

**T.C.
ERCIYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**YAPAY ZEKA TABANLI YÖNTEMLER KULLANILARAK
MİKROARRAY GEN VERİLERİNİN SINIFLANDIRILMASI**

**Hazırlayan
Mustafa Turan ARSLAN**

**Danışman
Prof. Dr. Adem KALINLI**

Yüksek Lisans Tezi

**Haziran 2016
KAYSERİ**

**T.C.
ERCIYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**YAPAY ZEKA TABANLI YÖNTEMLER KULLANILARAK
MİKROARRAY GEN VERİLERİNİN SINIFLANDIRILMASI
(Yüksek Lisans Tezi)**

**Hazırlayan
Mustafa Turan ARSLAN**

**Danışman
Prof. Dr. Adem KALINLI**

Bu çalışma; Erciyes Üniversitesi Bilimsel Araştırma Projeleri Birimi tarafından FYL-2015-6095 kodlu proje ile desteklenmiştir.

**Haziran 2016
KAYSERİ**

BİLİMSEL ETİĞE UYGUNLUK

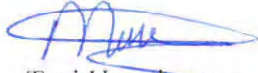
Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi belirtirim.

Mustafa Turan ARSLAN



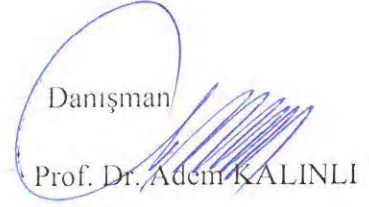
YÖNERGEYE UYGUNLUK

Yapay Zeka Tabanlı Yöntemler Kullanılarak Mikroarray Gen Verilerinin Sınıflandırılması adlı Yüksek Lisans tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi 'ne uygun olarak hazırlanmıştır.



Tezi Hazırlayan

Mustafa Turan ARSLAN



Danışman
Prof. Dr. Adem KALINLI



Bilgisayar Mühendisliği ABD Başkanı
Prof. Dr. Derviş KARABOĞA

Prof. Dr. Adem KALINLI danışmanlığında **Mustafa Turan ARSLAN** tarafından hazırlanan “**Yapay Zeka Tabanlı Yöntemler Kullanılarak Mikroarray Gen Verilerinin Sınıflandırılması**” adlı bu çalışma, jürimiz tarafından Erciyes Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalında **Yüksek Lisans** tezi olarak kabul edilmiştir.

03/06/2016

JÜRİ:

Başkan : Prof. Dr. Adem KALINLI

Üye : Doç. Dr. Mehmet KARAKÖSE

Üye : Doç. Dr. Alper BAŞTÜRK

ONAY:

Bu tezin kabulü, Enstitü Yönetim Kurulunun 14/06/2016 tarih ve 2016/27-24 sayılı kararı ile onaylanmıştır.

Prof. Dr. Mehmet AKKURT
Enstitü Müdürü

ÖNSÖZ / TEŞEKKÜR

Çalışmalarım boyunca farklı bakış açıları ve bilimsel katkılarıyla beni aydınlatan, yakın ilgi ve yardımlarını esirgemeyen ve bu günlere gelmemde en büyük katkı sahibi sayın hocam Prof. Dr. Adem KALINLI 'ya teşekkürü bir borç bilirim.

Bu tez çalışmasına maddi destek veren Erciyes Üniversitesi Bilimsel Araştırma Projeleri Birimi'ne (Proje No: FYL-2015-6095) teşekkür ederim.

Çalışmalarım sırasında sabır göstererek beni daima destekleyip yanımda olan Canım Aileme en içten teşekkürlerimi sunarım.

Mustafa Turan ARSLAN

Kayseri, Haziran 2016

YAPAY ZEKA TABANLI YÖNTEMLER KULLANILARAK MİKROARRAY GEN VERİLERİNİN SINIFLANDIRILMASI

Mustafa Turan ARSLAN

Erciyes Üniversitesi, Fen Bilimleri Enstitüsü

Yüksek Lisans Tezi, Haziran 2016

Danışman: Prof. Dr. Adem KALINLI

ÖZET

Teknolojinin hızla ilerlemesiyle mikroarray çalışmalarına olan ilgi de günden güne artmaktadır. DNA mikroarray çalışmaları, moleküler biyoloji ve tıp alanlarında kullanılan çok kapsamlı bir teknolojidir. DNA mikroarray veri analizi; kanser gibi genlerle ilişkili hastalıkların belirlenmesinde önemli rol oynamaktadır. Hastalık türüne ilişkin ilgili genler belirlenerek, herhangi bir bireyin hasta ya da sağlıklı olduğu yüksek olasılıklarda hesaplanabilir. Bunun için mikroarray verilerinde yüksek performanslı sınıflama yöntemleri oldukça önemlidir.

DNA mikroarray teknolojisindeki hızlı gelişmeye bağlı olarak mikroarray'leri sınıflandırmak için birçok sınıflandırma metodu kullanılmaya başlanmıştır. Destek Vektör Makineleri, Naive Bayes, k-En Yakın Komşu, Karar Ağacı, Torbalama ve Rastgele Orman gibi istatistiksel yöntemler yaygın bir şekilde kullanılmaktadır. Fakat bu yöntemler, çok boyutlu problemler olan mikroarray verilerini sınıflandırmada bazen başarısız olabilmektedir. Bu yüzden mikroarray problemleri üzerinde başarılı sonuçlar verebilen yapay zekâ tabanlı yöntemler ile ilgili son yıllarda araştırmalar yapılmaktadır. Bu çalışmada bizde istatistiksel yöntemlerin dışında yapay zekâ tabanlı yöntemler olan Tek Katmanlı Algılayıcılar, Çok Katmanlı Algılayıcılar, Radyal Tabanlı Yapay Sinir Ağları, Karınca Koloni Algoritması ve Bulanık k-NN yöntemleri kullanılmıştır.

Sınıflandırma yöntemlerinin doğrulukları 10-katlı çapraz doğrulama yöntemi ile test edildi. Sonuçlardan elde edilen bilgilere göre, genel olarak yapay zekâ tabanlı yöntemler, istatistiksel yöntemlere göre daha başarılı sonuçlar vermiştir.

Anahtar Kelimeler: DNA mikroarray, Sınıflandırma, 10-katlı Çapraz Doğrulama, İstatistiksel Yöntemler, Yapay Zekâ tabanlı Yöntemler

CLASSIFICATION OF MICROARRAY GENE DATA USING ARTIFICIAL INTELLIGENCE BASED METHODS

Mustafa Turan ARSLAN

Erciyes University, Graduate School of Natural and Applied Sciences

M.Sc. Thesis, June 2016

Supervisor: Prof. Dr. Adem KALINLI

ABSTRACT

The interest in working with the rapid advancement of microarray technology is increasing day by day. DNA microarray studies is a comprehensive technology used in molecular biology and medicine. DNA microarray data analysis, the identification of genes associated with diseases such as cancer, plays an important role. It can be calculated in high probability that any individual is ill or healthy by identifying disease-related genes. Therefore, high performance classification methods are very important for microarray datasets.

Depending on the rapid development of DNA microarray technology, many classification methods has been used to classify microarray data. SVM, Naive Bayes, k-NN, Decision Trees, Bagging, Random Forest method are used widely. However, these methods sometimes cannot be successful on microarray cancer datasets. Therefore, it has been studied on artificial intelligence-based methods which can be successful on microarray datasets in recent years. In this study, we have used Single Layer Perceptron, Multi-Layer Perceptron, Radial Basis Function Network, Ant Colony Optimization Algorithm and Fuzzy k-Nearest Neighbor approaches that artificial intelligence based classification methods excluding conventional methods.

In addition; we have tested accuracy of the classification method using 10-fold cross validation. Artificial Intelligence-based methods have given better results than the statistical methods.

Keywords: DNA microarray, Classification, 10-fold Cross Validation, Statistical Methods, Artificial Intelligence-based Methods

İÇİNDEKİLER

YAPAY ZEKA TABANLI YÖNTEMLER KULLANILARAK MİKROARRAY GEN VERİLERİNİN SINIFLANDIRILMASI	
BİLİMSEL ETİĞE UYGUNLUK SAYFASI	i
YÖNERGEYE UYGUNLUK SAYFASI.....	ii
KABUL VE ONAY SAYFASI	iii
TEŞEKKÜR.....	iv
İÇİNDEKİLER	vii
KISALTMALAR VE SİMGELER.....	xi
TABLolar LİSTESİ.....	xii
ŞEKİLLER LİSTESİ	xiv
 GİRİŞ	 1

1. BÖLÜM

GENEL BİLGİLER

1.1. Veri Madenciliğinin Tarihçesi	4
1.2. Veri Madenciliği	4
1.3. Veri Madenciliği ve Disiplinler Arası ilişki.....	6
1.4. Veri Madenciliğinin Uygulama Alanları.....	7
1.5. Veri Madenciliği Süreci.....	8
1.5.1. Problemin Belirlenmesi ve Verilerin Anlaşılması.....	9
1.5.2. Verilerin Toplanması	10
1.5.3. Verilerin Hazırlanması.....	10

1.5.3.1. Verilerin temizlenmesi.....	11
1.5.3.2. Verilerin kaynaştırılması ve dönüştürülmesi.....	11
1.5.3.3. Verilerin kesikli hale getirilmesi ve kavramsal hiyerarşi oluşturma.....	11
1.5.4. Modelin Kurulması.....	12
1.5.5. Modelin Yorumlanması	12
1.6. Veri Madenciliği Modelleri.....	12
1.7. Veri Madenciliğinde Öğrenme Yöntemleri.....	12
1.7.1. Denetimli Öğrenme (Supervised Learning)	12
1.7.2. Denetimsiz Öğrenme (Unsupervised Learning).....	13
1.7.3. Destekli Öğrenme (Reinforcement Learning).....	13
1.8. Veri Madenciliği Teknikleri.....	13
1.8.1. Sınıflandırma Problemi.....	14
1.8.2. Kümeleme Problemi	14
1.9. Sınıflandırma Modeli	15
1.9.1. Karar Ağaçları.....	15
1.9.2. Yapay Sinir Ağları.....	17
1.9.3. Evrimsel Algoritmalar.....	18
1.9.4. k-En Yakın Komşu(k-NN)	19
1.9.5. Bayes Sınıflandırıcılar	19
1.9.6. Sürü Zekâsı	20
1.9.7. Durum Tabanlı Nedenleme.....	20
1.9.8. Kaba Küme Yaklaşımı	20
1.9.9. Bulanık Küme Yaklaşımı.....	21
1.10. Sınıflandırma Veri Madenciliği.....	21
1.10.1. Sınıflandırma Kuralları	21
1.10.2. Sınıflandırma Kural Çıkarımı.....	22

1.10.3. Sınıflandırma Metotlarının Değerlendirilmesinde Temel Ölçütler	24
1.10.4. Sınıflandırma Kurallarının Mikro ve Makro Değerlendirilmesi.....	26
1.10.5. Kural Gösterimi.....	30
1.10.6. Uygunluk Fonksiyonu	31
1.11. Gen Ekspresyonu (İfade)	31
1.12. Mikroarray Teknolojisi	32
1.12.1. Mikroarraylerin Üretilmesi ve Çalışma Mantığı.....	34
1.12.2. Mikroarraylerin Kullanım Alanları	38
1.12.3. Mikroarray Teknolojisinin Avantajları	38

2. BÖLÜM

YÖNTEM

2.1. Weka Programı	39
2.2. Açık Kaynak Kodlu Ant-Miner Programı.....	43
2.3. Öznitelik Seçimi.....	44
2.3.1. KTÖS öznitelik seçimi.....	44
2.4. Arff Dosya Formatı.....	45
2.5. Sınıflandırma Yöntemleri.....	47
2.5.1. Naive Bayes.....	47
2.5.2. Destek Vektör Makineleri (DVM).....	47
2.5.3. Bagging.....	49
2.5.4. One-R.....	49
2.5.5. J48 Karar Ağacı.....	50
2.5.6. Bulanık k- En Yakın Komşu (Bulanık k-NN)	50
2.5.7. Tek Katmanlı Algılayıcılar (TKA)	52
2.5.8. Çok Katmanlı Algılayıcılar (ÇKA)	54

2.5.9. Radyal Tabanlı Yapay Sinir Ağları (RTYSA)	55
2.5.10. Karınca Koloni Algoritması.....	57
2.6. Mikroarray Gen Ekspresyon (İfade) Verileri	59
2.7. k-Katlamalı Çapraz Doğrulama Yöntemi Ve Başarı Ölçümü.....	60
2.8. Mikroarray Gen Ekspresyon (İfade) Veri Setlerine Sınıflandırma Yöntemlerinin Uygulanması.....	60

3. BÖLÜM

BULGULAR

3.1. Sınıflandırma Yöntemlerinin Performanslarının Karşılaştırılması İle İlgili Yapılmış Benzer Çalışmalar	63
3.2. Parametreler, Uygulama ve Sonuçlar	67

4. BÖLÜM

TARTIŞMA, SONUÇ VE ÖNERİLER

4.1. Tartışma, Sonuç ve Öneriler	80
KAYNAKÇA	100
ÖZGEÇMİŞ.....	109

KISALTMALAR VE SİMGELER

<u>Sembol</u>	<u>Anlamı</u>	<u>Birimi</u>
VM	Veri Madenciliği	-
YSA	Yapay Sinir Ağları	-
DVM	Destek Vektör Makineleri	-
GA	Genetik Algoritma	-
GP	Genetik Programlama	-
EP	Evrimsel Programlama	-
ES	Evrimsel Strateji	-
SZ	Sürü Zekâsı	-
DTN	Durum Tabanlı Nedenleme	-
DNA	Deoksiribo nükleik asit	-
RNA	Ribo nükleik asit	-
tRNA	transfer RNA	-
mRNA	mesajcı RNA	-
KK	Kaba Küme Yaklaşımı	-
BK	Bulanık Küme Yaklaşımı	-
kNN	k-En Yakın Komşu	-
KTÖS	Korelasyon Tabanlı Özellik Seçimi	-
One-R	Bir Kural	-
TKA	Tek Katmanlı Algılayıcılar	-
ÇKA	Çok Katmanlı Algılayıcılar	-
RTYSA	Radyal Tabanlı Yapay Sinir Ağları	-
Bulanık k-NN	Bulanık Tabanlı k-En yakın Komşu	-
SBO	Sistemin Sınıflandırma Başarı Oranı	-
dsvs	Doğru Sınıflandırılan Veri Sayısı	-
tvsv	Toplam Veri Sayısı	-
RBFNetwork	Radyal Tabanlı Yapay Sinir Ağları	-

TABLolar LİSTESİ

Tablo 1.1.	2x2 Durumsallık Tablosu [53].....	27
Tablo 2.1.	Çalışmada Kullanılan Mikroarray Kanseri Veri Setleri.....	62
Tablo 3.1.	Karaciğer Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri	69
Tablo 3.2.	Karaciğer Kanseri Yöntemlerinin Optimal Kontrol Parametreleri	70
Tablo 3.3.	RNA Doku Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri	70
Tablo 3.4.	RNA Doku Kanseri Yöntemlerinin Optimal Kontrol Parametreleri.....	71
Tablo 3.5.	Prostat Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri	71
Tablo 3.6.	Prostat Kanseri Yöntemlerinin Optimal Kontrol Parametreleri.....	72
Tablo 3.7.	Merkezi Sinir Sistemi Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri.....	72
Tablo 3.8.	Merkezi Sinir Sistemi Kanseri Yöntemlerinin Optimal Kontrol Parametreleri	73
Tablo 3.9.	Beyin Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri	73
Tablo 3.10.	Beyin Kanseri Yöntemlerinin Optimal Kontrol Parametreleri	74
Tablo 3.11.	Rahim Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri	74
Tablo 3.12.	Rahim Kanseri Yöntemlerinin Optimal Kontrol Parametreleri	75
Tablo 3.13.	Göğüs Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri	75
Tablo 3.14.	Göğüs Kanseri Yöntemlerinin Optimal Kontrol Parametreleri	76
Tablo 3.15.	Mesane Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri	76
Tablo 3.16.	Mesane Kanseri Yöntemlerinin Optimal Kontrol Parametreleri	77
Tablo 3.17.	Farklı Doku Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri	77
Tablo 3.18.	Farklı Doku Kanseri Yöntemlerinin Optimal Kontrol Parametreleri	78

Tablo 3.19. Kolon Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri.	78
Tablo 3.20. Kolon Kanseri Yöntemlerinin Optimal Kontrol Parametreleri.....	79
Tablo 4.1. Mikroarray Kanser Verilerinin Sınıflandırılmasında Literatürde Yaygın Kullanılan Kontrol Parametre Değerleriyle İstatistiksel Yöntemlerin Performansları.	81
Tablo 4.2. Mikroarray Kanser Verilerinin Sınıflandırılmasında Literatürde Yaygın Kullanılan Kontrol Parametre Değerleriyle Yapay Zeka Tabanlı Yöntemlerin Performansları.....	82
Tablo 4.3. Mikroarray Kanser Verilerinin Sınıflandırılmasında Optimal Kontrol Parametre Değerleriyle İstatistiksel Yöntemlerin Performansları.	83
Tablo 4.4. Mikroarray Kanser Verilerinin Sınıflandırılmasında Optimal Kontrol Parametre Değerleriyle Yapay Zeka Tabanlı Yöntemlerin Performansları.	83

ŞEKİLLER LİSTESİ

Şekil 1.1.	İlişkiler [10]	5
Şekil 1.2.	Veri madenciliği ile disiplinler arası ilişki [13]	6
Şekil 1.3.	Bilgi keşfi süreci	8
Şekil 1.4.	Veri madenciliği süreci adımları.....	9
Şekil 1.5.	Örnek bir karar ağacı yapısı	16
Şekil 1.6.	YSA yapısı	18
Şekil 1.7.	DNA mikroyarray işlem basamakları [63].....	34
Şekil 1.8.	DNA mikroyarray'in görünümü [63].....	35
Şekil 1.9.	Örneklerin hazırlanması, karıştırılması ve hibridizasyon [63].....	36
Şekil 1.10.	Etiketlerin uyarılması ve floresans [63].	36
Şekil 1.11.	Mikroçip yönteminin şematik gösterimi [63].	37
Şekil 2.1.	Weka kullanıcı ara yüzü.....	40
Şekil 2.2.	Explorer seçeneğindeki sekmeler	41
Şekil 2.3.	Weka "Classify" sekmesi.....	41
Şekil 2.4.	"Test Options" başlığı.....	42
Şekil 2.5.	"Select Attributes" sekmesi	43
Şekil 2.6.	cAnt-Miner arayüzü	43
Şekil 2.7.	KTÖS öznitelik seçimindeki elemanlar [75]	45
Şekil 2.8.	Arff dosya yapısı.....	46
Şekil 2.9.	İki boyutlu uzayda doğrusal ayrılabilen verilerin görünümü [81]	48
Şekil 2.10.	İki sınıfı birbirinden ayıran en uygun hiper düzlem [81].....	49
Şekil 2.11.	Üç girdi ve bir çıktılı perseptron modeli [86]	53
Şekil 2.12.	Çok katmanlı ileri beslemeli yapay sinir ağı modeli	54
Şekil 2.13.	Radyal tabanlı YSA yapısı [88]	55
Şekil 2.14.	Karınca koloni algoritmasının adımları [104].....	57
Şekil 2.15.	cAnt-Miner algoritması [104]	59
Şekil 3.1.	Arff dosya yapısı	68
Şekil 4.1.	Karaciğer kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	85
Şekil 4.2.	RNA doku kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	86

Şekil 4.3.	Prostat kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	87
Şekil 4.4.	Merkezi Sinir Sistemi kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	89
Şekil 4.5.	Beyin kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	90
Şekil 4.6.	Rahim kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	91
Şekil 4.7.	Göğüs kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	92
Şekil 4.8.	Mesane kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	94
Şekil 4.9.	Farklı doku kanserleri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	95
Şekil 4.10.	Kolon kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları	97
Şekil 4.11.	Mikroarray kanser veri setleri üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin ortalama performansları	98
Şekil 4.12.	Mikroarray kanser veri setleri üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin grup ortalama performansları	99

GİRİŞ

Eski çağlardan bu yana insanlık tarafından ilgilenilen bilim dalları temel bilimler olarak adlandırılmış ve karşılaşılan problemler bu bilim dalları arasında sınıflandırılarak çözümler aranmıştır. Zamanla temel bilim dallarının birkaçının birlikte çözüm bulacağı sorunlar ortaya çıkmış ve disiplinler arası çalışma önem kazanmıştır. Aynı disiplinler arası çalışılması gereken konuların artışıyla yeni bilim dalları oluşmuştur. Canlı organizmalarda gerçekleşen tüm yaşamsal ve ölüm sonrası faaliyetler doğal olarak biyoloji biliminin ilgi alanına girmektedir. Bu faaliyetler bazı farklı temel bilimlerin kural ve kaidelerine göre gerçekleşmektedir. Bu kural ve kaidelerin temel alındığı bilimleri kullanarak canlıların gizemlerini çözmeye çalışan iki yeni bilim dalı göze çarpmaktadır. Bunlar fizik yasaları ile canlılığı açıklayan biyofizik ve canlıları ilgilendiren temel kimyasal tepkimeleri ele alan biyokimyadır [1].

Bilim dünyasındaki gelişmeler sayesinde mikroarray alanındaki çalışmalara olan ilgi günden güne artmaktadır. DNA mikroarray çalışmaları, moleküler biyoloji ve tıp alanlarında kullanılan çok kapsamlı bir teknolojidir. DNA mikroarray veri analizi; kanser gibi genlerle ilişkili hastalıkların belirlenmesinde önemli rol oynamaktadır. Hastalık türüne ilişkin ilgili genler belirlenerek, herhangi bir bireyin hasta ya da sağlam olduğu yüksek olasılıklarda hesaplanabilir. Bunun için mikroarray verilerinde yüksek performanslı sınıflandırma yöntemleri oldukça önemlidir [2].

Sınıflandırma; veri tabanındaki gizli kalıpları ortaya çıkarmakta kullanılan bir yöntemdir. Sınıflama ile veri tabanı belli özelliklere göre küçük homojen gruplara ayrılır. Sınıflandırma, yeni gelen bir verinin hangi sınıfa ait olduğunu gösteren bir analiz tekniğidir ve bir öğrenme algoritmasına dayanmaktadır. Bu algoritmanın amacı; bir sınıflandırma modeli oluşturarak, hangi sınıfa ait olduğu bilinmeyen bir veri için sınıf belirlemektir. Burada iyi belirlenmiş değişkenler kilit rolü oynamaktadır. Literatürde

çeşitli sınıflandırma yöntemleri bulunmaktadır. Bunlardan bazıları istatistiksel sınıflandırma yöntemleri iken bazıları ise yapay zekâ tabanlı olan yapay sinir ağlarıdır [2].

Mikroarray gen ekspresyon verilerinin sınıflandırılmasında da bu istatistiksel sınıflandırma yöntemleri ve yapay sinir ağları kullanılmaktadır. Bu çalışmada ise son zamanlarda araştırmacılar tarafından üzerinde yeni çalışılmaya başlanan ve literatüre yeni girmeye başlayan optimizasyon algoritmaları ve hibrid algoritmalarla kullanılarak sınıflandırma yapılmıştır.

Son zamanlarda mikroarray gen ekspresyon verilerini sınıflandırmada sıkça kullanılan istatistiksel yöntemler; Destek Vektör Makineleri (DVM), Bayes Ağları, k-En Yakın Komşu (k-NN), Bagging ve One-R iken yapay zekâ tabanlı yöntemler ise Çok Katmanlı Algılayıcılar (ÇKA) ve Radyal Tabanlı Yapay Sinir Ağları (RTYSA)'dır. Bu çalışmada yeni yaklaşımlar olan optimizasyon algoritmalarından Karınca Koloni Algoritması (KKA) ve yine yapay zeka tabanlı, hibrid bir sınıflandırma yöntemi olan Bulanık k-En yakın Komşu (Bulanık k-NN) algoritmaları da kullanılarak mikroarray verileri sınıflandırılmıştır.

Literatürde yaygın olarak sınıflandırmada kullanılan yöntemlerin dışında bu yeni önerdiğimiz yaklaşımlarda kullanılarak mikroarray gen ekspresyon verilerini sınıflandırmada yeni yaklaşımlar getirilerek bu yeni yaklaşımlar ile sık kullanılan yöntemlerin başarımları birbirleriyle kıyaslanmıştır.

Tezin organizasyonu şu şekildedir; 1. bölümde Veri madenciliği ile ilgili teorik bilgiler verilmiş daha sonra bu tez çalışmasında kullanılacak verilerin mikroarray gen ekspresyon verileri olduğu üzerinde durulmuş ve bu veriler ile ilgili olarak teorik bilgi verilmiştir. Üzerinde çalışılacak olan mikroarray verilerinin normalizasyon işlemlerinden geçtikten sonra veri madenciliği yöntemlerinden olan sınıflandırma yöntemi ile işleneceği ortaya konulmuştur. Ayrıca sınıflandırma ile ilgili teorik bilgilere yer verilmekle birlikte hangi tür sınıflandırma yöntemlerinin bu çalışmada kullanılacağı açıklanmıştır.

Tezin 2. bölümünde çalışmada kullanılan sınıflandırma araçları Weka ve Ant-Miner yazılımlarına değinilmiş ve bu programlar hakkında bilgi verilmiştir. Daha sonra ise çalışmada kullanılacak olan mikroarray gen ekspresyon verileri yüksek boyutlu veriler

olduđu için bu verilerin boyutlarını azaltmak için, Korelasyon tabanlı öznitelik seçim yöntemi hakkında bilgi verilmiştir. Boyutları azaltılan mikroarray verilerini sınıflandırmada kullanılacak olan Naive Bayes, DVM, Bagging, One-R, J48 Karar Ağacı, Tek Katmanlı Algılayıcılar (TKA), Çok Katmanlı Algılayıcılar (ÇKA), Radyal Tabanlı Yapay Sinir Ağları (RTYSA), Karınca Koloni Algoritması (KKA) ve Bulanık k-En Yakın Komşu (Bulanık k-NN) yöntemleri hakkında bilgi verilmiş olup böylece çözüm stratejisine temel oluşturulmuştur. Daha sonra uygulanacak modeller ve probleme uygun düzenlemeler üzerinde durulmuştur.

3. bölümde çalışmada kullanılan parametreler hakkında bilgi verilerek yapılan çalışmaya benzer çalışmalar hakkında bilgi verilmiştir. Son bölümde ise elde edilen sonuçlar daha önce yapılmış çalışmalarla karşılaştırılıp önerilerde bulunulmuştur.

1. BÖLÜM

GENEL BİLGİLER

1.1. Veri Madenciliğinin Tarihçesi

1950’li yıllarda matematikçiler veri madenciliği teknikleri üzerine çalışarak mantık ve bilgisayar bilimleri alanlarında yapay zekâ ve makina öğrenme alanlarını ortaya çıkarmışlardır. 1960’lı yıllarda istatistikçiler veri madenciliğinin ilk adımlarını oluşturan regresyon analizi ve en büyük olabilirlik kestirimini keşfetmişlerdir. Daha sonraki 20 yıllık süreçte önce verilerin sınıflara ayrılması ardından bu sınıflar arasında ilişkisel bağlantıların kurulması ile veri tabanı kavramı ortaya çıkarılmıştır. 1990’lı yıllara gelindiğinde ise veri tabanında bilgi keşfinin ilk adımları oluşturulmuş ve bununla birlikte büyük veri tabanları için veri ambarı geliştirilmiştir. Aynı zamanlarda yeni teknolojilerle beraber veri madenciliği yaygın olarak kullanılmaya başlanmıştır.

1.2. Veri Madenciliği

Veri madenciliği veri tabanlarından yapılandırılmış bilgilerin otomatik olarak çıkarılma sürecidir. Bu süreç veri tabanlarından bilgi keşfi olarak adlandırılan genel sürecin özel bir parçasıdır [3].

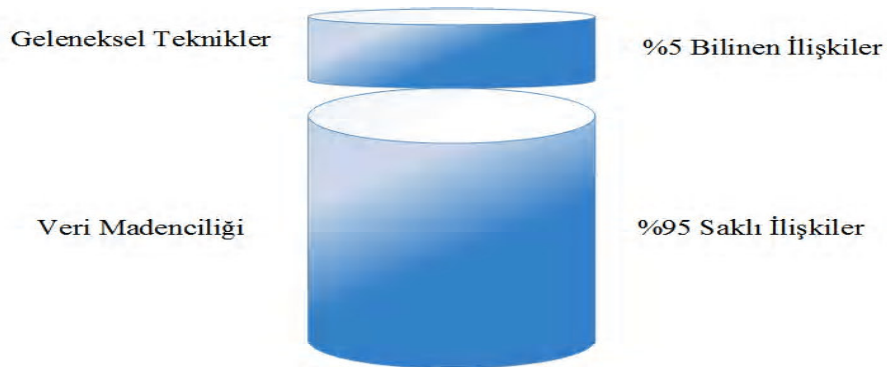
Veri madenciliği, çok sayıdaki verinin depolandığı veri tabanlarından elde edilen modeller, örüntüler, ilişkiler, sapmalar ve anlamlı yapılar gibi ilginç bilgilerin keşfedilmesi sürecidir [4].

Fayyad ve Ark. Veri madenciliğini; kabul edilebilir etkinlik sınırlarına sahip bilgisayar tekniklerini ihtiva eden, veri üzerinden olağandışı örüntü ve model sıralamaları üreten süreç olarak tanımlanan veri tabanlarından bilgi keşfi sürecinin bir adımı olarak tanımlamışlardır [5].

Son yıllarda bilgi toplama çok daha kolay hale gelmiştir, ancak eldeki bilgiler arasındaki ilgi parçacıklarını ortaya çıkarmak için ihtiyaç duyulan çaba özellikle büyük ölçekli veri tabanlarında büyük artış göstermiştir [6]. Veri toplama ve depolama teknolojilerindeki hızlı artışı veri tabanı, veri ambarı veya www gibi diğer depolama türlerindeki verilerin aşırı genişlemesine sebep olmuştur [7]. Buna karşılık bilim adamlarının, mühendislerin ve analistlerin sayısı değişmemektedir [8].

Örnek olarak; klinik tedavi alanında, büyüyen hacimli verilerden bilgi keşfetmede zorluklarla karşılaşmaktadır. Bugünlerde gözlem altında bulunan hastaların fizyolojik parametrelerinin sürekli olarak toplanması devasa hacimlerdeki bilgilerin ortaya çıkmasına sebep olmaktadır. Büyüyen miktarlardaki veri, elle analiz yapan tıp uzmanlarının görevlerini yapmalarını engellemektedir. Birçok saklı ve potansiyel olarak faydalı ilişkiler analist tarafından fark edilememektedir [9].

Geleneksel teknikler kendi hipotezinizin doğruluğunu kanıtlamanıza izin verir. Şekil 1.1.'de görüldüğü gibi bütün ilişkilerin yaklaşık %5'i bu yolla bulunabilir. Veri madenciliği ise kalan %95'lik ilişkilere açılan bir kapıdır [10].



Şekil 1.1. İlişkiler [10]

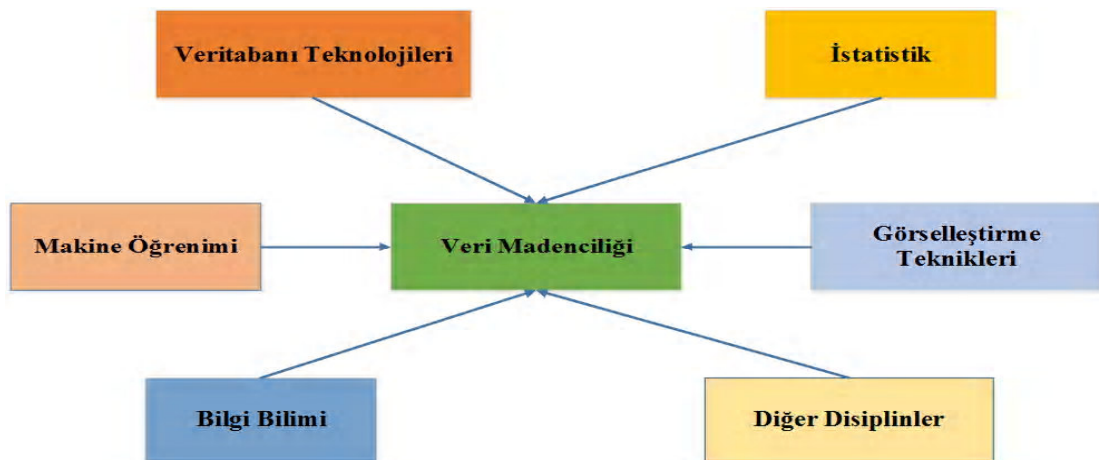
Birçok zaman büyük veri tabanları önceden varlığı bilinmeyen veya görülmemiş ilişkiler, eğilimler ve örüntüler için araştırılır. Bu ilişkiler veya eğilimler genelde mühendis, analist veya pazar araştırmacıları tarafından varsayılır, ancak bu ilişkilerin elde edildikleri veriler tarafından ispat edilmesine ihtiyaç duyulmaktadır. Yeni bilgiler

kullanıcıların yaptıkları işi daha iyi yapmalarına yardımcı olur [11]. Genel olarak veri madenciliğine ilginin artması aşağıdaki faktörlerle açıklanabilir [12];

1. 1980’lerde şirketler, müşterileri, rakipleri, ürünleri ile ilgili verilerden oluşan veri tabanları oluşturmuşlardır. Veri madenciliği algoritmaları tipik olarak, veri tabanının alt gruplarında ya da “uygun” kümelerde belirginleşir.
2. Bilgisayarlarda ağ kullanımı gelişmeye devam etmektedir. Bu durumda veri tabanı ile bağlantı kurmak kolaylaşır. Böylece demografik verili dosya ile müşteri dosyası arasında bağlantı kurulabilir ve belirli popülasyon gruplarının kimliklerinin belirlenmesi sağlanabilir.
3. Müşteri ile hizmet veren arasındaki ilişki, kişisel bilgileri hizmet verenin masasındaki bilgisayardan merkezi bilgi sistemlerine gönderir.

1.3. Veri Madenciliği ve Disiplinler Arası ilişki

VM, çok disiplinli bir yaklaşımdır ve birçok tekniği bünyesinde barındırmaktadır. Veri madenciliği ile makine öğrenimi, istatistik ve veri tabanı teknolojileri arasındaki yakın bağ kolaylıkla görülebilir. Bu disiplinler, veri içindeki ilginç birliktelikleri ve örüntüleri bulmayı amaçlar. Şekil 1.2 veri madenciliği ile disiplinler arası ilişkiyi göstermektedir [13].



Şekil 1.2. Veri madenciliği ile disiplinler arası ilişki [13]

1.4. Veri Madenciliğinin Uygulama Alanları

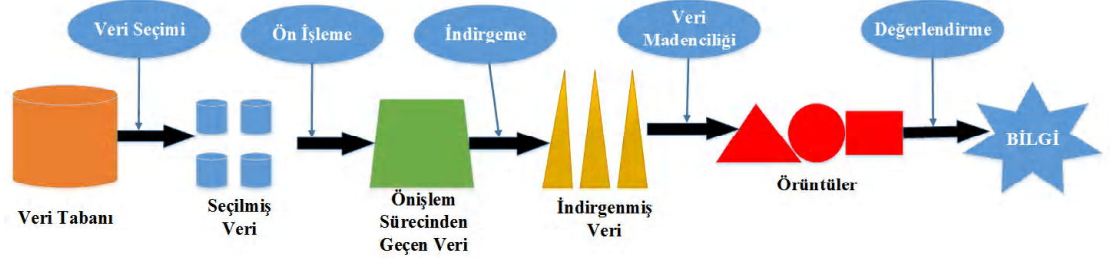
Veri madenciliğinin uygulama alanlarına kısaca değinilecek olursa [14];

- **Pazarlama:** Müşterilerin satın alma alışkanlıkları, demografik bilgileri, kampanya ürünleri belirleme, mevcut müşterileri kaybetmeden yeni müşteriler kazanma, market sepeti analizi, müşteri ilişkileri yönetimi ve satış tahmini alanları en yaygın veri madenciliği uygulama alanlarıdır.
- **Banka ve Sigortacılık:** Farklı finansal göstergeler arasında korelasyon tespitinde, kredi kartı dolandırıcılıklarının tespitinde, kredi taleplerinin değerlendirilmesinde, kredi kartı harcamalarına göre müşteri profili belirlenmesinde, sigorta dolandırıcılıklarının tespitinde ve yeni poliçe talep edecek müşterilerin tahmininde yoğun olarak kullanılmaktadır.
- **Biyoloji, Tıp ve Genetik:** Bitki türleri ıslahı, gen haritasının analizi ve genetik hastalıkların tespiti, kanserli hücrelerin tespiti, yeni virüs türlerinin keşfi ve sınıflandırılması, fizyolojik parametrelerin analizi ve değerlendirilmesinde kullanılmaktadır.
- **Kimya:** Yeni kimyasal moleküllerin keşfi ve sınıflandırılması, yem ve ilaç türlerinin keşfinde kullanılmaktadır.
- **Görüntü Tanıma ve Robot Görüş Sistemleri:** Çeşitli sensörler aracılığı ile tespit edilen görüntülerden yola çıkarak engel tanıma, yol tanıma, yüz tanıma ve parmak izi tanıma gibi tekniklerde kullanılmaktadır.
- **Uzay Bilimleri ve Teknolojisi:** Gezegen yüzey şekilleri ve gezegen yerleşimleri, yeni galaksiler keşfi ve yıldızların konumlarına göre gruplandırılmasında kullanılmaktadır.
- **Metin Madenciliği:** Çok büyük ve anlamsız metin yığınları arasından anlamlı ilişkiler elde etmekte kullanılmaktadır.
- **Bilimsel, Mühendislik ve Sağlık Bakım Verileri:** Günümüzde bilimsel veriler, iş sahası verilerinden daha da karmaşık hale gelmişlerdir.

1.5. Veri Madenciliği Süreci

Veri madenciliği teknikleriyle verilerden bilgi keşfi süreci aşağıdaki adımlardan oluşmaktadır.

1. Verilerin Toplanması
2. Verilerin Temizlenmesi
3. Verilerin Bütünleştirilmesi
4. Verilerin Dönüştürülmesi
5. Veri Madenciliği
6. Desen Değerlendirme
7. Bilgi Sunumu [15].



Şekil 1.3. Bilgi keşfi süreci

Veri madenciliğini daha detaylı bir şekilde ele almak istersek aşağıdaki adımları izlememiz gerekmektedir.

- İşin (Problemin) Analizi ve Verilerin Anlaşılması
- Veri seçimi
- Veri analizi ve hazırlığı

- Veri azaltma ve dönüştürme
- Önemli özelliklerin (Attributes) seçimi
- Veri kümesi boyutunun küçültülmesi
- Normalleştirme
- Birleştirme
- Veri madenciliği metodunun seçimi
- Veri madenciliği süreci
- Görselleştirme
- Değerlendirme
- Bilgi kullanımı ve sonuçların hedefe uygun olarak değerlendirilmesi

Veri madenciliğinin dar ve geniş anlamları dikkate alınarak oluşturulan genel deneysel prosedür Şekil 1.4’de belirtilen 5 adımdan oluşmaktadır.



Şekil 1.4. Veri madenciliği süreci adımları

1.5.1. Problemin Belirlenmesi ve Verilerin Anlaşılması

Problemin belirlenmesi ve verilerin anlaşılması kısmı veri madenciliği uygulamasının ilk aşamasını oluşturmaktadır. Problemin belirlenmesi ve verilerin anlaşılması adımının en iyi şekilde gerçekleştirilebilmesi için aşağıda belirtilen kurallara uyulması sayesinde ulaşılabilir:

- Problem elle tutulur faydalar sağlayacak şekilde açıkça tanımlanmalı.
- Muhtemel sonuç belirlenmeli
- Elde edilecek sonucun nasıl kullanılacağı belirlenmeli.
- Problem ve veriler olabildiğince anlaşılmalı.
- Problem modele dönüştürülmeli
- Varsayımlar belirlenmeli.
- Model döngüsel olarak iyileştirilmeli.
- Model mümkün olan en basit hale getirilmeli.
- Modelin kararsızlığı tanımlanmalı.
- Modelin belirsizliği tanımlanmalı.

Bu belirtilen kurallara uyulduktan sonra problemin ve verilerin niteliğine bağlı olarak ilgili alanda uzman desteğine ihtiyaç duyulabilir ve uzman ile veri madenciliği çalışmalarını yürüten kişilerle yapılan iş birliği sayesinde süreç daha başarılı işlenebilir [16].

1.5.2. Verilerin Toplanması

Bu aşama verilerin nasıl toplanacağı ile ilgilidir. Verilerin oluşturulması yani toplanması sürecinde iki farklı yaklaşım vardır. Eğer süreç uzman kontrolünde yapılırsa tasarlanmış deney; uzman kontrolü olmadan yapılırsa gözlemsel yaklaşım olarak adlandırılır. Kullanılan verilerin aynı bilinmeyen örneklemde gelmesi, modelin kurulması, test edilmesi ve uygulanması açısından önemlidir.

1.5.3. Verilerin Hazırlanması

“Veri madenciliği sürecinde zamanın en fazla harcandığı yer olan verilerin hazırlanması kısmı sürecin başarılı olmasında %75 ila %90 arasında bir katkı sağlar. Veri kümesinin zayıf veya var olmayan verilerden hazırlanması sürecin başarısızlığında %100 sorumludur [17]”.

Verilerin hazırlanma süreci, veri madenciliği sürecinin diğer adımlarından bağımsız değildir. “Veri madenciliğinin her adımında yapılan tüm işlemler birlikte yeni ve daha gelişmiş bir veri kümesi elde etmemize yardımcı olur [16]”. Verilerin hazırlanması aşaması; gerçek hayattan toplanan eksik, geniş dağılımlı, çelişkili verilerin temizlenmesi, verilerin kaynaştırılması ve dönüşümü, veri hacminin küçültülmesi, verilerin kesikli hale getirilmesi işlemlerinden oluşur.

1.5.3.1. Verilerin temizlenmesi

Verilerin temizlenmesi aşaması, eksik değerlerin doldurulması, uyumsuz verilerin tanımlanması, çelişkili verilerin doğrulanması veya veri setinden çıkartılması gibi işlemlerden oluşur. “Bu aşamada eksik verilerin yerine yenileri belirlenerek konulması için aşağıda belirtilen yöntemler kullanılabilir:

- a) Eksik değer içeren kayıtlar veri kümesinden atılabilir.
- b) Kayıp değerlerin yerine bir genel sabit kullanılabilir.
- c) Değişkenin tüm verileri kullanarak ortalaması hesaplanır ve eksik değer yerine bu değer kullanılabilir.
- d) Değişkenin tüm verileri yerine, sadece bir sınıfa ait örneklerin değişken ortalaması hesaplanarak eksik değer yerine kullanılabilir” [18].

1.5.3.2. Verilerin kaynaştırılması ve dönüştürülmesi

Verilerin kaynaştırılması aşamasında farklı kaynaklardan elde edilen verilerin uyumlu bir forma sokulması için bazı işlemlerin yapılması gerekmektedir. Bu işlemler meta data, korelasyon analizi, veri çatışması tespiti ve semantik uyumsuzluğun giderilmesi gibi aşamalardan oluşur. Verilerin dönüştürülmesi kısmı ise verilerin, veri madenciliği uygulamasında daha iyi sonuç verecek forma sokulması işlemidir [16].

1.5.3.3. Verilerin kesikli hale getirilmesi ve kavramsal hiyerarşi oluşturma

Verilerin kesikli hale getirilmesi sürekli değişkenler üzerinde uygulanan bir durumdur. Verilerin kesikli hale getirilmesi için şu teknikler kullanılmaktadır; birleştirme, histogram analizi ve entropi’dir.

Verilerin dönüştürülmesi aşamasında sayısal tabanlı algoritma kullanılacak ise kesikli verilerin sürekli hale getirilmesi; kategorik tabanlı algoritmalar kullanılacaksa sürekli verilerin kesikli hale getirilmesi gerekmektedir.

1.5.4. Modelin Kurulması

Tanımlanan problem için uygun modelin belirlenmesi aşamasında, mümkün olduğunca çok sayıda model kurularak ve bu kurulan modellerin denenmesi ile gerçekleşir. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele ulaşmaya kadar yinelenen bir aşamadır.

1.5.5. Modelin Yorumlanması

Model elde edilmesi beklenen hedefleri karşılamaya yeterli bulunduğu zaman, süreç bazlı daha geniş bir perspektiften değerlendirme yapılır. Bu değerlendirme sürecinde modelin doğru kurulup kurulmadığı, gelecekte kullanılabilecek farklı verilerin neler olabileceği, modelin genişletilmesi gibi konuları içerir.

1.6. Veri Madenciliği Modelleri

Veri madenciliği modelleri kestirime dayalı ve tanımlayıcı olarak ikiye ayrılır:

- **Kestirime dayalı modeller:** sınıflandırma, eğri uydurma, zaman serileri
- **Tanımlayıcı modeller:** demetleme, özetleme, bağıntı kuralları

1.7. Veri Madenciliğinde Öğrenme Yöntemleri

Veri madenciliğinde 3 farklı öğrenme şekli vardır:

1.7.1. Denetimli Öğrenme(Supervised Learning)

Denetimli öğrenmede amaç; girdi ve çıktı değerlerinin arasındaki ilişkiyi öğrenmektir. Böylelikle yeni bir girdi değeri için bir çıktı değeri tahminlenebilir. Sınıflandırma algoritmaları, özellikle regresyonun dâhil olduğu geleneksel istatistik teknikleri, yapay sinir ağları (YSA), Karar ağaçları (Decision Tree), Kural çıkarımı (Rule Induction), K-ortalamalar kümeleme (K-Means Clustering), En yakın k komşuluk (k-Nearest-Neighbor) ve destek vektör makineleri (SVM) denetimli öğrenime örnektir.

1.7.2. Denetimsiz Öğrenme (Unsupervised Learning)

Denetimsiz öğrenimde amaç, girdi değerlerine (verilere) en uygun gösterim şeklini bulmaktır. Birçok farklı yaklaşım vardır. Bilgi maksimizasyonu, minimum çapraz entropi, minimum yeniden yapılandırma hatası gibi. Veri sıkıştırma, dağılım tahmini, veri kaynağının ayırımı, veri görselleştirme gibi denetimsiz öğrenimin çeşitli uygulamaları, denetimli öğrenim için ön işlemler (preprocessing) olarak kullanılabilirler. Kümeleme en temel denetimsiz öğrenim yöntemidir [19]. Ayrıca kendi kendini düzenleyen haritalar (Self organized maps) da başka bir denetimsiz öğrenim yöntemidir.

1.7.3. Destekli Öğrenme (Reinforcement Learning)

Destekli öğrenim, hayvanların öğrenme şeklini modelleyen öğrenim yöntemidir. Öğrenici eylemde bulunur, eylemi nedeniyle çevresindeki değişimlerden sadece geri besleme (feedback) ile bilgi alır. Dinamik kaynak paylaşımı (dynamic resource allocation) oyun oynama (game playing) ve geçici uyumsuzluk öğrenimi (temporal difference learning) destekli öğrenime örnektir.

1.8. Veri Madenciliği Teknikleri

Veri madenciliği teknikleri özellikle işletmelerde çeşitli alanlarda başarı ile kullanılmaktadır. Başlıca kullanım alanları pazarlama, bankacılık, sigortacılık, borsa, telekomünikasyon, sağlık ve ilaç, endüstri, bilim ve mühendislik uygulamalarıdır. Veri madenciliği tekniklerinden bazıları aşağıdaki gibidir [20];

- Birliktelik Analizi
- Sınıflandırma
- Kümeleme Analizi
- Tanımlama ve Ayrımlama
- Sıradışılık Analizi
- Evrimsel Analiz

Büyük boyutlarda ve hızlı bir şekilde toplanan verilerin çeşitli analizler sonucunda anlamlı bilgilere dönüştürülmesi süreci olarak tanımlanan veri madenciliğinin, günümüzde en çok kullanılan teknikleri **sınıflandırma** ve **kümeleme** problemleridir.

1.8.1. Sınıflandırma Problemi

Sınıflandırma problemi, nesnelerin her bir nitelik kümesi ve önceden tanımlanmış olan sınıf etiketlerine atanmasından oluşur. Veri kümesindeki her bir veri için nitelik sınıfı ve sınıf etiketi bilgisi bulunmaktadır. Bu verilere göre elde edilen sınıflandırıcı model, devamında gelen kayıtları sınıflandırmak için kullanılabilecek kısa ve anlamlı veriler türetir. Denetimli sınıflandırma problemlerinde verilerin sınıf etiketleri mevcuttur. Buradaki amaç, sınıf etiketi mevcut olan nesneler üzerinde belirli bir amaca uygun modeller türetmektir [21]. Sınıflandırma modeli, "Tanımlama", Tahmin", Birliktelik Analizi", Kümeleme Analizi" gibi veri madenciliği amaçları için kullanılmaktadır [22].

- **Tanımlama:** Sınıflandırma modeli, farklı sınıfların objelerini ayırt etmek için açıklayıcı bir araç olarak da hizmet edebilir. Örneğin, hem biyologlar hem de diğer kişiler için vücut sıcaklığı, deri, doğurganlık gibi tür özelliklerin bir omurgalıyı memeli, sürüngen, kuş veya balık olarak tanımladığını açıklayan, tanımlayıcı bir modele sahip olmak yararlı olacaktır [22].
- **Tahmin:** Sınıflandırma, evet/hayır, memeli/sürüngen/kuş gibi kesikli çıktılar ile ilgilenir. Tahmin ise sürekli değerler alan çıktılar ile ilgilenir. Tahmin, bazı girdi verileri verildiğinde gelir düzeyi, oy miktarı, gelecek dönem satış tahmini gibi bilinmeyen sürekli değişkenlere ilişkin değerlerin bulunması için kullanılır [22].
- **Birliktelik Analizi:** Birliktelik analizi, belirli türlerdeki veri ilişkilerini tanımlayan bir modeldir. Herhangi bir ürün alındığında bu ürünün yanında bir başka ürünün de satın alınması bir birliktelik kuralı verir. Örneğin, bir süpermarkette yapılan alışverişlerin incelenip hangi ürünün hangi ürünle birlikte satın alındığının belirlenmesi birliktelik kurallarını ilgilendirir [23].

1.8.2. Kümeleme Problemi

Kümeleme analizi veri madenciliğinin en önemli alanlarından birisidir; amacı, nesneleri birbirine olan benzerliklerine göre benzeyenler bir kümeye, benzemeyenler ise bir başka

kümeye toplamaktır. Benzersizlikler ise nesneleri tanımlayan özelliklerin değerleri temel alınarak belirlenir. Birbirine benzer nesne gruplarının işaretlenmesi ya da başka gruplarla olan farklılıklarının bulunması ile kümeler oluşturulur.

Makine öğrenimi açısından, her bir küme gizli bir örüntüyü temsil eder ve uygulanan öğrenme ise bir denetimsiz öğrenmedir. İstatistikte çok değişkenli istatistiksel tahmin, ses ve resim tanınması, DNA analizi, coğrafi bilişim sistemleri ve bunlarla ilgili alanlarda kullanılmaktadır [23, 20].

1.9. Sınıflandırma Modeli

Sınıflandırma, yüksek-düzeyle veri ile tanımlanan nesneler kümesi gibi karmaşık olayları, daha iyi kontrol veya ifade etmek için hizmet eden küçük ve tanımlayıcı birimlere, sınıflara, alt yapılara veya parçalara ayırarak ve mevcut bilgi temelinde yeni durumları bu sınıflardan birine atayarak ilerleyen temel zihinsel bir yetenektir [24].

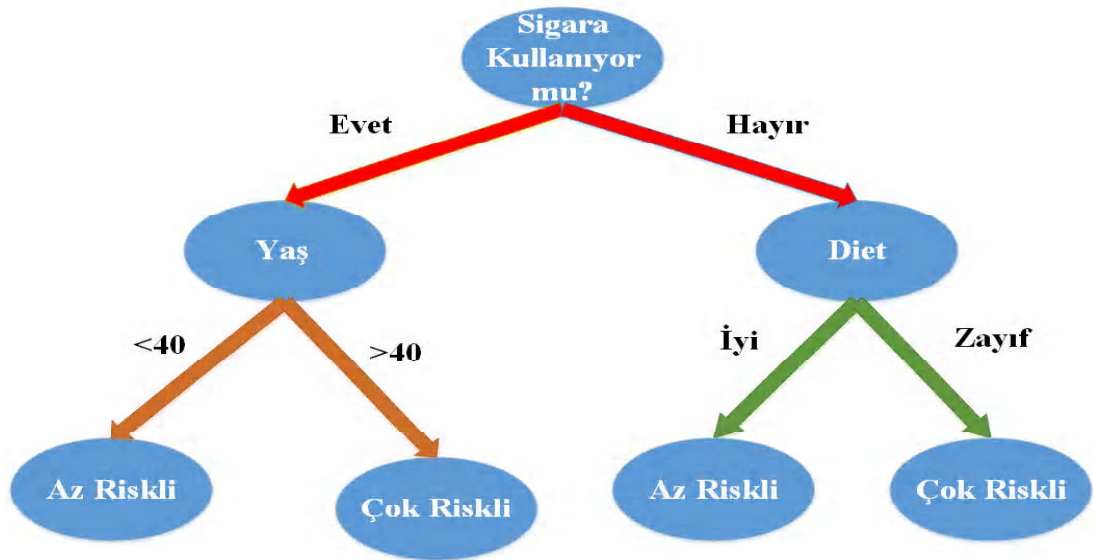
Sınıflandırmada kullanılan başlıca algoritmalar şöyle sıralanabilir;

- Karar ağaçları,
- Yapay sinir ağları
- Evrimsel Algoritmalar (EA)
- K-en yakın komşu
- Bayes sınıflandırıcılar
- Sürü zekâsı (SZ) teknikleri
- Durum-tabanlı nedenleme (DTN)
- Kaba küme (KK) yaklaşımı
- Bulanık küme yaklaşımı (BK)

1.9.1. Karar Ağaçları

Karar ağaçları veri madenciliğinde genellikle veriyi tanımlamak, tahminde bulunmakta kullanılacak ağacı ve ağacın kurallarını indirgemekte kullanılan tekniklerdir. Karar

ağacı akış diyagramına benzer bir ağaç yapısında olup, her bir dal bir testin sonucunu, yaprak düğümleri ise sınıfları temsil eder. Bilinmeyen bir örneği sınıflandırmak için nitelik değerleri karar ağacı karşısında test edilir. Değerlendirme yapılırken yeni bir örnek, ağacın kök bölümünden girer. Kökte test edilen bu yeni örnek, test sonucuna göre bir alt düğüme gönderilir. Bu süreç, yeni örnek herhangi bir yaprak düğüme ulaşana kadar devam eder. Ağacın belirli bir yaprağına gelen bütün örnekler aynı şekilde sınıflandırılırlar. Kökten her bir yaprağa giden sadece tek bir yol vardır. Bu yol, örnekleri sınıflandırmak için kullanılan bir kuralı tanımlamaktadır [25]. Şekil 1.5 örnek bir karar ağacı yapısını göstermektedir.



Şekil 1.5. Örnek bir karar ağacı yapısı

Bir karar ağacı, veri keşfine şu şekilde yardımcı olmaktadır [26];

- Temel özellik olarak korunan ve doğru bir özet sunan veriyi daha sıkıştırılmış bir hale sokarak, verinin hacmini azaltır.
- Veri iyi dağıtılmış sınıf nesneleri içersin veya içermesin karar ağacı keşifte bulunur, öyle ki sınıflar anlamlılık teorisi kavramında doğru bir şekilde yorumlanabilir.

- Veriyi bir ağaç formunda haritalar, böylece ağacın dallarından köküne doğru geri gidilerek tahmin değerleri üretilebilir. Bu değerler, yeni bir veri veya sorgunun çıktısını tahmin etmekte kullanılabilir.

Temel karar ağaçları iki gruba ayrılmaktadır [27]. Bunlar;

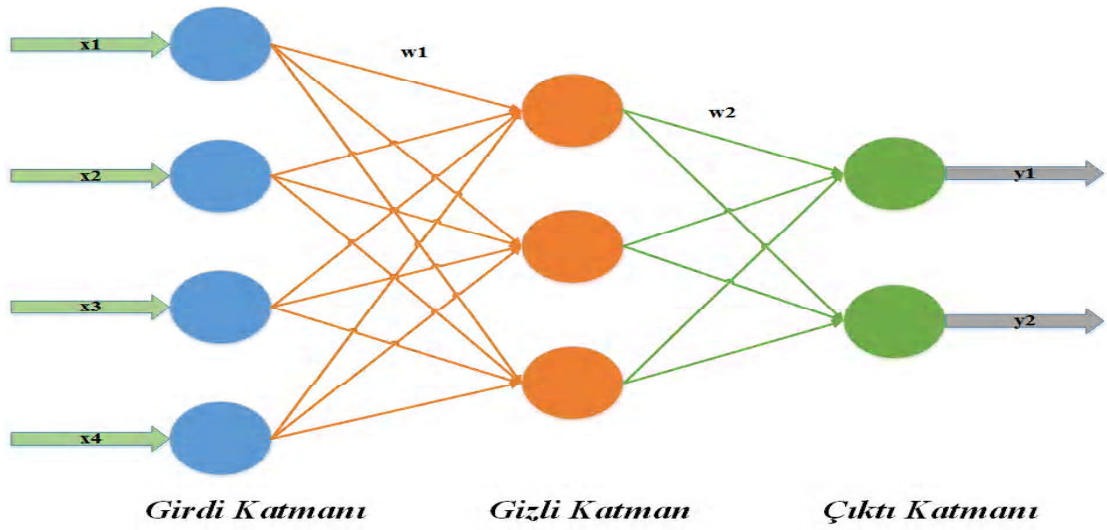
- Makine öğrenme topluluğundan sınıflandırıcılar: ID3, C4.5, CART
- Büyük veri tabanları için sınıflandırıcılar: SLIQ, SPRINT, SONAR, RainForest.

Karar ağaçlarının kolaylıkla sınıflandırma kurallarına dönüştürülebilmesi en büyük avantajlarından birisidir. Diğer avantajları ise şöyle sıralanabilir;

- Kuruluşlarının ucuz olması,
- Gürültülü verilerde etkili çalışması,
- Sınıfların ayırıcı özelliklerini keşfetmesi,
- Veri tabanı sistemlerine kolayca entegre edilebilmeleri,
- Yorumlanmalarının kolay olması,
- Güvenilirliklerinin iyi olması.

1.9.2. Yapay Sinir Ağları

Tahmin ve sınıflandırma amacıyla kullanılan bir başka veri madenciliği tekniği yapay sinir ağlarıdır. Sinir ağları, bir ağa bağlı çok sayıda nöron bileşiminin matematiksel bir modelini sunarak, biyolojik sinir sistemini taklit etmeye çalışan sinyal işleme sistemleridir [28, 29]. YSA' da amaç, insan beyninin davranışlarını taklit etmektir [30]. Nöronlar, giriş ve çıkış katmanlarında ve eğer varsa gizli katman (lar) da bulunur. Bir nöron önemli olarak belirlendiğinde, bu nöronla bağlantılı ağırlıklar değiştirilir. Bu da ilgili nöronun kendisiyle aynı düzeyde bulunan diğer nöronlara göre daha etkin olacağı anlamına gelir. Yapay sinir ağları nöronlar arasındaki ağırlıkların ayarlanması sayesinde öğrenirler. Şekil 1.6' da basit bir yapay sinir ağının yapısı gösterilmektedir.



Şekil 1.6. YSA yapısı

Yapay sinir ağlarının veri madenciliği açısından olumlu ve olumsuz tarafları şöyledir;

- Çok geniş uygulama alanları mevcuttur,
- Karmaşık durumlarda daha iyi sonuçlar üretirler,
- Sürekli ve kategorik veriler üzerinde işlem yapabilirler,
- Elde ettikleri sonuçları açık bir şekilde ifade edemezler,
- Elde edilen sonucun en iyi sonuç olduğunun garantisi yoktur.

1.9.3. Evrimsel Algoritmalar

Doğal evrim sürecinden esinlenerek geliştirilen evrimsel algoritmalar, veri madenciliğinde sınıflandırma ve kural çıkarımında yaygın olarak kullanılmaktadır. Radyan tabanlı tekniklerin aksine evrimsel algoritmalar, çoklu aday çözümlerinin performanslarını eşzamanlı olarak değerlendirerek zeki bir şekilde arama uzayını tararlar ve küresel en iyiye yaklaşırlar [31].

Genel olarak evrimsel algoritmalar, popülasyon ve seçim tabanlı olan, arama uzayında yeni bir arama noktası üreten genetik operatörlere sahip bütün algoritmaları içine alır. Bu algoritmalar genetik algoritmalar (GA), genetik programlama (GP), evrimsel

programlama (EP) ve evrimsel stratejilerdir (ES). Bu yaklaşımlar, kullandıkları operatörlere, uygulandıkları modellere, seçim metotlarına ve uygunluk fonksiyonlarına göre birbirlerinden ayrılırlar. GA ve GP genetik seviyede evrimsel modellerdir. Evrimsel stratejilerde kullanılan optimizasyonda, popülasyondaki bireylerin yapıları en iyilenmeye çalışılır. Bireylerin çeşitli davranışsal özellikleri parametrik hale getirilir ve en iyilenme süresinde bu değerler geliştirilir. Evrimsel programlamada ise çeşitli türlerin davranışsal özelliklerinin adaptasyonu üzerinde durularak en üst düzey soyutlama kullanılır [32].

1.9.4. k-En Yakın Komşu

k-en yakın komşu, örnekleme yoluyla öğrenmeye dayanan en eski tekniklerden biridir. Bu teknikte tüm örneklemeler n boyutlu uzayda saklanır [33]. Algoritma, bilinmeyen bir örneklemin hangi sınıfa dâhil olduğunu belirlemek için örüntü uzayını araştırarak bilinmeyen örnekleme en yakın olan k örneklemini bulur. Yakınlık, genellikle Öklid uzaklığı ile tanımlanır. Hedef fonksiyon kesikli veya reel değerli olabilir. Daha sonra, bilinmeyen örneklem, k en yakın komşu içinden en çok benzediği sınıfa atanır. Burada k harfi araştırılan komşuların sayısıdır. 5-en yakın komşuluğunda, 5 komşuya ve 1-en yakın komşuluğunda 1 komşuya bakılır [25].

k-en yakın komşu algoritmasında uzaklık fonksiyonu ve dâhil edilecek komşu sayısı parametresi olan k 'nın değerleri ile ilgili kararlar en önemli seçimlerdir [34].

1.9.5. Bayes Sınıflandırıcılar

Bayes sınıflandırıcılar istatistiksel sınıflandırıcılardır. Özel bir sınıfa ait olarak verilen bir olasılık gibi, sınıf üyeliklerine ait olasılıkları önceden söyleyebilirler [25]. Bunun yanı sıra büyük veri tabanları üzerinde işlem yapan Bayes sınıflandırıcılar hız ve doğruluk yönünden oldukça yüksek performansa sahiptirler. Bayes sınıflandırıcılar, verinin olasılık dağılımı verildiğinde en düşük hata oranıyla etkin bir şekilde çalışabilirler [35].

Bayes sınıflandırıcılar verinin tek bir kez taranmasını gerektiren basit sınıflandırıcılardır. Bu nedenle büyük veri yığınlarında yüksek doğruluk ve hız elde edebilirler. Performansları karar ağaçları ve sinir ağları ile rekabet edebilecek ölçüdedir [33].

1.9.6. Sürü Zekâsı

SZ, özerk yapıdaki basit bireyler grubunun kolektif bir zekâ geliştirmesi olarak tanımlanır [36]. SZ tanımı ilk olarak 1989 yılında Gerardo Beni ve Jing Wang [37] tarafından hücrel robotik sistemler kavramı içinde kullanılmıştır. SZ, merkezi olmayan, kendi kendini yöneten sistemlerde toplu davranış çalışmasına dayanmaktadır. Bu tip sistemlere örnekler doğada mevcuttur. Bunlara örnek olarak karınca kolonileri, kuş sürüleri, bakteri küfleri, arı kolonileri ve balık sürüsü vb. verilebilir.

Sürü zekâsı, karınca koloni optimizasyonu ve parça sürü optimizasyonu (PSO) olmak üzere farklı algoritmalar içermektedir. Bu algoritmalar optimizasyonda başarı ile kullanılmaktadır. Günümüzde sürü zekâsı teknikleri veri madenciliği alanında da kullanılmaya başlanmıştır ve yapılan uygulamalar bu tekniklerin sınıflandırmada iyi sonuçlar elde edebildiğini göstermektedir.

1.9.7. Durum Tabanlı Nedenleme

İlk olarak Schank [38] tarafından ortaya atılan DTN sınıflandırıcıları, örnek tabanlı sınıflandırıcılardır [25]. DTN, deneyimle öğrenir ve verilen problemle daha önce karşılaşılan problemlerin benzerlik ve farklılıklarından yararlanır. Eğitim örneklerini öklid uzayında noktalar olarak saklayan en yakın komşu sınıflandırıcılarından farklı olarak, DTN tarafından saklanan örnekler veya durumlar karmaşık sembolik tanımlamalardır. Yapay sinir ağlarından farklı olarak ise durum tabanlı nedenlemeye dayalı sınıflandırıcılar genelleme yapmazlar.

DTN'ye yeni yaklaşımlar iyi bir benzerlik ölçüsünün bulunması, eğitim durumlarını indekslemek için etkin tekniklerin geliştirilmesi ve çözümlerin birleştirilmesini içerir. Mevcut bazı durum seçim metotları şunlardır [40]; özetlenmiş en yakın komşu, durum genişletme veya budama ile örnek tabanlı öğrenme (IB3 gibi), nitelik ağırlıklandırma ile örnek tabanlı öğrenme (IB4 gibi).

1.9.8. Kaba Küme Yaklaşımı

İlk olarak Pawlak [41] tarafından ortaya atılan KK yaklaşımı sınıflandırmada kesin olmayan, gürültülü verideki yapısal ilişkileri bulmakta kullanılır ve kesikli değerli değişkenlere uygulanır. Kaba küme yaklaşımı, klasik kümedeki, kümenin yalnızca elemanları ile tanımlandığı ve kümenin elemanları hakkında ilave hiçbir bilginin

bulunmadığı yaklaşımın aksine, bir kümenin tanımlanması için başlangıçta uzay hakkında bazı bilgilere gereksinim olduğu varsayımına dayanır. KK yaklaşımının temelini ayırt edilememe ilişkisi oluşturur. Bilginin temelini oluşturan ve aynı nesnelerin kümesi olan kümeye “elemanter” küme denir. Elemanter kümelerin herhangi bir birleşimi ise “kesin” küme olarak adlandırılır, aksi halde bir küme “kaba” olarak ifade edilir. Her kaba kümenin kesinlikle kümenin kendisinin veya tümleyen kümesinin elemanları olarak sınıflandırılmayan elemanları vardır, bunlara “sınır hattı elemanları” denir [42].

Kaba küme yaklaşımı, ele alınan bir eğitim verisinde denklik sınıflarının kurulumuna dayanır. Bir denklik sınıfını oluşturan tüm veri örnekleri ayırt edilemez niteliktedir. Verilen bir sınıf için kaba küme tanımı iki küme ile tahmin edilir, bunlar “alt yaklaşım” ve “üst yaklaşım”dır. Alt yaklaşım kesin olarak sınıfa ait olan tüm nesnelerden oluşur. Üst yaklaşım ise sınıfa ait olması olası bütün nesneleri içerir. Alt ve üst yaklaşımlar arasındaki fark sınır bölgesini oluşturur [43]. Karar kuralları her bir sınır için oluşturulur. Genel olarak kaba küme yaklaşımında, kuralları göstermekte karar tabloları kullanılır [25].

1.9.9. Bulanık Küme Yaklaşımı

Zadeh tarafından ortaya atılan ve bulanık mantığa dayanan bulanık küme yaklaşımı, belirsizlik ile ilgilenir. Bulanık mantık, karmaşık, iyi tanımlanmamış ya da matematiksel olarak kolay analiz edilemeyen sistemlerin davranışlarını tanımlamak için hemen hemen doğru ve etkin yöntemler sunar [44]. Diğer bir deyişle, bulanık mantık, belirsiz ve kesin olmayan bilginin kullanılması için bir platform oluşturur. Kategoriler arasında kesin bir ayırmadan ziyade, 0-1 arasında bir doğruluk değeri kullanırlar.

1.10. Sınıflandırma Veri Madenciliği

1.10.1. Sınıflandırma Kuralları

Sınıflandırma, sınıf etiketi bilinmeyen nesnelerin sınıflarını tahmin etmek için elde edilen modeli kullanmak amacıyla, veri sınıf ve kavramlarını tanımlayan modeller veya fonksiyonlar kümesinin bulunması sürecidir [25]. Sınıflandırma işlemi genellikle veri tabanlarından sınıflandırma modeli oluşturmak için danışmanlı öğrenme metotlarını kullanır. Çıktı sınıfı bilinen bir örnekler kümesi verildiğinde, sınıflandırmanın amacı

değişkenler ve sınıflar arasındaki gizli ilişkilerin keşfedilmesidir [45]. Sınıflandırmada karar sınırları, farklı sınıflara ait örnekleri ayırt etmek için oluşturulur [33].

Sınıflandırma kuralları veri madenciliği uygulamalarında, çıktının gösterimi için en çok tercih edilen yöntemlerden biridir. Bunun nedeni kullanıcının kuralları anlamasının ve yorumlamasının kolay olmasıdır. Karar ağaçlarının düğümlerdeki testleri gibi bir kuralın önde gelen kısmı test serilerini içerir, sonra gelen kısmı ise o kuralca kapsanan sınıf veya sınıfları ifade eder [46].

Koşullu ifade olarak bilinen bir kuralın genel yapısı “EĞER koşul(lar) O HALDE sonuç” yapısındaki bir önermedir. Bu önermede, kuralın önde gelen (koşullar) kısmı tahminleyici değişkenlerin mantıksal kombinasyonlarını içerir ve kuralın sonra gelen (sınıf/sonuç) kısmı, kuralın önde gelen kısmını sağlayan durumlar için tahmin edilen sınıfı içerir.

Sınıflandırma kuralları tanımlama (characterization) kuralları ve ayırt etme (discrimination) kuralları olmak üzere ikiye ayrılır [47]. Tanımlama kurallarında amaç bir kavramın özelliklerini tanımlayan kuralları bulmaktır ve elde edilen kurallar “EĞER kavram O HALDE özellik” formundadır. Ayırt etme kurallarında ise amaç bir kavrama (sınıfa) ait örneklerin (veri kayıtlarının) kalan örnekler (veri kayıtları veya sınıfları) içinden seçilmesini (ayırt edilmesini) sağlayan kuralların bulunmasıdır. Ayırt etme kuralları “EĞER özellik O HALDE kavram” formunda bulunurlar. Tanımlama kurallarının tersi ayırt etme kuralları değildir.

Bir özelliğin kavram niteliği değişkenler kümesi ve bunların verilen bir kavram için belirleyici olan ilgili değerleri ile tanımlanır. Sınıf (karar) değişkeni olarak ifade edilen özel bir dc değişkeni içeren, n kayıttan oluşan bir veri kümesi ele alındığında, dc değişkeni veri kümesindeki kayıtları, sınıflar olarak ifade edilen ve kayıtları sınıflandıran, kayıtları sınıf değişkeninin değerine göre tanımlanan ayrık alt kümelere ayıran parçalara böler. Burada ayrık alt küme sayısı, veri kümesinde mevcut sınıf sayısına eşittir.

1.10.2. Sınıflandırma Kural Çıkarımı

Sınıflandırma kural çıkarımı, doğru bir sınıflandırıcı oluşturmak için veri tabanında küçük bir kurallar kümesi keşfetmeyi amaçlar [27, 48]. Kural çıkarımı süreci, bir veri

kümesinden sembolik, sürekli veya kesikli bilginin, veri kümesinde saklı olan bilgiyi yeterli doğrulukta tanımlayan sembolik önermeli mantık kurallarına çevrilmesi olarak düşünülebilir [49,50].

Kural çıkarımında temel amaç, verideki gizli bilgiyi ortaya çıkarıp anlaşılır şekilde ifade etmek, daha önce bilinmeyen ilişkileri ortaya çıkarmak, nedenleme ve tanımlama kabiliyeti sağlamaktır [51]. Sınıflandırma kural çıkarımı ile elde edilen kurallar şu özelliklere sahip olmalıdır [47];

- Anlaşılır olmalı,
- Kısa, basit, açık ve temiz olmalı
- Çıkarıldığı veriyi doğru tanımlamalı,
- Kurallar tekrarlanmamalı,
- İfade ettikleri bilgi kullanışlı olmalı,
- Verideki bilgiyi özetlemeli.

Veri tabanlarından kural çıkarımı, sadece kural tabanlı sınıflandırma için hizmet etmez aynı zamanda keşfedilen dilsel bilgi ile ele alınan probleme bakış açısı sağlar ve daha iyi anlaşılmasına olanak tanır [52].

Bir veri kümesinde sınıflara ait örnekleri sınıflandırma işlemini gerçekleştiren bir kural çoğu zaman “sınıflandırıcı” olarak adlandırılır. Sınıflandırma kural çıkarım teknikleri kural tabanlı metotlar ve kural tabanlı olmayan metotlar olarak ikiye ayrılır [53].

- **Kural tabanlı metotlar:** Kural tabanlı sınıflandırma metotları veriden gizli bilgiyi doğrudan çıkarırlar ve kullanıcılar bu bilgileri kolaylıkla anlayabilirler. C4.5 karar ağacı, karar tabloları vb. kural tabanlı metotlara örnek olarak verilebilir.
- **Kural tabanlı olmayan metotlar:** Kural tabanlı olmayan sınıflandırma metotları genellikle kural tabanlı sınıflandırma metotlarına göre daha doğru sonuçlar verirler fakat kara kutu gibi davrandıklarından dolayı elde ettikleri bilgileri kullanıcılar anlayacağı biçimde sunamazlar. Genel olarak yapay sinir ağları gibi kural tabanlı

olmayan sınıflandırıcılar çok iyi sınıflandırma doğrulukları elde edebilmelerine rağmen anlaşılabilirlik yönünden rekabetçi değildir. Destek vektör makineleri, yapay sinir ağları, doğrusal genetik programlama kural tabanlı olmayan metotlara örnek olarak verilebilir.

Bir veri kümesini sınıflandırmak için kural kümesi oluşturulurken, yerel, küresel veya yerel-küresel bir melez algoritma eğitim ve test etme olmak üzere iki aşamada kullanılır. Birinci aşamada sınıflandırılmış kayıtlardan oluşan bir veri kümesinden sınıflandırma kuralları öğrenilir. Bu adımda kullanılan veri kümesine “eğitim veri kümesi” denir. Bu tip öğrenmede öğrenilene ise “kavram” denir. Dolayısıyla bu süreç “kavram öğrenme” olarak adlandırılır [54]. İkinci aşamada, birinci aşamada öğrenilen kurallar bir test veri kümesine uygulanır ve kuralların doğruluğu değerlendirilir. Eğitim veri kümesi gibi, test veri kümesi de sınıf değerlerini içerir.

1.10.3. Sınıflandırma Metotlarının Değerlendirilmesinde Temel Ölçütler

Veri madenciliğinde belki de en yaygın kullanıma sahip teknik sınıflandırmadır [55]. Sınıflandırmada kategorik bir hedef değişken vardır. Veri madenciliği modeli bu hedef değişkenle ilgili bilgileri ve aynı zamanda girdi değişkenlerini de içeren çok sayıda kayıtlar kümesini ele alır. Aşağıda sıralanan ölçütler sınıflandırma metotlarının değerlendirilmesinde yaygın olarak kullanılmaktadır [54].

- Tahminleyici doğruluk,
- Kural anlaşılabilirliği,
- Hız ve ölçeklenebilirlik,
 - Modeli kurmak için gerekli süre,
 - Modeli kullanmak için gerekli süre.
- Sağlamlık,
 - Gürültüyü ve boş değerleri ele alma
- İlgi çekicilik.

Tahminleyici doğruluk: Tahminleyici doğruluğun belirlenmesi, sınıflandırıcının daha önce görülmemiş örnekleri hangi doğrulukta keşfedeceğini belirleyeceğinden dolayı çok önemlidir. “Direnme (holdout)”, “çapraz-doğrulama (crossvalidation)”, “önyükleme (bootstrapping)” ve “bir-dışarda-bırak (leave-one-out)” doğruluk tahmininde kullanılan tekniklerdir [25].

Direnme ve çapraz-doğrulama, sınıflandırıcı doğruluğunu değerlendirmede en yaygın kullanıma sahip yöntemlerdir. Bu tekniklerde ele alınan veri rastgele parçalara ayrılır. Direnme metodunda, veri eğitim ve test kümeleri olmak üzere birbirinden bağımsız iki parçaya ayrılır. Genellikle verinin üçte ikisi eğitim kümesi, kalan üçte biri ise test kümesi olarak değerlendirilir. Eğitim kümesi sınıflandırıcı elde etmek için kullanılır ve sınıflandırıcının doğruluğu test kümesi ile elde edilir. Başlangıç verisinin sadece bir kısmı sınıflandırıcı elde etmek için kullanıldığından dolayı direnme tekniği kötümser bir tekniktir.

k-katlı çapraz-doğrulama tekniğinde veri eşit büyüklükte rastgele k parçaya ayrılır. Eğitim ve test işlemleri k kez tekrarlanır. Her bir tekrarda k parçadan farklı bir tanesi test kümesi, kalan k-1 parça ise eğitim kümesi olarak kullanılır. Doğruluk tahmini k iterasyondan elde edilen doğru sınıflandırmaların toplam sayısının başlangıç verisi örnek sayısına bölünmesi ile hesaplanır. Genel olarak 10-katlı çapraz-doğrulama, düşük eşik ve varyansından dolayı sınıflandırıcı doğruluğunun belirlenmesinde önerilen bir tekniktir. Çapraz doğrulama tekniğinde kat olarak en çok 10 kullanılmasının nedeni, farklı öğrenme teknikleri ile çok sayıda veri kümesi üzerinde yapılan detaylı testlerde 10 sayısının en az hatayı veren, doğru sayı olarak çıkmasıdır. Ayrıca bunu destekleyecek teorik bilgiler de mevcuttur [46].

Önyükleme metodu ise değiştirmeli örneklemeye dayanan istatistiksel bir yöntemdir. Daha önce anlatılan yöntemlerde eğitim ve test veri kümeleri değiştirme olmaksızın oluşturulurken yani aynı örnek bir kere seçildikten sonra tekrar seçilemezken önyüklemadaki temel fikir eğitim kümesini oluşturmak için değiştirme ile veri kümesinin örneklenmesidir.

Bir-dışarda-bırak yöntemi, n veri kümesindeki örnek sayısını göstermek üzere, basitçe n-katlı-çapraz-doğrulamayı ifade eder. Bu yöntemde sırasıyla her bir örnek dışarıda bırakılır ve öğrenme metodu kalan örneklerle eğitilir. Test işlemi dışarıda bırakılan

örnek üzerinde yapılır, doğruluk değeri ya pozitif ya negatif çıkacaktır. Her bir örnek üzerindeki testin sonucu, yani n testin sonucu ortalanır ve bu ortalama değer son doğruluk değerini ifade eder.

Kural anlaşılabilirliği: Sınıflandırma kural çıkarımında önemli bir diğer değerlendirme kriteri modelin anlaşılabilirliğidir. Modelin anlaşılabilir olması, kullanıcılar tarafından kolayca anlaşılıp yorumlanabilmesi anlamına gelmektedir.

Hız ve ölçeklenirlik: Hız ve ölçeklenirlik sınıflandırma kural çıkarımında elde edilen kuralların değerlendirilmesinde yararlanılan ölçütlerdir.

Sağlamlık: Sağlamlık, sınıflandırma metotlarının değerlendirilmesinde kullanılan diğer bir ölçüttür ancak doğruluk ve anlaşılabilirlik kadar sık kullanılmamaktadır. Sağlamlık, eğitim verisinde veya başlangıç alan bilgisindeki bozukluklara duyarlılığının ölçüsüdür.

İlgi çekicilik: Yukarıda açıklanan ölçütlerin yanı sıra, sınıflandırma kural çıkarımında dikkate alınması gereken bir diğer ölçüt ilgi çekiciliktir. Son kullanıcı için [56];

- Kurallar kullanıcının bilgi ve beklentilerine ters düşüyorsa (beklenmedik kurallar),
- Kullanıcılar kurallarla bir şeyler yapabiliyor ve fayda sağlayabiliyorlarsa(kullanılabilir),
- Kullanıcının daha önceki bilgisine bilgi ekliorsa (yeni) ilgi çekicidir.

1.10.4. Sınıflandırma Kurallarının Mikro ve Makro Değerlendirmesi

Kuralların değerlendirilmesi sınıflandırma sürecinde büyük önem taşır. Mevcut kural öğrenme algoritmalarının çoğu tekil kural değerlendirme ölçütüne göredir. Bununla birlikte kural çıkarım sisteminin performansı ve sınıflandırma süreci kurallar kümesinin değerlendirilmesinde dikkate alınır. Bu da tekil kural ve kural kümesi değerlendirme ölçütlerinin birleştirilmesini gerektirir. Kural değerlendirme ölçütleri 2×2 durumsallık tablosunda, kuralın önde gelen ve sonra gelen kısmı arasındaki ilişki analiz edilerek belirlenir. Tablo 1.1 örnek bir durumsallık tablosunu göstermektedir.

Tablo 1.1. 2x2 Durumsallık Tablosu [53]

	Sınıf	Yanlış Sınıf	Toplam
Önde gelen	tp	fp	Önde gelen sağlanan
Yanlış önde gelen	fn	tn	Önde gelen sağlanmayan
Toplam	Sınıf sağlanan	Sınıf sağlanmayan	Veri kümesi

Bir kural, verilen bir eğitim örneğini sınıflandırmakta kullanıldığında; pozitif doğru (tp), pozitif yanlış (fp), negatif doğru (tn) ve negatif yanlış (fn) olmak üzere dört durum ortaya çıkar [53]. Pozitif doğru ve negatif doğru, doğru sınıflandırmaları; pozitif yanlış ve negatif yanlış, yanlış sınıflandırmaları ifade eder.

- Pozitif doğru (*tp*): Kural sınıfın pozitif olduğunu tahmin eder ve verilen örneğin sınıfı da pozitiftir.
- Negatif doğru (*tn*): Kural sınıfın negatif olduğunu tahmin eder ve verilen örneğin sınıfı da negatiftir.
- Pozitif yanlış (*fp*): Kural sınıfın pozitif olduğunu tahmin eder fakat verilen örneğin sınıfı negatiftir.
- Negatif yanlış (*fn*): Kural sınıfın negatif olduğunu tahmin eder fakat verilen örneğin sınıfı pozitiftir.

Yao ve Zhou [57], kural değerlendirme ölçütlerini değerlendirilen kural sayısına göre makro ve mikro değerlendirme olmak üzere ikiye ayırmışlardır. İkinci aşamada ise makro değerlendirme ölçütlerini, bir kümede kurallar arasındaki ilişkiye göre çakışan ve çakışmayan kurallar olarak iki gruba sınıflandırmışlardır.

Mikro değerlendirme tekil kurallar üzerinedir. Mevcut birçok kural değerlendirme ölçütü mikro değerlendirme için önerilmiştir. Bu ölçütler kural üretiminde durdurma kriterinin belirlenmesinde ve sınıflandırma amacıyla yüksek kaliteli kurallar çıkarılmasında kullanılır. Bununla birlikte, tekil kurallara göre değerlendirme çakışan sonuçların ortaya çıkmasına neden olabilir.

Makro değerlendirme bir kural kümesi üzerinedir. Bir karar vermek için birden çok kural olduğundan dolayı mikro değerlendirmeye göre daha karmaktır. Makro değerlendirmede, ele alınan uzayda eğer bir nesne kural kümesinde en fazla bir kuralı sağlıyorsa kurallar çakışmayan kurallardır, eğer birden fazla kuralı sağlıyorsa çakışan kurallardır. Çakışan kuralları da tutarlı ve çelişen kurallar olmak üzere ikiye ayırmışlardır. Ele alınan uzayda eğer bir nesne bir veya daha fazla kuralı aynı sınıfla sağlıyorsa kurallar tutarlıdır, eğer birden çok kuralla en az iki farklı sınıfı sağlıyorsa çelişen kurallardır.

Mikro değerlendirme ölçütleri tekil kuralların gücünü göstermek için tasarlanmıştır. En çok kullanılan mikro performans değerlendirme ölçütlerine ilişkin formüller (1.1.)-(1.4.)'de verilmiştir.

$$\text{Doğruluk} = \frac{tp+tn}{tp+tn+fp+fn} \quad (1.1)$$

Doğruluk ele alınan kuralın veri kümesini ne derecede yansıttığını ölçen bir ölçüttür ve veri kümesindeki toplam örnek sayısının doğru sınıflandırılan örnek sayısına bölünmesi ile bulunur.

$$\text{Güvenilirlik} = \frac{tp}{tp+fp} \quad (1.2)$$

Sınıflandırmada güvenilirlik, kuralın önde gelen kısmını sağlayan nesneler içinde hem sınıfı hem de önde gelen kısmı sağlayan nesnelerin oranını ifade eder ve 0 ile 1 arasında değerler alır.

$$\text{Destek} = \frac{tp}{tp+fn} \quad (1.3)$$

Destek ölçütü bir kuralın kabul edilebilirliğinin ölçüsüdür ve 0 ile 1 arasında değerler alır. Sınıflandırma problemlerinde destek sınıfı doğru sağlayan nesneler içinde önde gelen kısmı da sağlayan nesnelerin oranını verir.

$$\text{Genellik} = \frac{tp+fp}{N} \quad (1.4)$$

Genellik, tüm veri kümesinde kuralın önde gelen kısmını sağlayan parçayı ifade eder. Güvenilirlik ve destek gibi genellik değerlendirme ölçütü de (0-1) arasında değerler alır. Mikro kurallarda karmaşıklık ölçütü olarak değişken sayısı kullanılabilir. Bu ölçüt ise kuralın önde gelen kısmında bulunan değişken sayısı ile belirlenir.

Makro değerlendirme, sistemin her bir tekil kuralının performansını değerlendirmek yerine tüm kural çıkarım sisteminin performansını değerlendirmeye odaklanır. Makro performans değerlendirme ölçütleri şu şekilde hesaplanabilir. Doğruluk;

$$\text{Doğruluk} = \frac{\text{Kural Kümesi tarafından doğru sınıflandırılan örnek sayısı}}{\text{Veri kümesindeki örnek sayısı}} \quad (1.5)$$

Makro değerlendirmede doğruluk kural kümesinin doğruluğunu ifade eder ve veri kümesindeki toplam örnek sayısının doğru sınıflandırılan örnek sayısına bölünmesi ile bulunur. Güvenilirlik;

$$\text{Güvenilirlik} = \frac{\text{Kural kümesi tarafından doğru sınıflandırılan örnek sayısı}}{\text{Kural kümesi tarafından sınıflandırılan örnek sayısı}} \quad (1.6)$$

Güvenilirlik kural kümesi tarafından sınıflandırılan nesneler içinde doğru sınıflandırılan nesnelerin oranını verir. Destek ise tüm nesneler içinde kural kümesi tarafından doğru sınıflandırılan nesnelerin oranını verir (Denklem 1.7);

$$\text{Destek} = \frac{\text{Kural kümesi tarafından doğru sınıflandırılan örnek sayısı}}{\text{Veri kümesindeki örnek sayısı}} \quad (1.7)$$

Bir kural kümesinin genelliği, veri kümesindeki nesneler içinde kural kümesi tarafından kapsanan nesnelerin oranını gösterir ve aşağıdaki formül ile hesaplanır;

$$\text{Genellik} = \frac{\text{Kural kümesi tarafından kapsanan örnek sayısı}}{\text{Veri kümesindeki örnek sayısı}} \quad (1.8)$$

Makro değerlendirmede karmaşıklık ölçütü olarak değişken sayısı kural kümesindeki değişken sayısı hesaplanarak bulunur. Kural sayısı ise kural kümesindeki kural sayısı ile ifade edilir.

1.10.5. Kural Gösterimi

Sınıflandırma kural çıkarımında ilk verilmesi gereken karar bir çözümde kaç kuralın kodlanacağıdır [58]. Bu problemle ilgili iki yaklaşım söz konusudur: *Michigan* yaklaşımı ve *Pittsburgh* yaklaşımı. Michigan yaklaşımında her bir çözüm dizisi bir kural içerirken Pittsburgh yaklaşımında birçok kural bir diziyi meydana getirir.

Pittsburgh yaklaşımı, bütün kural kümesini değerlendirme yeteneğinden dolayı sınıflandırmaya daha uygundur ve bunun sonucu olarak kural etkileşimlerini dikkate alır. Bu yaklaşımla elde edilen kuralların kalitesinin değerlendirilmesi kolaydır.

Michigan yaklaşımı, karmaşık olmayan çözümler içerdiğinden dolayı düşük işlem zamanı avantajına sahiptir. Bununla birlikte, bir anda bir kuralı ele almasından dolayı kural etkileşimlerini dikkate almaz ve bu da kural kümesinin tahminleyici doğruluğunun temel engelidir.

Çözüm dizisinde tek bir kural mı yoksa çoklu kurallar mı kodlanacağına karar verildikten sonra, dizilerin nasıl kodlanacağına karar verilmelidir. Yüksek düzeyli kural gösterimi kullanılabileceği gibi, ikili kodlama en çok kullanılan ve en basit kodlama şeklidir [54]. Sembolik değişkenler (kesikli değerler içeren değişkenler) için bir yaklaşım her bir değere bir bit kullanmaktır. Bu durumda N değer içeren bir değişken N bitlik bir alt dizi ile gösterilecektir. Örnek olarak, {kırmızı, beyaz, siyah, sarı} değerlerini içeren bir renk değişkeni ele alınırsa “1010” alt dizisi rengin kırmızı veya siyah olduğunu temsil edecektir. Bu yaklaşımda, bir değişkenin eğer tüm alt dizilerinin bit değerleri 1 ise, bu değişkenin test neticesi sürekli doğru dönecektir, başka bir ifade ile bu değişken kuralın önde gelen kısmının bir parçası olmayacaktır. Bu kuralın genelliği dikkate alındığında kritik bir faktördür. Bazı değişkenlerin çıkarılması daha basit kurallar ortaya çıkarır ve çakışmaları engeller.

İkili kodlama için diğer bir yaklaşım ise bir değişkenin indeks değerini ikili formda göstermektir. Bu gösterimle, özellikle çok sayıda eleman içeren değişkenler için kısa alt diziler elde edilir. Bu yaklaşımda kuraldan değişken çıkarmak için önemsizlik biti gibi farklı bir teknik kullanmak gerekir.

Sürekli (sayısal) değişkenler için genellikle bir kesiklendirme tekniği gerekir. Eğer değişken değerleri tamsayı veya ondalıklı kısımlarda sabit sayı indeksli ise değişkenin

ikili değere doğrudan dönüştürülmesi mümkündür. Kullanılabilecek çok basit kesiklendirme teknikleri olmasına rağmen, kesiklendirme işlemi çok karmaşık olabilir ve kuralların başarısını etkileyebilir. Danışmanlı ve Danışmansız kesiklendirme olmak üzere iki tür kesiklendirme vardır. Danışmansız kesiklendirmede sınıf bilgisi kullanılmadan değişken değerleri kesiklendirilir, danışmanlı kesiklendirmede ise sınıf değerleri de kesiklendirme işlemine dâhil edilerek değişken kesiklendirilir [46].

1.10.6. Uygunluk Fonksiyonu

Sınıflandırmada verilmesi gereken ikinci en önemli karar kullanılacak olan uygunluk yani amaç fonksiyonudur. Sınıflandırma kural çıkarımı için uygunluk fonksiyonu çoğu zaman sınıflandırma sürecinin amacına ve kullanılan kural gösterim yapısına göre seçilir. Tahminleyici doğruluğun en büyüklenmesi çoğu zaman ilk amaçtır ve tahminleyici doğruluk basit bir şekilde, doğru sınıflandırılan örneklerin eğitim kümesinde yer alan örneklere oranı olarak hesaplanır. Spears ve De Jong [59], doğru sınıflandırılan örneklere doğrusal olmayan bir eşik sunan aşağıdaki uygunluk fonksiyonunu önermiştir.

$$\text{Uygunluk (dizi } i) = \text{doğru sınıflandırılan örnekler yüzdesi}^2 \quad (1.9)$$

Ancak bu fonksiyon tekil kuralların neden olduğu yanlış sınıflandırmaları etkin bir şekilde cezalandıramaz ve bu da performans problemine neden olabilir. Bu, özellikle kurallar Michigan yaklaşımı ile gösterildiğinde doğrudur, çünkü kurallar arasında etkileşimler dikkate alınmaz. Bu problemi azaltmak için, sadece pozitif doğru ve negatif doğruyu (doğru sınıflandırılan örnekler) değil, aynı zamanda pozitif yanlış ve negatif yanlış (yanlış sınıflandırılan örnekler) değerlerinin farklı varyasyonlarını dikkate alan uygunluk fonksiyonları kullanılabilir.

İlgi çekiciliğin ölçülmesi genellikle daha karmaşıktır. Mesela Noda vd. [60], kullanıcıların atadığı ağırlıklara bağlı olarak kuralın önde gelen kısmında yer alan her bir değişkenden bilgi kazancı hesaplayan bir yöntem önermiştir.

1.11. Gen Ekspresyonu (İfade)

İnsan vücudundaki hücrelerin hepsi aynı genetik materyali içerse de her hücrede aynı genler etkin değildir. Hangi genin aktif olup hangisinin olmadığı bilgisi biyologlara, normalde bu hücrelerin nasıl işlediği ve bazı genler doğru çalışmadığı zaman hücrenin

bundan nasıl etkileneceği bilgisini vermektedir. Bu aktiflik bilgisine hücrenin gen ifadesi denmektedir [61].

Her hücremiz aynı genetik bilgiyi içermektedir. Ancak, deri hücreleri, böbrek, ciğer, kan, beyin hücreleri birbirinden farklıdır. Bu farklılıklar genlerin farklı hücrelerde farklı farklı ifade edilmelerinden kaynaklanmaktadır [62].

Geçmişte biyologlar birkaç genin gen ifade verisini aynı anda ölçebilirken, DNA (Deoksiribonükleik asit) mikroçip teknolojisinin gelişimiyle binlerce genin gen ifade verisi eşzamanlı olarak ölçülebilmeye başladı [61].

İnsan genom projesiyle birlikte ortaya çıkarılan gen dizilerinden sonra, yeni amaç; bu genlerin nasıl ifade gösterdiklerini bulmak yani mRNA profillerini çıkarmak diğer genlerle beraber nasıl bir ilişki içinde olduklarını göstermek ve böylece belirli hastalıklarda hangi genlerin rol oynadığını ortaya çıkarmak olmuştur. Bir hücrenin tipi veya içinde bulunduğu evre o hücrenin mRNA ifadesi ile ilgilidir. Daha önceden tanımlanmamış genlerin ifade düzeyleri incelenerek ve diğer bilinen genlerin mRNA ifadeleri ile karşılaştırılarak o genlerin işlevleri hakkında bilgiler edinilmeye çalışılmaktadır [62].

1.12. Mikroarray Teknolojisi

Bilgisayar teknolojisinin moleküler biyolojiye paralel olarak hızla gelişmesi, iki disiplini birbirine yaklaştırmıştır. Böylece, biyoteknolojinin kavramsal olarak ulaşabileceği son noktalardan biri olan mikroarray (gen çip) ortaya çıkmıştır. Mikroarray tekniğinin ilk girişimleri Shalon ve Schena tarafından gerçekleştirilmiştir [63].

Mikroarray teknolojisi, geleneksel yöntemlerin aksine aynı anda pek çok genin araştırılmasına olanak sağlayan yeni analitiksel metotlar sunmaktadır. Bu teknik özellikle tıp alanı olmakla birlikte; biyoloji, mikrobiyoloji, genetik gibi pek çok alanda da kullanılmaktadır. Farklı amaçlara uygun pek çok mikroarray tipi mevcuttur [64].

DNA mikroarray teknolojisi; binlerce genin ifade düzeylerini (ekspresyon seviyesi) ve DNA değişimlerini eşzamanlı olarak inceleme yöntemidir [65].

Moleküler biyolojideki geleneksel metotlarda genellikle “bir deneyde bir gen” ilkesi geçerlidir. Yani; aynı çalışmada gen fonksiyonlarının tamamını görmek geleneksel yöntemlerle zordur. Gen çip teknolojisi olarak da adlandırılan yeni yöntemler ile bütün genomun görüntülenmesi sağlanır ve bu metot aynı anda binlerce genin birbirleriyle olan etkileşimlerinin görülmesine olanak tanır. Binlerce genin tek bir çalışmada incelenmesi esasına dayanan mikroarray teknolojilerinin ilk girişimleri 1990’ların başında Schena tarafından gerçekleştirilmiştir. Bu teknoloji tek seferde çok sayıda genin ekspresyon düzeyinin incelenmesine olanak tanıyan bir teknolojidir ve binlerce DNA’nın aynı anda analiz edilebilmesini sağlamaktadır [63].

DNA mikroarray teknolojisi, birçok genin aktivitesinin aynı zamanda izlenebilmesi; hızlı bir yöntem olması; hasta ve sağlıklı hücrelerdeki genlerin aktivitelerinin karşılaştırılmasını sağlaması ve hastalıkları alt-gruplar halinde kategorize edebilme gibi avantajlarının yanı sıra, tek bir seferde çok fazla veri analizi yapıldığından, tüm sonuçların analizinin zaman alması; sonuçların yorumlamak için çok kompleks olabilmesi; sonuçların yeterince kantitatif olmaması ve oldukça pahalı bir teknoloji olması gibi bazı dezavantajlara da sahiptir [66].

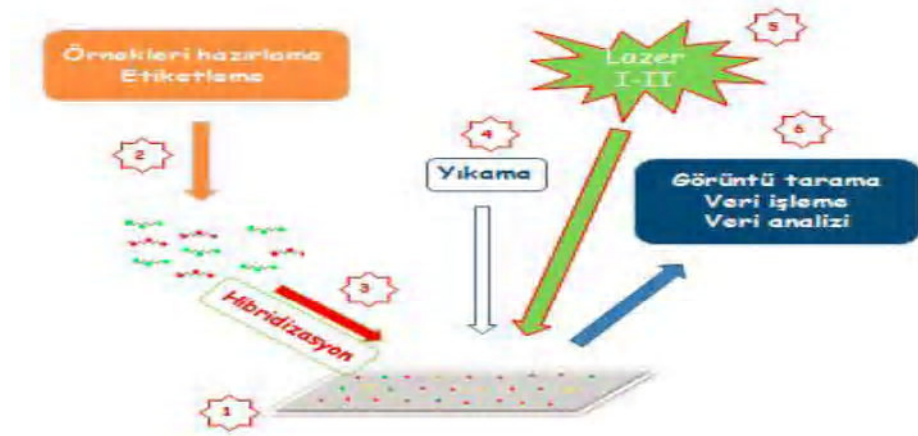
Gen ekspresyonunun mikroarray ’ler kullanılarak ölçülmesi biyoloji ve tıbbın pek çok alanında uygulanabilirlik arz etmektedir. Örneğin microarray ’ler hastalıklı ve normal bir hücrede gen ekspresyonunun karşılaştırılması suretiyle hastalık ile alakalı genlerin belirlenmesinde kullanılabilir. Diğer bir çalışmada beyin dokusundan alınan postmortem prefrontal kortekslere major depresyon için DNA mikroarray analizi yapılmıştır. Major depresyon hastaları ve kontrol grubunun gen ekspresyon modelleri karşılaştırılmış ve major depresyonda 99 genin farklı şekilde eksprese olduğu görülmüştür [67].

DNA mikroarray (gen çipi, DNA çipi ya da biyoçip olarak da bilinir) ekspresyon profili oluşturmak, başka bir ifade ile binlerce genin ekspresyon düzeyini aynı anda izlemek amacı ile cam, plastik veya silikon çip gibi katı bir yüzeye tutturularak sıralı bir şekilde (**array**) oluşturulmuş mikroskobik DNA spotlarıdır [67].

Yöntem ve ortaya çıkış metotlarında bazı küçük farklar olmasına rağmen “DNA array”, “DNA çip” ve “mikroçip” gibi tanımlamalar da benzer uygulamaları ifade etmek için kullanılmaktadır [68].

1.12.1. Mikroarray'lerin Üretilmesi ve Çalışma Mantığı

90'lı yılların ortalarında geliştirilmeye başlanan, günümüzde hücre ve dokulardaki gen ekspresyonlarının topluca incelenmesine olanak sağlayan yeni ve güçlü bir teknolojidir. Bir mikroorganizmanın tüm genleri mikroskop lamı ölçeğinde bir alana yerleştirilebilir ve binlerce genin ekspresyon seviyeleri tek bir deneyde eş zamanlı olarak çalışılabilir. DNA mikroarray analizi aşağıdaki basamaklardan oluşmaktadır (Şekil 1.7):

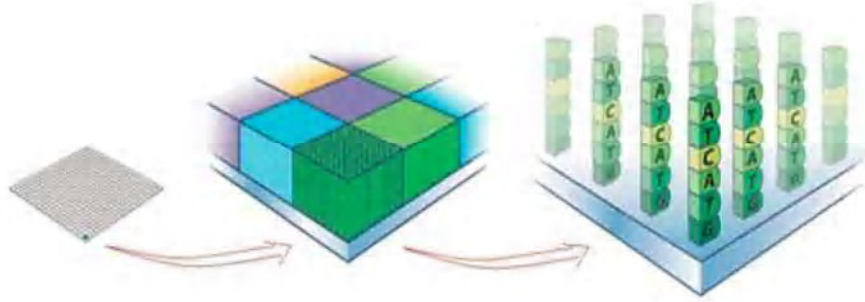


Şekil 1.7. DNA mikroarray işlem basamakları [63]

1. Mikroarray 'in hazırlanması
2. Test edilecek örneklerin hazırlanıp etiketlenmesi
3. Hibridizasyon
4. Yıkama
5. Etiketlerin uyarılması
6. Görüntü tarama / Veri işleme-analizi

1. Çipin Hazırlanması: Mikroarray'de destek malzemesi olarak genellikle cam, plastik silikon vb katı yüzeyler kullanılır. Bu katı yüzeyler, elektrostatik etkileşimi arttırarak nükleik asitlerin bağlanmasını kolaylaştırmak amacıyla spotlama öncesi işlem

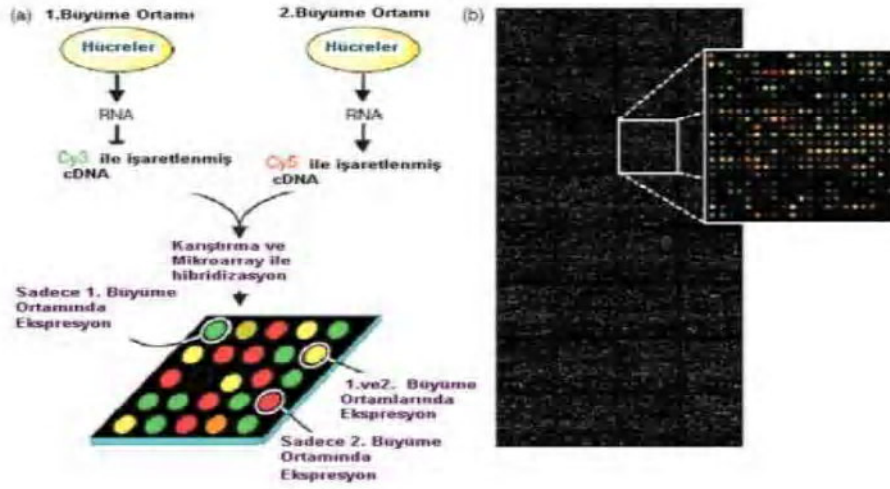
geçirilir. DNA mikroarray çipleri *cDNA (komplementer DNA) çipler*dir. *cDNA çipler*, cDNA klon kütüphanesinden elde edilmiş 500-2000 baz çifti uzunluğundaki cDNA'ların veya Expressed Sequenced Tagged (EST) klonlarının özel yazıcılar aracılığıyla (çoklu uçlu mekanik yazıcılar, ink-jet yazıcılar vb.) slaytlara spotlanması ile oluşturulur. cDNA çipler daha çok ekspresyon analizleri için kullanılır. Çipler hangi yolla elde edilirse edilsin, sonuçta her probunda çok sayıda homolog DNA'nın bulunduğu, farklı problemlerin farklı DNA zincirlerini içerdiği bir platform elde edilmiş olur ve uygun şekilde hazırlanmış örnekler ile hibridizasyona hazırdır (Şekil 1.8).



Şekil 1.8. DNA mikroarray'in görünümü [63]

2. Örneklerin Hazırlanması ve Etiketlenmesi: DNA mikroarray yönteminde örnek hazırlanması önemli bir aşamadır ve çalışma için her iki durum örneğinden izole edilen hedef mRNA'lar, revers transkriptaz aracılığı ile cDNA 'ya çevrilir. Elde edilen cDNA'lar radyoaktif veya floresan belirteçler ile etiketlenir. Radyoaktif işaretleme için genellikle ^{33}P gibi radyoizotoplar kullanılırken floresan işaretleme için Siyanin (Cy3, Cy5) gibi siyanin boya ları kullanılır. Bunlardan yeşil renk veren Cy3 ve kırmızı renk veren Cy5 çeşitli avantajları nedeniyle en sık kullanılan boyalardır.

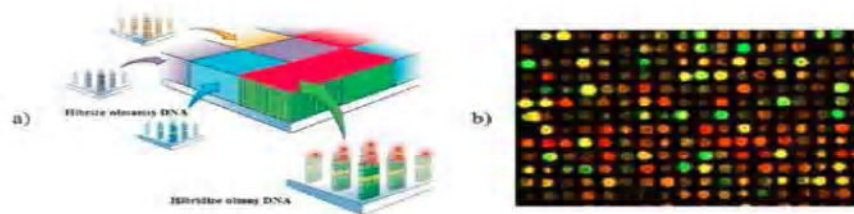
3. Hibridizasyon: Etiketlenen cDNA' lardan oluşan karışım hibridizasyonun sağlanması için mikroarray üzerine inkübe edilir (Şekil 1.9-a). Örnekte mikroarray problemleri ile komplementer dizinler bulunuyorsa bu süreç sonunda problemlere hibridize olurlar (Şekil 1.9-b).



Şekil 1.9. Örneklerin hazırlanması, karıştırılması ve hibridizasyon [93]

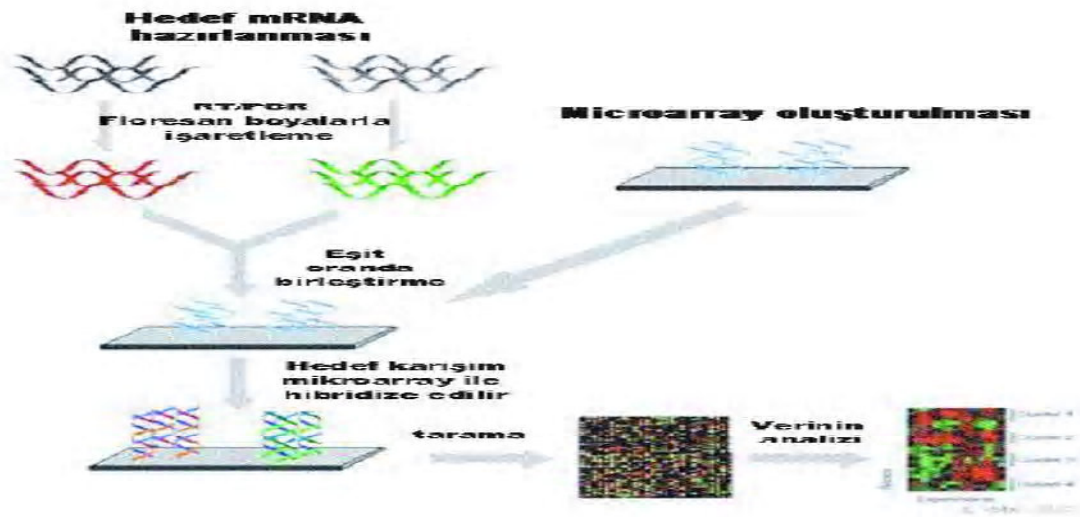
4. Yıkama: Hibridizasyon aşaması sonrası ortamda bulunan ve prob ile bağlanmamış dizinler, nonspesifik sinyal odakları gibi yapıların uzaklaştırılması için yıkama işlemi gereklidir. Reaksiyonun doğru değerlendirilebilmesi için önemli bir aşamadır.

5. Etiketlerin uyarılması: Örnekteki dizinler ile hibridize olmuş, yıkama işlemi ile bağlı olmayan materyalden temizlenmiş mikroarray'de hibridizasyonun görünür hale getirildiği aşamadır. Amaç mikroarray'e hibridize olan dizinlerin görünür veya değerlendirilebilir hale getirilmesidir (Şekil 1.10-a). Süreç, kullanılan etiketin niteliğine ve tarayıcıların türüne uygun uyarı kaynakları kullanılarak gerçekleştirilir.



Şekil 1.10. Etiketlerin uyarılması ve floresans [63]

6. Görüntü Tarama / Veri İşleme-Analizi: Mikroarray üzerindeki floresan ya da radyoaktif sinyalin toplandığı aşamadır. Bu işlem için, mikroarray spotlarındaki ışık yoğunluğunu ölçen konfokal veya charge coupled device (CCD) gibi floresan sinyal detektörleri ya da radyoaktif sinyaller için phosphorimager dedektörler kullanılır. Tüm detektörler özel bir bilgisayara, yazılım programı ile bağlantılıdır ve detektörlerden gelen çok sayıdaki veri bu yazılımlar tarafından değerlendirilir. Tarama sırasında detektörler, hibridize olmuş örneklerin oluşturduğu her mikroarray probundaki sinyal yoğunluğunu belirler. Tarama sonuçları yazılım aracılığıyla işlenerek anlamlı veriler haline getirilir. Eğer örnek içerisindeki Cy3 ile işaretli olan dizinler, hedef prob’ da ki cDNA’ya hibridize olduysa o prob yeşil renk, Cy5 ile işaretli olan dizinler hibridize oldu ise o prob kırmızı renk yayacak, eğer hem Cy3 hem de Cy5 işaretli dizinler eşit miktarda hibridize olduysa prob sarı renk yayacaktır (Şekil 1.10-b). Verilerin işlenmesi ve analizi, normalizasyon, filtreleme, kümeleme, örüntü tanımlama gibi çeşitli süreçleri içermektedir [63]. Mikroçip yönteminin şematik gösterimi ise Şekil 1.11’de gösterilmiştir.



Şekil 1.11. Mikroçip yönteminin şematik gösterimi [63]

1.12.2. Mikroarray'ların Kullanım Alanları

- Gen ifade profillerinin çıkarılması
- Polimorfizm analizleri
- Mutasyon analizleri
- DNA dizi analizi
- Evrimsel çalışmalar
- Potansiyel terapötik ajanların bulunması, geliştirilmesi, optimizasyonu ve klinik değerlendirmesi

Microarray tekniğinin en göze çarpan kullanım alanı gen ifadesindeki farklılıkların ölçülmesidir. Genomik DNA'dan transkripsiyona uğrayan genlerin tamamına transkriptom veya gen ekspresyon profili adı verilmektedir. Genom hücreden hücreye sabit olmasına karşın gen ekspresyon profili hücrenin içinde bulunduğu koşullara göre hızla değişebilen bir yapıdadır. Çeşitli koşullarda genlerin ekspresyon düzeylerindeki değişimler takip edilerek bu genlerin kodladığı proteinlerin fonksiyonları hakkında önemli ipuçlar elde edilebilir.

DNA mikroarray, kanser hücrelerindeki gen ekspresyon farklılaşmasının karakterize edilmesinde oldukça kapsamlı bir şekilde kullanılır. Örneğin, insan akciğer epitelyum hücreleri tarafından ekspresyonu yapılan yaklaşık 5500 gen, akciğer kanserli doku genleri ile karşılaştırılabilmektedir. Böylece kanserleşme sürecinde rol oynayan gen takımları hakkında bilgi edinmek mümkündür. Kanser tedavisindeki bir diğer önemli rolü, gen ekspresyon durumlarına göre kanserleşen hücrelerin sınıflandırılabilmesidir.

1.12.3. Mikroarray Teknolojisinin Avantajları

Gen ifadesi modellerinin genel bir görüntüsünü elde etmeyi mümkün kılar. Belli bir hücre türünün belirli bir ortam için gen ifadesi profili belirlenip, bu profil farklı hücre tipleri ve farklı çevre koşullarındaki gen ifadesi profilleriyle bu yöntemle karşılaştırılabilir. Kısa sürede ve oldukça pratik olarak birkaç bin genin analizini yapmak mümkündür. Otomasyona dayalı bir sistem olduğu için insan kaynaklı hataların ortaya çıkma ihtimali oldukça düşüktür [63].

2. BÖLÜM

YÖNTEM

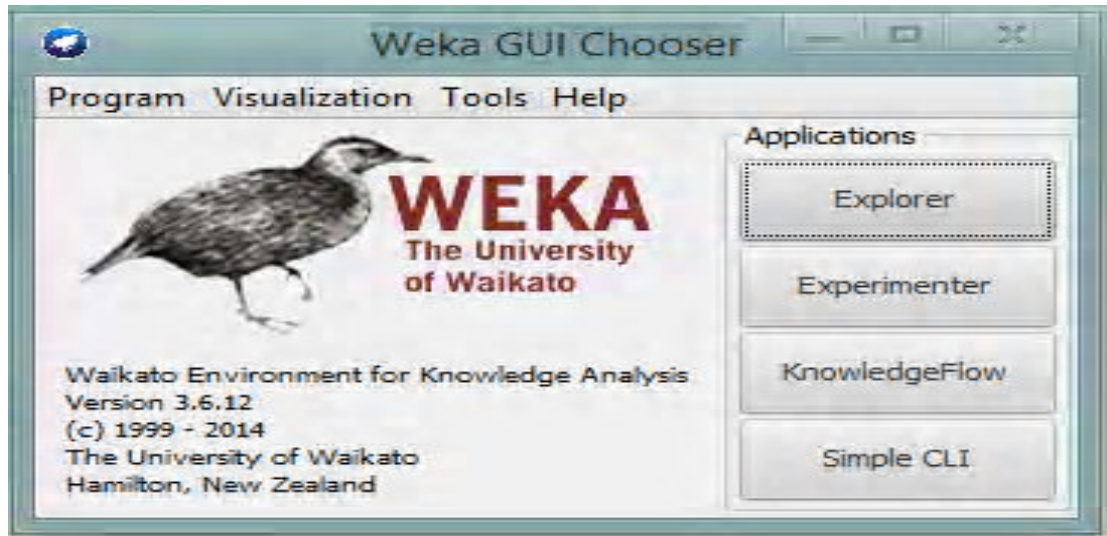
Çalışmada mikroarray gen ekspresyon verilerinin boyutları, öznelilik seçme yöntemleri kullanılarak azaltılmış böylece düşük boyutlu mikroarray gen ekspresyon verilerinin sınıflandırılması için açık kaynak kodlu Weka programı ve Ant-Miner programları kullanılmıştır. Bu programlar sayesinde mikroarray verilerini sınıflandırmak için istatistiksel ve yapay zekâ tabanlı yöntemlerden yararlanılmış ve bu yöntemlerin mikroarray gen verileri üzerinde başarımları birbirleriyle karşılaştırılmıştır.

Böylece literatürde yaygın kullanılan istatistiksel yöntemlerin yanında yeni yöntemlerden olan yapay zekâ tabanlı sınıflandırma yöntemleri de bu çalışmada kullanılmıştır.

2.1. Weka Programı

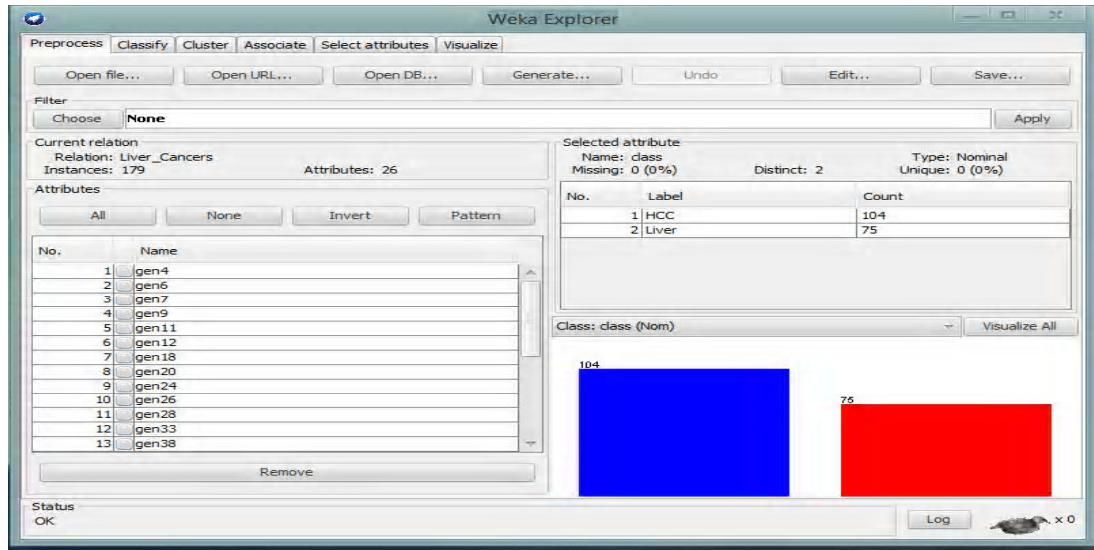
Weka, Yeni Zelanda'daki Waikato Üniversitesi tarafından geliştirilmiş, makine öğrenme algoritmalarını bir arada barındıran, işlevsel bir grafik arabirimine sahip, açık kaynak kodlu, Java programlama diliyle geliştirilmiş bir veri madenciliği programıdır [69]. Weka çeşitli veri ön işleme, sınıflandırma, regresyon, kümeleme, ilişkilendirme kuralları ve görselleştirme araçları içerir. Algoritmalar veri kümesine doğrudan veya Java kodundan çağrılarak uygulanabilir [70,71]. Aynı zamanda yeni makine öğrenme algoritmaları geliştirmek için de uygundur.

Weka, ham verinin işlenmesi, öğrenme metotlarının veri üzerinde istatistiksel olarak değerlendirilmesi, ham verinin ve ham veriden öğrenilerek çıkarılan modelin görsel olarak izlenmesi gibi veri madenciliğinin tüm basamaklarını destekler. Geniş bir öğrenme algoritmaları yelpazesine sahip olduğu gibi pek çok veri ön işleme filtreleri içerir. Explorer, Experimenter, Knowledge Flow ve Simple CLI adı verilen 4 temel uygulamayı barındırır.



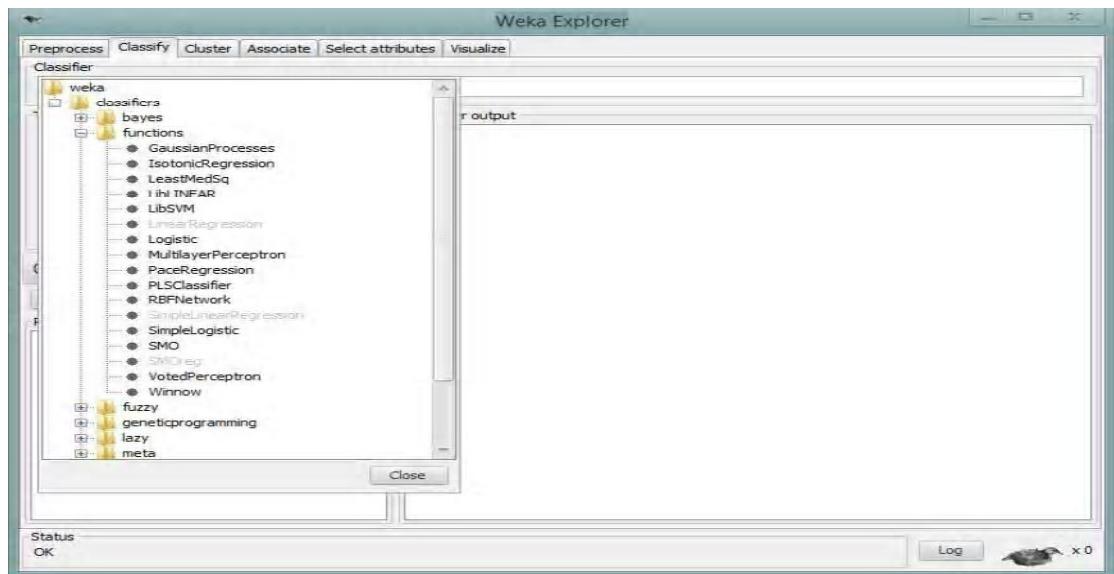
Şekil 2.1. Weka kullanıcı ara yüzü

Program çalıştırıldığında Şekil 2.1'deki kullanıcı ara yüzü ekrana gelir. Bu ara yüz ekranında “Program”, “Visualization”, “Tools” ve “Help” menülerinden oluşan ana menü ve “Explorer”, “Experimenter”, “Knowledge Flow” ve “Simple CLI” kısımlarından oluşan “Applications” bölümleri bulunmaktadır. “Applications” bölümünde yer alan “Explorer” seçeneği mevcut veri üzerinde yapılabilecek uygulamaları içeren Genel bir grafiksel kullanıcı ara yüzünü içerir. “Experimenter” seçeneği ise bir veya daha çok veri kümesi üzerinde bir veya daha çok algoritmanın uygulanabilmesine ve gözlemlenebilmesine olanak sağlayan bir kullanıcı ara yüzüdür. “KnowledgeFlow” seçeneği ise Matlab içindeki Simulink veya National Instruments firmasına LabView programı gibi sürükle bırak özelliğine sahip olan “Explorer” penceresi gibi çalışmaktadır. Kullanıcı tercihinine bağlı olarak “Explorer” ya da “Knowledge Flow” seçeneklerini kullanabilir. Son seçenek olan “Simple CLI” ise komut ekranı aracılığıyla işlem yapmayı sağlar.



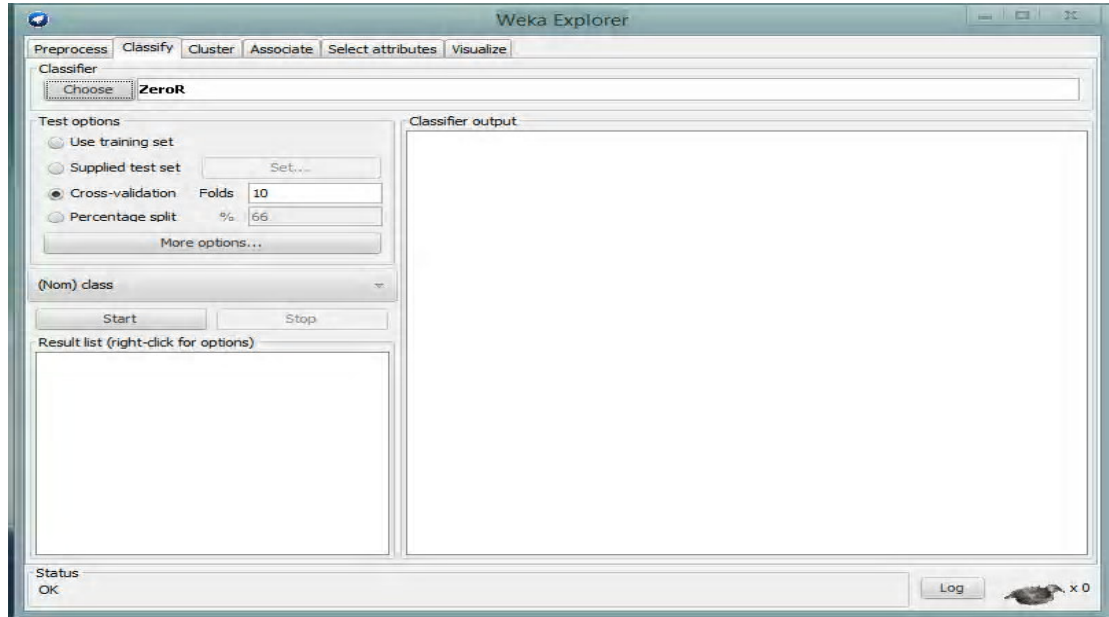
Şekil 2.2. Explorer seçeneğindeki sekmeler

Şekil 2.2’de Weka veri madenciliği programının “Explorer” seçeneğindeki sekmeler görülmektedir. Bu sekme altında sınıflandırma, kümeleme, ilişkilendirme, özellik seçme ve görselleştirme gibi menüler bulunmakta ve ayrıca bu sayfada verinin attribute’leri ve sınıfları hakkında bilgi vermektedir.



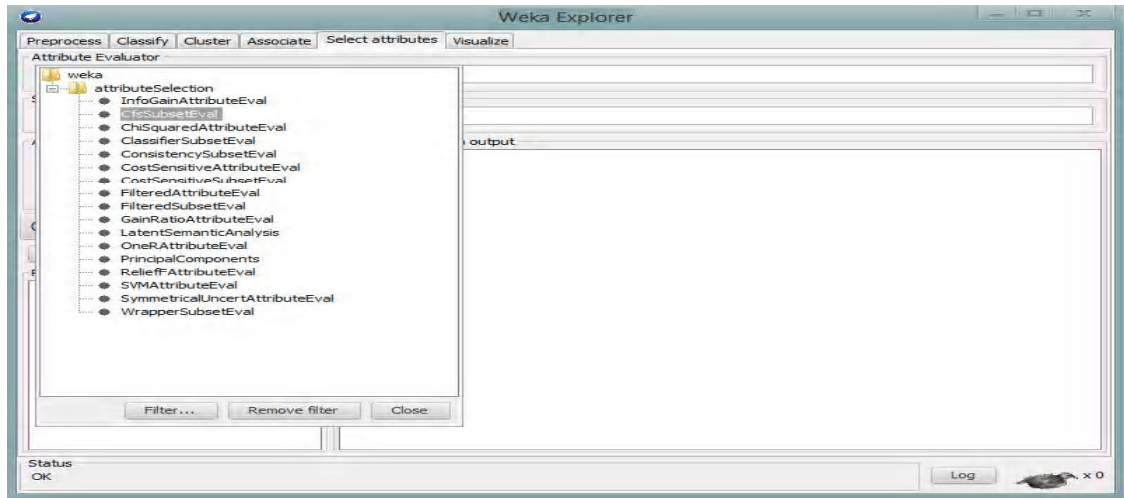
Şekil 2.3. Weka “Classify” sekmesi

Şekil 2.3’de “Classify” sekmesinde çeşitli sınıflandırma algoritmalarını barındıran bir kullanıcı ara yüzü ekranı gelir. Bu ara yüzün altındaki “Bayes Net”, “LibSVM”, “IBk”, “OneR” ve “J48” gibi birçok sınıflandırıcı algoritmasından birisi seçilir.



Şekil 2.4. “Test Options” başlığı

Şekil 2.4’de “Test Options” başlığında eğitim setinin tamamının (Use training set), ayrı bir veri setinin (Supplied test set), parçalara ayrılarak birinin (Cross-validation) ve belli bir oranının (Percentage split) test amaçlı kullanılmasına olanak sağlayan seçenekler bulunmaktadır.

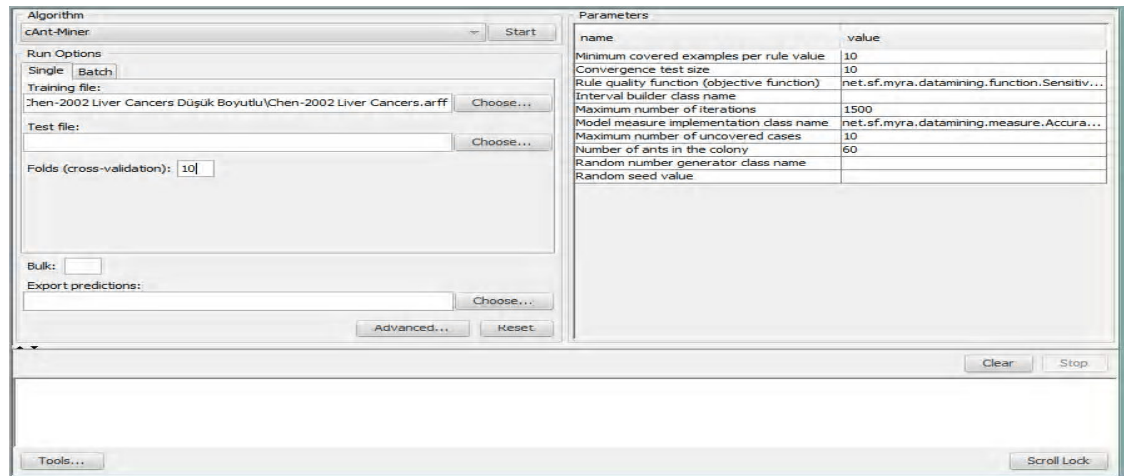


Şekil 2.5. “Select Attributes” sekmesi

Şekil 2.5’de ise “Select attributes” sekmesindeki öz nitelik seçme yöntemlerinden biri kullanılarak boyut indirilmesi yapılır.

2.2. Açık Kaynak Kodlu Ant-Miner Programı

Java’da yazılmış açık kaynak kodlu karınca koloni algoritmasının veri madenciliğine uyarlanmış bir Framework’üdür ve sınıflandırma problemlerinde Karınca Koloni Algoritmasını (KKA) destekleyen özel bir veri madenciliği katmanı sağlar.



Şekil 2.6. cAnt-Miner arayüzü

Şekil 2.6’da cAnt-Miner’in ara yüzünde görüldüğü üzere sağ tarafta ayarlanabilir parametreleri ve sol tarafta ise Training dosyasının ve Test dosyasının seçileceği alan görülmektedir. Eğer cross-validation seçilirse veri seti alt kümelere bölünüp kümelerin bir kısmı eğitim bir kısmı da test için kullanılacağından direk “Training files” bölümünde bütün veriyi seçtirip “Test file” kısmını boş bırakabiliriz.

2.3. Öznitelik Seçimi

Öznitelik seçimi tüm öznitelikler içinden azaltılmış ve muhtemelen daha iyi sınıflandırma performansına sahip bir öznitelik alt kümesinin bulunması işidir. mRNA bilgisi kullanılarak doğruluk oranı yüksek ve güvenilir kanser sınıflandırma sonuçları elde etmek söz konusu olduğunda öznitelik seçiminin kritik bir gereklilik olduğu kanıtlanmıştır. Bu çalışma için benzer problemlerdeki başarılarını da göz önünde bulundurarak pek çok öznitelik seçme algoritması değerlendirilmiş ve aralarından Korelasyon Tabanlı Öznitelik Seçim Yöntemi (KTÖS) kullanılmıştır [72,73,74].

2.3.1. KTÖS öznitelik seçimi

Diğer öznitelik seçimi algoritmalarının birçoğunda olduğu gibi KTÖS yöntemi de öznitelik alt kümelerinin bilgi değerlerini ölçen bir fonksiyonun yanında bir arama algoritması kullanır. KTÖS’ ün, öznitelik altkümelerinin değerini ölçerken kullandığı sezgisel yaklaşım her özneliğin sınıf etiketlerini tahmin etmedeki bireysel yeteneklerinin yanı sıra aralarındaki iç korelasyonu da dikkate alır. Bu yaklaşımın temel aldığı hipoteze göre; iyi öznitelik altkümeleri ilgili sınıf ile yüksek, birbirleri ile ise düşük korelasyona sahip özniteliklerden oluşurlar [75].

$$G_s = \frac{k\bar{r}_{ci}}{\sqrt{k + k(k-1)\bar{r}_{ii}}} \quad (2.1)$$

Eşitlik 2.1, çıkış noktası toplanmış varlıkların bulunduğu bir testin güvenilirliğinin tekil varlıkların güvenilirliğine göre ölçülmesinde kullanıldığı test teorisidir [30]. Örneğin bir insanın eğitimindeki başarısının ölçülmesinde birçok farklı becerisinin ölçüldüğü birden fazla farklı testin bileşkesi, sınırlı sayıda becerinin aynı anda ölçüldüğü tek bir testten daha doğru bir ölçüt olabilmektedir. Bu eşitlikte k , veri alt setindeki öznitelik sayısı, r_{ci} ortalama öznitelik korelasyonu ve r_{ii} ortalama öznitelik iç korelasyonudur.

Eşitlik 2.1 aslında tüm değişkenlerin standartlaştırıldığı ‘**Pearson korelasyonu**’dur. Eşitlikteki payın bir grup özneliğin sınıf üzerindeki tahmin yeteneğini, paydanın ise bu özneliklerin arasındaki fazlalığı temsil ettiğini düşünebiliriz. Sezgisel iyilik ölçümü ile alakasız öznelikler sınıf tahmininde kötü oldukları için elenirken, fazlalık öznelikler ise bir veya daha fazla öznelikle yüksek korelasyona sahip oldukları için eleneceklerdir. Şekil 2.7, KTÖS öznelik seçimindeki elemanları göstermektedir [75].



Şekil 2.7. KTÖS öznelik seçimindeki elemanlar [75]

2.4. Arff Dosya Formatı

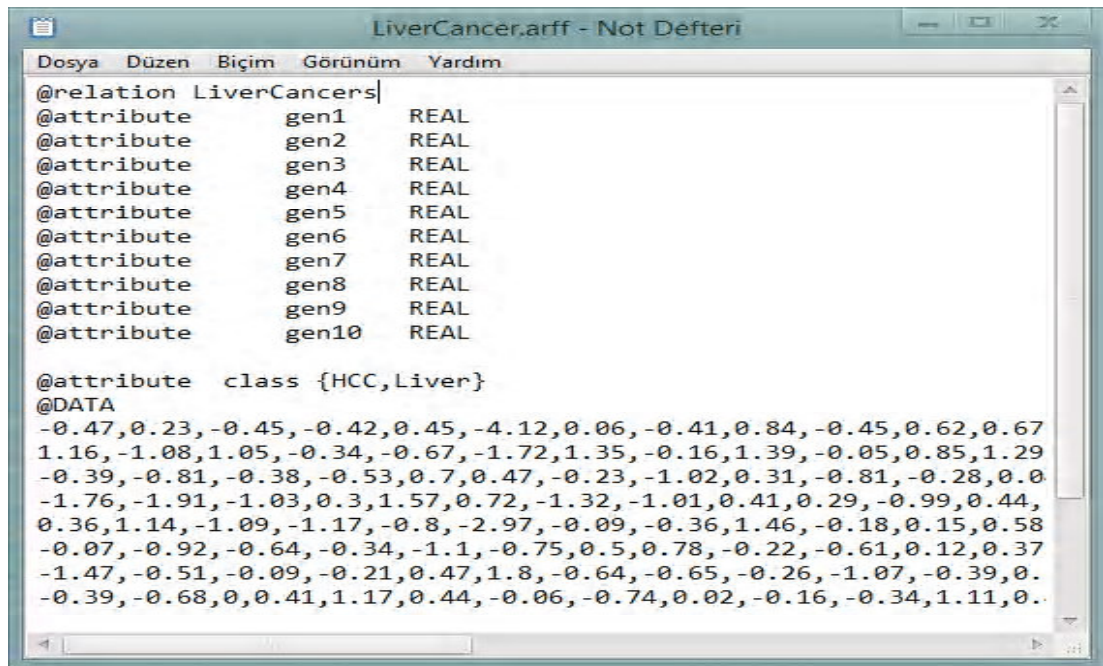
Arff formatı weka programının temel veri kaynağı olup temelde bir metin belgesidir. Metin belgesi içerisinde “%” karakteri ile başlayan satırlar yorum satırlarıdır. Yorum satırları program tarafından değerlendirmeye alınmayan bilgi amaçlı yazılmış kontrol cümleleridir. Genel itibarı ile arff dosyaları üç bölümden oluşur. İlk bölüm dosya içerisindeki veriyi tanımlayan isimlendirme bölümüdür. @relation kelimesi ile veri kaynağına isim verilir.

İkinci bölüm veriyi tanımlayan verinin hangi niteliklerden oluştuğunun tanımının yapıldığı bölümdür. Her nitelik @attribute kelimesinden sonra nitelik adı ve tipi

yazılarak tanımlanır. Veri tipi katar ise string, sayı ise numeric, tarih ise date veri tipleri kullanılır. Nominal değerler için niteliğin alacağı değerler küme parantezi “{ }” içerisinde listelenir.

Üçüncü ve son bölüm ise verilere ait nitelik değerlerinin verildiği veri bölümüdür. Veri bölümü @data belirteci ile başlar ve metin dosyanın sonuna kadar devam eder. Her satırda bir nesneye ait tüm niteliklerin değerleri virgülle ayrılmış şekilde yer alır. Katar tipindeki niteliklerin değerleri çift tırnak (“ ”) içerisinde yazılırlar.

Şekil 2.8’de görüldüğü gibi bir Arff uzantılı dosya ve bu dosya içerisinde verinin nasıl alınacağı görülmektedir. @relation komutuyla Arff uzantılı dosyadaki veri setinin adı belirtiliyor. @attribute komutuyla ise verideki her bir boyutun adı belirtiliyor. REAL komutuyla ise boyutların ne tür bir veri olduğunu simgeliyor. @DATA komutuyla ise veri değerlerinin buradan itibaren başladığını gösteriyor. Her satırda bir nesneye ait tüm öznitelikler, aralarında virgül bulunacak şekilde yer alır.



Şekil 2.8. Arff dosya yapısı

2.5. Sınıflandırma Yöntemleri

2.5.1. Naive Bayes (NB)

Naive Bayes sınıflandırma algoritması; Bayes teorisine dayanan, kolay anlaşılan ve hızlı çalışan bir yöntemdir. Naive Bayes; tüm değişkenlerin birbirinden bağımsız ve hepsinin aynı öneme sahip olduğu varsayımlarına dayanan bir algoritmadır [76, 77, 78].

Naive Bayes; hem tahmin edici hem de tanımlayıcı bir sınıflama tekniğidir. Bağımsızlık varsayımı gerçekte çok nadir görülse de, yani bu varsayım çoğu zaman gerçek dışı olsa da, Naive Bayes'in sınıflamadaki başarısı ve diğer bazı sınıflama algoritmalarından üstünlüğü çeşitli çalışmalarda ortaya konmuştur. Bağımsızlık varsayımı her bir değişkenin tek tek öğrenilmesine olanak vermektedir. Böylece, çok değişkene sahip olan verilerde bile sınıflama işleminin hızlı olmasına olanak sağlamaktadır [78,79].

Naive Bayes sınıflama algoritması; bayes kuralına göre sınıflandırılacak örneğe, en yüksek olasılıkla benzerlik gösteren sınıf seçimi ile yapılır. Bu seçim hesaplanırken önsel olasılıklardan yararlanılır. Naive Bayes sınıflaması, doküman sınıflaması, medikal teşhis gibi konularda sıklıkla kullanılmaktadır.

Naive Bayes sınıflandırması, sınıfları belirli örneklerin özelliklerinin birbirlerinden şartlı bağımsız oldukları varsayımı üzerine dayanır. Bu varsayım, özelliklerin birbirleriyle güçlü bir şekilde bağımlı olduğu gerçek dünya problemleri için uygun olmasa da, problemi basitleştirerek çok boyutluluğun etkisini azaltmaya yardımcı olmaktadır. Özellik vektörü $(x_1, \dots, x_n)$ olan bir X örneği verildiğinde, Naive Bayes sınıflandırıcısı,

$$P(X|C)=P(x_1, \dots, x_n|C) \quad (2.2)$$

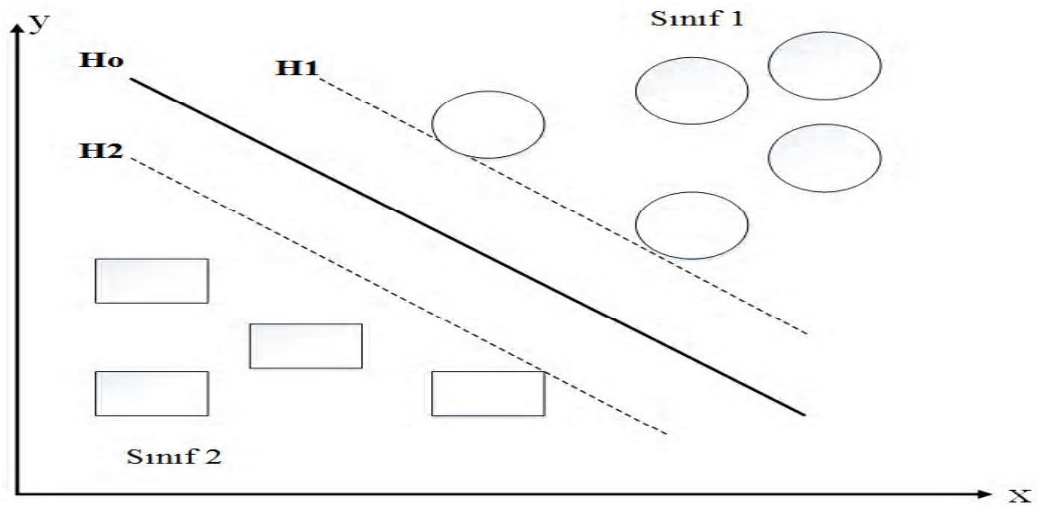
denklemini kullanarak benzerliği en yüksek yapan bir C sınıf etiketi arar [80].

2.5.2. Destek Vektör Makineleri (DVM)

Destek vektör makinesi (DVM), performansı ile dikkat çeken makine öğrenme yöntemlerinde birisidir. DVM, veriyi birbirinden ayıran en uygun doğrusal ya da doğrusal olmayan bir fonksiyon yardımıyla sınıflandırma yapmaktadır.

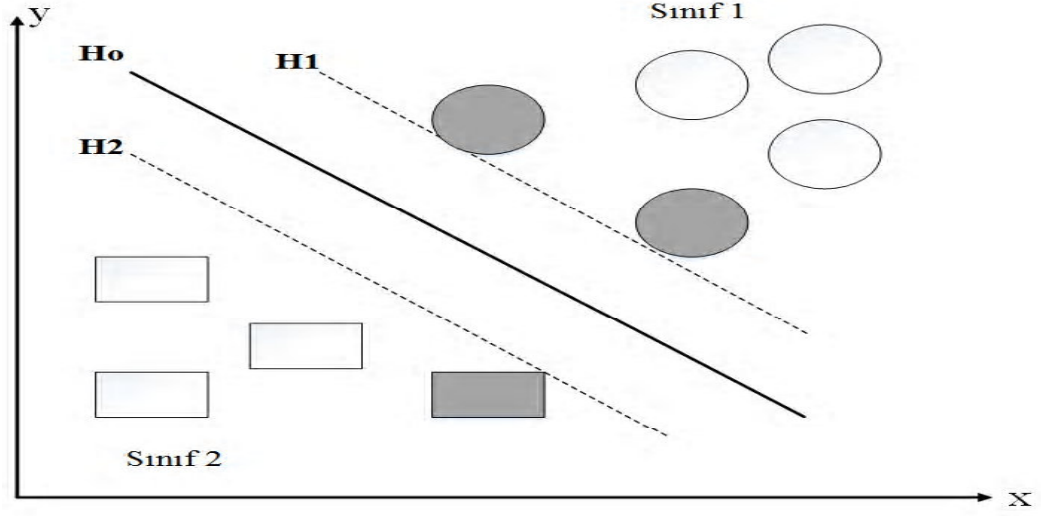
Pek çok makine öğrenmesi veya istatistiksel veri analizi yönteminde olduğu gibi DVM de önceden var olan verilerden öğrenme yöntemini kullanır. DVM'ler, danışmanlı veya yarı danışmanlı çalışabilen bir sınıflandırma yöntemidir. DVM'ler diğer makine öğrenme yöntemlerinin aksine yerel en iyiye takılma, ezberleme gibi problemleri aşabilmektedir.

Destek vektör makinesinin amacı her biri $y=\{+1,-1\}$ ile gösterilen sınıflardan birine ait olan n elemanlı bir veri kümesinde, sınıfları birbirinden ayıran alt düzlem bulmaktır. Şekil 2.9'da iki boyutlu düzlemde birbirinden doğrusal ayrılabilen iki sınıfa ait verilerin dağılımı ve bu verilerin birbirinden çok sayıda doğru ile ayrılabilirdiği görülmektedir. Çok boyutlu uzayda veriler hiper düzlemler ile ayrılabilir.



Şekil 2.9. İki boyutlu uzayda doğrusal ayrılabilen verilerin görünümü [81]

Verileri birbirinden ayıran en uygun hiper düzlem, birbirine en uzak iki hiper düzlem bulunarak elde edilir. Şekil 2.9'da birbirine en uzak H_1 ve H_2 düzlemleri arasından geçen H_0 hiper düzlemi, iki sınıfı birbirinden ayıran en uygun hiper düzlem olarak seçilmektedir. H_0 , düzlemine optimal ayırma hiper düzlemi adı verilir. H_1 ve H_2 , düzlemleri üzerindeki her bir veri destek vektörü olarak adlandırılır.



Şekil 2.10. İki sınıfı birbirinden ayıran en uygun hiper düzlem [81]

Şekil 2.10'da iki ayrı sınıf +1 ve -1 ile ifade edilir ise bu iki sınıfı birbirinden ayıran hiper düzlem $H_0 = wx + b$ fonksiyonu ile gösterilebilir. Verinin ayrılabilir olduğu varsayılırsa n elemanlı bir kümede w ağırlık değeri ile bu fonksiyon hesaplanabilir. Şekil 2.10'da görüldüğü gibi bu veri kümesinde noktalardan bazıları $wx + b = 1$ durumunu sağlarken bazıları da $wx + b = -1$ durumunu sağlamaktadır. H_0 fonksiyonu, düzlemin geçtiği noktalar $\sum_{i=1}^n w_i x_i + b$ cinsinden şeklinde de yazılabilir. Burada w , ağırlık vektörünü $w = \{w_1, w_2, \dots, w_n\}$ ve n , nitelik sayısını göstermektedir [81].

2.5.3. Bagging

Bagging kestirimleyicinin birden çok versiyonunu üretmek için bir metottur ve bunları kümelenmiş kestirimleyici elde etmek için kullanır. Kümeleme sayısal çıktıyı kestirirken bu versiyonların ortalamasını alır ve sınıfı kestirirken oy çokluğu (plurality vote) prensibini uygular. Öğretim kümesinin karışması yapılandırılan kestirimleyicide önemli değişikliklere yol açıyorsa, bagging doğruluğu artırılabilir [82].

2.5.4. One-R

One-R veya "One Rule (Bir Kural)" R. C. Holt tarafından önerilen basit bir algoritmadır. Bu algoritma eğitim verilerinde her bir özellik için bir kural üretmektedir ve sonra One-R' sine göre en küçük hata oranına sahip kural seçilmektedir [83].

2.5.5. J48 Karar Ağacı

Quinlan (1993) budanmış veya budanmamış C4.5 karar ağacı üretmek için sınıf tanımladı, budama uygulayan ve ağaçları kurallara çeviren metotlar anlattı. Bu algoritma başlangıç ağacı yapılandırmak için kullanılır. Eksik öznitelik değerleri gibi çeşitli problemler ele alır. Karar ağacı algoritmaları durumlar veya örnekler kümesiyle başlar ve yeni durumları sınıflandırabilmek için kullanılan ağaç veri yapısı yaratır. Her durum sayısal veya sembolik değer alabilen öznitelikler kümesi ile tanımlanır. Karar ağacının her bir iç düğümü test içerir, testin sonuçları o düğümden hangi dalın takip edileceğine karar vermek için kullanılır. Yaprak düğümleri testler yerine sınıf etiketleri içerir. Sınıflandırma modu test durumunda, etiket yoksa yaprak düğümüne ulaşır, C4.5 orada bulunan etiketi kullanarak sınıflandırır. Bu alanların çoğu için, C4.5 tarafından üretilen ağaçlar küçük ve doğrudurlar, hızlı ve güvenilir sınıflandırıcılar meydana getirirler. Bu özellikler karar ağaçlarını sınıflandırma için değerli ve popüler araçlar yapar (Quinlan, 1986). Böl ve fethet algoritması her yaprak tek sınıf durumu içerene kadar veya iki durumun farklı sınıflara ait olmalarına rağmen her bir öznitelik için aynı değerlere sahip olduğu daha fazla bölünmenin imkânsız olduğu duruma kadar veriyi parçalara ayırır. Sonuç olarak, çelişkili durumlar içermiyorsa, karar ağacı tüm eğitim durumlarını doğru sınıflandıracaktır. Aşırı uyma, bazı eğitim durumları kümelerini alt bölümlere ayılmaktan koruyan durdurma kriterleri ile veya üretildikten sonra karar ağacının bir kısım yapısını kaldırarak önlenabilir [84].

2.5.6. Bulanık k- En Yakın Komşu (Bulanık k-NN)

Bulanık mantık kavramı ilk kez 1965 yılında Prof. Lotfi Asker Zadeh tarafından ortaya atılmıştır. Bu mantık insanların günlük hayatta kullandıkları ifadelerin, bilgisayar ortamında matematiksel olarak modellenmesi prensibine dayalı bir yaklaşımdır. Klasik yöntemlerde 1 ya da 0'larla gösterilen kararların, keskin geçişleri, bulanık mantık ile yumuşatılmış ve ara değerlerle ifade edilebilir hale gelmiştir.

Bulanık k-en yakın komşuluk yöntemi de k-en yakın komşu yöntemi gibi bir sınıflandırma algoritması olmasına rağmen sonuçlarının ifadesi itibariyle k-en yakın komşu yönteminden ayrılır. Bulanık k en yakın komşuluk algoritması, vektörü belirli bir sınıfa atamak yerine, sınıf üyeliğini bir örnek vektöre atar. Bir elemanın bir kümeye veya bir sınıfa ait olması klasik küme kavramında ya aittir (üyelik=1) veya ait değildir

(üyelik=0) şeklinde tanımlanmaktadır. Gerçekte bir eleman bir kümeye ne tam aittir ne de değildir. Yani bu elemanın o küme veya sınıf için bir aitlik derecesi (üyelik değeri) olmalıdır. Bu üyelik değeri 0 ile 1 arasında sonsuz değer alabilmektedir. Bulanık algoritmalarda, test edilecek örnek sınıflandırırken örneğin sınıfını belirlemenin yanında o sınıfa ne kadar ait olduğuna dair bir bilgi de verilmektedir. Bu bilgi, örneğin o sınıfa olan üyelik değeri olmaktadır. Bulanık k-en yakın komşu yönteminin, k-en yakın komşu yöntemine göre avantajı Bulanık k-en yakın komşu yönteminin daha fazla bilgi içermesidir.

Bulanık k-en yakın komşu yönteminde test edilecek örnek için belirlenen üyelik değerleri sonuçta oluşan sınıflandırma için bir güvence seviyesi sunar. Örnek olarak; elimizde iki adet sınıf olduğunu düşünürsek, test edilecek örnek için üyelik değerlerinin bir sınıfa 0.89 üyelik değeri ile ve diğer sınıfa 0.11 üyelik değeri ile üyelik değerlerinin hesaplandığı varsayılırsa; 0.89 üyelik değerinin belirlendiği sınıfın test edilecek örneğin sınıfı olduğu kararına üyelik değerlerinin sayılarına bakarak kolayca karar verebiliriz. Farklı bir örnek için; elimizde üç adet sınıfın olduğu varsayılırsa, eğer test edilecek örneğin birinci sınıfa üyeliği 0.55 üyelik değeri, ikinci sınıfa üyeliği 0.44 üyelik değeri ve üçüncü sınıfa üyeliği 0.01 üyelik değeri olarak hesaplanmış ise test edilecek örneğin sınıflandırılmasında kesin bir karara varmak mümkün olmayabilir. Fakat üçüncü sınıfa ait olmadığı konusunda da emin olabiliriz. Böyle bir durumda, sınıfının anlaşılabilmesi için test edilecek örnek için farklı yöntemler de denenerek incelenebilir, çünkü test edilecek örnek her iki sınıfta da yüksek üyelik derecesi göstermektedir.

$W=\{x_1, x_2, \dots, x_n\}$ n adet eğitim sınıfında yer alan sınıfları belirlenmiş örneğin kümesi olsun. Aynı zamanda, $u_i(x)$; x vektörünün üyeliği (hesaplanacak olan), ve u_{ij} de sınıfı belirlenmiş örnek kümesinin j 'ninci vektörünün i 'ninci sınıf içindeki üyeliği olsun.

$$u_i(x) = \frac{\sum_{j=1}^k u_{ij}(1/\|x-x_j\|^{2/(m-1)})}{\sum_{j=1}^k (1/\|x-x_j\|^{2/(m-1)})} \quad (2.3)$$

Eşitlik 2.3'de görüldüğü gibi x' in atanmış üyelikleri en yakın komşulara olan uzaklığın tersiyle ve onların sınıf üyeliklerinden etkilenmektedir. Ters uzaklık bir vektörün üyeliğine, vektörden olan uzaklık azsa daha fazla ağırlık verir, uzaklık çoksa daha az ağırlık verir. Sınıfı bilinen örnekler için çok çeşitli şekilde üyelikler atanabilir. Bilinen

sınıflarında tam üyelik verilip diğer bütün sınıflarda hiç üyelik verilmeyebilir. Bu yöntemin yerine bulanık temelli alternatif metotların da kullanılması mümkündür.

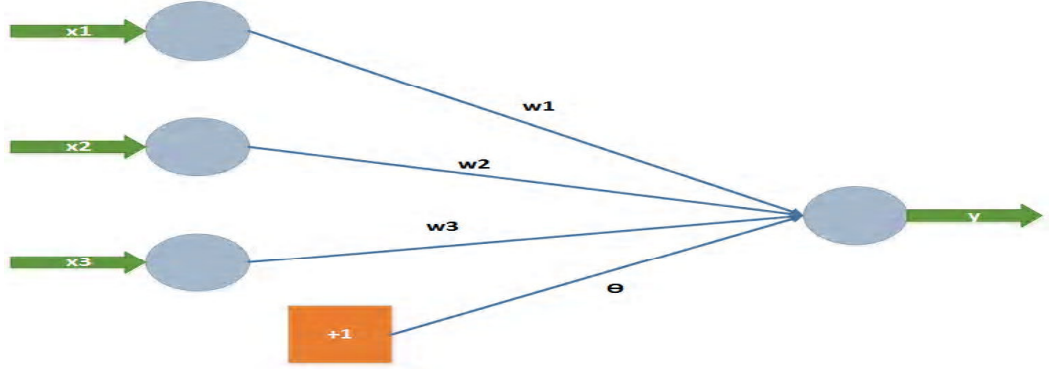
m değişkeni her bir komşunun üyelik değerine katkısını hesaplarken mesafenin ne kadar ağırlıkta verildiğini belirler. m değeri 2 ise her komşu noktanın katkısı, sınıflanan noktadan olan uzaklığın karşıtıyla ağırlıklandırılır. m arttıkça komşular daha eşit bir şekilde ağırlıklandırılır, ve sınıflandırılan noktadan olan göreceli uzaklıklarının etkileri daha az olur. m değeri 1'e yaklaştıkça yakın olan komşular uzaktaki komşulardan çok daha fazla ağırlıklandırılırlar ve bunun etkisi sınıflanan noktanın üyelik değerine katkıda bulunan nokta sayısını azaltan şekilde olmaktadır.

Sonuç olarak; Bulanık k -en yakın komşu algoritmasının, k -en yakın komşu algoritmasından en büyük farkı bilinmeyen bir test örneğini sınıflamak yerine, test örneğinin belirli sınıfa ne kadar ait olduğu sorusuna yönelik verdiği cevaptır. Bulanık k -en yakın komşu algoritmasında ise örnek için bir sınıfa ait olma ya da olmama bilgisine ek olarak o sınıfa ne kadar ait olduğuna ilişkin değeri de hesaplanır. Bu değer kullanılarak örneğin sınıflandırılması yapılır [85].

2.5.7. Tek Katmanlı Algılayıcılar (TKA)

Perseptron (Algılayıcı), Rosenblatt (1958) tarafından örüntü sınıflandırıcı olarak ortaya atılmış basit bir YSA modelidir. Tek katmanlı perseptron, bir girdi ve bir çıktı katmanı içerir. Perseptron'da hem girdi katmanı hem de çıktı katmanı ikili (0 ve 1) birimlerden oluşmaktadır. Bununla beraber perseptron'un çıktı birimlerinde aktivasyon fonksiyonu olarak eşik değer fonksiyonu kullanılmaktadır.

En basit hali ile üç girdi ve tek çıktı bir perseptron modeli Şekil 2.11'de verilmektedir. Şekil 2.11'de gösterildiği gibi perseptron'da her zaman yan değeri "+1" olarak alınmaktadır.



Şekil 2.11. Üç girdi ve bir çıktılı perseptron modeli

Perseptron modelinde girdi katmanındaki nöronların aldığı değerler, ilgili bağlantıları ile çarpılarak net sinyal değeri hesaplanır ve bu değer çıktı katmanı nöronlarının girdisini oluşturur. Herhangi bir çıktı katmanı nöronu için sözü edilen bu net sinyal, ilgili çıktı katman nöronunun kendisine bağlı her bir girdi katman nöronlarının ilettiği sinyal değerlerinin ağırlık değerleri ile çarpımlarının toplamına yan değer ağırlığının eklenmesi ile elde edilir. Çıktı katmanı nöronlarının dış dünyaya ilettiği bilgi ise her bir çıktı nöronuna gelen net sinyale karşılık eşik değer fonksiyonunun verdiği sonuçtur. Buna göre j'ninci çıktı katmanı nöronunun girdisi ve bu girdiye karşılık gelen çıktısı sırasıyla (2.4) ve (2.5) ile verilmektedir.

$$net_j = \sum_{i=1}^m w_{ij} x_i + \theta_j \quad (2.4)$$

$$y_j = f(net_j) = f\left(\sum_{i=1}^m w_{ij} x_i + \theta_j\right) \quad (2.5)$$

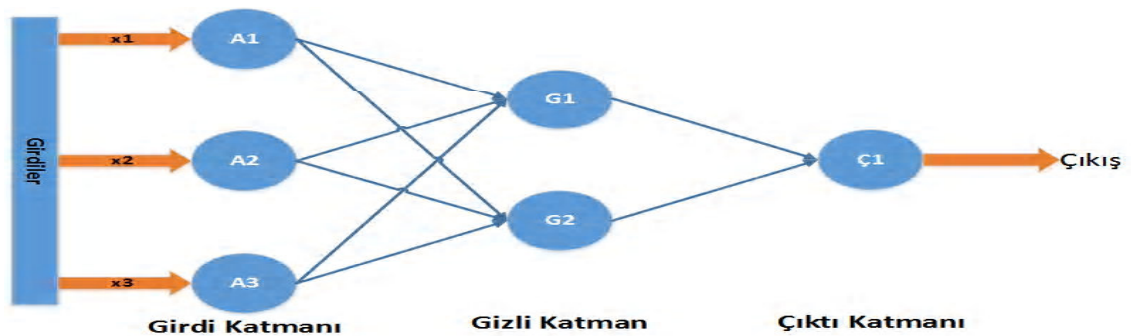
(2.4) ve (2.5)'te yer alan x_i , i' inci girdi katmanı nöronunun değerini, net_j , j'ninci çıktı katman nöronunun net girdisini, w_{ij} , i'ninci girdi nöronu ile j'ninci çıktı nöronu arasındaki bağlantı ağırlığını, θ_j , eşik değer ile j'ninci çıktı nöronu arasındaki bağlantı ağırlığını ve y_j ise j'ninci çıktı nöronunun ürettiği çıktıyı ifade etmektedir. Çıktı katmanı nöronlarının kullandığı aktivasyon fonksiyonu ise f ile gösterilmiştir ve denklem (2.6) ile tanımlanmaktadır.

$$f(net_j) = \begin{cases} 1, & net_j > 0 \\ -1, & net_j \leq 0 \end{cases} \quad (2.6)$$

Denklem (2.6) ile verilen formülden anlaşılacağı gibi tek katmanlı perseptron, “-1” ya da “+1” sonuçlarını üretmektedir. Bu hali ile perseptron, verilen örüntüleri iki kümeye ayırma problemlerinde kullanılabilir. İki kümenin ayrılma sınırını ise denklem (2.6) ile verilen eşik değer fonksiyonundan da anlaşılacağı gibi net_j değeri belirler [86].

2.5.8. Çok Katmanlı Algılayıcılar (ÇKA)

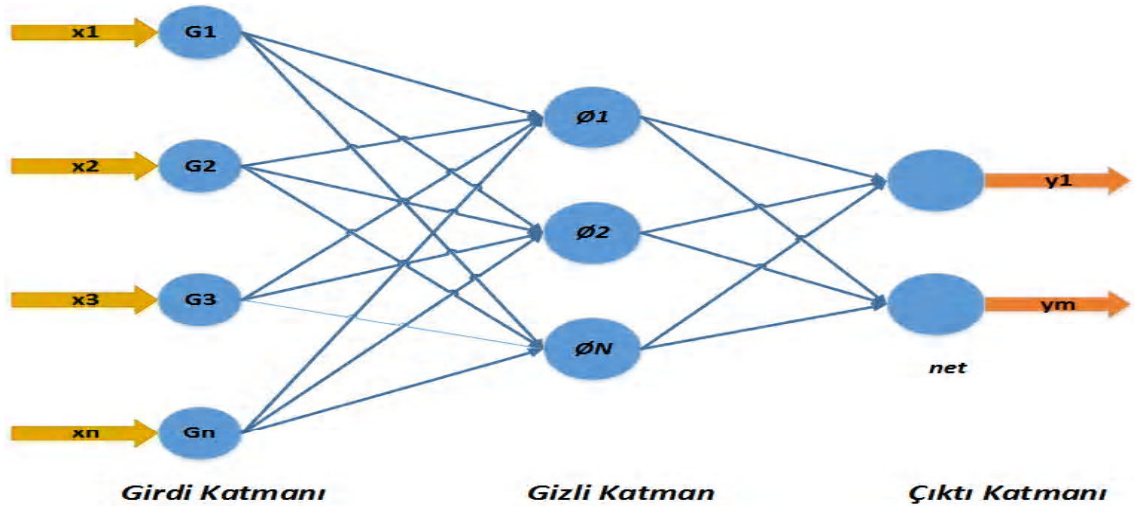
Rumelhart ve arkadaşları tarafından geliştirilen bu modele hata yayma modeli veya geriye yayılım modeli (Back Propagation network) de denilmektedir. ÇKA modeli yapay sinir ağlarına olan ilgiyi çok hızlı bir şekilde arttırmış ve YSA tarihinde yeni bir dönem başlatmıştır. Bu ağ modeli özellikle mühendislik uygulamalarında en çok kullanılan sinir ağı modeli olmuştur. Birçok öğretim algoritmasının bu ağı eğitmede kullanılabilir olması, bu modelin yaygın kullanılmasının sebebidir. Bir ÇKA modeli, bir giriş katmanı, bir veya daha fazla ara katman ve bir de çıkış katmanından oluşur (Şekil 2.12). Bir katmandaki bütün işlem elemanları bir üst katmandaki bütün işlem elemanlarına bağlıdır. Bilgi akışı ileri doğru olup geri besleme yoktur. Bunun için ileri beslemeli sinir ağı modeli olarak adlandırılır. Giriş katmanında herhangi bir bilgi işleme yapılmaz. Buradaki işlem elemanı sayısı tamamen uygulanan problemlerin giriş sayısına bağlıdır. Ara katman sayısı ve ara katmanlardaki işlem elemanı sayısı ise, deneme-yanılma yolu ile bulunur. Çıkış katmanındaki eleman sayısı ise yine uygulanan probleme dayanılarak belirlenir. Bu ağ modeli, özellikle sınıflandırma, tanıma ve genelleme yapmayı gerektiren problemler için çok önemli bir çözüm aracıdır [87].



Şekil 2.12. Çok katmanlı ileri beslemeli yapay sinir ağı modeli

2.5.9. Radyal Tabanlı Yapay Sinir Ağları (RTYSA)

Radyal Tabanlı Yapay Sinir Ağları (RTYSA), biyolojik sinir hücrelerinde görülen etki-tepki davranışlarından esinlenilerek 1988 yılında geliştirilmiş ve filtreleme problemine uygulanarak YSA tarihine girmiştir. RTYSA modellerinin eğitimi çok boyutlu uzayda eğri uydurma yaklaşımı olarak görmek mümkündür. Bu nedenle RTYSA modelinin eğitim performansı, çıktı vektör uzayındaki verilere en uygun yüzeyi bulma ve dolayısıyla bir interpolasyon problemine dönüşmektedir. RTYSA modelleri genel YSA mimarisine benzer şekilde giriş katmanı, gizli katman ve çıktı katmanı olmak üzere 3 katman halinde tanımlanmaktadır (Şekil 2.13). Ancak, klasik YSA yapılarından farklı olarak RTYSA'larda, girdi katmanından gizli katmanına geçişte radyal tabanlı aktivasyon fonksiyonları ve doğrusal olmayan bir kümeleme (cluster) analizi kullanılmaktadır. Gizli katman ile çıktı katmanı arasındaki yapı ise diğer YSA türlerinde olduğu gibi işleyişini sürdürmekte olup asıl eğitim burada gerçekleştirilmektedir.



Şekil 2.13. Radyal tabanlı YSA yapısı [88]

RTYSA modellerinde ağın ürettiği çıktı (y) ise denklem (2.7) yardımıyla hesaplanabilmektedir.

$$y_i = \sum_{k=1}^N w_{ik} \phi_k(x, c_k) = \sum_{k=1}^N w_{ik} \phi_k(\|x - c_k\|_2), \quad i = 1, 2, 3, \dots, m \quad (2.7)$$

Burada $x \in R^{n \times 1}$ ağıın girdi vektörünü; $\phi_k \in R^+$ radyal tabanlı aktivasyon fonksiyonunu; $c_k \in R^{n \times 1}$ girdi vektör uzayının bir alt setinden seçilen radyal tabanlı merkezleri; $\|\cdot\|_2$ girdi vektörünün merkezden ne kadar uzak olduğunun bir ölçütü olan Öklidyen normunu; w_{ik} çıktı katmanındaki ağırlıkları; N ise gizli katmanda bulunan hücre sayısını göstermektedir.

Burada x girdi vektörünü, c_k merkezleri göstermektedir. σ ise standart sapma değerini simgelemekte olup; YSA terminolojisinde, RTYSA modelinin performansını önemli ölçüde etkileyen dağılma (spread) parametresi olarak da anılmaktadır.

RTYSA modellerinin eğitimi, hücre merkezlerinin bulunması ve çıktı katmanındaki ağırlıkların optimize edilmesi olmak üzere iki aşamada gerçekleşmektedir. Literatürde hücre merkezlerini (c_k) ve çıkış ağırlıklarını (w_{ik}) bulabilmek için farklı yöntemler kullanılmaktadır. Hücre merkezlerini bulabilmek için en sık kullanılan yöntem k -ortalamalar (K -means) ve Kohonen kümeleme yöntemleridir. Çıkış ağırlıklarını bulmakta kullanılan yöntemler ise En Küçük Ortalamalı Kareler (LMS) ve Moore-Penrose Sözde Ters (Pseudo-inverse) yöntemleridir. Dağılma parametresi ise genellikle bütün hücreler için sabit alınmaktadır. RTYSA modellerinde dağılma parametresi için yaklaşık denklikler olmakla birlikte, bu parametre deneme-yanılma yöntemiyle de belirlenebilmektedir.

Çalışmada RTYSA modeline ait hücre merkezleri (c_k) K -ortalamalar yöntemi ile belirlenmiştir. Hücre merkezlerinin belirlenmesinin ardından, belli bir Q eğitim seti için ağıın çıktısı (y) aşağıdaki gibi hesaplanmaktadır.

$$y(q) = \sum_{k=1}^N w_{ik} \phi_k(x(q), c_k), \quad q = 1, 2, \dots, Q \quad (2.8)$$

Bu aşamadan sonra, ağıın ürettiği çıktı değerleri ile beklenen çıktı değerleri arasındaki farklar karşılaştırılmakta ve denklem (2.8)'de verilen performans fonksiyonunun minimize edilmesi amaçlanmaktadır [88].

2.5.10. Karınca Koloni Algoritması

Karıncalar, yuvaları ile yiyecek kaynağı arasındaki en kısa yolu bulma kabiliyetine sahiptirler ve ayrıca çevredeki değişimlere adapte olabilmektedirler. Örneğin, yuva ile yiyecek arasındaki en kısa yol belirli bir zamanda keşfedilir ve bir yiyeceğe giden yolda herhangi bir problem meydana gelmesi (bir engelin ortaya çıkması gibi) ve yolun kullanılamaz olması durumunda, yeniden en kısa yolu bulurlar. Karıncaların en kısa yolu bulma kabiliyetleri, birbirleri arasındaki kimyasal haberleşmenin bir sonucudur. Karıncalar birbirleriyle haberleşmede feromon olarak adlandırılan kimyasal bir madde kullanmaktadır. Karıncalar ilerlerken yolları üzerine bir miktar feromon maddesi bırakır ve her bir karınca yuva ya da yiyecek bulmak için bir doğrultuyu seçer. Bir yönün seçilme ihtimali, bu yön üzerindeki feromonun miktarına bağlıdır. Her karıncanın, ortalama aynı hızda ve aynı miktarda feromon bıraktığı göz önüne alınırsa, daha kısa yollarda birim zamanda daha çok feromon bulunacaktır. Dolayısıyla, karıncaların büyük çoğunluğu hızla en kısa yolları seçecektir. Bu geri besleme işlemi otokatalitik işlem olarak bilinir. Otokatalitik işlem araştırmanın çok hızlı bir şekilde optimal yola yakınsamasına neden olur ve sonuçta bir alt optimal yola takılmadan yuva ve yiyecek arasındaki en kısa yol bulunur [89].

Karınca kolonilerinin yuvaları ile yiyecekleri arasındaki bu davranışları Dorigo tarafından problem çözme tekniği olarak kullanılmıştır [89]. Bunun için geliştirilen algoritmaların genel adımları Şekil 2.14’de gösterilmiştir.

```

Procedure AntColonyAlgorithm
Begin
  İlk feromon miktarının hesaplanması
  while (not durdurma kriteri)
    Aday çözümler oluştur
    Lokal arama gerçekleştir
    Feromonları güncelle
  end while
  return bulunan en iyi çözüm
End

```

Şekil 2.14. Karınca koloni algoritmasının adımları [104]

Bütün karıncalar başlangıç safhasında bir ilk çözüm oluşturur. Oluşan bu çözümler, feromon miktarları göz önünde bulundurularak geliştirilir. Her iterasyonun sonunda

feromon miktarı güncellenir. Çünkü doğal ortamda karıncaların yolları üzerine bıraktıkları feromon miktarı zamanla azalmakta ya da artmaktadır. İterasyonlar önceden belirlenen durdurma şartı sağlanana kadar devam edecektir.

Karınca algoritmalarının sınıflandırma kural madenciliğinde kullanımı çok yenidir ve yakın zamanda Parpineli ve arkadaşları tarafından Ant-Miner [90] önerilmiştir. Daha sonra da Liu ve arkadaşları [91] tarafından aynı mantıkla, ancak farklı sezgisel ve feromon güncelleme stratejisiyle Ant-Miner geliştirilmiştir. Karınca tabanlı aramanın geleneksel metotlardan daha esnek ve sağlam olduğunu savunmuşlar ve daha global arama yaptıklarını belirtmişlerdir. Ayrıca greedy algoritmalarla göre nitelik etkileşimleriyle Genetik algoritmalar gibi iyi baş edebilir. Başka bir özellik de karınca algoritması uygulamalarının problem alanıyla minimum bilgi gerektirmesidir.

Bu algoritmada kurallar EĞER <şartlar> O ZAMAN <sınıf> yapısındadır. <şartlar> ya da ata kısmı, bir nitelik ve bunun alanında özel bir değer mesela $\text{nitelikA}=\text{deger}_1$ olan terimlerin mantıksal kombinasyonudur. <sınıf> ya da sonuç kısmı ise, özel terimlerin kısmi kombinasyonlarına atanan sınıf etiketini içerir.

Ant-Miner algoritması ‘ayır-ve-yönet’ mantığıyla çalışır. Önce tüm eğitim verisiyle başlar. Eğitim verisinin bir alt kümesini kapsayan bir ‘en iyi’ kural oluşturur ve bu en iyi kuralı **Kesfedilen_kurallar_listesi**’ ne ekler. Sonra bu kural tarafından kapsanan örnekleri eğitim verisinden kaldırıp azaltılmış bir eğitim kümesiyle tekrar başlar. Bu iş eğitim verisinde sadece birkaç örnek kalana (**KapsananMaksOrn**’ten az) kadar devam eder. Bu aşamada, bu kalan örnekleri kapsayacak varsayılan bir kural oluşturulur. Ant-Miner algoritması ayrık değerler ile çalışırken Ant-Miner ’in sürekli değerler ile çalışan versiyonu ise cAnt-Miner algoritmasıdır. cAnt-Miner algoritması sadece Ant-Miner algoritmasının sürekli değerler için genişletilmiş halidir. Bizimde mikroyarray gen ifade verileri sürekli değerler içerdiği için çalışmamızda cAnt-Miner algoritmasını kullanmış bulunmaktayız. Şekil 2.15’te algoritmanın adımları gösterilmektedir.

```

Keşfedilen_kurallar_listesi = [ ]
// Başlangıçta boş
Eğitim kümesi= tüm eğitim örnekleri
while (Eğitim kümesindeki kapsanmayan örnek sayısı>
KapsananMaxÖrnek) // Karınca Koşması
    i=0
    repeat //iterasyon
        i = i+1
        Karınca (i) artımsal olarak
            sınıflandırma kuralını oluştur
        Kuralın sonucunu ata
        Yeni oluşturulan kuralı buda
        Karınca (i) tarafından takip edilen yolun feromonunu güncelle
        until (i>=KarıncaSayısı) or (Karınca (i) bir önceki
MaksKuralYakınsa-1 karınca ile aynı kuralı oluşturursa)
        Oluşturulan tüm kuralların en iyisini seç
        Kuralı Keşfedilen_kurallar_listesi' ne ekle
        Seçilen kural tarafından doğru olarak kapsanan eğitim
örneklerini Eğitim Kümesi'nden kaldır
    end while
Varsayılan kuralı oluştur
Keşfedilen_kurallar_listesi'ni ver

```

Şekil 2.15. cAnt-Miner algoritması [104]

2.6. Mikroarray Gen Ekspresyon (İfade) Verileri

DNA mikroarray veri analizi; kanser gibi genlerle ilişkili hastalıkların belirlenmesinde önemli rol oynamaktadır. Hastalık türüne ilişkin ilgili genler belirlenerek, herhangi bir bireyin hasta ya da sağlam olduğu yüksek olasılıklarda hesaplanabilir [92].

Moleküler biyolojideki geleneksel metotlarda genellikle “Bir deneyde bir gen” ilkesi geçerlidir. Bu demektir ki gen fonksiyonlarının “bütün resmini” görmek geleneksel yöntemlerle zordur. Gen çip teknolojisinin büyük bir ilgi ile karşılanmasının sebebi, bütün genomun basit bir çip üzerinde görüntülenmesini vaat etmesi ve bu sayede bilim adamlarının aynı anda binlerce genin birbirleriyle olan etkileşimlerini görmesine olanak tanımasıdır [93].

Mikroarray’ler hastalıklı ve normal bir hücrede gen ekspresyonunun karşılaştırılması suretiyle hastalık ile alakalı genlerin belirlenmesinde kullanılabilmektedir [94].

Mikroarray gen ekspresyon verileri, hem maliyeti hem de ölçümlerin tekrar edilmesindeki zorluklar nedeniyle çok fazla hasta üzerinden bu veriler

toplanamamaktadır. Bu nedenle az sayıdaki hastaya ait binlerce gen verisi üzerinden bazı yorumlara ulaşılmak durumunda kalınmıştır.

2.7. k-Katlamalı Çapraz Doğrulama Yöntemi ve Başarı Ölçümü

Veri madenciliği çalışmalarında, uygulanan yöntemin başarısının sınanması için, veri kümesi eğitim ve test kümeleri olmak üzere ikiye ayrılmaktadır. Bu ayırma işlemi çeşitli şekillerde yapılabilir.

Herhangi bir amaç için oluşturulan sistemlerin yürütülmesi sonucunda elde edilen sonuçların doğruluklarının ölçülmesi suretiyle sistemlerin başarı değerlendirilmesi yapılabilmektedir. Bu amaçla kullanılan en yaygın yöntemlerden biri de k-katlı çapraz doğrulama (**k-Fold Cross Validation**) yöntemidir. Bu yöntemde A olarak temsil edilen veri kümesi, her birinde yaklaşık aynı sayıda veri olacak şekilde $A_1, A_2, \dots, A_k$ biçiminde gösterilen rastgele k sayıda alt kümeye ayrılır. Yöntemin veri üzerinde sınanmasında k sayıda alt kümeden her seferinde bir küme “test kümesi” olarak, geri kalanlar ise “eğitim kümesi” olarak ele alınır.

Deneysel çalışmalar k-katlı çapraz doğrulama yönteminde k için optimum değer 10 olduğunu göstermiştir [95]. Fakat veri kümesinin az sayıda olduğu bazı durumlarda bu sayı 2 veya 5 olabilmektedir. Bu tez çalışmasında 10-katlı çapraz doğrulama yöntemi kullanılmıştır.

Sistemin Sınıflandırma Başarı Oranı (SBO), doğru sınıflandırılan veri sayısının (dsvs) toplam veri sayısına (tvs) oranlanmasıyla elde edilir. Eşitlik 2.9’da sırasıyla sistemin SBO ve ortalama SBO yer almaktadır [96].

$$SBO = \frac{dsvs}{tvs}, \text{ ortalama SBO} = \frac{\sum_{i=1}^k SBO_i}{k} \quad (2.9)$$

2.8. Mikroarray Gen Ekspresyon (İfade) Veri Setlerine Sınıflandırma Yöntemlerinin Uygulanması

Veri madenciliğinde güvenilirliğin artırılması için, veriye ön işlem sürecinin uygulanması gerekir. Aksi halde hatalı veya eksik girdi verileri bizi hatalı çıktıya veya başarımı daha düşük sonuçlara götürecektir. Veri ön işleme süreci, zaman isteyen ve aynı zamanda önemli bir veri madenciliği aşamasıdır. Veri ön işleme süreci aşağıdaki sebeplerden dolayı verilere uygulanır:

- 1.Veriler üzerinde herhangi bir analiz türünün uygulanmasını engelleyecek veri problemlerinin çözümü
- 2.Verilerin doğasının anlaşılması ve anlamlı veri analizinin başarılması
- 3.Verilen bir veri kümesinden daha anlamlı bilginin çıkarılması. Birçok uygulamada, veri ön işlemenin bir türünden daha fazlasına ihtiyaç duyulmaktadır. Veri ön işleme türünün belirlenmesi bu sebeple önemli bir iştir [97].

Yukarıda sayılan veri ön işleme teknikleri, verilerin kalitesini artıracaktır. Kaliteli kararların kaliteli veriler gerektirdiği bir gerçektir. Bu sebeple, veri ön işleme süreci, veri madenciliğinin önemli bir adımıdır. Bu yüzden tez çalışmasında öncelikli olarak veriler bir ön işlem sürecinden geçirilerek kararlı ve kaliteli bir veri haline getirildi.

Kararlı ve kaliteli hale gelen veri setleri, sınıflandırma yöntemleri ile işlenmeye hazır hale getirildi. Aşağıda görüldüğü gibi 10 farklı mikroarray kanser veri setinde farklı doku tümörleri görülmektedir (Tablo2.1).

Daha sonra ön işlemde geçirilen yüksek boyutlu mikroarray gen ekspresyon kanser verileri üzerinde sınıflandırma yöntemlerinin başarılı sonuçlar verebilmesi için boyut azaltma yöntemlerinden biri olan “korelasyon tabanlı öznelik seçim (KTÖS)” yöntemi uygulandı. Verilere KTÖS yöntemi uygulandıktan sonra veri üzerinde uygulanacak olan sınıflandırma yöntemlerinin başarımlarını düşürecek olan ve sınıflandırmaya etkisi olmayan genler elenmiş oldu. Böylece yüksek boyutlu mikroarray gen ekspresyon verileri daha düşük boyutlu hale getirilerek sınıflandırma yöntemleri için daha ideal bir hale getirildi.

Boyutları azaltılan her mikroarray gen ekspresyon veri setine k-katlamalı çapraz doğrulama yöntemi uygulandı. Bu yöntemde k değeri 10 seçildi. Böylece mikroarray gen ekspresyon verileri 10 adet alt kümeye ayrıldı. Alt kümelerden biri test kümesi diğerleri eğitim kümesi olacak şekilde bu işlem 10 defa yapılarak verilere sırasıyla yukarıda belirtilen istatistiksel yöntemler ve yapay zekâ tabanlı yöntemler uygulandı. Bu yöntemlerin mikroarray gen ekspresyon veri setleri üzerindeki başarımları birbiriyle karşılaştırıldı.

Tablo 2.1. Çalışmada Kullanılan Mikroarray Kanser Veri Setleri

Veri Setleri	Sınıf	Toplam Gen	Gen	Örnek
Karaciğer Kanseri (Chen-2002)	2	22699	85	179
RNA doku Kanserleri (Chowdary-2006)	2	22283	182	104
Prostat Kanseri (Singh-2002)	2	12600	339	102
Merkezi Sinir Sistemi Kanseri (Pomeroy-2002-v1)	2	7129	857	34
Beyin Kanseri (Nutt-2003-v2)	2	12625	1070	28
Rahim Kanseri (Risinger-2003)	4	8872	1771	42
Göğüs Kanseri (West-2001)	2	7129	1198	49
Mesane Kanseri (Dyrskjot-2003)	3	7129	1203	40
Farklı doku Kanserleri (Su-2001)	10	12533	1571	174
Kolon Kanseri (Laiho-2007)	2	22883	2202	37

3. BÖLÜM

BULGULAR

3.1. Sınıflandırma Yöntemlerinin Performanslarının Karşılaştırılması İle İlgili Yapılmış Benzer Çalışmalar

Veri madenciliği fonksiyonlarından biri olan ‘sınıflandırma’ tekniği günümüzde çoğu alanda yaygın olarak kullanılmaktadır. Önceden birçok alanda verileri sınıflandırmak için klasik istatistiksel sınıflandırma yöntemlerinden (lojistik regresyon, varyans analizi, doğrusal regresyon analizi) faydalanılmaktaydı. Fakat bu yöntemlerin zamanla sınıflandırmada yetersiz kalmasıyla beraber son zamanlarda daha iyi sonuçlar veren istatistiksel yöntemler geliştirilmiş ve kullanılmaya başlamıştır. Geliştirilen bu istatistiksel yöntemlerden bazıları ise Destek Vektör Makineleri (DVM), Karar Ağaçları, Bayes Ağları, İlişki tabanlı sınıflandırıcılar, k-En yakın komşu yöntemi (k-NN) gibi sınıflandırıcılardır.

Bu istatistiksel sınıflandırma yöntemlerinin son zamanlarda birçok alanda kullanılmasıyla birlikte en çok kullanıldığı alanlar mühendislik ve özellikle tıp alanıdır. Çünkü tıp alanında son zamanlarda mikroyarray gen çalışmaları sayesinde verilerin boyutlarında çok fazla artış olmuş ve bu verileri sınıflandırmada güncel istatistiksel yöntemler kullanılmaya başlanmıştır.

Günümüzde yoğun bir şekilde kullanılan istatistiksel yöntemlerin haricinde araştırmacılar son zamanlarda veri madenciliğinde başarıyı daha yüksek sınıflandırma yapabilmek için yapay zekâ alanına merak salmışlardır. Araştırmacıların yapay zekâ tabanlı sınıflandırma yöntemleri arayışına girmelerinin sebebi ise istatistiksel yöntemlere alternatif olabileceklerini düşünmeleri ve ayrıca sınıflandırmada istatistiksel yöntemlerden daha iyi sonuçlar verebilirler mi? sorusudur.

Bizim çalışmamız da son zamanlarda dikkat çekici şekilde üzerinde çalışılan yapay zekâ tabanlı sınıflandırıcılar ile ilgilidir. Literatürdeki araştırmalardan görüldüğü kadarıyla istatistiksel yöntemler ve yapay zekânın alanı içerisine giren yapay sinir ağları sınıflandırmada vazgeçilmez yöntemler olmuşlardır ve genellikle bu sınıflandırma yöntemlerinin başarımlarının birbirleriyle kıyaslanması ile ilgili çalışmalar yapılmıştır.

Coşkun ve Baykal SEER (Surveillance Epidemiology and End Results) veri kaynağını kullanarak bu veriler üzerinde sınıflandırma yöntemlerinin başarımlarını karşılaştırmışlardır. SEER veri tabanı farklı kanser gruplarını içeren ve bilimsel araştırmalarda son derece önemli bir yer tutan, güvenilir, dokümente edilmiş, eşine az rastlanır bir veri kümesidir. National Cancer Institute (NCI)'in sağladığı Amerika Birleşik Devletleri'nin belli başlı coğrafi bölgelerini kapsayan, nüfusunun %26'sını ilgilendiren ve bu kanser vakaları hakkında istatistiksel önem taşıyan bilgiler içerir. Yıllık olarak güncellenen bu veri tabanı bilimsel çalışma yapanlara, sağlık sektöründe çalışanlara, halk sağlığı konusunda görevli kurumlara açık bir veri kaynağı olup, binlerce bilimsel çalışmada kaynak olarak kullanılmıştır. Çalışmada SEER verileri üzerinde bir karar ağacı algoritması olan ve temeli ID3 ve C4.5 algoritmalarına dayanan J48, istatistiksel bir algoritma olan Bayes sınıflandırma algoritmalarından Naive-Bayes, regresyon tabanlı algoritmalarından lojistik regresyon ve örnek tabanlı sınıflandırma algoritmalarından Kstar algoritmaları kullanılarak modeller oluşturulmuş ve oluşturulan modellerin başarımlarını karşılaştırılmıştır. Çalışmada göğüs kanseri hastalarının kayıtları incelenmiş, hastaların hayatta olup olmadıkları, hayatta değil iseler ne kadar süre hayatta kaldıkları ve ölüm sebepleri göz önünde tutularak herhangi bir hastanın hastalığı yenip yenemeyeceği sınıflandırılarak ileriye dönük tahminlerde bulunabilme amacı ile farklı algoritmalarla oluşturulan modellerin başarımlarını karşılaştırılmıştır. Algoritmaları çalıştırırken test yöntemi olarak “10-katlı çapraz doğrulama” yöntemi kullanılmıştır. Bu yöntemle veri kaynağı 10 bölüme ayrılır ve her bölüm bir kez test kümesi, kalan diğer 9 bölüm öğrenme kümesi olarak kullanılır. Çalışma sonuçları incelendiğinde, J48 algoritmasının model testine ait %86.36 doğruluk derecesiyle en iyi sonucu ürettiği söylenebilir. Doğruluk ölçütü oldukça basit ve önemli bir kriterdir. Bu ölçüte göre J48 algoritmasını sırasıyla KStar, Lojistik Regresyon ve Naive Bayes algoritmaları izlemektedir [98].

Makinacı ve Güneşer Wisconsin Madison Üniversitesi, Klinik Bilimler Merkezi'nin göğüs kanseri veri tabanını kullanarak sınıflandırma yöntemlerinin başarımları ölçülmüştür. Veri tabanında, ince iğne aspirasyon yöntemi ile elde edilmiş toplam 699 örnek bulunmaktadır. Bunlardan 241'i kötü huylu (kanser), 458'i de iyi huylu tümör örnekleridir. Her örnek uzman tarafından değerlendirilerek dokuz farklı fiziksel özelliği 1-10 arasında puanlanmıştır. Bunlar örneklerin öznitelik vektörünü oluşturmaktadır. Örnekler, kanser olup olmama durumuna göre uzman tarafından ayrıca etiketlenmiştir. Çalışmada YSA, DVM ve k-NN algoritmaları kullanılmıştır ve Wisconsin göğüs kanseri verilerini iyi/kötü huylu olarak ayırmak için kullanılan sınıflandırma yöntemlerinin başarımları ölçülmüştür. Algoritmaları çalıştırırken test yöntemi olarak "10-katlı çapraz doğrulama" metodu kullanılmıştır. Çalışma sonuçları incelendiğinde en iyi genel başarımlar oranı (%98.63) ileri beslemeli yapay sinir ağı sınıflandırıcısı ile sağlanmıştır. DVM sınıflandırıcısının genel başarımlar oranı ise %96.20 olarak hesaplanmıştır. k-en yakın komşu sınıflandırıcı %96.60 genel başarımlar oranı vermiştir [99].

Daş ve Türkoğlu, National Center for Biotechnology (NCBI) sitesi gen bankasından gerçek veriler alınarak DNA dizilimlerindeki nükleotit çiftlerinin frekans değerlerine göre farklı sınıflandırma yöntemleri ile başarımlarının karşılaştırılmıştır. Alınan *Escherichia coli*, *Bacillus cereus*, *Buchnera aphidicola* ve *Enterobacter cloacae* gibi 4 bakteri türü sınıflandırılma için kullanılmıştır. Çalışmada sınıflandırma yöntemlerinden Yapay Sinir Ağları (YSA), Destek Vektör Makineleri (DVM) ve k-En yakın komşu algoritması kullanılmıştır. Bu çalışma, gen bankasından alınan *Escherichia coli*, *Bacillus cereus*, *Buchnera aphidicola* ve *Enterobacter cloacae* gibi 4 bakteri türünden oluşan toplam 154 örnek üzerinde gerçekleştirilmiştir. Bu çalışmada, bakteri türlerinin farklı uzunluklardaki DNA dizilimleri alınarak, bu dizilimlerde tekrar eden nükleotid çiftlerin frekansı bulunmuş ve elde edilen bu frekans değerlerine YSA, DVM ve k-NN yöntemleri uygulanarak bir sınıflandırılma yapılmıştır. Elde edilen sınıflandırma sonucunda k-NN yönteminin, YSA ve DVM yöntemlerinden başarımlarının daha yüksek olduğu sonucuna varılmıştır [100].

Daş ve Türkoğlu, National Center for Biotechnology (NCBI) sitesi gen bankasından gerçek veriler alınarak DNA dizilimlerinin sınıflandırılmasında karar ağacı algoritmalarının başarımları karşılaştırılmıştır. Alınan *Escherichia coli*, *Bacillus cereus*,

Buchnera aphidicola ve *Enterobacter cloacae* gibi 4 bakteri türü sınıflandırılma için kullanılmıştır. Çalışmada sınıflandırma yöntemlerinden J48, LMT ve Rastgele Orman karar ağacı algoritmaları kullanılmıştır. Bu çalışma, gen bankasından alınan *Escherichia coli*, *Bacillus cereus*, *Buchnera aphidicola* ve *Enterobacter cloacae* gibi 4 bakteri türünden oluşan toplam 154 örnek üzerinde gerçekleştirilmiştir. Bu çalışmada, bakteri türlerinin farklı uzunluklardaki DNA dizilimleri alınarak, bu dizilimlerde tekrar eden nükleotid çiftlerin frekansı bulunmuş ve elde edilen bu frekans değerlerine Karar Ağacı algoritmalarından J48, LMT ve Rastgele Orman algoritmaları uygulanarak bir sınıflandırma yapılmıştır. Sınıflandırma sonucunda, Rastgele Orman algoritmasının J48 ve LMT algoritmalarının başarımından daha yüksek olduğu ve hatta %100 doğruluk oranıyla sınıflandırma yaptığı sonucuna varılmıştır [101].

Pirooznia ve arkadaşları, Mikroarray gen ekspresyon verilerini kullanarak bu veriler üzerinde sınıflandırma yöntemlerinin başarımlarını birbirleriyle karşılaştırmışlardır. Mikroarray gen teknolojisi, genomik bir ölçek üzerinde genlerin diziliş şeklini bilim adamlarının gözlemlemesine izin verir. Böylece gen ekspresyon seviyesinde kanser sınıflandırması ve tanı olasılığını artırır. Bu çalışmada mikroarray gen ekspresyon verileri üzerinden kansere sebep olan genleri bularak hangi genlerin hastalık yapıcı olduğunu keşfederek ilerde bir hastanın kanser olup olmayacağı fikrini elde etmişlerdir. Çalışma için 8 adet mikroarray kanser veri seti kullanılmıştır. Bu veri setleri öncelikli olarak ön işleminden geçirilmiş daha sonra yüksek boyutlara sahip olan veri setleri, boyut azaltma tekniklerinden Support Vector Machine Recursive Feature Elimination (SVM-RFE), Chi Squared ve KTÖS kullanılmış ve bu öz nitelik seçme yöntemlerinin de başarımları kıyaslanmıştır. Lymphoma veri seti, 7129 gene sahip olan normal ve kötü huylu plazma hücrelerinden gelen 25 örneği içermektedir. Göğüs kanseri, 1753 gene sahip olan normal ve tümör alt tiplerinin 84 örneğini içerirken Kolon kanseri, 7464 gene sahip Epithelial, normal ve tümör hücrelerinin 45 örneğini içermektedir. Akciğer kanseri, 917 gene sahip olan normal ve kötü huylu hücrelerden gelen 72 örneği bulundururken Adenocarcinoma, 5377 gene sahip olan akciğer adenocarcinomas' lardan gelen 86 örneği bünyesinde bulundurmaktadır. Farklı bir Lymphoma veri seti ise 4027 gene sahip olan DLBCL1 ve DLBCL2 hücrelerinin 96 örneğini; Melanoma, 8067 gen içeren normal ve kötü huylu hücrelerin 38 örneğini; Yumurtalık kanseri ise, 7129 gene sahip olan normal ve kötü huylu hücrelerin 39 örneğini içerir. Algoritmaları

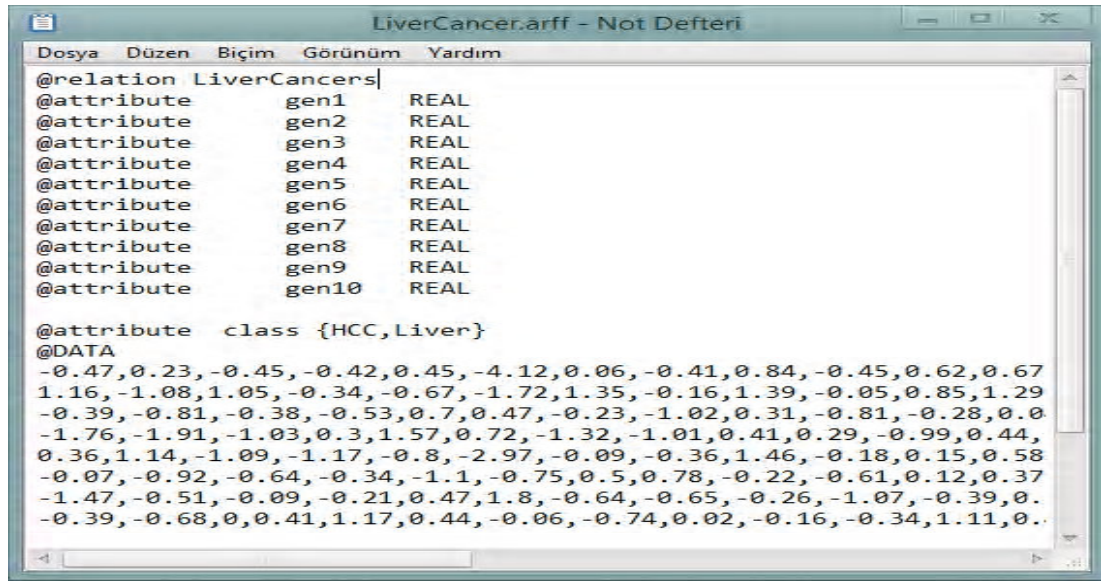
çalıştırırken test yöntemi olarak “10-katlı çapraz doğrulama” yöntemi kullanılmıştır. Çalışmada sınıflandırma yöntemlerinden Destek Vektör Makineleri (DVM), Bayes Ağları, J48 Karar Ağacı, Rastgele Orman, Id3, Bagging, Çok Katmanlı Algılayıcılar (ÇKA) ve Radyal Tabanlı Yapay Sinir Ağları (RTYSA) kullanılmıştır. Sınıflandırma sonucunda DVM, ÇKA, RTYSA ve Bayes Ağlarının daha iyi sonuçlar verdiği görülmüştür [102].

Hu ve arkadaşları, ön işlem den geçirilmiş 7 adet mikroarray gen ekspresyon veri seti üzerinde 5 farklı sınıflandırma yöntemi uygulayıp başarımlarını karşılaştırmışlardır. Algoritmaları çalıştırırken yöntem olarak “10-katlı çapraz doğrulama” yöntemi kullanılmıştır. Çalışmada sınıflandırma yöntemlerinden C4.5, Rastgele Orman, AdaBoost, Bagging ve DVM kullanılmıştır. Sınıflandırma sonucunda Rastgele Orman ve AdaBoost algoritmalarının diğer algoritmalara göre daha iyi sonuçlar verdiği görülmüştür [103].

3.2. Parametreler, Uygulama ve Sonuçlar

Sınıflandırma algoritmaları; açık kaynak kodlu, Java programlama dili ile yazılan Weka ve Ant-Miner programları ile uygulanmıştır. Ayrıca programlar i7 işlemcili 2.90 GHz hızlı 12 GB RAM ve 64 bit işletim sistemine sahip Notebook bilgisayarda koşulmuştur.

Çalışmada 10 adet farklı mikroarray gen ekspresyon kanser veri seti kullanıldı. Bu veri setleri öncelikli olarak ön işlem süreçlerinden geçirilerek sınıflandırma başarımını etkileyecek olumsuzluklar ortadan kaldırıldı. Daha sonra veriler üzerinde normalizasyon işlemleri gerçekleştirildi. Diğer aşamada ise Weka ve Ant-Miner programlarının desteklediği arff uzantılı dosyalarda bu veriler programların anlayacağı duruma getirildi. Bu dosyadaki attribute’ler genleri temsil etmektedir. Yani her veri setindeki gen sayısı verinin boyutunu temsil etmektedir. Şekil 3.1’de arff uzantılı dosya ve bu dosyada bir mikroarray gen ekspresyon kanser verisi görülmektedir.



```

LiverCancers.arff - Not Defteri
Dosya  Düzen  Biçim  Görünüm  Yardım
@relation LiverCancers
@attribute      gen1      REAL
@attribute      gen2      REAL
@attribute      gen3      REAL
@attribute      gen4      REAL
@attribute      gen5      REAL
@attribute      gen6      REAL
@attribute      gen7      REAL
@attribute      gen8      REAL
@attribute      gen9      REAL
@attribute      gen10     REAL

@attribute class {HCC,Liver}
@DATA
-0.47,0.23,-0.45,-0.42,0.45,-4.12,0.06,-0.41,0.84,-0.45,0.62,0.67
1.16,-1.08,1.05,-0.34,-0.67,-1.72,1.35,-0.16,1.39,-0.05,0.85,1.29
-0.39,-0.81,-0.38,-0.53,0.7,0.47,-0.23,-1.02,0.31,-0.81,-0.28,0.0
-1.76,-1.91,-1.03,0.3,1.57,0.72,-1.32,-1.01,0.41,0.29,-0.99,0.44,
0.36,1.14,-1.09,-1.17,-0.8,-2.97,-0.09,-0.36,1.46,-0.18,0.15,0.58
-0.07,-0.92,-0.64,-0.34,-1.1,-0.75,0.5,0.78,-0.22,-0.61,0.12,0.37
-1.47,-0.51,-0.09,-0.21,0.47,1.8,-0.64,-0.65,-0.26,-1.07,-0.39,0.
-0.39,-0.68,0,0.41,1.17,0.44,-0.06,-0.74,0.02,-0.16,-0.34,1.11,0.

```

Şekil 3.1. Arff dosya yapısı

Daha sonra yüksek boyutlu olan bu veriler üzerinde boyut azaltma tekniklerinden Korelasyon tabanlı Öznitelik Seçim yöntemi (KTÖS) kullanıldı. Mikroarray verileri üzerinde boyut azaltma tekniklerinin kullanılmasının sebebi ise sınıflandırmada başarıyı düşürecek olan veya sınıflandırmaya katkısı olmayan genleri elemek hem de sınıflandırma yöntemlerinin zorlanmadan daha az boyutla daha fazla başarı vermesini sağlamaktır.

Boyutları azaltılan mikroarray gen ekspresyon kanser verilerine sırasıyla istatistiksel yöntemlerden Naive Bayes, DVM, Bagging, OneR ve J48 Karar Ağacı yöntemleri; yapay zekâ tabanlı yöntemlerden ise Tek Katmanlı Algılayıcı (TKA), Çok Katmanlı Algılayıcı (ÇKA), RTYSA, cAnt-Miner ve Bulanık k-NN yöntemleri uygulanmıştır. Hem istatistiksel hem de yapay zekâ tabanlı sınıflandırma yöntemleri farklı veri setleri üzerinde farklı parametre değerleri ile denenmiştir. Sınıflandırma yöntemlerinin kontrol parametreleri belirlenirken literatürde yaygın olarak kullanılan parametre değerleri kullanılmıştır.

Kontrol parametre değerleri, sınıflandırma yapan yöntemlerin başarıyı üzerinde önemli bir etkiye sahiptir. Ancak bu tür yöntemler için optimal parametre değerlerinin belirlenmesi, başlı başına ayrı bir araştırma konusunu oluşturmaktadır. Bu yüzden genellikle araştırmacılar tarafından yaygın olarak önerilen kontrol parametre değerleri

kullanılmaktadır. Ayrıca bu tür yöntemlerin kontrol parametreleri, probleme özgü de önemli değişiklikler göstermektedir. Bu nedenle; tez çalışmasında dikkate alınan sınıflandırma problemleri için kullanılan yöntemlerin kontrol parametrelerinin optimal değerlerinin araştırılmasına yönelik çalışma da yapılmıştır. Kontrol parametreleri probleme özgü olarak önemli değişiklikler gösterdiği için her problem için uygulanan kontrol parametreleri aşağıdaki tablolarda gösterilmiştir.

Aşağıdaki Tablo 3.1-3.20’de literatürde yaygın olarak kullanılan kontrol parametreleri ve çalışmalar sonucu literatürdeki parametrelerden daha iyi sonuçlar veren optimal parametreler verilmiştir.

Bu sınıflandırma yöntemlerinin performanslarını test etmek için k-katlamalı çapraz doğrulama tekniği kullanılmış ve k değeri 10 seçilmiştir. Sınıflandırma yöntemlerinin optimal sonuçlar vermesi için her problem üzerine araştırmalar sonucu elde edilen optimal kontrol parametreleri uygulanmış ve daha iyi sonuçların elde edildiği gözlenmiştir.

Tablo 3.1. Karaciğer Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

Karaciğer Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10, Başlangıç (seed) değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=13 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3,Bulanıklık parametresi=3.0

Tablo 3.2. Karaciğer Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

Karaciğer Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=200, Başlangıç (seed) değeri=2
OneR	Minimum bucket sayısı=3
J48	Yaprak başına minimum örnek sayısı=5
TKA	Öğrenme Oranı=0.3, İterasyon sayısı=500
ÇKA	Gizli katman sayısı=1, Gizli katmandaki nöron sayısı=4 Öğrenme Oranı=0.3, İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=3 Minimum Standart Sapma=0.4
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=2000 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=5, Bulanıklık parametresi=3.0

Tablo 3.3. RNA Doku Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

RNA doku Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10 Başlangıç (seed) değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500
ÇKA	Gizli katmandaki nöron sayısı=11 Gizli katman sayısı=1 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.4. RNA Doku Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

RNA doku Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=lineer
Bagging	İterasyon sayısı=10 Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=11 Öğrenme Oranı=0.3, İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.5. Prostat Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

Prostat Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10, Başlangıç (seed) değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=0
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=13 Öğrenme Oranı=0.3 İterasyon Sayısı=500, Seed=1
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.6. Prostat Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

Prostat Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=lineer
Bagging	İterasyon sayısı=30 Başlangıç (seed) değeri=2
OneR	Minimum bucket sayısı=10
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1, İterasyon sayısı=500, Seed=2
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=13 Öğrenme Oranı=0.3 İterasyon Sayısı=500, Seed=0
RTYSA	k-Means için üretilecek küme sayısı=3 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=150
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.7. Merkezi Sinir Sistemi Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

Merkezi Sinir Sistemi Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10, Başlangıç (seed) değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1, İterasyon sayısı=500, Seed=0
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=16 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum iterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.8. Merkezi Sinir Sistemi Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

Merkezi Sinir Sistemi Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=lineer
Bagging	İterasyon sayısı=10 Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=8
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=2
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=16 Öğrenme Oranı=0.3, İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum iterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=40
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.9. Beyin Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

Beyin Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10, Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1, İterasyon sayısı=500
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=19 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.10. Beyin Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

Beyin Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=lineer
Bagging	İterasyon sayısı=20 Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=4
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=19 Öğrenme Oranı=0.3, İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.4
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=20 Yakınsama test sayısı=20 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=20
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.11. Rahim Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

Rahim Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10 Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=0
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=28 Öğrenme Oranı=0.3, İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.12. Rahim Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

Rahim Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=lineer
Bagging	İterasyon sayısı=20 Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=8
J48	Yaprak başına minimum örnek sayısı=10
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=2
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=28 Öğrenme Oranı=0.3, İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=5 Yakınsama test sayısı=5 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=5, Karınca sayısı=350
Bulanık k-NN	k=1, Bulanıklık parametresi=3.0

Tablo 3.13. Göğüs Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

Göğüs Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10 Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=0
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=21 Öğrenme Oranı=0.3, İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.14. Göğüs Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

Göğüs Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=lineer
Bagging	İterasyon sayısı=20 Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=2
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=21 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=40
Bulanık k-NN	k=1, Bulanıklık parametresi=3.0

Tablo 3.15. Mesane Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

Mesane Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10 Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=0
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=23 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.16. Mesane Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

Mesane Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=polinomsal
Bagging	İterasyon sayısı=20 Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=1
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=50 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=4 Minimum Standart Sapma=0.3
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum İterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=20
Bulanık k-NN	k=10, Bulanıklık parametresi=2.0

Tablo 3.17. Farklı Doku Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

Farklı doku Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10, Başlangıç (seed)değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=0
ÇKA	Gizli katman sayısı=1, Gizli katmandaki nöron sayısı=10, Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum iterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.18. Farklı Doku Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

Farklı doku Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=lineer
Bagging	İterasyon sayısı=30 Başlangıç (seed)değeri=2
OneR	Minimum bucket sayısı=5
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=5
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=10 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.3
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=5 Yakınsama test sayısı=5 Maksimum iterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=5, Karınca sayısı=200
Bulanık k-NN	k=5, Bulanıklık parametresi=2.0

Tablo 3.19. Kolon Kanseri Yöntemlerinin Literatürdeki Yaygın Kontrol Parametreleri

Kolon Kanseri	Literatürdeki Yaygın Parametreler
DVM	Çekirdek tipi=radyal
Bagging	İterasyon sayısı=10 Başlangıç (Seed) değeri=1
OneR	Minimum bucket sayısı=6
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=0
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=26 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=2 Minimum Standart Sapma=0.1
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=10 Yakınsama test sayısı=10 Maksimum iterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=10, Karınca sayısı=60
Bulanık k-NN	k=3, Bulanıklık parametresi=3.0

Tablo 3.20. Kolon Kanseri Yöntemlerinin Optimal Kontrol Parametreleri

Kolon Kanseri	Optimal Parametreler
DVM	Çekirdek tipi=lineer
Bagging	İterasyon sayısı=50 Başlangıç (Seed) değeri=1
OneR	Minimum bucket sayısı=10
J48	Yaprak başına minimum örnek sayısı=2
TKA	Öğrenme Oranı=0.1 İterasyon sayısı=500, Seed=6
ÇKA	Gizli katman sayısı=1 Gizli katmandaki nöron sayısı=26 Öğrenme Oranı=0.3 İterasyon Sayısı=500
RTYSA	k-Means için üretilecek küme sayısı=3 Minimum Standart Sapma=0.3
cAnt-Miner	Kural başına kapsanan minimum örnek sayısı=5 Yakınsama test sayısı=5 Maksimum iterasyon sayısı=1500 Eğitim kümesindeki kapsanmayan örneklerin maksimum sayısı=5, Karınca sayısı=250
Bulanık k-NN	k=1, Bulanıklık parametresi=3.0

4. BÖLÜM

TARTIŞMA, SONUÇ VE ÖNERİLER

4.1. Tartışma, Sonuç ve Öneriler

Verilerin sınıflandırılmasında birçok farklı yöntem bulunmakta ve bu sınıflandırma yöntemleri birçok alanda etkin bir şekilde kullanılmaktadır. Günümüz teknolojisindeki gelişmeye bağlı olarak mühendislik ve özellikle de tıp alanında bu sınıflandırma yöntemlerine rağbet çok fazla olmaktadır. DNA'larda bulunan genlerin teknolojik araçlar vasıtasıyla ortaya çıkarılmasıyla birlikte bu genler üzerindeki veriler sınıflandırma yöntemleri ile işlenerek kişilerin hastalık geni taşıyıp taşımadığı ortaya çıkarılmakta ve bu kişinin ileride hangi hastalığa yakalanıp yakalanmayacağı tahmin edilebilmektedir.

Literatürde mikroarray gen ekspresyon verileri üzerine birçok çalışma yapılmıştır. Fakat bu çalışmaların çoğunda istatistiksel yöntemler kullanılmıştır. Bu tez çalışmasında ise istatistiksel yöntemlerin dışında farklı türden yapay sinir ağları, bir optimizasyon algoritması olan ve doğadaki gerçek karıncaların davranışlarından esinlenilerek ortaya çıkarılan karınca koloni algoritması ve bulanık mantıkla istatistiksel yöntemin çalışması birleştirilerek ortaya atılan hibrid bir algoritma olan Bulanık mantık tabanlı k-en yakın komşu (Bulanık k-NN) algoritması kullanılmıştır. Literatürde mikroarray gen verilerini sınıflandırmak için bu kadar farklı sınıflandırma algoritması problemleri çözmek için bir arada kullanılmadığından dolayı bu çalışma diğer çalışmalarla tam anlamıyla karşılaştırılamamaktadır. Ayrıca bu tez çalışmasında diğer çalışmalardan farklı olarak daha fazla veri seti üzerinde çalışılmış ve 10 adet farklı mikroarray gen ekspresyon kanser veri seti kullanılmıştır. Sınıflandırma yöntemlerinin farklı problemler üzerindeki başarımlarını ayrıntılı olarak test edebilmek için birçok farklı kanser verisi kullanılmıştır. Tablo 4.3'de araştırmacılar tarafından yaygın olarak önerilen kontrol parametreleri, 10 farklı mikroarray gen ekspresyon kanser verisi üzerinde kullanılarak

problemlere çözüm aranmıştır. Bu kontrol parametreleri optimize edilmemiştir ve sadece literatürde yaygın kullanılan kontrol parametreleri ile mikroarray problemlerine çözüm aranmıştır. Literatürde yaygın olarak kullanılan kontrol parametreleri ile elde edilen sınıflandırma yöntemlerinin başarımları Tablo 4.1-Tablo 4.2’de verilmiştir.

Tablo 4.1. Mikroarray Kanser Verilerinin Sınıflandırılmasında Literatürde Yaygın Kullanılan Kontrol Parametre Değerleriyle İstatistiksel Yöntemlerin Performansları

Veri Setleri	NB	DVM	Bagging	OneR	J48
Karaciğer Kanseri	91.62	96.65	88.27	74.86	84.92
RNA doku Kanserleri	98.08	59.62	97.12	97.12	96.15
Prostat Kanseri	94.12	50.98	87.25	73.53	86.27
Merkezi Sinir Sistemi Kanseri	94.12	73.53	82.35	73.53	85.29
Beyin Kanseri	89.29	39.29	82.14	82.14	71.43
Rahim Kanseri	85.71	76.19	80.95	47.62	73.81
Göğüs Kanseri	97.96	51.02	87.76	91.84	89.80
Mesane Kanseri	80.00	50.00	92.50	80.00	85.00
Farklı doku Kanserleri	95.98	16.09	86.21	39.66	79.89
Kolon Kanseri	94.59	78.38	89.19	86.49	89.19
Ortalama Başarı (%)	92.15	59.18	87.37	74.68	84.18

Tablo 4.2. Mikroarray Kanser Verilerinin Sınıflandırılmasında Literatürde Yaygın Kullanılan Kontrol Parametre Değerleriyle Yapay Zeka Tabanlı Yöntemlerin Performansları

Veri Setleri	TKA	ÇKA	RTYSA	cAnt Miner	Bulanık kNN
Karaciğer Kanseri	89.39	93.30	94.97	85.05	96.65
RNA doku Kanserleri	97.12	98.08	98.08	96.19	98.08
Prostat Kanseri	92.16	90.20	90.20	85.45	90.20
Merkezi Sinir Sistemi Kanseri	91.18	97.06	94.12	82.50	88.24
Beyin Kanseri	53.57	100	92.86	80.33	92.86
Rahim Kanseri	83.33	92.86	85.71	59.40	73.81
Göğüs Kanseri	89.80	97.96	100	87.60	83.67
Mesane Kanseri	80.00	87.50	87.50	81.00	90.00
Farklı doku Kanserleri	83.33	95.98	91.95	74.20	87.36
Kolon Kanseri	91.89	97.30	94.59	80.17	75.68
Ortalama Başarı (%)	85.18	95.02	93.00	81.19	87.66

Sınıflandırma problemleri üzerine uygulanan yöntemler için kontrol parametreleri önem arz eden bir konudur ve sınıflandırma yöntemlerinin başarımları üzerinde önemli bir etkiye sahiptir. Ayrıca bu tür yöntemlerde kontrol parametreleri probleme özgü değişiklikler göstermektedir. Bu nedenle çalışmada kontrol parametrelerinin optimal değerlerinin araştırılmasına yönelik bir çalışma da yapılmış ve optimize edilecek kontrol parametreleri literatürde yaygın olarak kullanılan değerler etrafında belirli bir uzayda araştırılmıştır. Her probleme özgü optimal kontrol parametrelerine sahip sınıflandırma yöntemleri, mikroarray kanser verilerine uygulanmıştır.

Farklı problemler üzerinde, her yöntem için 30 farklı deneme yapılmış ve elde edilen optimal kontrol parametre değerleriyle yöntemlerin problemler üzerindeki başarımları incelenmiştir. Elde edilen sonuçlar Tablo 4.3-Tablo 4.4’ te verilmiştir.

Tablo 4.3. Mikroarray Kanser Verilerinin Sınıflandırılmasında Optimal Kontrol Parametre Değerleriyle İstatistiksel Yöntemlerin Performansları

Veri Setleri	NB	DVM	Bagging	OneR	J48
Karaciğer Kanseri	91.62	96.65	91.62	75.98	86.03
RNA doku Kanserleri	98.08	96.15	97.12	97.12	96.15
Prostat Kanseri	94.12	89.22	92.16	78.43	86.27
Merkezi Sinir Sistemi Kanseri	94.12	97.06	82.35	79.41	85.29
Beyin Kanseri	89.29	96.43	85.71	82.14	82.14
Rahim Kanseri	85.71	83.33	83.33	66.67	76.19
Göğüs Kanseri	97.96	89.80	89.80	91.84	89.80
Mesane Kanseri	80.00	92.50	95.00	80.00	85.00
Farklı doku Kanserleri	95.98	91.38	87.93	40.80	79.89
Kolon Kanseri	94.59	97.30	91.89	91.89	89.19
Ortalama Başarı (%)	92.15	92.98	89.69	78.43	85.60

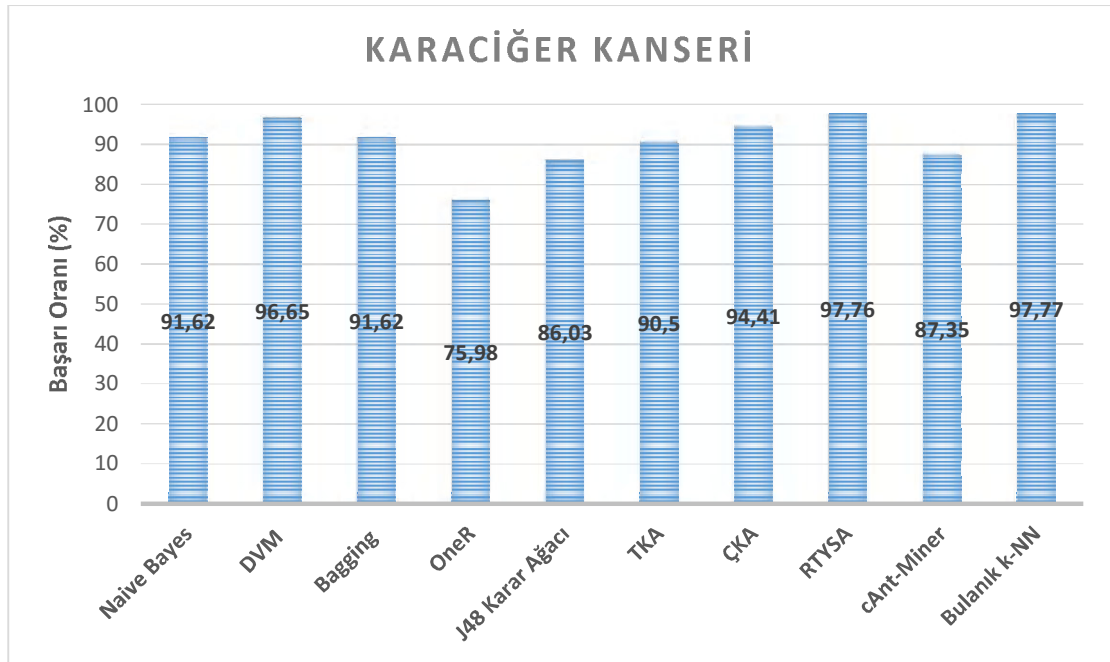
Tablo 4.4. Mikroarray Kanser Verilerinin Sınıflandırılmasında Optimal Kontrol Parametre Değerleriyle Yapay Zeka Tabanlı Yöntemlerin Performansları

Veri Setleri	TKA	ÇKA	RTYSA	cAnt Miner	Bulanık kNN
Karaciğer Kanseri	90.50	94.41	97.76	87.35	97.77
RNA doku Kanserleri	97.12	98.08	98.08	96.19	98.08
Prostat Kanseri	94.12	91.18	94.12	86.04	90.20
Merkezi Sinir Sistemi Kanseri	97.06	97.06	94.12	86.42	88.24
Beyin Kanseri	53.57	100	96.43	88.83	92.86
Rahim Kanseri	88.10	92.86	85.71	70.60	76.19
Göğüs Kanseri	91.84	97.96	100	92.03	89.80
Mesane Kanseri	90.00	92.50	92.50	85.00	92.50
Farklı doku Kanserleri	87.93	95.98	93.10	78.62	87.93
Kolon Kanseri	94.59	97.30	100	84.33	83.78
Ortalama Başarı (%)	88.48	95.73	95.18	85.54	89.74

Kontrol parametreleri optimize edildikten sonra bu parametre değerleriyle sınıflandırma yöntemlerinin problemlere uygulanması sonucu elde edilen bilgilere göre optimize edilen parametrelerle yapılan başarımın, optimize edilmeyen parametrelerle elde edilen başarımından çok daha fazla olduğu görülmektedir (Tablo 4.3-Tablo 4.4). Neredeyse bütün sınıflandırma yöntemlerinin başarımlarının, kontrol parametreleri optimize edildikten sonra arttığı görülmektedir (Tablo 4.3-Tablo 4.4).

Deney sonuçlarından elde edilen bilgilere göre; genel anlamda yapay zekâ yöntemleri, istatistiksel yöntemlere göre sınıflandırma problemlerinde daha başarılı bir tablo sergilemiştir. Çok katmanlı Algılayıcılar (ÇKA), farklı sınıflandırma problemleri içerisinde ortalama olarak en iyi başarıya ulaşan yöntem olmuş ve arkasından ise yine yapay zeka tabanlı Radyal Tabanlı Yapay Sinir Ağları (RTYSA), sınıflandırma problemlerinde ortalama olarak ikinci başarılı yöntem olmuştur (Tablo 4.4). İstatistiksel yöntemler içerisinde en başarılı sonuçları veren yöntemler; Naive Bayes ve DVM olmuştur (Tablo 4.3). Yapay zekâ tabanlı yöntemler içerisinde ise MLP ve RTYSA diğer yapay zekâ tabanlı yöntemler olan TKA, cAnt-Miner ve Bulanık k-NN'den daha iyi sonuçlar vermiştir (Tablo 4.4).

Çalışmada kullanılan sınıflandırma yöntemlerinin başarısı, her veri seti üzerinde algoritmaların başarımını optimal yapacak parametreler ile ölçülmüş ve her sınıflandırma algoritması, veri setleri üzerinde farklı optimal parametreler ile sınanmıştır. Çalışmada kullanılan veri setleri üzerinde sınıflandırma algoritmalarının optimal kontrol parametre değerleriyle başarımları aşağıdaki Şekil 4.1 - 4.10 arasındaki grafiklerde gösterilmiştir.

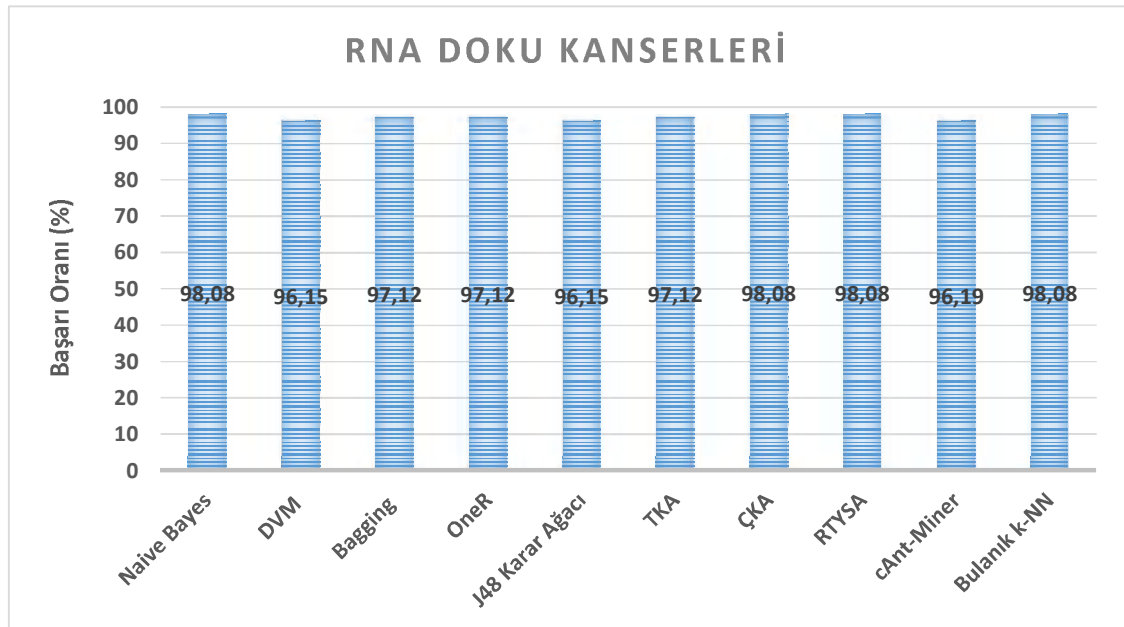


Şekil 4.1. Karaciğer kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

Karaciğer kanser veri seti, karaciğer dokularından alınan 179 mikroarray gen ekspresyon profili içerir. Her gözlem 22699 genden oluşur. Bu 22699 genden 85 tanesi anlamlı ve ilgili bulunmuştur. Veri kümesi kanserli hücreler ve normal hücreler olmak üzere 2 sınıf ile temsil edilmektedir. Bu iki sınıftan biri tümörlü dokulardan alınan HCC hastalığına sahip hücreler iken diğeri ise normal dokulardan alınan hücrelerdir. HCC hastalığına sahip sınıftaki gözlem sayısı 104 sağlıklı hücrelerin bulunduğu normal sınıftaki gözlem sayısı ise 75' tir.

Şekil 4.1'de; Karaciğer kanser verisi istatistiksel ve yapay zeka tabanlı yöntemler ile sınıflandırılmıştır. Bu sınıflandırıcıların doğruluk başarımı ise 10-katlı çapraz doğrulama yöntemiyle yapılmıştır. Kanser verisinin sınıflandırılması sonucunda Bulanık k-NN, %97.77'lik oran ile en başarılı sınıflandırıcı olmuştur. Buna rağmen OneR algoritmasının başarımının ise karaciğer kanseri mikroarray gen ifade verilerini sınıflandırmada %75.98 ile diğer sınıflandırma yöntemlerinden düşük olduğu gözlemlenmiştir. İstatistiksel yöntemler ile yapay zekâ tabanlı yöntemleri kendi aralarında değerlendirecek olursak, istatistiksel yöntemlerden en başarılı sınıflandırmayı %96.65'lik oranla DVM yaparken, yapay zekâ tabanlı yöntemlerden ise en başarılı

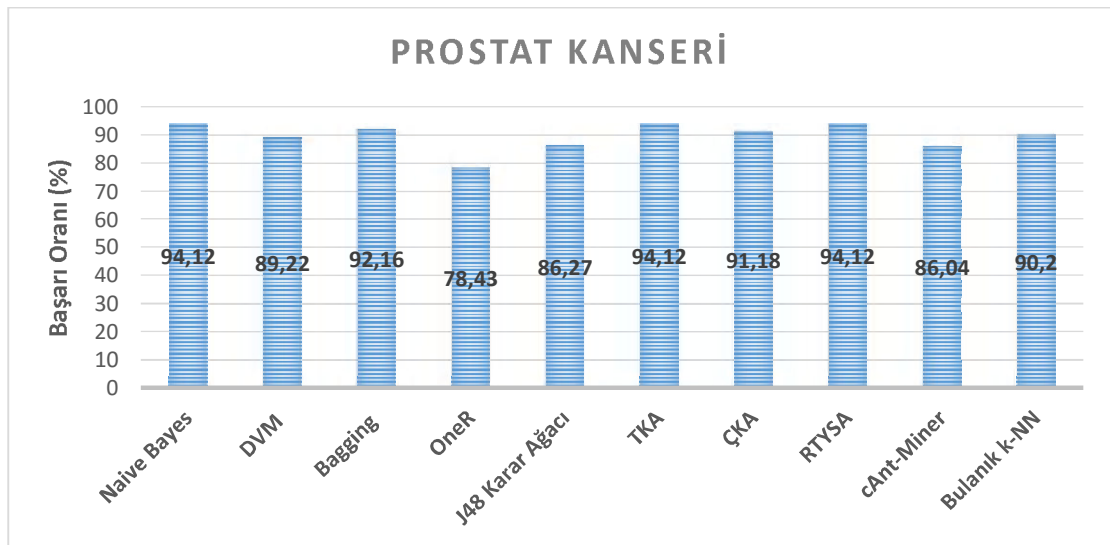
sınıflandırmayı %97.77 doğruluk oranıyla Bulanık k-NN yönteminin yaptığı gözlenmiştir. İstatistiksel yöntemler içerisinde en düşük sınıflandırmayı %75.98 ile OneR yaparken yapay zekâ tabanlı yöntemler içerisinde en düşük sınıflandırmayı %87.35 ile cAnt-Miner'in yaptığı görülmüştür. Genel anlamda karaciğer kanser verisini sınıflandırmada yapay zekâ tabanlı yöntemlerin istatistiksel yöntemlerden daha başarılı olduğu görülmüştür.



Şekil 4.2. RNA doku kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

RNA doku kanseri veri setinde, göğsün RNA dokusundan çıkarılan 62 lymph node-negative göğüs tümörü ve kolonun RNA dokusundan çıkarılan 42 kolon tümörü olmak üzere toplamda 104 örnek bulunmaktadır. Her örnek 22283 mikroarray gen ekspresyon değeri içerir. Fakat kalite açısından bu mikroarray gen ekspresyon değerlerinin sadece 182 tanesi kullanılmıştır. Veri seti, göğüs kanseri tümörü ve kolon kanseri tümörü olmak üzere 2 sınıfla temsil edilmektedir.

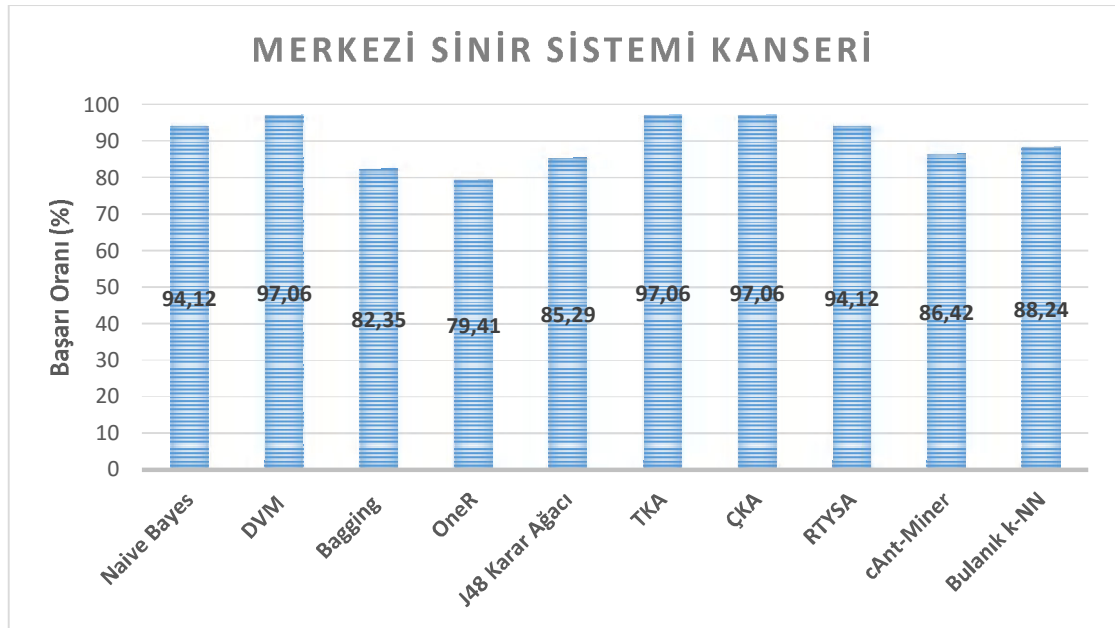
Şekil 4.2’de; RNA doku kanseri veri setinin sınıflandırılmasında yapay zeka ve istatistiğe dayanan sınıflandırma algoritmalarının başarımlarının doğruluğu 10-katlı çapraz doğrulama sistemi ile test edilmiştir. Test edilen sınıflandırma algoritmalarından veri üzerinde en başarılı olanlar Naive Bayes, ÇKA, RTYSA ve Bulanık k-NN algoritmalarıdır. Bu algoritmaların veri setini sınıflandırmadaki başarımları aynıdır ve oranları %98.08’dir. Veri kümesi üzerinde en düşük başarılı sınıflandırmayı ise %96.15 doğruluk oranıyla DVM ve J48 Karar Ağacı’nın yaptığı görülmektedir. İstatistiksel metotlar içinden en başarılı yöntem %98.08’lik doğruluk derecesiyle Naive Bayes olurken yapay zekâ metotlarından ise %98.08 doğruluk derecesiyle ÇKA, RTYSA ve Bulanık k-NN’ nin olduğu görülmektedir. Ayrıca istatistiksel yöntemlerden olan DVM ve J48 Karar Ağacı %96.15 başarımla ile diğer istatistiksel yöntemlerin başarımlarının gerisinde kalırken yapay zekâ yöntemlerinde ise cAnt-Miner %96.19 başarımla derecesiyle diğer yapay zekâ yöntemlerinin başarımlarının gerisinde kalmıştır. Buna rağmen veriyi sınıflandırmada bütün sınıflandırıcılar başarılı olmuştur. Çünkü yöntemler içerisinde en düşük sınıflandırma derecesine sahip yöntem bile %96.15 başarı göstermiştir. Özetleyecek olursak, veri üzerinde hem istatistiğe hem de yapay zekâyâ dayanan yöntemler başarılı olmuş fakat yapay zekâ tabanlı sınıflandırma algoritmalarının daha iyi sonuçlar verdiği görülmüştür.



Şekil 4.3. Prostat kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

Prostat kanser veri seti, sağlıklı ve hastalıklı prostat dokularından alınan hücrelerden oluşmuştur. Hastalıklı prostat dokularından (PR) 52, sağlıklı prostat dokularından (N) ise 50 adet örnek alınarak toplamda 102 gözlem bulunmaktadır. Bu gözlemlerin her birinden ise 12600 gen çıkarılmış olmakla beraber bu genlerden sadece 339 tanesi kullanılmıştır. Bu prostat kanseri veri seti, prostat tümörü bulunan hücreler (PR) ve normal hücreler (N) olmak üzere 2 sınıf ile temsil edilmektedir.

Şekil 4.3’de; Prostat kanser veri seti üzerindeki çalışma sonuçları incelendiğinde 10-katlı çapraz doğrulama yöntemi ile Naive Bayes, TKA ve RTYSA algoritmalarının %94.12 başarıyla en iyi sonuçları verdiği söylenebilir. Oysa OneR algoritmasının ise %78.43 ile sınıflandırma başarısının diğer algoritmaların başarılarından daha düşük olduğu gözlenmiştir. Bu veri seti üzerinde istatistiksel yöntemler ile yapay zekâ tabanlı sınıflandırma yöntemlerini ayrı değerlendirecek olursak; istatistiksel algoritmalarından en başarılısı %94.12 başarıyla Naive Bayes olurken sınıflandırma başarımı en düşük algoritma ise %78.43 sınıflandırma derecesiyle OneR olmuştur. Yapay zekâ algoritmalarından ise en başarılı olanlar %94.12’lik doğrulukla TKA ve RTYSA iken cAnt-Miner ise %86.04 doğruluk oranıyla sınıflandırma yeteneği diğer yapay zekâ algoritmalarından daha düşüktür. Özetle söylenecek olursa, yapay zekâ tabanlı sınıflandırma yöntemlerinin prostat kanser verisi üzerinde daha başarılı sınıflandırma yaptığı söylenebilir.

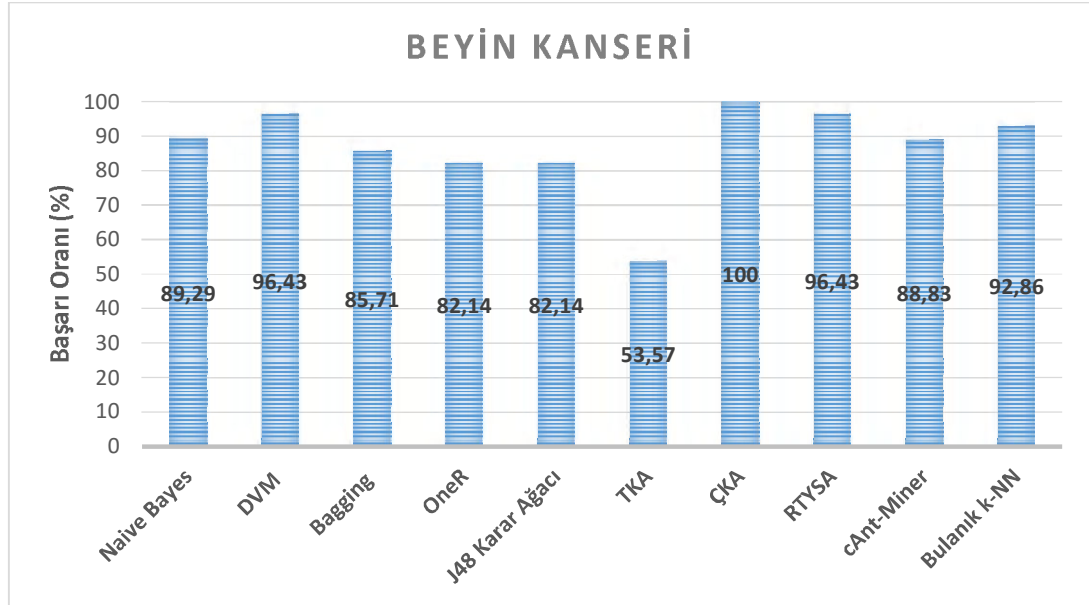


Şekil 4.4. Merkezi Sinir Sistemi kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

Merkezi Sinir Sistemi kanser veri seti, beyin dokularından alınan iki farklı kanser tümörü içerir. Bu veri seti 25 CMD tümöründen ve 9 da DMD tümöründen olmak üzere toplamda 34 örnek içermektedir. Bu tümörlerden ise toplamda 7129 gen çıkarılmaktadır. Fakat çıkarılan bu genlerin sadece 857 tanesi üzerinden işlem yapılmaktadır. Bu veri seti, CMD ve DMD olmak üzere iki farklı kanser türü ile temsil edilmektedir.

Şekil 4.4’de; Merkezi Sinir Sistemi kanser verisi, 10-katlı çapraz doğrulama yöntemi sayesinde istatistiksel ve yapay zeka yöntemleri kullanılarak sınıflandırılmıştır. Bu yöntemler içerisinde en başarılı sınıflandırmayı %97.06 doğruluk ile DVM, TKA ve ÇKA’nın yaptığı görülmüştür. Bu verinin sınıflandırılmasında ise OneR %79.41, Bagging ise %82.35 doğruluk derecesiyle en düşük başarıya sahiptir. İstatistiksel yöntemler içerisinde DVM %97.06 başarı oranıyla en yüksek performansa sahipken OneR %79.41 oranla en düşük performansa sahiptir. Yapay sinir ağlarının bir çeşidi olan TKA ve ÇKA %97.06 sınıflandırmayla en yüksek başarıya sahipken sezgisel bir algoritma olan karınca koloni algoritmasından esinlenilerek elde edilen cAnt-Miner’in ise %86.42 doğruluk derecesiyle en düşük başarıya sahip olduğu görülmüştür. Genel

olarak, yapay zekâ tabanlı sınıflandırıcıların merkezi sinir sistemi kanser verisini daha başarılı sınıflandırdıkları söylenebilir.

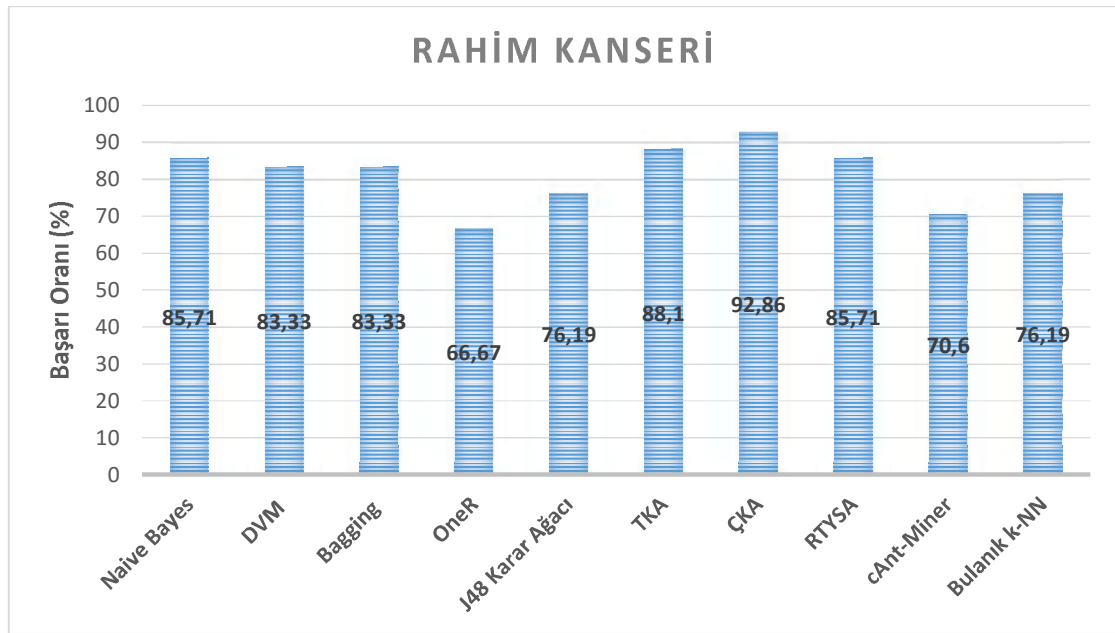


Şekil 4.5. Beyin kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

Beyin kanser veri seti, 28 hastanın beyin dokularından alınan mikroarray gen ekspresyon verisini içermektedir. Bu veri setinde, CG ve NG olmak üzere beyin kanserinin iki farklı alt türü bulunmaktadır. Bu sınıfların her biri 14 adet CG ve NG tümörü içermektedir. Tümörlerin her biri 12625 genin ifade değerini içermektedir. Bu genlerin ise sadece 1070 tanesi kullanılmaktadır.

Şekil 4.5’de görüldüğü üzere; Beyin kanser verisi üzerinde çeşitli sınıflandırma yöntemleri optimal kontrol parametre değerleri ile 10 katlı-çapraz doğrulama tekniğiyle uygulanmıştır. Veriyi sınıflandırmada en çok kullanılan sınıflandırma algoritmaları seçilmiş ve bu yöntemlerin başarımları birbirleriyle karşılaştırılmıştır. Beyin kanser verisini sınıflandırma da istatistiksel yöntemlerden Naive Bayes, DVM, Bagging, OneR ve J48 Karar Ağacı kullanılmış ve bunlar içerisinde en başarılısı %96.43 ile DVM olmuştur. Diğer taraftan bu yöntemlerin yanında yapay sinir ağları, sezgisel bir algoritma olan karınca koloni algoritması ve bulanık mantıkla istatistiksel bir yöntem

olan kNN' nin birleşiminden olan Bulanık k-NN algoritması kullanılmıştır. Bu yöntemler içerisinde ÇKA, test verilerinin tamamını doğru sınıflandırmıştır ve %100 başarılı olmuştur. Beyin kanser verisi üzerinde kullanılan algoritmalar içerisinde ise en başarılı sınıflandırma algoritması %100 doğruluk başarıyla ÇKA olurken bütün sınıflandırma algoritmalarının başarım olarak gerisinde kalan ise %53.57 oranla TKA olmuştur.

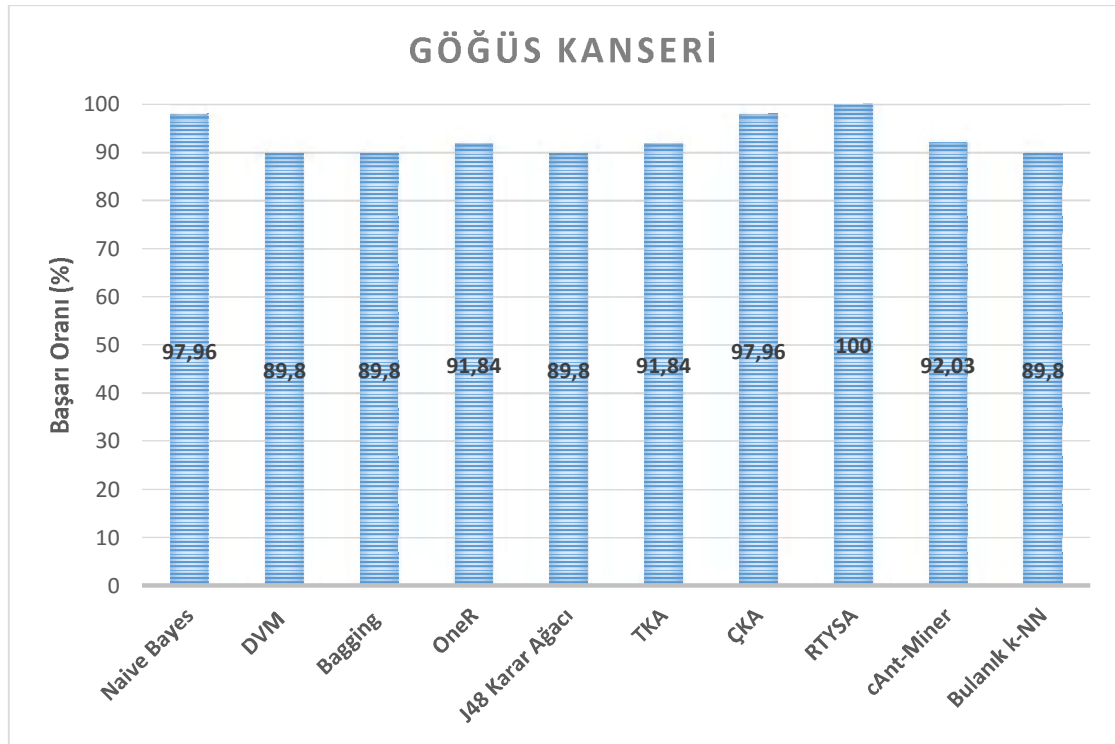


Şekil 4.6. Rahim kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

Rahim kanser veri seti, 42 tane hastanın rahim dokusundan alınan farklı kanser tümörlerinden oluşmaktadır. Bu veri setinde, 13 PS,3 CC,19 E ve 7 N tümörü olmak üzere toplamda 42 adet tümör ve 4 farklı kanser sınıfı bulunmaktadır. Bu kanserli dokuların her birinden ise toplam 8872 gen çıkarılmış ve bu genlerin yalnızca 1771 tanesi kullanılarak bilgi elde edilmeye çalışılmıştır.

Şekil 4.6'da görüldüğü üzere; 42 hastadan alınan ve 4 farklı rahim kanseri tümörü barındıran rahim kanseri verisi üzerinde 10-katlı çapraz doğrulama yöntemi kullanarak yapılan sınıflandırma sonuçlarına göre ÇKA'nın %92.86 doğruluk ölçüsüyle en başarılı

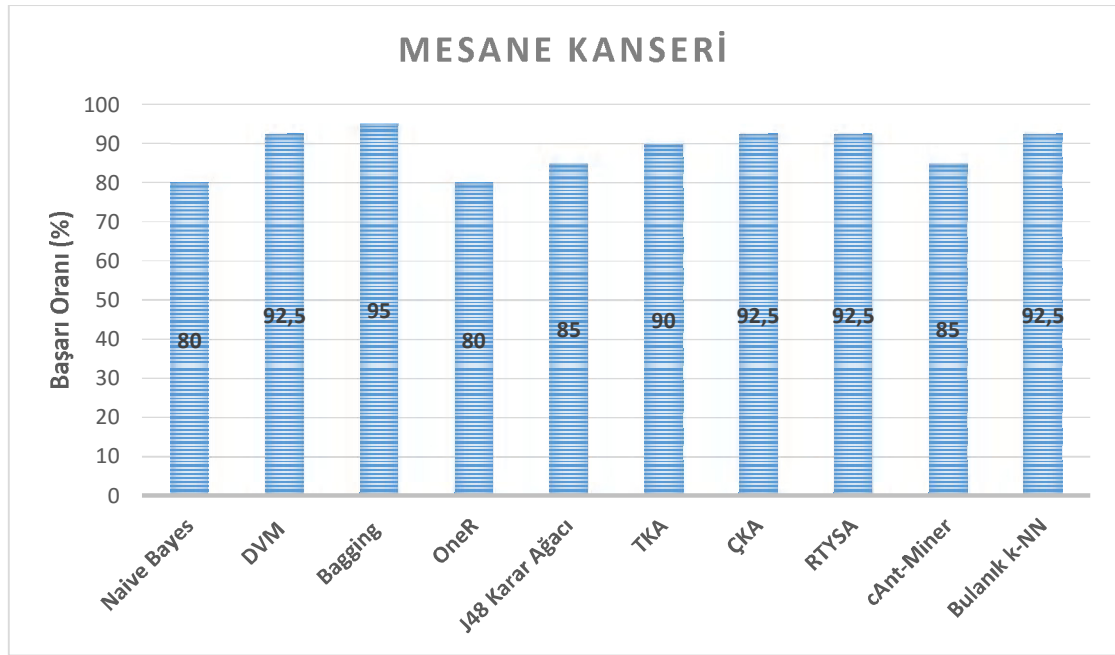
sınıflandırma metodu olduğu görülmüştür. Buna rağmen bu verileri sınıflandırmada OneR %66.67 sınıflandırma derecesiyle diğer yöntemlerin gerisinde kalmıştır. İstatistiksel yöntemler içinde en başarılı %85.71 ile Naive Bayes olurken en düşük başarıyı ise OneR'in %66.67 ile gösterdiği gözlenmiştir. Yapay zekâ algoritmaları içinde en yüksek başarı %92.86 ile ÇKA'ya aitken en düşük başarının ise %70.60 ile cAnt-Miner'e ait olduğu görülmüştür. Elde edilen sonuçlardan anlaşılabacağı üzere, rahim kanseri verisini sınıflandırmada yapay zekâ tabanlı yöntemlerin istatistiksel yöntemlerden daha iyi sonuçlar verdiği söylenebilir. Ayrıca yöntemlerin rahim kanseri problemini çözmede çok başarılı olduklarını söylemek mümkün değildir. Çünkü diğer veri setleri üzerinde özellikle istatistiksel yöntemler ortalama bir başarı aralığında iken bu problem üzerindeki performansları daha düşük kalmıştır.



Şekil 4.7. Göğüs kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

Göğüs kanser veri seti, hastaların göğüs dokularından alınan toplam 49 tümörden oluşmaktadır. Bu tümörlerden 25 tanesi Estrogen-Receptor-Positive (ER+) tümörü ve 24 tanesi ise Estrogen-Receptor-Negative (ER-) tümörüdür. Bu veri seti, 2 farklı göğüs kanser sınıfı ile temsil edilmektedir. Bu kanserli tümörlerin her biri 7129 genin ifade değerini içermektedir. Ancak kalite açısından ve tümörlerden daha sağlıklı ve düzgün bilgi alabilmek için bu genlerden sadece 1198 tanesi kullanılmış diğerleri ise göz ardı edilmiştir.

49 hastadan alınan göğüs kanser veri setine sırasıyla Naive Bayes, DVM, Bagging, OneR, J48 Karar Ağacı, TKA, ÇKA, RTYSA, cAnt-Miner ve Bulanık k-NN metotları optimal kontrol parametreleri kullanılarak uygulanmış ve bu metotlardan birkaç tanesinin bu problem üzerindeki performanslarının aynı olduğu görülmüştür (Şekil 4.7). RTYSA yönteminin göğüs kanser verisini %100 başarımla sınıflandırdığı görülmüştür. Buna rağmen göğüs kanseri problemini sınıflandırmada başarısı en düşük yöntemler ise %89.80 başarımla DVM, Bagging, J48 Karar Ağacı ve Bulanık k-NN'dir. İstatistiksel ve yapay zekâya dayanan sınıflandırma yöntemlerini kendi aralarında değerlendirecek olursak, istatistiksel yöntemlerden en başarılı sınıflandırmayı %97.96 ile Naive Bayes yaparken, başarımı en düşük sınıflandırmayı ise %89.80 ile DVM, Bagging ve J48 Karar Ağacı yapmıştır. Yapay zekâya dayanan yöntemlerden ise RTYSA verileri eksiksiz olarak %100 başarımla sınıflandırarak en başarılı yöntem olurken Bulanık k-NN %89.80 ile verileri doğru sınıflara ayırmada diğer yöntemlerin gerisinde kalmıştır. Özetleyecek olursak, göğüs kanserini sınıflandırmada yapay zekâ tabanlı yöntemlerin istatistiksel yöntemlerden daha iyi sonuçlar verdiği söylenebilir.

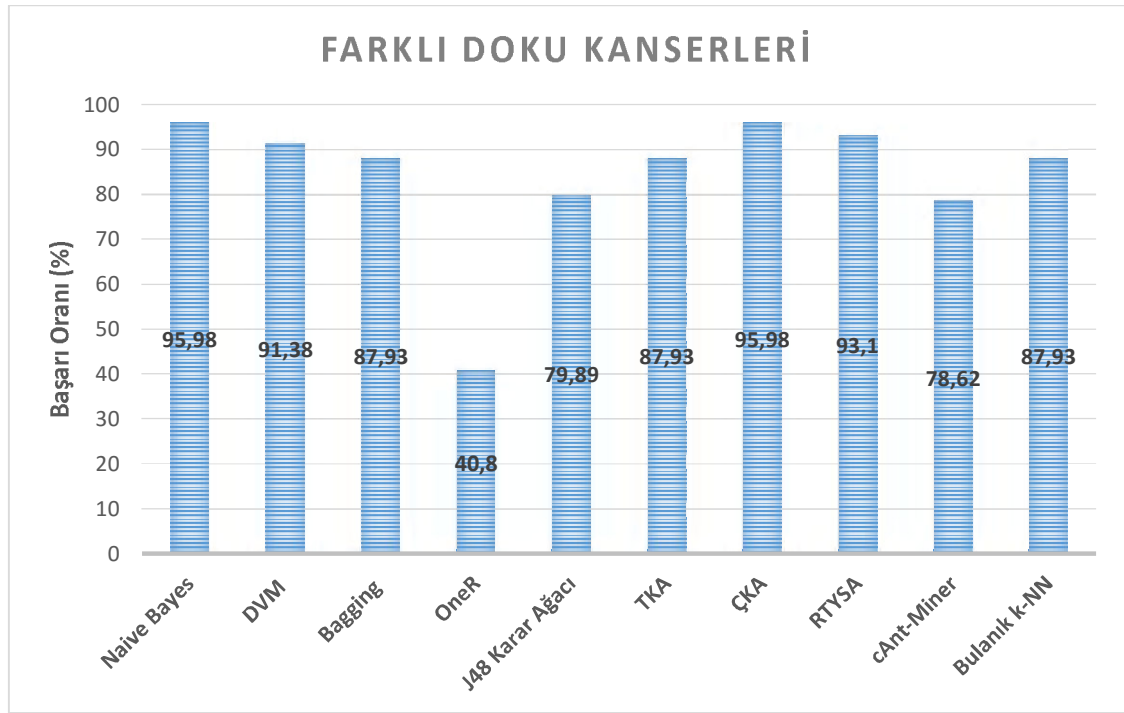


Şekil 4.8. Mesane kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

Mesane kanser veri setinde, mesane kanseri olan 40 farklı hastadan alınan örnekler bulunmaktadır. Bu örneklerden 9 tanesi T2+, 20 tanesi TA ve 11 tanesi T1 olmak üzere 3 farklı mesane kanseri türüne neden olan tümöre sahiptir. Bu tümörlerin her birinden ise 7129 adet gen çıkarılmış ve bu genlerden seçim kriterine bağlı olarak sadece 1203 genden faydalanılmış diğer genler elenmiştir. Veri seti, T2+, TA ve T1 gibi farklı mesane tümörleri olmak üzere 3 sınıfla temsil edilmektedir.

Şekil 4.8’de; mesane kanserinin 40 mikroyarray gen ifade verisi üzerinde farklı sınıflandırma yöntemleri ile en uygun parametre değerleri kullanılarak 10-katlı çapraz doğrulama yöntemi ile yapılan analizler sonucunda sınıflandırma yöntemlerinin farklı sonuçlar ürettiği gözlenmiştir. Mesane kanserinin 3 farklı alt türünü içeren bu veri kümesini sınıflandırmada en başarılı yöntemin %95.00 sınıflandırma doğruluğuyla istatistiksel bir yöntem olan Bagging olduğu gözlenirken en düşük başarıya sahip yöntemlerin ise %80.00 doğruluk derecesiyle yine istatistiksel yöntemlerden Naive Bayes ve OneR olduğu görülmüştür. Yapay zekâ tabanlı sınıflandırıcılar içerisinde en başarılı %92.50 başarıyla ÇKA, RTYSA ve Bulanık k-NN olurken, istatistiksel sınıflandırıcılar içinde ise veriyi sınıflandırmada en başarılı %95.00 ile Bagging

olmuştur. Yapay zekâ yöntemleri içerisinde veriyi sınıflara ayırmada ise %85.00 ile cAnt-Miner diğer yapay zekâ tabanlı sınıflandırıcıların gerisinde kalırken istatistiksel yöntemlerde %80.00 başarımla Naive Bayes ve OneR diğer istatistiksel yöntemlerin gerisinde kalmıştır. Özetlemek gerekirse, mesane kanser verilerini sınıflara ayırma probleminde en başarılı yöntem, istatistiksel bir sınıflandırıcı olan Bagging olurken ortalama olarak baktığımızda ise yapay zekâ tabanlı sistemlerin bu problem üzerinde daha fazla performans gösterdikleri söylenebilir.

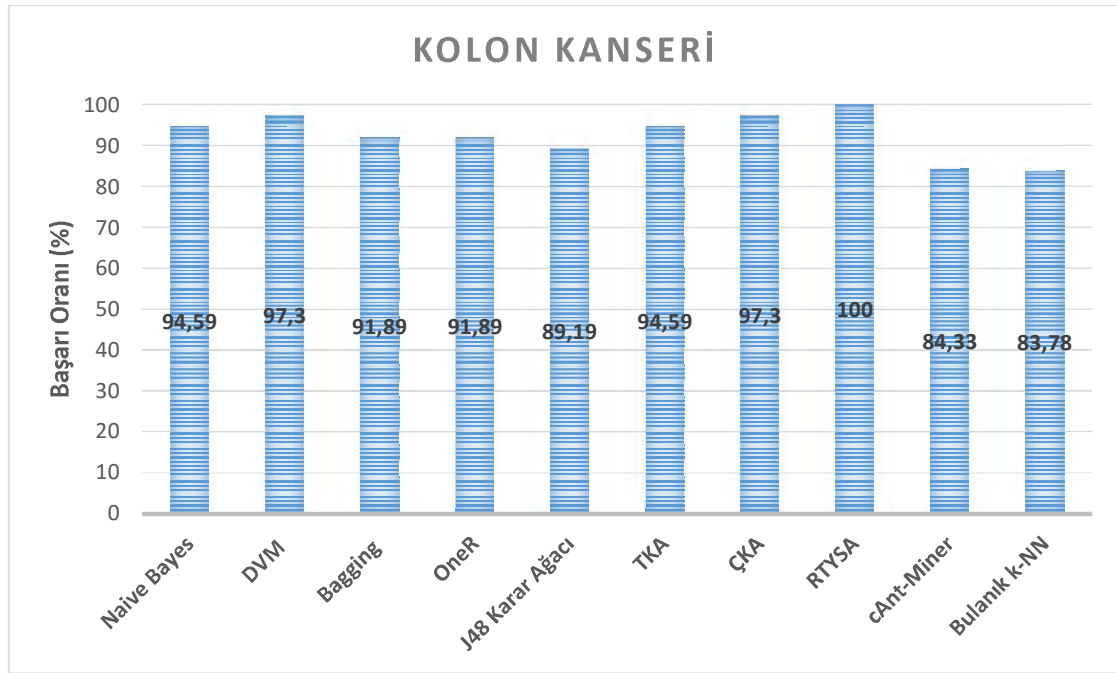


Şekil 4.9. Farklı doku kanserleri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

Farklı dokulardan alınan kanser veri setinde toplamda 174 hastadan alınan örnekler veya tümörler bulunmaktadır. Bu veri seti, 26 tanesi kanserli prostat dokusundan alınan tümörlerden (PR), 8 tanesi kanserli mesane dokusundan alınan tümörlerden (BL), 26 tanesi kanserli göğüs dokusundan alınan tümörlerden (BR), 23 tanesi kanserli kolon dokusundan alınan tümörlerden (CO), 12 tanesi kanserli mide dokusundan alınan

tümörlerden (GA), 11 tanesi kanserli böbrek dokusundan alınan tümörlerden (KI), 7 tanesi kanserli karaciğer dokusundan alınan tümörlerden (LI), 27 tanesi kanserli yumurta dokusundan alınan tümörlerden (OV), 6 tanesi kanserli pankreas dokusundan alınan tümörlerden (PA) ve 28 tanesi de kanserli akciğer dokusundan alınan tümörlerden (LU) oluşmaktadır. Ayrıca bu veri seti farklı dokulardan alınan tümörlerin oluşturduğu 10 farklı sınıfla temsil edilmektedir. Bu tümörlerden 12533 gen çıkarılmış ve sadece kansere neden olan 1571 aktif gen kullanılmıştır.

Şekil 4.9'da; 174 hastadan alınan ve birçok farklı kanser tümörü içeren veri setini sınıflandırmak için birbirinden farklı yöntemlere başvurulmuştur. Farklı sınıflandırma yöntemlerinin problemi çözmedeki başarısının ne kadar etkili olduğunu görebilmek için birçok yöntem kullanılmıştır. Yöntemlerin sınıflandırma problemlerindeki performanslarını test etmek için 10-katlı çapraz doğrulama tekniği kullanılmıştır. Bu teknikte, probleme uygulanan sınıflandırıcıların performansları analiz edilmiştir. Bu veri seti farklı kanser tümörlerini içermesi yönünden zengindir. Veri setindeki verileri en doğru sınıflara atayarak sınıflandırmada optimuma yakın bir başarı elde etmek için sınıflandırma yöntemleri optimal kontrol parametre değerleri ile test edilmiştir. Problem üzerinde test edilen bu yöntemler içinde en yüksek performans sergileyenler %95.98 doğruluk derecesiyle istatistiksel yöntemlerden Naive Bayes ve yapay zekâ yöntemlerinden ise ÇKA olmuştur. Bu iki yöntem, uygulanan problem üzerinde aynı performansı göstermiş ve en başarılı yöntem olmuşlardır. Problem üzerindeki en düşük performansı ise OneR yöntemi göstermiş ve veriyi %40.80 doğruluk derecesiyle sınıflara atayabilmiştir. Yapay zekâ tabanlı yöntemler içerisinde ise en düşük başarıya sahip sınıflandırıcının %78.62 ile cAnt-Miner olduğu görülmüştür. Özetlenecek olursa, yapay zekâyâ dayanan yöntemlerin istatistiksel yöntemlerden daha iyi sonuçlar verdiği gözlenmiştir.

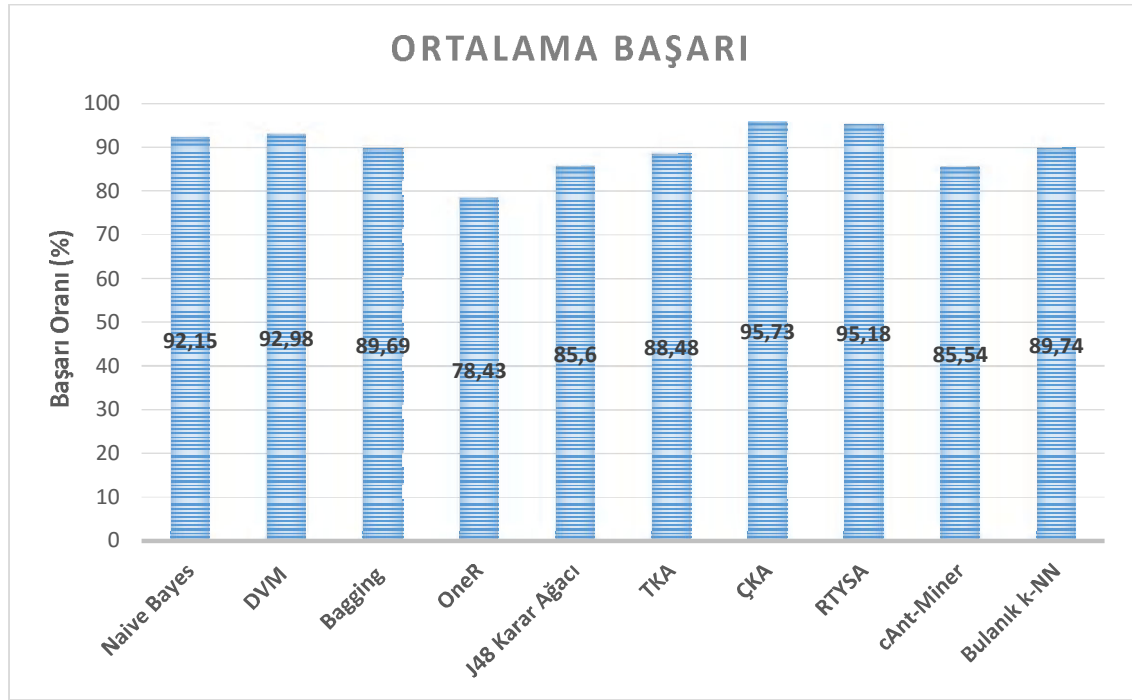


Şekil 4.10. Kolon kanseri veri seti üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin performansları

Kolon kanser veri seti, kanserli kolon dokusuna sahip 37 hastadan alınan örnekler içermektedir. Bu örneklerden 8 tanesi Serrated CRC ve 29 tanesi ise Conventional CRC tümörü olmak üzere kolon kanser veri seti, kolon kanserinin iki farklı alt türünü içermektedir. Bu tümörlerin her birinden toplamda 22883 gen çıkarılmıştır. Fakat bu genlerden 2202 tanesi kanserle ilgilidir bu yüzden diğer genler dikkate alınmamıştır. Bu veri seti, SCRC ve CCRC tümörleri olmak üzere 2 farklı sınıfla temsil edilmektedir.

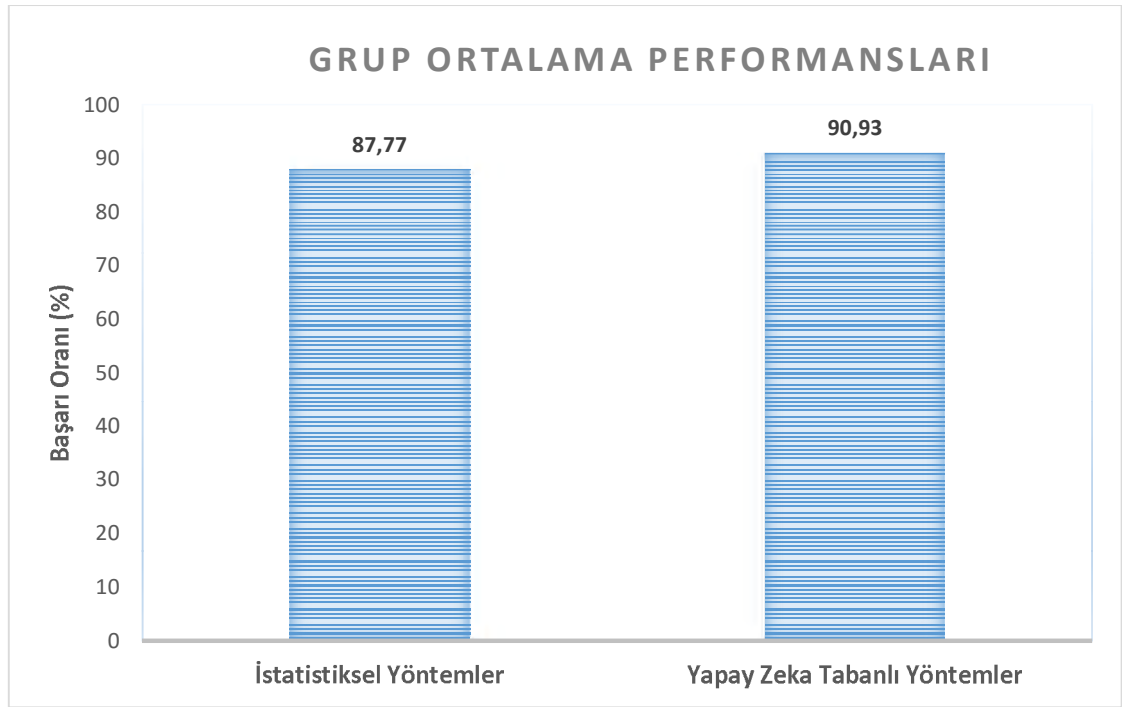
Kolon kanser verisi, 37 hastadan alınan 2 farklı kolon kanseri türünü içermektedir. Kolon kanserinin türlerini 10-katlı çapraz doğrulama yöntemi kullanarak sınıflandırma neticesinde RTYSA, iki sınıfa dâhil olan test verilerinin tamamını doğru şekilde sınıflandırmış ve veri kümesini sınıflandırmada en başarılı yöntem olmuştur. Yine yapay zekâ yöntemlerinden olan Bulanık k-NN ise veriyi sınıflandırmada %83.78 oranla diğer yöntemlere göre en düşük başarılı sınıflandırıcı olmuştur. İstatiksel yöntemleri kendi aralarında değerlendirdiğimiz zaman en başarılı sınıflandırıcı %97.30 başarımla DVM olurken en düşük başarılı yöntemin ise %89.19 performans ile J48 Karar Ağacı'nın olduğu gözlenmiştir. Yapay zekâ yöntemlerinden ise en başarılı yöntem, %100 başarı oranıyla RTYSA olurken en düşük başarımla ise %83.78 ile Bulanık k-NN'in olduğu görülmüştür (Şekil 4.10). Özetleyecek olursak; Kolon kanseri

verisinin sınıflandırılmasında istatistiksel yöntemlerin performanslarının yapay zekâ tabanlı yöntemlerin performanslarından daha iyi olduğu söylenebilir.



Şekil 4.11. Mikroarray kanser veri setleri üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin ortalama performansları

Optimal kontrol parametre değerlerine sahip sınıflandırma yöntemlerinin problemler üzerindeki ortalama performanslarının karşılaştırılması sonucu elde edilen grafik yukarıda görülmektedir (Şekil 4.11). Elde edilen bilgilere göre, doğru sınıflandırılmış en düşük gözlem sayısının %78.43 ile OneR'e ait olduğu görülürken doğru sınıflandırılmış en yüksek gözlem sayısının ise %95.73 ile ÇKA' ya daha sonra ise %95.18 ile RTYSA yöntemine ait olduğu gözlenmiştir. İstatiksel yöntemler içerisinde en düşük başarılı yöntem OneR olurken en başarılı yöntemin ise %92.98 ile DVM olduğu gözlenmiştir. Yapay zekâ tabanlı yöntemler içerisinde %85.54 ile cAnt-Miner'in başarısı diğer sınıflandırma yöntemlerinin başarısının gerisinde kalırken ÇKA ise %95.73 ile en başarılı yapay zekâ yöntemi olmuştur.



Şekil 4.12. Mikroarray kanser veri setleri üzerinde optimal kontrol parametrelerine sahip sınıflandırma yöntemlerinin grup ortalama performansları

Optimal parametrelili istatistiksel ve yapay zekâ tabanlı yöntemlerin problemler üzerindeki grup ortalama performanslarının karşılaştırılması sonucu elde edilen grafik yukarıda görülmektedir (Şekil 4.12). Yukarıdaki şekilden elde edilen bilgilere göre, yapay zekâ tabanlı yöntemlerin verileri sınıflandırmadaki grup ortalama performanslarının istatistiksel yöntemlerinkinden daha iyi olduğu görülmektedir.

Gelecekte elde edilecek sonuçlar açısından daha detaylı bir çalışma yaparak mikroarray gen ifade verilerinin sınıflandırılmasında farklı yapay zekâ tabanlı yöntemlerin gücü denenebilecektir. Ayrıca farklı yapay zekâ tabanlı veya hibrid yöntemler kullanılarak sonuçların başarımları daha da artırılabilir ve bu yeni yaklaşımlar mevcut yöntemlerin yerini alabileceklerdir.

KAYNAKÇA

1. Batbat, T., (2014). Protein Yapısının Yapay Arı Koloni Algoritması ile Tahmini. Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Kayseri, 88 s.
2. Korkem, E., (2013). Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naive Bayes Sınıflama Yöntemleri Yaklaşımı. Hacettepe Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Programı, Yüksek Lisans Tezi, Ankara, 65 s.
3. Fayyad, U., (1997). Data mining and knowledge discovery in databases: implications for scientific databases. (pp. 2-11). Ninth International Conference on Scientific and Statistical Database Management, August 11-13, 1997, Olympia, WA, IEEE, 10 pp.
4. Han, J., Cheng, H., Xin, D., Yan, X., (2007). Frequent pattern mining: current status and future directions. **Data Mining and Knowledge Discovery**, **15**(1), 55-86.
5. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., (1996). From data mining to knowledge discovery in databases. **AI magazine**, **17**(3), 37.
6. De Falco, I., Della Cioppa, A., Tarantino, E., (2002). Discovering interesting classification rules with genetic programming. **Applied Soft Computing**, **1**(4), 257-269.
7. Zhou, Z. H., (2003). Three perspectives of data mining, pp. 139-146. In: Artificial Intelligence (Eds: Z. H, Zhou). Elsevier Science Publishers.
8. Vahaplar, A., İnceoğlu, M. M., (2001). Veri madenciliği ve elektronik ticaret.Türkiye’de İnternet Konferansları, Harbiye İstanbul, 1-3.
9. Tan, K. C., Yu, Q., Heng, C. M., Lee, T. H., (2003). Evolutionary computing for knowledge discovery in medical diagnosis. **Artificial Intelligence in Medicine**, **27**(2), 129-154.
10. Distilleries, D., (1999). Introduction to Data Mining-Discover Hidden Value in Your Databases. Amsterdam, The Netherlands, 7-9
11. Kamrani, A., Rong, W., Gonzalez, R., (2001). A genetic algorithm methodology for data mining and intelligent knowledge acquisition. **Computers & Industrial Engineering**, **40**(4), 361-377.

12. Telcioğlu, M. B., (2007). Veri Madenciliğinde Genetik Programlama Temelli Yeni Bir Sınıflandırma Yaklaşımı ve Uygulaması. Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Kayseri, 145 s.
13. Özbakır, L., Baykasoğlu, A., Kulluk, S., (2008). Rule extraction from neural networks via ant colony algorithm for data mining applications. **In Learning and Intelligent Optimization** (pp. 177-191). Springer Berlin Heidelberg.
14. Han, J., Kamber, M., (2001). Data mining: concepts and techniques. (Web Page: <http://www.ict.griffith.edu.au/~vlad/teaching/kdd.d/6117cit.htm>), (Date accessed: December 2015).
15. Han, J., Kamber, M., (2000). Data mining: concepts and techniques (the Morgan Kaufmann Series in data management systems).
16. Kantardzic, M., (2003). Chapter 9: Artificial Neural Networks Chapter 1-1.4.DataMining Concepts, Models, Methods and Algorithms, (Eds: J. Wiley , Sons).
17. Goldberg, D. E. "Genetic algorithms in search, optimization, and machine learning, addison-wesley, reading, ma, 1989." K. Ohno, K. Esfarjani, and Y Kawazoe: Computational Materi.
18. Özkan, Y., (2008). "Veri Madenciliği Yöntemleri". Papatya Yayıncılık, İstanbul, 216 s.
19. Herbrich, R., Graepel, T., Obermayer, K., (1999). Support vector learning for ordinal regression, pp. 97-102. 9th International Conference on Artificial Neural Networks: ICANN '99, September 7-10 , 1999, IET Digital Library, 5 pp.
20. Işık, M., (2006). Bölünmeli Kümeleme Yöntemleri İle Veri Madenciliği Uygulamaları. Marmara Üniversitesi ,Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, İstanbul, 73 s.
21. Rastogi, R., Shim, K., (2000). PUBLIC: A decision tree classifier that integrates building and pruning. **Data Mining and Knowledge Discovery**, 4(4), 315-344.
22. Tan, P. N., Steinbach, M. Kumar ,V., (2006). Introduction to Data Mining. (Web page: <http://www-users.cs.umn.edu/~kumar/dmbook/index.php>), (Date accessed: December 2015).
23. Silahtaroglu, G., (2008). Kavram ve Algoritmalarıyla Temel Veri Madenciliği. Papatya Yayıncılık Egitim, İstanbul, 304 s.

24. Bock, H. H., (2002). Data mining tasks and methods: Classification: the goal of classification. **In Handbook of data mining and knowledge discovery**(pp. 254-258). Oxford University Press, Inc..
25. Han, J., Kamber, M., (2001). Data mining: concepts and techniques. (Web Page: <http://www.ict.griffith.edu.au/~vlad/teaching/kdd.d/6117cit.htm>), (Date accessed: December 2015).
26. Murthy, S. K., (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. **Data mining and knowledge discovery**, 2(4), 345-389.
27. Quinlan, R. J., (1993). 4.5: Programs for machine learning. Morgan Kaufmann Publishers Inc. San Francisco, USA, 302 pp.
28. Haykin, S., Lippmann, R., (1994). Neural Networks, A Comprehensive Foundation. **International Journal of Neural Systems**, 5(4), 363-364.
29. Horne, B. G., (1993). Progress in supervised neural networks. **Signal Processing Magazine, IEEE**, 10(1), 8-39.
30. Setiono, R., Leow, W. K., & Thong, J. Y., (2000). Opening the neural network blackbox: An algorithm for extracting rules from function approximating neural networks. pp. 176-186. In 21st International Conference on Information Systems (ICIS), October, 2000, Brisbane, Australia.
31. Michalewicz, Z., (2013). Genetic algorithms+ data structures= evolution programs. Springer Science & Business Media, 363 pp.
32. Wong, M. L., Leung, K. S., (2006). Data mining using grammar based genetic programming and applications (Vol. 3). Springer Science & Business Media, 196 pp.
33. Mitra, S., Acharya, T., (2005). Data mining: multimedia, soft computing, and bioinformatics. John Wiley & Sons, 399 pp.
34. Lewis-Beck, M., (1980). Applied regression: An introduction (Vol. 22). Sage publications, 77 pp.
35. Sydow, A., (1977). Tou, JT/Gonzalez, RC, Pattern Recognition Principles, London-Amsterdam-Dom Mills, Ontario-Sydney-Tokyo. Addison-Wesley Publishing Company. 1974. 378 S., \$19, 50. **ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik**, 57(6), 353-354.
36. Bonabeau, E., Théraulaz, G., (2000). Swarm smarts, pp. 72-79. Scientific American.

37. Beni, G., Wang, J., (1993). Swarm intelligence in cellular robotic systems, pp. 703-712. In: *Robots and Biological Systems: Towards a New Bionics?* (Eds: G. Beni., J. Wang) Springer Berlin Heidelberg.
38. Schank, R. C., (1982). *Dynamic Memory: a theory of learning in computer and people*. Cambridge University, Press, New York.
39. Kolodner, J. L., (1997). Educational implications of analogy: A view from case-based reasoning. **American psychologist**, **52**(1), 57.
40. Aha, D. W., Kibler, D., Albert, M. K., (1991). Instance-based learning algorithms. **Machine learning**, **6**(1), 37-66.
41. Pawlak, Zdzislaw., (1991). *Rough sets-theoretical aspect of reasoning about data*. 4. Institute of Computer Science. Springer Netherlands.
42. Binay, H. S., (2002). *Yatırım Kararlarında Kaba Küme Yaklaşımı*. Ankara Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, Ankara, 339 s.
43. Z. Pawlak, A. Skowron., (1999). *Rough set rudiments*. Bulletin of the International Rough Set. Society, Springer Berlin Heidelberg, 736 pp.
44. Zadeh, L. A., (1965). Fuzzy sets. *Information and control*, **8**(3), 338-353.
45. Tan, K. C., Yu, Q., Heng, C. M., Lee, T. H., (2003). Evolutionary computing for knowledge discovery in medical diagnosis. **Artificial Intelligence in Medicine**, **27**(2), 129-154.
46. Witten, I. H., Frank, E., (2005). *Data Mining: Practical Machine Learning Tools And Techniques*. Morgan Kaufmann, 525 pp.
47. Nel, G. M., (2004). *A Memetic Genetic Program For Knowledge Discovery*. University of Pretoria, Faculty of Engineering, Master of Science, Pretoria, 179 pp.
48. Brieman, L., Friedman, J. H., Olshen, R., Stone, C., (1984). *Classification and regression trees* (Belmont, CA: Wadsworth). UC San Diego, Chapman and Hall/CRC, 368 pp.
49. De Falco, I., Della Cioppa, A., Tarantino, E., (2002). Discovering interesting classification rules with genetic programming. **Applied Soft Computing**, **1**(4), 257-269.
50. Rouwhorst, S. E., Engelbrecht, A. P., (2000). *Searching The Forest: A Building Block Approach to Genetic Programming for Classification Problems in Data*

- Mining, Vrije Universiteit Amsterdam, Department of Mathematics and Informatics, Master of Thesis, Amsterdam, 156 pp.
51. Parpinelli, R. S., Lopes, H. S., Freitas, A. A., (2002). An ant colony algorithm for classification rule discovery, pp. 191-208.
 52. Tan, K. C., Yu, Q., Ang, J. H., (2006). A coevolutionary algorithm for rules discovery in data mining. **International Journal of Systems Science**, **37**(12), 835-864.
 53. Tan, K. C., Yu, Q., Ang, J. H., (2006). A dual-objective evolutionary algorithm for rules extraction in data mining. **Computational optimization and applications**, **34**(2), 273-294.
 54. Uran, G., (2005). A Hybrid Heuristic Model for Classification Rule Discovery. Pace University New York, NY, USA, 110 pp.
 55. Larose, D. T. (2005). An introduction to data mining. Traduction et adaptation de Thierry Vallaud.
 56. Kaur, H., Wasan, S. K., Al-Hegami, A. S., Bhatnagar, V., (2006). A unified approach for discovery of interesting association rules in medical databases. In *Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining* (pp. 53-63). Springer Berlin Heidelberg.
 57. Yao, Y., Zhou, B., (2008). Micro and macro evaluation of classification rules. In *Cognitive Informatics*, pp. 441-448. 7th IEEE International Conference on IEEE, August ,2008.
 58. Bojarczuk, C. C., Lopes, H. S., Freitas, A. A., Michalkiewicz, E. L., (2004). A Constrained-Syntax Genetic Programming System For Discovering Classification Rules: Application To Medical Data Sets. **Artificial Intelligence in Medicine**, **30**(1), 27-48.
 59. De Jong, K. A., Spears, W. M., Gordon, D. F., (1994). Using genetic algorithms for concept learning (pp. 5-32). In: *Genetic Algorithms for Machine Learning* (Eds: K. A., De Jong, W. M., Spears, D. F., Gordon). Springer US.
 60. Romão, W., Freitas, A. A., Gimenes, I. M. D. S., (2004). Discovering interesting knowledge from a science and technology database with a genetic algorithm. **Applied soft computing**, **4**(2), 121-137.
 61. Özcan, G. S., (2014). Bütünleştirici Modül Ağlarıyla Gen Düzenleme Analizi. Başkent Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara, 62 s.

62. Lüleyp, H. Ü., (2008). Moleküler Genetiğin Esasları. Nobel Kitabevi, 437 s.
63. Bal, S. H., Budak, F., (2012). Mikroarray teknolojisi. **Uludağ Üniversitesi Tıp Fakültesi Dergisi** **38**(3), 227-233.
64. Şimşek, D. Ö., (2013). Mikroarray teknolojisi ve diş hekimliği'nde kullanımı. **Atatürk Üniversitesi Diş Hekimliği Fakültesi Dergisi**, **2013**(7).
65. Shakya, K., Ruskin, H. J., Kerr, G., Crane, M., Becker, J., (2010). Comparison of microarray preprocessing methods. **In Advances in Computational Biology** (pp. 139-147). Springer New York.
66. Liu, H., Bebu, I., Li, X., (2010). Microarray probes and probe sets. **Frontiers in bioscience (Elite edition)**, **2**, 325.
67. İpekdağ, K., 2011. Mikroarray teknolojisi. (Web sayfası: http://yunus.hacettepe.edu.tr/~mergen/sunu/s_mikroarrayandecology.pdf) (Erişim Tarihi: Nisan 2015)
68. Bier, F. F., von Nickisch-Rosenegk, M., Ehrentreich-Förster, E., Reiß, E., Henkel, J., Strehlow, R., Andresen, D., (2008). DNA microarrays, pp. 433-453. In: Biosensing for the 21st Century (Eds: F. F., Bier, M., von Nickisch-Rosenegk, E., Ehrentreich-Förster, E., Reiß, J., Henkel, R., Strehlow, D., Andresen). Springer Berlin Heidelberg.
69. Witten, I. H., Frank, E., (2005). Data Mining: Practical Machine Learning Tools And Techniques. Morgan Kaufmann, 560 pp.
70. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., (2009). The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, **11**(1), 10-18.
71. Patterson, D., Liu, F., Turner, D., Concepcion, A., Lynch, R., (2008). Performance comparison of the data reduction system. pp. 1-8. In Proceedings of the SPIE Symposium on Defense and Security, March, 2008, Orlando, FL.
72. Saeys, Y., Inza, I., Larrañaga, P., (2007). A review of feature selection techniques in bioinformatics. **bioinformatics**, **23**(19), 2507-2517.
73. Guyon, I., Weston, J., Barnhill, S., Vapnik, V., (2002). Gene selection for cancer classification using support vector machines. **Machine learning**, **46**(1-3), 389-422.
74. Hall, M. A., (1999). Correlation-based feature selection for machine learning. (Doctoral dissertation, The University of Waikato).

75. Hall, M. A., & Smith, L. A. (1998). Practical feature subset selection for machine learning.
76. Kohavi, R., (1996). Scaling up the accuracy of Naive-Bayes classifiers: A Decision-Tree hybrid. pp. 202-207. In KDD, August, 1996, Shoreline Blvd.
77. Rish, I., (2001). An empirical study of the Naive Bayes classifier, pp. 41-46. In IJCAI 2001 workshop on empirical methods in artificial intelligence, August, 2001, New York, IBM, 6 pp.
78. Zhang, H., (2004). The optimality of naive Bayes. **AA**, **1**(2), 3.
79. Aydoğan, E., (2008). Veri Madenciliğinde Sınıflandırma Problemleri İçin Evrimsel Algoritma Tabanlı Yeni Bir Yaklaşım: Rough-Mep Algoritması. Gazi Üniversitesi, Doktora tezi, Ankara, 137 s.
80. John, G. H., Langley, P., (1995). Estimating continuous distributions in Bayesian classifiers, pp. 338-345. In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. August, 1995. Morgan Kaufmann Publishers Inc , (7): 338-345.
81. Yıldız, O., Bilge, H. Ş., Akcayol, M. A., Güler, İ., (2012). Meme kanseri sınıflandırması için veri füzyonu ve genetik algoritma tabanlı gen seçimi. **Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi**, **27**(3).
82. Breiman, L., (1996). Bagging predictors. **Machine learning**, **24**(2), 123-140.
83. Novakovic, J., Minic, M., Veljovic, A., (2010). Genetic search for feature selection in rule induction algorithms, (pp. 1109-1112). In 18th Telecommunications forum TELFOR, November 23-25, 2010, Serbia, Belgrade, 4 pp.
84. Kääriäinen, M., Malinen, T., Elomaa, T., (2004). Selective Rademacher penalization and reduced error pruning of decision trees. **The Journal of Machine Learning Research**, **5**, 1107-1126.
85. Eren, Ö., (2008). Alerjen Proteinlerin Otomatik Sınıflandırılması. Başkent Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara, 92 s.
86. Akbilgiç, O., Danışman, G., Bozdoğan, H., (2011). Hibrit Radyal Tabanlı Fonksiyon Ağları İle Değişken Seçimi Ve Tahminleme: Menkul Kıymet Yatırım Kararlarına İlişkin Bir Uygulama. İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, Doktora Tezi, İstanbul, 165 s.
87. Elmas, Ç., (2007). Yapay Zeka Uygulamaları. Seçkin Yayıncılık, Ankara, 424 s.

88. Okkan, U., Dalkılıç, H. Y., (2012). Radyal tabanlı yapay sinir ağları ile kemer barajı aylık akımlarının modellenmesi. **DMO Teknik Dergi**, 5957-5966.
89. Dorigo, M., Maniezzo, V., Coloni, A., Maniezzo, V., (1991). Positive feedback as a search strategy. pp. 1-22. Politecnico Di Milano, June, 1991, Milano, 22 pp.
90. Parpinelli, R. S., Lopes, H. S., Freitas, A. A., (2002). An ant colony algorithm for classification rule discovery. pp. 191-208. In: Data mining: A heuristic approach(Eds: Parpinelli, R. S., Lopes, H. S., Freitas, A. A). University of New South Wales, Australia, Idea Group Publishing.
91. Liu, B., Abbass, H. A., McKay, B., (2003). Classification rule discovery with ant colony optimization. Technology, October, 2003, Australia. IEEE. , 83 pp.
92. Korkem,E.,2013. Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naive Bayes Sınıflama Yöntemleri Yaklaşımı. Hacettepe Üniversitesi, Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, Ankara, 65 s.
93. Mergen, H., Mikroarray teknolojisi. (Web sayfası: http://yunus.hacettepe.edu.tr/~mergen/derleme/d_microarray.pdf), (Erişim tarihi: Eylül 2015).
94. İpekdağ, K., Mikroarray teknolojisi. Ekolojik ve evrimsel çalışmalarda kullanımı.(Web sayfası: http://yunus.hacettepe.edu.tr/~mergen/sunu/s_mikroarrayandecology.pdf), (Erişim tarihi: Ağustos 2015).
95. Olshen, L. B. J. F. R., Stone, C. J. (1984). Classification and regression trees. **Wadsworth International Group**, 93(99), 101.
96. Erpolat, S. "Kanser verilerinin sınıflandırılmasında yapay sinir ağları ile destek vektör makinelerinin karşılaştırılması. 71-83. Aydın Üniversitesi, 13 s."
97. Famili, F., Shen, W. M., Weber, R., Simoudis, E., (1997). Data pre-processing and intelligent data analysis. **International Journal on Intelligent Data Analysis**, 1(1).
98. Coşkun, C., Baykal, A., (2011). Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması, pp. 1-8. Akademik Bilişim,2011, Malatya, 8 s.
99. Makinacı, M., Güneşer, C., (2007). Göğüs kanseri verilerinin sınıflandırılması. pp. 1-2. Elektrik-Elektronik-Bilgisayar Mühendisliği 12. Ulusal Kongresi,Kasım 14, 2007, 2 s.

100. Daş, B., Türkoğlu, İ., (2014). DNA dizilimlerindeki nükleotit çiftlerinin frekans değerlerine göre farklı sınıflandırma yöntemleri ile karşılaştırılması. 191-194. Tıp Teknolojileri Ulusal Kongresi (TIPTEKNO'2014), 25-27 Eylül, 2014, Kapadokya, 4 s.
101. Daş, B., Türkoğlu, İ., (2014). DNA dizilimlerinin sınıflandırılmasında karar ağacı algoritmalarının karşılaştırılması. 381-383. Eleco 2014 Elektrik-Elektronik-Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu, 27-29 Kasım, 2014, Bursa, Eleco, 3 s.
102. Pirooznia, M., Yang, J., Yang, M. Q., Deng, Y. (2008). A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics*, **9** (Suppl 1), S13.
103. Hu, H., Li, J., Plank, A., Wang, H., Daggard, G., (2006). A comparative study of classification methods for microarray data analysis, (pp. 33-37). In Proceedings of the fifth Australasian conference on Data mining and analytics-Volume 61, November, 2006, Australian Computer Society, Inc.. 5 pp.
104. Alataş, B., Akın, E., (2004). Sınıflandırma kurallarının karınca koloni algoritmasıyla keşfi. Eleco'2004 Elektrik-Elektronik-Bilgisayar Mühendisliği Sempozyumu, 8-12 Aralık, 2004, Eleco, 5 s.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı, Soyadı : Mustafa Turan ARSLAN
 Uyuğu : Türkiye (TC)
 Doğum Tarihi ve Yeri: 18 Ekim 1986, Kilis
 Medeni Durumu : Evli
 Tel : +90 553 349 95 24
 E-mail : mustafaturanarslan@gmail.com
 Yazışma Adresi : Mustafa Kemal Üniversitesi Kırıkkhan Meslek Yüksekokulu
 Kırıkkhan/HATAY

EĞİTİM

Derece	Kurum	Mezuniyet Tarihi
Yüksek Lisans	ERÜ Fen Bilimleri Enstitüsü	-
Lisans	ERÜ MF Bilgisayar Mühendisliği	2011
Lise	Kilis(Süper) Lisesi	2007

İŞ DENEYİMLERİ

Yıl	Kurum	Görev
2016-Halen	Mustafa Kemal Üniversitesi Kırıkkhan Meslek Yüksekokulu	Öğretim Görevlisi
2013 - 2016	Orman Genel Müdürlüğü	Bilgisayar Mühendisi
2012 - 2013	Mustafa Kemal Üniversitesi Mühendislik Fakültesi	Araştırma Görevlisi

YABANCI DİL

İngilizce