gösterilmiştir. Kriterlere göre en iyi sınıflandırma sonuçları kalın şekilde belirtilmiştir.

Tablo 4.2 Yöntemlerin genel karşılaştırma sonuçları

	ACC	*GM*	*ERR*	*SENS*	*SPEC*	*F-M*	*MCC*	*YI*	*Kappa*	*FPR*	*ROC*
BA	**0,887**	**0,882**	**0,113**	**0,887**	**0,976**	**0,887**	**0,862**	**0,863**	**0,828**	**0,024**	**0,970**
YSA	0,828	0,826	0,172	0,828	0,963	0,828	0,787	0,791	0,741	0,037	0,953
LR	0,819	0,817	0,181	0,819	0,960	0,819	0,772	0,778	0,727	0,040	0,951
NB	0,799	0,800	0,201	0,799	0,940	0,799	0,736	0,739	0,693	0,060	0,951
J48	0,804	0,804	0,196	0,804	0,958	0,804	0,752	0,762	0,705	0,042	0,930
DVM	0,828	0,826	0,172	0,828	0,962	0,828	0,784	0,790	0,741	0,038	0,916
KStar	0,838	0,835	0,162	0,838	0,963	0,838	0,794	0,801	0,756	0,037	0,952
k-NN	0,809	0,808	0,191	0,809	0,958	0,809	0,756	0,766	0,715	0,042	0,943

Genel sonuçlar incelendiğinde Bayes Ağının diğer yöntemlere göre daha başarılı sonuçlar ürettiği görülmüştür. Bayes Ağ sınıflandırma sonuçlarının diğer yöntemlere göre küçük farklarla daha iyi olduğu ortaya konmuştur. Öte yandan, diğer makine öğrenmesi yöntemlerinin sonuçlarının birbirine yakın olduğu gözlemlenmiştir. DVM için Polykernel çekirdeği kullanılmıştır. k-NN için en başarılı sonucun en yakın 1 komşu ile elde edildiği görülmüştür.

Genel karşılaştırma tablosu incelendiğinde Bayes Ağının %88,7 doğruluk üretmiştir. Diğer yöntemlerin başarı oranlarının yaklaşık %80 olduğu görülmektedir. Genel karşılaştırma tablosunda Duyarlılık ve Kesinlik değerleri aynı olduğu için kesinlik değerleri tabloya dâhil edilmemiştir. Sonuçlara olan güveni ifade eden özgüllük değeri, pozitif olarak sınıflandırılan değişkenleri doğru olarak göstermektedir [134] ve bu oran tüm yöntemlerde yüksek değerler vermiştir. Ancak en düşük yanlış pozitif sınıflandırma oranı (0.024) Bayes Ağları ile elde edilmiştir. Aynı sonuçlar ROC değeri için de geçerlidir.



Şekil 4.3 Yöntemlerin genel karşılaştırma sonuçlarının grafiksel gösterimi

Sonuçların grafiksel karşılaştırması Şekilde verilmiştir. Grafik incelendiğinde karşılaştırılan makine öğrenmesi yöntemlerinin birbirine yakın olduğu ve Bayes Ağlarının karşılaştırılan diğer makine öğrenmesi algoritmalarından daha iyi sonuçlar ürettiği görülmektedir.

Yöntemlerin genel olarak karşılaştırılmasının yanında alt sınıflar için de ayrı ayrı karşılaştırma yapılmıştır. Her bir alt sınıf için elde edilen karşılaştırma sonuçları Tablo 4.3'te verilmiştir. Buna göre Bayes Ağlarının ALS tahmininde diğer yöntemlere göre daha başarılı sonuçlar verdiği görülmüştür. ALS hasta grubundaki tüm bireylerin doğru sınıflandırıldığı gözlemlendi. DVM ve YSA ile

elde edilen sonuçlar da bu değerlere yakındır. ALS hastalarını öngörmede tüm yöntemlerin başarılı sonuçlar verdiği ileri sürülebilir.

Tablo 4.3 Yöntemlerin ALS, kontrol, nörolojik kontrol ve Parkinson hastalığı alt grupları bazında karşılaştırılması

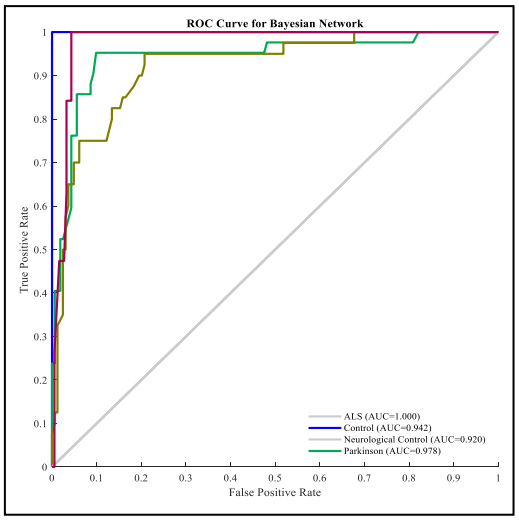
		ACC	GM	ERR	PREC	SENS	SPEC	F-M	MCC	YI	Kappa
ALS	BA	1,000	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	YSA	0,985	0,985	0,015	1,000	0,971	1,000	0,985	0,971	0,971	0,971
	LR	0,975	0,975	0,025	1,000	0,951	1,000	0,975	0,952	0,951	0,951
	NB	0,971	0,970	0,029	0,962	0,981	0,960	0,971	0,941	0,941	0,941
	J48	0,980	0,980	0,020	1,000	0,961	1,000	0,980	0,962	0,961	0,961
	DVM	0,980	0,980	0,020	1,000	0,961	1,000	0,980	0,962	0,961	0,961
	KStar	0,975	0,976	0,025	0,990	0,961	0,990	0,975	0,951	0,951	0,951
	k-NN	0,956	0,956	0,044	0,990	0,922	0,990	0,955	0,914	0,912	0,912
Kontrol	BA	0,917	0,874	0,083	0,791	0,810	0,944	0,800	0,747	0,754	0,747
	YSA	0,882	0,813	0,118	0,714	0,714	0,926	0,714	0,640	0,640	0,640
	LR	0,868	0,845	0,132	0,642	0,810	0,883	0,716	0,638	0,692	0,631
	NB	0,882	0,854	0,118	0,680	0,810	0,901	0,739	0,668	0,711	0,664
	J48	0,887	0,902	0,113	0,661	0,929	0,877	0,772	0,718	0,805	0,700
	DVM	0,892	0,897	0,108	0,679	0,905	0,889	0,776	0,719	0,794	0,706
	KStar	0,907	0,888	0,093	0,735	0,857	0,920	0,791	0,735	0,777	0,732
	k-NN	0,912	0,891	0,088	0,750	0,857	0,926	0,800	0,746	0,783	0,744
Nörolojik Kontrol	BA	0,902	0,816	0,098	0,778	0,700	0,951	0,737	0,678	0,651	0,677
	YSA	0,848	0,738	0,152	0,615	0,600	0,909	0,608	0,513	0,509	0,513
	LR	0,848	0,668	0,152	0,655	0,475	0,939	0,551	0,471	0,414	0,462
	NB	0,809	0,568	0,191	0,519	0,350	0,921	0,418	0,317	0,271	0,309
	J48	0,819	0,532	0,181	0,571	0,300	0,945	0,393	0,320	0,245	0,299
	DVM	0,843	0,650	0,157	0,643	0,450	0,939	0,529	0,449	0,389	0,439
	KStar	0,853	0,670	0,147	0,679	0,475	0,945	0,559	0,485	0,420	0,474
	k-NN	0,833	0,661	0,167	0,594	0,475	0,921	0,528	0,432	0,396	0,428
Parkinson	BA	0,956	0,903	0,044	0,727	0,842	0,968	0,780	0,759	0,810	0,756
	YSA	0,941	0,869	0,059	0,652	0,789	0,957	0,714	0,686	0,746	0,682
	LR	0,946	0,898	0,054	0,667	0,842	0,957	0,744	0,721	0,799	0,715
	NB	0,936	0,840	0,064	0,636	0,737	0,957	0,683	0,650	0,694	0,648
	J48	0,922	0,832	0,078	0,560	0,737	0,941	0,636	0,600	0,677	0,593
	DVM	0,941	0,842	0,059	0,667	0,737	0,962	0,700	0,669	0,699	0,667
	KStar	0,941	0,920	0,059	0,630	0,895	0,946	0,739	0,721	0,841	0,707
	k-NN	0,917	0,857	0,083	0,536	0,789	0,930	0,638	0,607	0,719	0,593

Kontrol grubu için sonuçlar incelendiğinde en yüksek ACC değerini 0,917 ile Bayes Ağlarının verdiği görülmüştür. J48 algoritması ise GM (0,902), SENS (0,929) ve YI (0,805) kriterlerine göre en iyi sonuçları vermiştir. Bununla birlikte, Bayes Ağları veriler ve öngörü sonuçları arasında Kappa değeri [135] ile en iyi uyumu (0,747) göstermiştir. Ayrıca kontrol grubu için diğer kriterler için en iyi sonuçlar Bayes Ağları tarafından üretilmiştir.

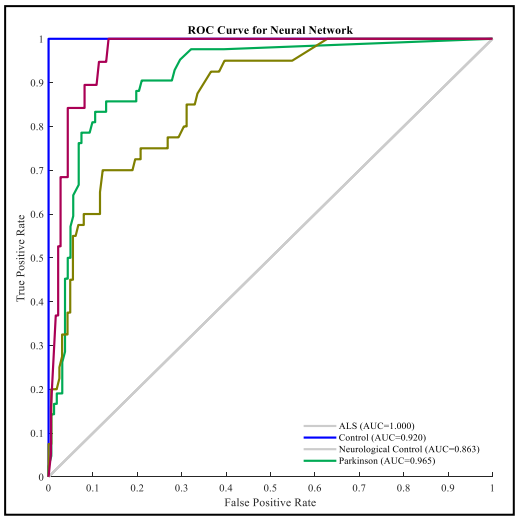
Bayes Ağları, ALS grubunda olduğu gibi Nörolojik Kontrol grubu için de tüm karşılaştırma kriterlerine göre diğer yöntemlerden daha başarılı sonuçlar üretmiştir. Bu grup için Bayes Ağlarının ACC değeri (0,902) olarak bulunmuştur. Diğer yöntemlerin Kappa değerleri bu grup için elde edilen sonuçların rastgele olduğunu gösterirken, Bayes Ağlarının Kappa değeri (0,677) bulunmuştur.

Son grup olan Parkinson hastaları için kontrol grubuna benzer sonuçlar elde edilmiştir. Kstar algoritması GM (0,920), SENS (0,895) ve YI (0,841) kriterlerine göre en iyi sonuçları vermiştir. Ancak Bayes Ağları için elde edilen sonuçların bu değerlere yakın olduğu görülmüştür.

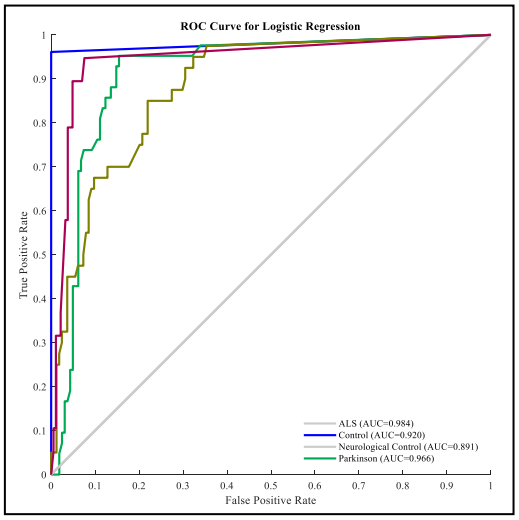
Yöntemlerin ROC eğrileri ve AUC değerleri Şekil 4.4'te verilmiştir. Bu değerlere göre Bayes Ağlarının her sınıf için AUC değeri diğer yöntemlere göre daha yüksektir. Bu sonuç Tablo 4.3'te gösterilen sonuçları desteklemektedir.


ROC Curve for Bayesian Network
True Positive Rate
False Positive Rate
ALS (AUC=1.000)
Control (AUC=0.942)
Neurological Control (AUC=0.920)
Parkinson (AUC=0.978)

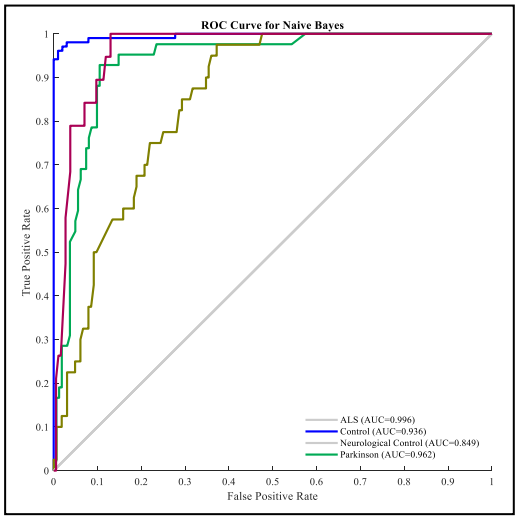
(a)


ROC Curve for Neural Network
True Positive Rate
False Positive Rate
ALS (AUC=1.000)
Control (AUC=0.920)
Neurological Control (AUC=0.863)
Parkinson (AUC=0.965)

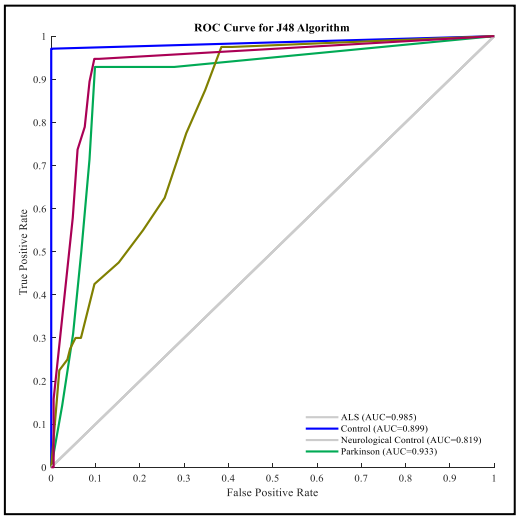
(b)


ROC Curve for Logistic Regression
True Positive Rate
False Positive Rate
ALS (AUC=0.984)
Control (AUC=0.920)
Neurological Control (AUC=0.891)
Parkinson (AUC=0.966)

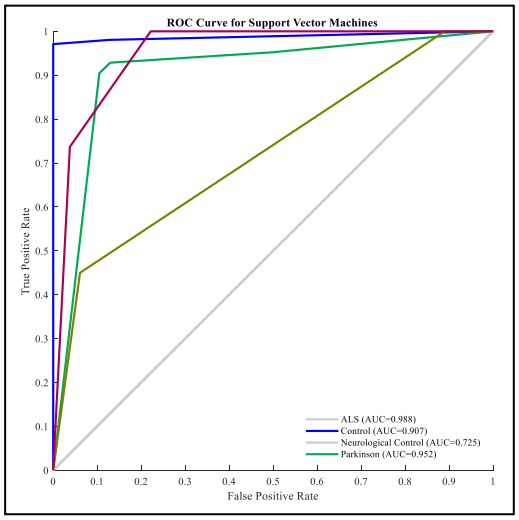
(c)


ROC Curve for Naive Bayes
True Positive Rate
False Positive Rate
ALS (AUC=0.996)
Control (AUC=0.936)
Neurological Control (AUC=0.849)
Parkinson (AUC=0.962)

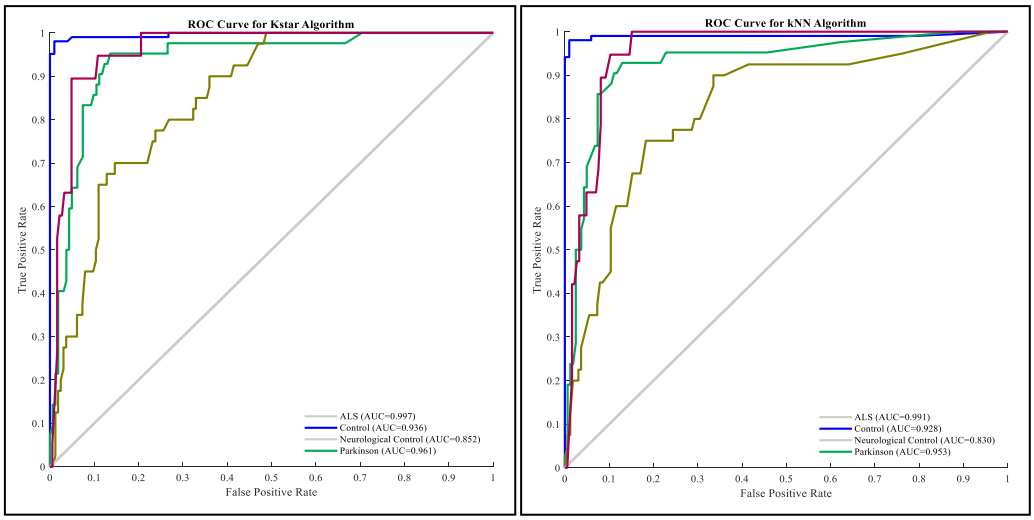
(d)


ROC Curve for J48 Algorithm
True Positive Rate
False Positive Rate
ALS (AUC=0.985)
Control (AUC=0.899)
Neurological Control (AUC=0.819)
Parkinson (AUC=0.933)

(e)


ROC Curve for Support Vector Machines
True Positive Rate
False Positive Rate
ALS (AUC=0.988)
Control (AUC=0.907)
Neurological Control (AUC=0.725)
Parkinson (AUC=0.952)

(f)



(g) **(h)**

Şekil 4.4 Yöntemlerin ROC analizi sonuçları; **(a)** Bayes Ağları **(b)** Yapay Sinir Ağları **(c)** Lojistik Regresyon **(d)** Naive Bayes **(e)** J48 Algoritması **(f)** DVM **(g)** Kstar **(h)** kNN.

4.3.1 Bayes Ağları Modeli ile Üretilen Sorgular

Bayes ağlarının en önemli özelliklerinden biri, mevcut bilgi ve verilerle sorgular oluşturularak tahminlerde bulunulabilmesidir [52]. Bilinen değişkenler kanıt olarak dâhil edilirken, tahmin edilen değişkenler hedef düğümler olarak alınır. Hastalık gruplarından birine yeni dâhil olan bir birey düşünüldüğünde diğer değişkenlerin durumu ile ilgili sorular Tablo 4.4'te verilmiştir.

Tabloda ilk satırda verilen değerler hastalık tipi ile ilgili herhangi bir bilginin olmadığı durumda kişilerin hastalık tiplerinin olasılıkları verilmektedir. Buna göre bir kişinin ALS hastası olma ihtimali P(Hasta Tipi = ALS) = 0,340; kontrol grubunda olma ihtimali P(Hasta Tipi = Control) = 0,252; Nörolojik kontrol grubunda olma ihtimali P (Hasta Tipi = Nörolojik Kontrol) = 0,257 ve Parkinson hastası olma ihtimali P (Hasta Tipi = Parkinson) = 0,151 olarak verilmiştir.

Verilen bu değerlerden yola çıkılarak ALS hastası olduğu bilinen bir kişinin cinsiyeti için elde edilen koşullu olasılık değeri denklem (4.1)'de gösterilmiştir.

$$P\left(CINSIYET = Kadın \mid Hasta\ Tipi = ALS\right) = 0,382$$
$$P\left(CINSIYET = Erkek \mid Hasta\ Tipi = ALS\right) = 0,618 \quad (4.1)$$

Hesaplanan olasılık değerleri incelendiğinde kişinin ALS hastası olduğu bilindiğinde cinsiyetin farklılaştığı ve kişinin hasta olma olasılıklarının cinsiyetten etkilendiği anlaşılmaktadır. Bu sonuca göre ALS hastalığının erkeklerde %62, kadınlarda ise %38 oranında görüldüğü tespit edilmiştir. ALS hastalığının erkeklerde kadınlardan daha fazla görüldüğü bilinmektedir. Buna göre ele alınan veri setinde de cinsiyetin hastalık üzerinde etkili bir değişken olarak bulunmuştur.

Tablo 4.4 Hasta tipi için Bayes ağı modelinin sorguları

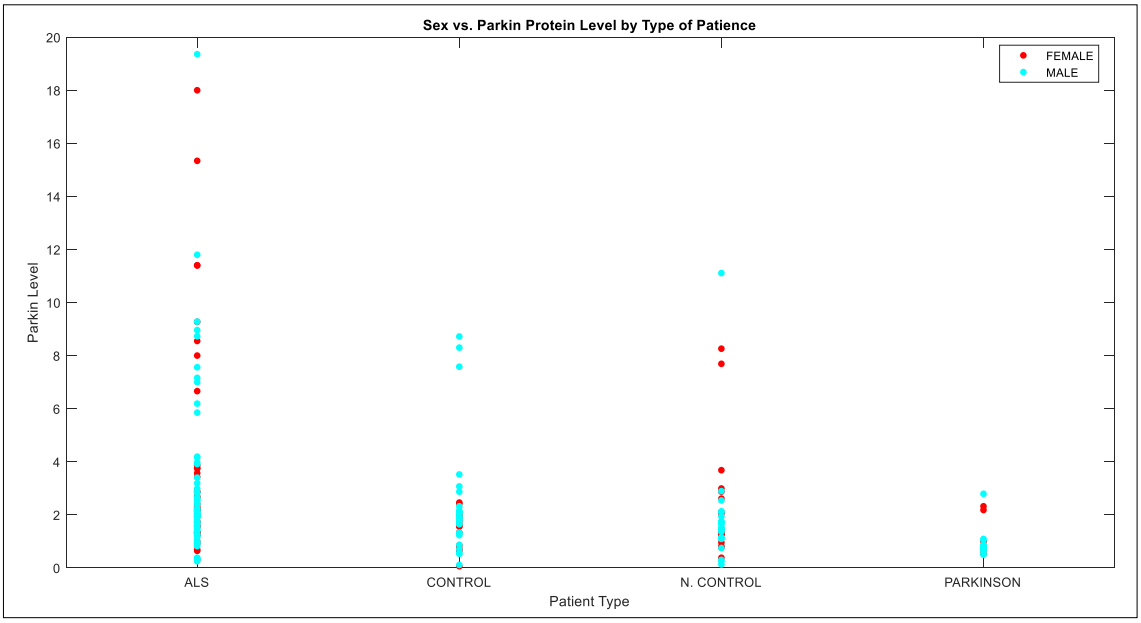
Hedef Düğüm	Hedef Değer	Kanıt (Hasta Tipi)			
		ALS	Kontrol	N.Kontrol	Parkinson
Yok	Yok	0,340	0,252	0,257	0,151
YAŞ	36'dan az	0,119	0,249	0,106	0,078
	36 – 52 arası	0,333	0,419	0,298	0,316
	52 – 67 arası	0,437	0,232	0,449	0,430
	67'den büyük	0,111	0,100	0,148	0,175
CİNSİYET	Kadın	0,382	0,342	0,401	0,452
	Erkek	0,618	0,658	0,599	0,548
PARKIN Seviyesi (ng/mL)	3.74'ten yüksek	0,244	0,107	0,098	0,114
	2.79 – 3.74 arası	0,078	0,072	0,100	0,084
	2.06 – 2.79 arası	0,192	0,200	0,159	0,131
	1.36 – 2.06 arası	0,238	0,279	0,340	0,107
	1.36'den az	0,247	0,343	0,303	0,563
UMN	Hayır	0,273	0,840	0,844	0,733
	Evet	0,727	0,160	0,156	0,267
LMN	Hayır	0,822	0,912	0,913	0,852
	Evet	0,178	0,088	0,087	0,148
BULBAR	Hayır	0,853	0,924	0,925	0,872
	Evet	0,147	0,076	0,075	0,128
Toplam Kronik Hastalık Sayısı	Yok	0,040	0,696	0,288	0,075
	Bir	0,704	0,189	0,595	0,731
	İki	0,140	0,060	0,059	0,101
	Üç	0,103	0,042	0,045	0,070
	Dört	0,008	0,008	0,007	0,013
	Beş	0,006	0,006	0,006	0,010

ALS hastalığı literatürde büyük oranda erişkin başlangıçlı bir hastalık olarak ifade edilmektedir [130]. Bu bölümde ALS hastalarının %88,1'inin 36 yaşından büyük olduğu belirlenmiştir. Ayrıca ALS hastalarının %54,8'inin ve Parkinson hastalarının %60,5'inin 52 yaşından büyük olduğu öngörülmüştür.

$$\begin{aligned} P\left(UMN = Hayır \mid Hasta\ Tipi = ALS\right) &= 0,273 \\ P\left(UMN = Evet \mid Hasta\ Tipi = ALS\right) &= 0,727 \end{aligned} \tag{4.2}$$

Denklem (4.2)'de verilen bilgiler incelendiğinde ALS hastalarının %72,7'sinde UMN tipi tutulumun olduğu tahmin edilmektedir. Ayrıca hastaların %82,2'sinde LMN'nin olmadığı ve %85,3'ünde BULBAR başlangıçlı tutulumun olmadığı anlaşılmaktadır. Ancak ALS hastalarının %3,8'inde tutulum olmadığı tahmin edilmiştir. Özetle, ALS hastalarında en az 1 tip başlangıç tutulumunun olma olasılığı %96,2 olarak tahmin edilmiştir.

Tabloda ALS hastalarının %25,7'sinin kendi hastalığı dışında en az 1 hastalığı bulunmaktadır. Parkinson hastalarında bu olasılık %19,4 olarak bulunmuştur. Bu olasılık kontrol grubunda %11,6, nörolojik kontrol grubunda %11,7 olarak hesaplanmıştır. Buna göre farklı nörolojik-kronik hastalıkların Parkinson veya ALS gibi nörolojik hastalıklarla ilişkili olduğu düşünülebilir.



Şekil 4.5 Cinsiyete karşı parkin düzeyinin (ng/mL) hastalık tipine göre gösterimi

Ayrıca ALS hastalarının %75,3'ünün Parkin düzeyinin 1,36'nın (ng/mL) üzerinde olduğu tahmin edilmektedir. Parkinson hastalarında bu değer oldukça farklıdır. Tablo incelendiğinde Parkinson hastalarının %56,3'ünün parkin düzeyi 1,36 (ng/mL)'nin altında olduğu tahmin edilmiştir. Ayrıca Parkin düzeyi dağılımı Şekil 4.5'te verilmiştir. Grupların protein düzeyi farklılıkları da grafikte gösterilmiştir. ALS hastalarında protein düzeyi en yüksektir, ancak Parkinson hastalarında bu düzey en düşüktür. Buna göre Parkin proteini düzeyinin hastalık tipi üzerinde anlamlı bir etkisi olabileceği anlaşılmaktadır.

Verilen bu bilgilere ek olarak ele alınan değişkenlerin etkileşimlerinin de irdelenebilmesi mümkündür. Özellikle ALS hastalarında tutulum görülme ihtimali yüksektir. Hastaların tutulum tipleri ele alınarak oluşturulan sorgu tablosu aşağıda verilmiştir.

Tablo 4.5 Bayes Ağları ile tutulum tiplerinin etkileşimleri bazında elde edilen sorgular

Kanıt Düğüm(ler)	Kanıt Bilgi	Hedef Değer (Hasta Tipi)			
		ALS	Kontrol	N. Kontrol	Parkinson
Yok	Yok	0,340	0,252	0,257	0,151
UMN, LMN	Hayır, Hayır	0,090	0,362	0,371	0,178
	Hayır, Evet	0,534	0,155	0,155	0,155
	Evet, Hayır	0,715	0,095	0,095	0,095
	Evet, Evet	0,375	0,208	0,208	0,208
UMN, BULBAR	Hayır, Hayır	0,105	0,355	0,364	0,176
	Hayır, Evet	0,493	0,169	0,169	0,169
	Evet, Hayır	0,703	0,099	0,099	0,099
	Evet, Evet	0,414	0,195	0,195	0,195
LMN, BULBAR	Hayır, Hayır	0,299	0,275	0,281	0,145
	Hayır, Evet	0,491	0,170	0,170	0,170
	Evet, Hayır	0,499	0,167	0,167	0,167
	Evet, Evet	0,283	0,239	0,239	0,239
UMN, LMN, BULBAR	Hayır, Hayır, Hayır	0,038	0,386	0,396	0,180
	Hayır, Hayır, Evet	0,521	0,160	0,160	0,160
	Hayır, Evet, Hayır	0,562	0,146	0,146	0,146
	Evet, Hayır, Hayır	0,748	0,084	0,084	0,084
	Hayır, Evet, Evet	0,302	0,233	0,233	0,233
	Evet, Hayır, Evet	0,438	0,187	0,187	0,187
	Evet, Evet, Hayır	0,390	0,203	0,203	0,203
	Evet, Evet, Evet	0,250	0,250	0,250	0,250

Bir önceki tabloda tutulum tiplerinin hastalık tipi bilindiği durumda tek başına görüldüğü hali göstermektedir. Bu tabloda ise tutulum tiplerinin birlikte görüldüğü durumda hastalık tipi tahmin edilmektedir. Bir kişide UMN görüldüğü ve LMN görülmediği durumda kişinin ALS hastası olma olasılığı %71,5'tir.

$$P\left(Hasta\ Tipi = ALS \mid\ UMN = Evet, LMN = Hayır\right) = 0,715 \tag{4.3}$$

Bu ifadenin olasılıksal gösterimi denklem (4.3)'te verilmiştir. Bir kişide UMN görüldüğü ve BULBAR görülmediği durumda ise kişinin ALS hastası olma olasılığı %70,3'tür. Ayrıca bir kişide UMN görülüyorken LMN ve BULBAR görülmediği durumda bu kişinin ALS hastası olma olasılığı %74,8 olarak hesaplanmıştır.

Bir kişide herhangi bir tutulum görülmediği halde ALS hastası olma olasılığı %3,8 olarak tahmin edilmektedir. Bir diğer durumda ise her üç tutulum tipinin de görülme olasılığı %25 olarak verilmiştir. Ancak bu olasılık diğer hastalıklar ile eşit dağıtılmış bir olasılık değeridir. Yani bir kişide her üç tutulum tipinin bulunma olasılığı sınıflara eşit dağıtılmış rastgele bir olasılık değeri olarak verilmiştir. Öyleyse bir kişide her üç tutulum tipinin görülmediği tahmin edilmektedir.

5
SONUÇ VE ÖNERİLER

Tıbbi karar verme süreçlerinde kişisel tıbbi kayıtlarla makine öğrenmesi yöntemlerinin kullanımı artmaktadır. Bu çalışmada klinik karar vermede en faydalı makine öğrenmesi yöntemlerinden biri olan Bayes Ağları, Parkin plazma protein düzeyi, tutulum tipleri, yaş, cinsiyet ve toplam hastalık sayısı değişkenlerine dayalı olarak ALS hastalığının tahmini için kullanılmıştır. Daha sonra elde edilen sonuçlar 7 popüler makine öğrenmesi algoritmasının sonuçları ile karşılaştırılmıştır. Literatürden elde edilen bilgilere dayanarak bu Bayes Ağı modeli ve modellemede kullanılan değişkenler ile ALS hastalığını tahmin etmeye yönelik makine öğrenmesi modellemeleri için ilk performans karşılaştırma çalışmasıdır.

Bayes Ağları, verileri en iyi temsil eden grafiksel bir yapı elde etmek için bir belirsizlik ölçüsü olarak olasılığı kullanan olasılıksal uzman sistemlerden biridir [29, 30]. Bayes Ağları modeldeki tüm değişkenleri kullandığından veri eksikliğinin olduğu durumlarda dahi kolaylıkla kullanılmaktadır [33, 122]. Bayes Ağları ile tanısal akıl yürütme ile çeşitli semptomlar gözlenerek hasta ve hastalık hakkında bir yargıya varılabilmektedir [63].

YSA, LR, DVM gibi çeşitli kural tabanlı makine öğrenmesi yöntemlerinden farklı olarak Bayes Ağları, bir çıkarım ve akıl yürütme yöntemidir. Belirtilen özellikleri ile Bayes Ağları elde edilen modelde değişkenler arasındaki neden-sonuç ilişkilerinin ortaya çıkarılması için çeşitli sorguların yapılmasını sağlar. Bayes Ağlarında sonsal olasılık değerleri elde edilen her yeni bilgi ile güncellenir [47]. Bu nedenle, BN'nin tahmin problemlerinde kullanılması birçok durumda etkili sonuçlar vermektedir [137].

Ağ yapısındaki tüm ilişkilerin şeffaf olması Bayes Ağlarını k-nn, YSA ve LR gibi diğer makine öğrenmesi yöntemlerine göre avantajlı hale getirmektedir. Ayrıca küçük veri setlerinde ve değişken sayısının çok fazla olduğu durumlarda Bayes

Ağları başarılı sonuçlar üretebilmektedir [138]. Ancak modelleme aşamasında değişkenlerin kesiklileştirilmesinin neden olduğu bilgi kaybı Bayes Ağlarının en büyük dezavantajıdır [139]. Bununla birlikte ayrık verilerle çalışmak, sınıflarla ilgili tahmin performansını arttırmaktadır [140]. Tüm bu özellikler Bayes Ağlarını klinik çalışmalarda tercih edilen bir yöntem haline getirmiştir [35, 122, 124, 127–129].

Bu çalışmada literatürde incelenen çalışmalardan farklı olarak üç kontrol grubu bulunmaktadır. Kontrol grubunda Parkinson hastaları, nörolojik kontrol grubu ve tamamen sağlıklı bireyler yer almaktadır. Bu şekilde, çok sayıda denek içeren farklı kontrol grupları ile karşılaştırmalı bir sonuç, çalışmanın pratikte uygulanabilirliğini arttırmaktadır.

Bu çalışmanın sonuçlarına göre ALS hastalığının erkeklerde görülme olasılığı kadınlara göre daha fazladır. Çeşitli araştırmalarda da belirtildiğine göre [112, 130, 131] cinsiyet diğer demografik faktörlerle birlikte ALS'yi etkileyen bağımsız bir değişkendir. Araştırmada literatüre benzer şekilde cinsiyetin etkili bir değişken olduğu ve ALS hastalığının erkeklerde daha sık görüldüğü doğrulanmıştır [114, 115, 132–134]. Cinsiyete bağlı olarak farklı mutasyonlara sahip ALS hastalarında hastalık başlangıcında farklılık olduğunu gösteren çalışmalar da mevcuttur [131, 135].

Elde edilen Bayes Ağı ile yaş arttıkça ALS hastası olma olasılığının yükseleceği belirlenmiştir. Bu bulgu önceki çalışmaların sonuçlarıyla da uyumludur [136, 137]. UMN, LMN ve Bulbar tipi tutulumlar, ALS hastalığında görülen tutulum türleridir. ALS hastalarında her bir tutulum türünden en az birinin bulunma olasılığı %96,2 olarak bulunmuştur. UMN ise ALS hastalığında en yaygın tutulum tipi olarak belirlenmiştir. LMN ve Bulbar tipi tutulumlar daha az görülmektedir. ALS hastalarında LMN, UMN veya her ikisi birlikte yaygın görülen fenotiptir [151]. ALS fenotiplerinin önceden belirlinen klinik ve demografik özellikleri geniş bir popülasyona dayalı epidemiyolojik ortamda gösterilmiştir. Belirli bir fenotipin farklı yaş ve cinsiyet gruplarında ortaya çıkma

olasılığı değişmektedir. Bulbar fenotipi iki cinsiyette neredeyse eşit oluşum oranlarıyla çoğunlukla yaşlı hastalarda görülmektedir [150].

Analiz edilen veri setinde özellikle ALS, çevresel ve genetik faktörlerden etkilenen çok faktörlü bir hastalık olarak kabul edilmektedir. İnsanların sahip olduğu diğer nörolojik hastalıkların ALS üzerinde küçük bir etkisi olabilir. Şizofreni [152], Alzheimer hastalığı, Parkinson hastalığı veya frontotemporal demans [140, 141] gibi diğer hastalıkların neden olduğu beyin hasarlarının ve mutasyonların [151] ALS ile ilişkili olduğu düşünülmektedir. Çalışmamızda ALS ile birden fazla hastalığa sahip olma olasılığı kontrol ve nörolojik kontrol gruplarına göre daha yüksektir. Parkinson hastalarında da benzer sonuçlar görülmüştür. Bu nedenle hastaların birden fazla hastalık durumu göz önünde bulundurularak tedavi edilmesi faydalı olacaktır.

Tüm bu sonuçlara göre bireylerin tutulum tipi, Parkin protein düzeyi, yaşı, çeşitli kronik hastalıkların sayısı gibi bilgiler göz önüne alındığında kullandığımız algoritmalar ve Bayes ağı yüksek doğruluk oranlarıyla doğru sınıfları tahmin edebilmektedir. Diğer makine öğrenme algoritmaları da yüksek başarı ile sonuçlar üretse de Bayes ağının bu konudaki en önemli avantajı yeni ek bilgilerle güncellenebilmesi ve bu yönüyle başarısını artırmasıdır. Bu açıdan kara kutu özelliği gösteren yapay sinir ağları gibi diğer makine öğrenmesi yöntemlerine göre daha kullanışlı sonuçlar sağlamaktadır.

İleriye dönük çalışmalarda örnek sayısı artırılarak ve daha fazla değişken kullanılarak sonuçların değişimi incelenmelidir. Klasik istatistiksel çalışmalardan ve hesaplamalı yöntemlerden elde edilen sonuçlar, nörogörüntüleme yöntemleri ile birleştirilerek kullanılması halinde daha faydalı olabilir [155]. Farklı beyin ağlarının görüntülerini alternatif veya ek görünümler olarak sınıflandırmak için benzer bir yaklaşım kullanılabilir. Ayrıca çalışmanın uygulama çerçevesi görüntülemeyi klinik, davranışsal ve hatta genetik çok boyutlu veriler gibi değişkenlerle birleştirilerek hastalıklar hakkında daha geniş karar verme sistemi oluşturulabilir.

KAYNAKÇA

[1] A. Oniśko, M. J. Druzdzel, ve H. Wasyluk, "Learning Bayesian network parameters from small data sets: application of Noisy-OR gates", *International Journal of Approximate Reasoning*, c. 27, sy 2, ss. 165-182, Ağu. 2001, doi: 10.1016/S0888-613X(01)00039-1.

[2] J. Galapero, S. Fernández, C. J. Pérez, F. Calle-Alonso, J. Rey, ve L. Gómez, "Identifying risk factors for ovine respiratory processes by using Bayesian networks", *Small Ruminant Research*, c. 136, ss. 113-120, Mar. 2016, doi: 10.1016/j.smallrumres.2016.01.017.

[3] M. Li ve L. Wang, "Feature fatigue analysis in product development using Bayesian networks", *Expert Systems with Applications*, c. 38, sy 8, ss. 10631-10637, Ağu. 2011, doi: 10.1016/j.eswa.2011.02.126.

[4] N. Jongsawat ve W. Premchaiswadi, "A SMILE web-based interface for learning the causal structure and performing a diagnosis of a Bayesian network", içinde *2009 IEEE International Conference on Systems, Man and Cybernetics*, Eki. 2009, ss. 376-382. doi: 10.1109/ICSMC.2009.5346198.

[5] Z. Liu, B. Malone, ve C. Yuan, "Empirical evaluation of scoring functions for Bayesian network model selection", *BMC Bioinformatics*, c. 13, sy 15, s. S14, Eyl. 2012, doi: 10.1186/1471-2105-13-S15-S14.

[6] D. Cattell ve P. Love, "Using Bayesian Networks to assess the risk appetite of construction contractors", içinde *38th Australian University Building Educators Association Conference*, 2013, ss. 1-11. Erişim: Haz. 23, 2021. [Çevrimiçi]. Erişim adresi: https://research.bond.edu.au/en/publications/using-bayesian-networks-to-assess-the-risk-appetite-of-constructi

[7] A. Aghaie ve A. Saeedi, "Using Bayesian Networks for Bankruptcy Prediction: Empirical Evidence from Iranian Companies", içinde *2009 International Conference on Information Management and Engineering*, Nis. 2009, ss. 450-455. doi: 10.1109/ICIME.2009.91.

[8] L. Sun ve P. P. Shenoy, "Using Bayesian networks for bankruptcy prediction: Some methodological issues", *European Journal of Operational Research*, c. 180, sy 2, ss. 738-753, Tem. 2007, doi: 10.1016/j.ejor.2006.04.019.

[9] S. Sarkar ve R. S. Sriram, "Bayesian Models for Early Warning of Bank Failures", *Management Science*, c. 47, sy 11, ss. 1457-1475, 2001.

[10] K. Masmoudi, L. Abid, ve A. Masmoudi, "Credit risk modeling using Bayesian network with a latent variable", *Expert Systems with Applications*, c. 127, ss. 157-166, Ağu. 2019, doi: 10.1016/j.eswa.2019.03.014.

[11] T. Pavlenko ve O. Chernyak, "Credit risk modeling using bayesian networks", *International Journal of Intelligent Systems*, c. 25, sy 4, ss. 326-344, 2010, doi: 10.1002/int.20410.

[12] W. Abramowicz, M. Nowak, ve J. Sztykiel, "Bayesian networks as a decision support tool in credit scoring domain", içinde *Managing Data Mining Technologies in Organizations: Techniques and Applications*, P. C. Pendharkar, Ed. IGI Global, 2003. doi: 10.4018/978-1-59140-057-8.

[13] C. K. Leong, "Credit Risk Scoring with Bayesian Network Models", *Comput Econ*, c. 47, sy 3, ss. 423-446, Mar. 2016, doi: 10.1007/s10614-015-9505-8.

[14] C. Shenoy ve P. P. Shenoy, "Bayesian Network Models of Portfolio Risk and Return", The MIT Press, 2000. Erişim: Haz. 20, 2021. [Çevrimiçi]. Erişim adresi: https://kuscholarworks.ku.edu/handle/1808/161

[15] M. E. De Giuli, M. A. Maggi, ve C. Tarantola, "Bayesian outlier detection in Capital Asset Pricing Model", *Statistical Modelling*, c. 10, sy 4, ss. 375-390, Ara. 2010, doi: 10.1177/1471082X0901000402.

[16] P. Wijayatunga, S. Mase, ve M. Nakamura, "Appraisal of companies with Bayesian networks", *International Journal of Business Intelligence and Data Mining*, Mar. 2006, Erişim: Haz. 24, 2021. [Çevrimiçi]. Erişim adresi: https://www.inderscienceonline.com/doi/abs/10.1504/IJBIDM.2006.009138

[17] I. E. Donaldson Soberanis, "An extended Bayesian network approach for analyzing supply chain disruptions", Doctor of Philosophy, University of Iowa, 2010. doi: 10.17077/etd.e4hw0b69.

[18] Y. Zuo ve E. Kita, "Stock price forecast using Bayesian network", *Expert Systems with Applications*, c. 39, sy 8, ss. 6729-6737, Haz. 2012, doi: 10.1016/j.eswa.2011.12.035.

[19] B. Baesens, M. Egmont-Petersen, R. Castelo, ve J. Vanthienen, "Learning Bayesian network classifiers for credit scoring using Markov chain Monte Carlo search", içinde *Object recognition supported by user interaction for service robots*, Ağu. 2002, c. 3, ss. 49-52 c.3. doi: 10.1109/ICPR.2002.1047792.

[20] J. Gemela, "Financial analysis using Bayesian networks", *Applied Stochastic Models in Business and Industry*, c. 17, sy 1, ss. 57-67, 2001, doi: 10.1002/asmb.422.

[21] J. Gemela, "Learning Bayesian networks using various datasources and applications to financial analysis", *Soft Computing*, c. 7, sy 5, ss. 297-303, Nis. 2003, doi: 10.1007/s00500-002-0216-4.

[22] Y. Wu, J. McCall, ve D. Corne, "Two novel Ant Colony Optimization approaches for Bayesian network structure learning", içinde *IEEE Congress on Evolutionary Computation*, Tem. 2010, ss. 1-7. doi: 10.1109/CEC.2010.5586528.

[23] P. Weber, G. Medina-Oliva, C. Simon, ve B. Iung, "Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas", *Engineering Applications of Artificial Intelligence*, c. 25, sy 4, ss. 671-682, Haz. 2012, doi: 10.1016/j.engappai.2010.06.002.

[24] J. Olbrys, "Forecasting Portfolio Return Based on Bayesian Network Model", 2009, doi: 10.13140/2.1.2349.4087.

[25] O. Chernyak ve Y. Chernyak, “Classification of Financial Conditions of the Enterprises in Different Industries of Ukrainian Economy Using Bayesian Networks”, içinde *Proceedings of the International Conference on Information and Communication Technologies*, Skiathos, 2011, s. 12.

[26] C. Tarantola, P. Vicard, ve I. Ntzoufras, “Monitoring and improving Greek banking services using Bayesian Networks: An analysis of mystery shopping data”, *Expert Systems with Applications*, c. 39, sy 11, ss. 10103-10111, Eyl. 2012, doi: 10.1016/j.eswa.2012.02.060.

[27] C.-W. Shen, “A bayesian networks approach to modeling financial risks of e-logistics investments”, *Int. J. Info. Tech. Dec. Mak.*, c. 08, sy 04, ss. 711-726, Ara. 2009, doi: 10.1142/S0219622009003594.

[28] A. Oniśko, M. Druzdzel, ve H. Wasyluk, “A Probabilistic Causal Model for Diagnosis of Liver Disorders”, içinde *Proceedings of the Intelligent Information Systems VII Workshop*, Malbork, Poland, Eki. 1998, ss. 379-387.

[29] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, ve D. C. Muchaluat Saade, “A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer′s disease and mild cognitive impairment”, *Computers in Biology and Medicine*, c. 51, ss. 140-158, Ağu. 2014, doi: 10.1016/j.compbiomed.2014.04.010.

[30] P. P. Rodrigues, D. Ferreira-Santos, A. Silva, J. Polónia, ve I. Ribeiro-Vaz, “Causality assessment of adverse drug reaction reports using an expert-defined Bayesian network”, *Artificial Intelligence in Medicine*, c. 91, ss. 12-22, Eyl. 2018, doi: 10.1016/j.artmed.2018.07.005.

[31] P. Arora, D. Boyne, J. J. Slater, A. Gupta, D. R. Brenner, ve M. J. Druzdzel, “Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine”, *Value in Health*, c. 22, sy 4, ss. 439-445, Nis. 2019, doi: 10.1016/j.jval.2019.01.006.

[32] D. Zhao ve C. Weng, “Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction”, *Journal of Biomedical Informatics*, c. 44, sy 5, ss. 859-868, Eki. 2011, doi: 10.1016/j.jbi.2011.05.004.

[33] S. Lee *vd.*, “Bayesian network ensemble as a multivariate strategy to predict radiation pneumonitis risk: Multivariate BN ensemble to predict RP risk”, *Med. Phys.*, c. 42, sy 5, ss. 2421-2430, Nis. 2015, doi: 10.1118/1.4915284.

[34] N. H. Orak, “A Hybrid Bayesian Network Framework for Risk Assessment of Arsenic Exposure and Adverse Reproductive Outcomes”, *Ecotoxicology and Environmental Safety*, c. 192, s. 110270, Nis. 2020, doi: 10.1016/j.ecoenv.2020.110270.

[35] M.-S. Ong, “A Bayesian Network Approach to Disease Subtype Discovery”, içinde *Bioinformatics and Drug Discovery*, c. 1939, R. S. Larson ve T. I. Oprea, Ed. New York, NY: Springer New York, 2019, ss. 299-322. doi: 10.1007/978-1-4939-9089-4_17.

[36] H. J. P. Marvin *vd.*, “Application of Bayesian networks for hazard ranking of nanomaterials to support human health risk assessment”,

Nanotoxicology, c. 11, sy 1, ss. 123-133, Oca. 2017, doi: 10.1080/17435390.2016.1278481.

[37] P. Antal, G. Fannes, D. Timmerman, Y. Moreau, ve B. De Moor, "Using literature and data to learn Bayesian networks as clinical models of ovarian tumors", *Artificial Intelligence in Medicine*, c. 30, sy 3, ss. 257-281, Mar. 2004, doi: 10.1016/j.artmed.2003.11.007.

[38] S. Sciarretta, F. Palano, G. Tocci, R. Baldini, ve M. Volpe, "Antihypertensive treatment and development of heart failure in hypertension: a Bayesian network meta-analysis of studies in patients with hypertension and high cardiovascular risk", *Arch Intern Med*, c. 171, sy 5, ss. 384-394, Mar. 2011, doi: 10.1001/archinternmed.2010.427.

[39] S. L. Mestizo-Gutiérrez, J. A. Jácome-Delgado, V. Y. Rosales-Morales, N. Cruz-Ramírez, ve G. E. Aranda-Abreu, "A Bayesian Network Model for the Parkinson's Disease: A Study of Gene Expression Levels", içinde *Current Trends in Semantic Web Technologies: Theory and Practice*, c. 815, G. Alor-Hernández, J. L. Sánchez-Cervantes, A. Rodríguez-González, ve R. Valencia-García, Ed. Cham: Springer International Publishing, 2019, ss. 153-186. doi: 10.1007/978-3-030-06149-4_7.

[40] C. Sa-ngamuang, P. Haddawy, V. Luvira, W. Piyaphanee, S. Iamsirithaworn, ve S. Lawpoolsri, "Accuracy of dengue clinical diagnosis with and without NS1 antigen rapid test: Comparison between human and Bayesian network model decision", *PLOS Neglected Tropical Diseases*, c. 12, sy 6, s. e0006573, Haz. 2018, doi: 10.1371/journal.pntd.0006573.

[41] R. C. Sato ve G. T. K. Sato, "Probabilistic graphic models applied to identification of diseases", *Einstein (São Paulo)*, c. 13, sy 2, ss. 330-333, Haz. 2015, doi: 10.1590/S1679-45082015RB3121.

[42] R. E. Neapolitan, *Probabilistic methods for bioinformatics: with an introduction to Bayesian networks*. Amsterdam ; Boston: Morgan Kaufmann/Elsevier, 2009.

[43] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, ve R. G. Cowell, "Bayesian Analysis in Expert Systems", *Statistical Science*, c. 8, sy 3, ss. 219-247, 1993.

[44] F. V. Jensen ve T. D. Nielsen, *Bayesian Networks and Decision Graphs*, 2nd ed. New York: Springer, 2007.

[45] D. Heckerman, "A Tutorial on Learning with Bayesian Networks", içinde *Innovations in Bayesian Networks: Theory and Applications*, D. E. Holmes ve L. C. Jain, Ed. Berlin, Heidelberg: Springer, 2008, ss. 33-82. doi: 10.1007/978-3-540-85066-3_3.

[46] J. Lu, C. Bai, ve G. Zhang, "Cost-benefit factor analysis in e-services using bayesian networks", *Expert Systems with Applications*, c. 36, sy 3, ss. 4617-4625, Nis. 2009, doi: 10.1016/j.eswa.2008.05.018.

[47] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Revised Second Printing. San Francisco (CA): Morgan Kaufmann, 2014. doi: 10.1016/B978-0-08-051489-5.50002-3.

[48] C. A. Pollino, O. Woodberry, A. Nicholson, K. Korb, ve B. T. Hart, "Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment", *Environmental Modelling & Software*, c. 22, sy 8, ss. 1140-1152, Ağu. 2007, doi: 10.1016/j.envsoft.2006.03.006.

[49] R. Nagarajan, M. Scutari, ve S. Lèbre, *Bayesian Networks in R: with Applications in Systems Biology*. New York: Springer-Verlag, 2013. doi: 10.1007/978-1-4614-6446-4.

[50] C. M. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[51] H. Olmuş, "Bayes Ağlar", Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, Ankara, Türkiye, 2001.

[52] D. Koller ve N. Friedman, *Probabilistic Graphical Models. Principles and Techniques. Cambridge, Massachusets*. MIT Press, 2009.

[53] S. L. Lauritzen ve D. J. Spiegelhalter, "Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems", *Journal of the Royal Statistical Society. Series B (Methodological)*, c. 50, sy 2, ss. 157-224, 1988.

[54] H. Olmuş, "Bayes Ağlar ve Markov Ağları Kullanan Kümeleme Yönteminin İncelenmesi ve Bir Uygulama", Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, Ankara, Türkiye, 2007.

[55] M. Scutari ve J.-B. Denis, *Bayesian Networks: With Examples in R*. CRC Press : Taylor & Francis Group, 2014.

[56] E. Dünder, "Bayesçi Ağlarda Öğrenme Algoritmalarının Karşılaştırılması", Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, Ondokuz Mayıs Üniversitesi, Samsun, Türkiye, 2013.

[57] T. Verma ve J. Pearl, "Equivalence and synthesis of causal models", içinde *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, USA, Tem. 1990, ss. 255-270.

[58] M. Scutari, "Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimized Implementations in the **bnlearn** *R* Package", *J. Stat. Soft.*, c. 77, sy 2, 2017, doi: 10.18637/jss.v077.i02.

[59] S. Yaramakala ve D. Margaritis, "Speculative Markov blanket discovery for optimal feature selection", içinde *Fifth IEEE International Conference on Data Mining (ICDM'05)*, Kas. 2005, s. 4 pp.-. doi: 10.1109/ICDM.2005.134.

[60] M. Scutari, C. E. Graafland, ve J. M. Gutiérrez, "Who Learns Better Bayesian Network Structures: Constraint-Based, Score-based or Hybrid Algorithms?", içinde *International Conference on Probabilistic Graphical Models*, Ağu. 2018, ss. 416-427. Erişim: Haz. 28, 2021. [Çevrimiçi]. Erişim adresi: http://proceedings.mlr.press/v72/scutari18a.html

[61] P. Larrañaga, C. M. H. Kuijpers, M. Poza, ve R. H. Murga, "Decomposing Bayesian networks: triangulation of the moral graph with genetic algorithms", *Statistics and Computing*, c. 7, sy 1, ss. 19-34, Mar. 1997, doi: 10.1023/A:1018553211613.

[62] B. Lerner ve R. Malka*, “Investigation of the K2 Algorithm in Learning Bayesian Network Classifiers”, *Applied Artificial Intelligence*, c. 25, sy 1, ss. 74-96, Oca. 2011, doi: 10.1080/08839514.2011.529265.
[63] K. B. Korb ve A. E. Nicholson, *Bayesian Artificial Intelligence*. CRC Press, 2010.
[64] U. B. Kjærulff ve A. L. Madsen, *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*, 2nd Edition., c. 22. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-5104-4.
[65] P. Spirtes, C. N. Glymour, ve R. Scheines, *Causation, prediction, and search*, 2nd ed. Cambridge, Mass: MIT Press, 2000.
[66] D. E. Holmes ve L. C. Jain, Ed., *Innovations in Bayesian Networks*, c. 156. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-85066-3.
[67] C. Vasilopoulou, A. P. Morris, G. Giannakopoulos, S. Duguez, ve W. Duddy, “What Can Machine Learning Approaches in Genomics Tell Us about the Molecular Basis of Amyotrophic Lateral Sclerosis?”, *Journal of Personalized Medicine*, c. 10, sy 4, Art. sy 4, Ara. 2020, doi: 10.3390/jpm10040247.
[68] J. T. Senders *vd.*, “Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review”, *World Neurosurgery*, c. 109, ss. 476-486.e1, Oca. 2018, doi: 10.1016/j.wneu.2017.09.149.
[69] E. A. Gerlein, M. McGinnity, A. Belatreche, ve S. Coleman, “Evaluating machine learning classification for financial trading: An empirical approach”, *Expert Systems with Applications*, c. 54, ss. 193-207, Tem. 2016, doi: 10.1016/j.eswa.2016.01.018.
[70] S. Marsland, *Machine Learning: An Algorithmic Perspective*, Second Edition. CRC press, 2015.
[71] E. Alpaydin, *Introduction to machine learning*, Third edition. Cambridge, Massachusetts: The MIT Press, 2014.
[72] Deo Rahul C., “Machine Learning in Medicine”, *Circulation*, c. 132, sy 20, ss. 1920-1930, Kas. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.
[73] K.-H. Yu, A. L. Beam, ve I. S. Kohane, “Artificial intelligence in healthcare”, *Nature Biomedical Engineering*, c. 2, sy 10, Art. sy 10, Eki. 2018, doi: 10.1038/s41551-018-0305-z.
[74] I. H. Witten, E. Frank, M. A. Hall, ve C. J. Pal, *Data mining: practical machine learning tools and techniques with Java implementations*, Fourth Edition. Morgan Kaufmann, 2017.
[75] M. Gevrey, I. Dimopoulos, ve S. Lek, “Review and comparison of methods to study the contribution of variables in artificial neural network models”, *Ecological modelling*, c. 160, sy 3, ss. 249-264, 2003, doi: 10.1016/S0304-3800(02)00257-0.
[76] H. A. Karaboğa, “Dalgacık Dönüşümü Kullanılarak Hisse Senedi Fiyat Tahmini Üzerine Bir Uygulama”, Yüksek Lisans Tezi, Fen Bilimleri Enstitüsü, Yıldız Teknik Üniversitesi, İstanbul, Türkiye, 2015.

[77] S. Dreiseitl ve L. Ohno-Machado, "Logistic regression and artificial neural network classification models: a methodology review", *Journal of biomedical informatics*, c. 35, sy 5-6, ss. 352-359, 2002.
[78] D. W. Hosmer Jr, S. Lemeshow, ve R. X. Sturdivant, *Applied logistic regression*, c. 398. John Wiley & Sons, 2013.
[79] B. G. Tabachnick ve L. S. Fidell, *Using Multivariate Statistics*, 6th edition. Boston: Pearson, 2012.
[80] D. G. Kleinbaum ve M. Klein, *Logistic Regression: A Self-Learning Text*. New York, NY: Springer New York, 2010. doi: 10.1007/978-1-4419-1742-3.
[81] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, ve B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models", *Journal of Clinical Epidemiology*, c. 110, ss. 12-22, Haz. 2019, doi: 10.1016/j.jclinepi.2019.02.004.
[82] G. John ve P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers. In proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence", San Francisco, CA, United States, Ağu. 1995, ss. 338-345. doi: 10.5555/2074158.2074196.
[83] W. Wei, S. Visweswaran, ve G. F. Cooper, "The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data", *J Am Med Inform Assoc*, c. 18, sy 4, ss. 370-375, Tem. 2011, doi: 10.1136/amiajnl-2011-000101.
[84] W. Jiang *vd.*, "A naive Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system", *International journal of cancer*, c. 142, sy 2, ss. 357-368, 2018.
[85] T. M. Mitchell, *Machine learning*. New York, NY: McGraw-Hill, 1997.
[86] M. Kuhn ve K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-6849-3.
[87] H. Zhang, "The Optimality of Naïve Bayes", içinde *In FLAIRS2004 conference*, Miami Beach, Florida, USA, 2004, ss. 562-567.
[88] W. Buntine, "Theory Refinement on Bayesian Networks", içinde *Uncertainty Proceedings 1991*, Elsevier, 1991, ss. 52-60. doi: 10.1016/B978-1-55860-203-8.50010-3.
[89] P. Domingos ve M. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss", *Machine Learning*, c. 29, sy 2, ss. 103-130, Kas. 1997, doi: 10.1023/A:1007413511361.
[90] L. Rokach ve O. Maimon, "Decision trees", içinde *Data mining and knowledge discovery handbook*, Springer, 2005, ss. 165-192. Erişim: Ara. 06, 2020. [Çevrimiçi]. Erişim adresi: https://link.springer.com/chapter/10.1007/0-387-25465-X_9
[91] G. Kaur ve A. Chhabra, "Improved J48 classification algorithm for the prediction of diabetes", *International Journal of Computer Applications*, c. 98, sy 22, ss. 13-17, 2014.

[92] J. R. Quinlan, “Improved use of continuous attributes in C4. 5”, *Journal of Artificial Intelligence Research*, c. 4, sy 1, ss. 77-90, 1996, doi: 10.1613/jair.279.
[93] A. K. Yadav ve S. Chandel, “Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model”, *Renewable Energy*, c. 75, ss. 675-693, 2015, doi: 10.1016/j.renene.2014.10.046.
[94] M. F. bin Othman ve T. M. S. Yau, “Comparison of different classification techniques using WEKA for breast cancer”, içinde *3rd Kuala Lumpur International Conference on Biomedical Engineering 2006*, 2007, ss. 520-523.
[95] J. Han, M. Kamber, ve P. Jian, *Data Mining: Concepts and Techniques*. Haryana, India; Burlington, MA: Elsevier, 2012.
[96] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufmann Publishers, 1993.
[97] B. Schölkopf ve A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, Mass: MIT Press, 2009.
[98] A. Shmilovici, “Support Vector Machines”, içinde *Data Mining and Knowledge Discovery Handbook*, O. Maimon ve L. Rokach, Ed. Boston, MA: Springer US, 2009, ss. 231-247. doi: 10.1007/978-0-387-09823-4_12.
[99] W.-Z. Lu ve D. Wang, “Learning machines: Rationale and application in ground-level ozone prediction”, *Applied Soft Computing*, c. 24, ss. 135-141, Kas. 2014, doi: 10.1016/j.asoc.2014.07.008.
[100] N. Cristianini, J. Shawe-Taylor, ve others, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
[101] M. Mohri, A. Rostamizadeh, ve A. Talwalkar, *Foundations of machine learning*, Second edition. Cambridge, Massachusetts: The MIT Press, 2018.
[102] C. Cortes ve V. Vapnik, “Support-vector networks”, *Mach Learn*, c. 20, sy 3, ss. 273-297, Eyl. 1995, doi: 10.1007/BF00994018.
[103] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*, Second edition. Sebastopol, CA: O’Reilly Media, Inc, 2019.
[104] D. Pinto, M. Tovar, D. Vilarino, B. Beltrán, H. Jiménez-Salazar, ve B. Campos, “BUAP: Performance of K-Star at the INEX’09 Clustering Task”, içinde *International Workshop of the Initiative for the Evaluation of XML Retrieval*, 2009, ss. 434-440.
[105] J. G. Cleary ve L. E. Trigg, “K*: An Instance-based Learner Using an Entropic Distance Measure”, içinde *Machine Learning Proceedings 1995*, A. Prieditis ve S. Russell, Ed. San Francisco (CA): Morgan Kaufmann, 1995, ss. 108-114. doi: 10.1016/B978-1-55860-377-6.50022-0.

[106] S. Painuli, M. Elangovan, ve V. Sugumaran, "Tool condition monitoring using K-star algorithm", *Expert Systems with Applications*, c. 41, sy 6, ss. 2638-2643, 2014, doi: 10.1016/j.eswa.2013.11.005.
[107] W. Wiharto, H. Kusnanto, ve H. Herianto, "Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm", *Healthc Inform Res.*, c. 22, sy 1, ss. 30-38, 2016, doi: 10.4258/hir.2016.22.1.30.
[108] C. K. Madhusudana, H. Kumar, ve S. Narendranath, "Condition monitoring of face milling tool using K-star algorithm and histogram features of vibration signal", *Engineering Science and Technology, an International Journal*, c. 19, sy 3, ss. 1543-1551, Eyl. 2016, doi: 10.1016/j.jestch.2016.05.009.
[109] S. Piramuthu ve R. T. Sikora, "Iterative feature construction for improving inductive learning algorithms", *Expert Systems with Applications*, c. 36, sy 2, ss. 3401-3406, Mar. 2009, doi: 10.1016/j.eswa.2008.02.010.
[110] S. Zhang, D. Cheng, Z. Deng, M. Zong, ve X. Deng, "A novel kNN algorithm with data-driven k parameter computation", *Pattern Recognition Letters*, c. 109, ss. 44-54, 2018, doi: 10.1016/j.patrec.2017.09.036.
[111] S. Russell ve P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th edition. Hoboken: Pearson, 2020.
[112] D. Ballabio, F. Grisoni, ve R. Todeschini, "Multivariate comparison of classification performance measures", *Chemometrics and Intelligent Laboratory Systems*, c. 174, ss. 33-44, 2018, doi: 10.1016/j.chemolab.2017.12.004.
[113] A. Tharwat, "Classification assessment methods", *Applied Computing and Informatics*, 2020, doi: 10.1016/j.aci.2018.08.003.
[114] C. Ferri, J. Hernández-Orallo, ve R. Modroiu, "An experimental comparison of performance measures for classification", *Pattern Recognition Letters*, c. 30, sy 1, ss. 27-38, Oca. 2009, doi: 10.1016/j.patrec.2008.08.010.
[115] L. I. Kuncheva, Á. Arnaiz-González, J.-F. Díez-Pastor, ve I. A. D. Gunn, "Instance selection improves geometric mean accuracy: a study on imbalanced data classification", *Prog Artif Intell*, c. 8, sy 2, ss. 215-228, Haz. 2019, doi: 10.1007/s13748-019-00172-4.
[116] S. Sakr *vd.*, "Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project", *BMC Med Inform Decis Mak*, c. 17, sy 1, s. 174, Ara. 2017, doi: 10.1186/s12911-017-0566-6.
[117] J. Akosa, "Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data", içinde *Proceedings of the SAS Global Forum*, Nis. 2017, ss. 2-5.
[118] M. Hossin ve M. N. Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations", *IJDKP*, c. 5, sy 2, ss. 01-11, Mar. 2015, doi: 10.5121/ijdkp.2015.5201.

[119] T. Fawcett, “An introduction to ROC analysis”, *Pattern Recognition Letters*, c. 27, sy 8, ss. 861-874, 2006, doi: 10.1016/j.patrec.2005.10.010.
[120] L. Cuadros-Rodríguez, E. Pérez-Castaño, ve C. Ruiz-Samblás, “Quality performance metrics in multivariate classification methods for qualitative analysis”, *TrAC Trends in Analytical Chemistry*, c. 80, ss. 612-624, Haz. 2016, doi: 10.1016/j.trac.2016.04.021.
[121] T. Hastie, R. Tibshirani, ve J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media, 2009.
[122] L. P. Rowland ve N. A. Shneider, “Amyotrophic lateral sclerosis”, *New England Journal of Medicine*, c. 344, sy 22, ss. 1688-1700, 2001, doi: 10.1056/NEJM200105313442207.
[123] O. Hardiman, L. H. Van Den Berg, ve M. C. Kiernan, “Clinical diagnosis and management of amyotrophic lateral sclerosis”, *Nature reviews neurology*, c. 7, sy 11, ss. 639-649, 2011, doi: 10.1038/nrneurol.2011.153.
[124] B. Swinnen ve W. Robberecht, “The phenotypic variability of amyotrophic lateral sclerosis”, *Nature Reviews Neurology*, c. 10, sy 11, s. 661, 2014, doi: 10.1038/nrneurol.2014.184.
[125] A. Al-Chalabi ve O. Hardiman, “The epidemiology of ALS: a conspiracy of genes, environment and time”, *Nature Reviews Neurology*, c. 9, sy 11, s. 617, 2013, doi: 10.1038/nrneurol.2013.203.
[126] T. Filippini *vd.*, “Clinical and Lifestyle Factors and Risk of Amyotrophic Lateral Sclerosis: A Population-Based Case-Control Study”, *International Journal of Environmental Research and Public Health*, c. 17, sy 3, Art. sy 3, Oca. 2020, doi: 10.3390/ijerph17030857.
[127] D. Mendonça *vd.*, “Neuroproteomics: an insight into ALS”, *Neurological Research*, c. 34, sy 10, ss. 937-943, 2012, doi: 10.1179/1743132812Y.0000000092.
[128] Z. R. Manjaly *vd.*, “The sex ratio in amyotrophic lateral sclerosis: a population based study”, *Amyotrophic Lateral Sclerosis*, c. 11, sy 5, ss. 439-442, 2010, doi: 10.3109/17482961003610853.
[129] J. A. Pape ve J. H. Grose, “The effects of diet and sex in amyotrophic lateral sclerosis”, *Revue Neurologique*, c. 176, sy 5, ss. 301-315, May. 2020, doi: 10.1016/j.neurol.2019.09.008.
[130] E. Longinetti ve F. Fang, “Epidemiology of amyotrophic lateral sclerosis: an update of recent literature”, *Current opinion in neurology*, c. 32, sy 5, s. 771, 2019, doi: 10.1097/WCO.0000000000000730.
[131] L. Le Gall, E. Anakor, O. Connolly, U. G. Vijayakumar, W. J. Duddy, ve S. Duguez, “Molecular and Cellular Mechanisms Affected in ALS”, *Journal of Personalized Medicine*, c. 10, sy 3, Art. sy 3, Eyl. 2020, doi: 10.3390/jpm10030101.
[132] B. R. Brooks, R. G. Miller, M. Swash, ve T. L. Munsat, “El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis”, *Amyotrophic lateral sclerosis and other motor neuron*

disorders, c. 1, sy 5, ss. 293-299, 2000, doi: 10.1080/146608200300079536.
[133] L. BayesFusion, "GeNIe modeler user manual", *BayesFusion, LLC, Pittsburgh, PA*, 2017.
[134] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation", *J. Mach. Learn. Technol*, c. 2, sy 1, ss. 37-63, 2011.
[135] A. J. Viera, J. M. Garrett, ve others, "Understanding interobserver agreement: the kappa statistic", *Fam med*, c. 37, sy 5, ss. 360-363, 2005.
[136] J.-H. Lin ve P. J. Haug, "Exploiting missing clinical data in Bayesian network modeling for predicting medical problems", *Journal of Biomedical Informatics*, c. 41, sy 1, ss. 1-14, Şub. 2008, doi: 10.1016/j.jbi.2007.06.001.
[137] R. Chen ve E. H. Herskovits, "Clinical diagnosis based on bayesian classification of functional magnetic-resonance data", *Neuroinformatics*, c. 5, sy 3, ss. 178-188, 2007, doi: 10.1007/s12021-007-0007-2.
[138] Y. Luo *vd.*, "Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis", *Radiotherapy and Oncology*, c. 123, sy 1, ss. 85-92, Nis. 2017, doi: 10.1016/j.radonc.2017.02.004.
[139] F. Nojavan A., S. S. Qian, ve C. A. Stow, "Comparative analysis of discretization methods in Bayesian networks", *Environmental Modelling & Software*, c. 87, ss. 64-71, Oca. 2017, doi: 10.1016/j.envsoft.2016.10.007.
[140] Y. Yang ve G. I. Webb, "A Comparative Study of Discretization Methods for Naive-Bayes Classifiers", içinde *In Proceedings of PKAW 2002: The 2002 Pacific Rim Knowledge Acquisition Workshop*, 2002, ss. 159-173.
[141] V. Rodríguez-López ve R. Cruz-Barbosa, "Improving Bayesian Networks Breast Mass Diagnosis by Using Clinical Data", içinde *Pattern Recognition*, Cham, 2015, ss. 292-301. doi: 10.1007/978-3-319-19264-2_28.
[142] S. Khanna, D. Domingo-Fernández, A. Iyappan, M. A. Emon, M. Hofmann-Apitius, ve H. Fröhlich, "Using Multi-Scale Genetic, Neuroimaging and Clinical Data for Predicting Alzheimer's Disease and Reconstruction of Relevant Biological Mechanisms", *Scientific Reports*, c. 8, sy 1, Art. sy 1, Tem. 2018, doi: 10.1038/s41598-018-29433-3.
[143] A. Palmieri *vd.*, "Female gender doubles executive dysfunction risk in ALS: a case-control study in 165 patients", *J Neurol Neurosurg Psychiatry*, c. 86, sy 5, ss. 574-579, May. 2015, doi: 10.1136/jnnp-2014-307654.
[144] F. Trojsi, G. D'Alvano, S. Bonavita, ve G. Tedeschi, "Genetics and Sex in the Pathogenesis of Amyotrophic Lateral Sclerosis (ALS): Is There a Link?", *International Journal of Molecular Sciences*, c. 21, sy 10, Art. sy 10, Oca. 2020, doi: 10.3390/ijms21103647.
[145] A. Chiò *vd.*, "ALS phenotype is influenced by age, sex, and genetics: A population-based study", *Neurology*, c. 94, sy 8, ss. e802-e810, Şub. 2020, doi: 10.1212/WNL.0000000000008869.

[146] C. Ingre, P. M. Roos, F. Piehl, F. Kamel, ve F. Fang, "Risk factors for amyotrophic lateral sclerosis", *Clin Epidemiol*, c. 7, ss. 181-193, Şub. 2015, doi: 10.2147/CLEP.S37505.
[147] F. Trojsi *vd.*, "Comparative Analysis of C9orf72 and Sporadic Disease in a Large Multicenter ALS Population: The Effect of Male Sex on Survival of C9orf72 Positive Patients", *Front. Neurosci.*, c. 13, 2019, doi: 10.3389/fnins.2019.00485.
[148] J. Rooney *vd.*, "C9orf72 expansion differentially affects males with spinal onset amyotrophic lateral sclerosis", *J Neurol Neurosurg Psychiatry*, c. 88, sy 4, s. 281.1-281, Nis. 2017, doi: 10.1136/jnnp-2016-314093.
[149] N. Atsuta *vd.*, "Age at onset influences on wide-ranged clinical features of sporadic amyotrophic lateral sclerosis", *Journal of the Neurological Sciences*, c. 276, sy 1, ss. 163-169, Oca. 2009, doi: 10.1016/j.jns.2008.09.024.
[150] A. Chiò, A. Calvo, C. Moglia, L. Mazzini, G. Mora, ve P. study Group*, "Phenotypic heterogeneity of amyotrophic lateral sclerosis: a population based study", *Journal of Neurology, Neurosurgery & Psychiatry*, c. 82, sy 7, ss. 740-746, Tem. 2011, doi: 10.1136/jnnp.2010.235952.
[151] O. Connolly, L. Le Gall, G. McCluskey, C. G. Donaghy, W. J. Duddy, ve S. Duguez, "A Systematic Review of Genotype–Phenotype Correlation across Cohorts Having Causal Mutations of Different Genes in ALS", *Journal of Personalized Medicine*, c. 10, sy 3, Art. sy 3, Eyl. 2020, doi: 10.3390/jpm10030058.
[152] M. A. van Es *vd.*, "Amyotrophic lateral sclerosis", *The Lancet*, c. 390, sy 10107, ss. 2084-2098, Kas. 2017, doi: 10.1016/S0140-6736(17)31287-4.
[153] H. P. Nguyen, C. Van Broeckhoven, ve J. van der Zee, "ALS Genes in the Genomic Era and their Implications for FTD", *Trends in Genetics*, c. 34, sy 6, ss. 404-423, Haz. 2018, doi: 10.1016/j.tig.2018.03.001.
[154] P. M. Andersen ve A. Al-Chalabi, "Clinical genetics of amyotrophic lateral sclerosis: what do we really know?", *Nat Rev Neurol*, c. 7, sy 11, ss. 603-615, Kas. 2011, doi: 10.1038/nrneurol.2011.150.
[155] M. Fratello *vd.*, "Multi-View Ensemble Classification of Brain Connectivity Images for Neurodegeneration Type Discrimination", *Neuroinform*, c. 15, sy 2, ss. 199-213, Nis. 2017, doi: 10.1007/s12021-017-9324-2.

A
EK BİLGİLER

Çalışmada kullanılan kan örneklerinin toplanmasında emeği geçen Prof. Dr. Atilla Halil Idrisoğlu'na teşekkür ederiz. Bu çalışmada kullanılan veriler için etik onaylar İstanbul Üniversitesi Tıp Fakültesi'nden alınmıştır (Dosya numarası: 2016/665, Toplantı sayısı: 10, Tarih: 24 Mayıs 2016). Araştırmaya katılan tüm gönüllüler onaylarını beyan etmiş ve ilgili etik belgeleri imzalamıştır.

TEZDEN ÜRETİLMİŞ YAYINLAR

İletişim Bilgisi: h.aykut.karaboga@amasya.edu.tr

Makaleler

[1] **H. A. Karaboga**, A. Gunel, S. V. Korkut, I. Demir, and R. Celik, "Bayesian network as a decision tool for predicting ALS disease," *Brain Sci.*, vol. 11, no. 2, p. 150, 2021.

[2] E. Sener, **H. A. Karaboga**, and I. Demir, "Bayesian network model of Turkish Financial Market from year-to-September 30th of 2016," *Sigma: Journal of Engineering & Natural Sciences / Mühendislik ve Fen Bilimleri Dergisi*, vol. 37, no. 4, pp. 1493–1507, 2019.