KANSERLİ HÜCRELERİN MİKROARRAY GEN İFADELERİNİN İNCELENMESİ VE VERİ MADENCİLİĞİ YÖNTEMLERİ KULLANARAK SINIFLANDIRILMASI

Murat AKIN

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TEMMUZ 2012
ANKARA

Murat AKIN tarafından hazırlanan "KANSERLİ HÜCRELERİN MİKROARRAY GEN İFADELERİNİN İNCELENMESİ VE VERİ MADENCİLİĞİ YÖNTEMLERİ KULLANARAK SINIFLANDIRILMASI" adlı bu tezin Yüksek Lisans tezi olarak uygunluğunu onaylarım.

Yrd.Doç.Dr. Hasan Şakir BİLGE
Tez Danışmanı, Bilgisayar Müh. Anabilim Dalı

Bu çalışma, jürimiz tarafından oy birliği ile Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans tezi olarak kabul edilmiştir.

Prof.Dr. M.Cengiz TAPLAMACIOĞLU …….
Elektrik-Elektronik Mühendisliği Anabilim Dalı

Yrd.Doç.Dr. Hasan Şakir BİLGE …….
Bilgisayar Mühendisliği Anabilim Dalı

Prof.Dr. M.Ali AKCAYOL …….
Bilgisayar Mühendisliği Anabilim Dalı

Tarih: 04/07/2012

Bu tez ile G.Ü. Fen Bilimleri Enstitüsü Yönetim Kurulu Yüksek Lisans derecesini onamıştır.

Prof. Dr. Bilal TOKLU …….
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

Murat AKIN

KANSERLİ HÜCRELERİN
MİKROARRAY GEN İFADELERİNİN İNCELENMESİ
VE VERİ MADENCİLİĞİ YÖNTEMLERİ KULLANARAK
SINIFLANDIRILMASI
(Yüksek Lisans Tezi)

Murat AKIN

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
Temmuz 2012

ÖZET

Biyoinformatik, bilgisayar bilimleri, matematik ve istatistiğin entegrasyonu sonucundan ortaya çıkan, biyolojik sistem ve olaylar hakkındaki bilgilerin derlenmesi ve değerlendirilmesini amaçlayan disiplinler arası bir bilimdir. Biyoinformatik bilim dalının hedefi biyolojik olaylardan doğan bilgileri elde etmek ve bu bilgilerin saklanması için veritabanları oluşturmak ve bilişim teknolojilerinin yardımıyla biyolojik olayların moleküler düzeyde açıklanmasına olanak sağlamaktır.

Veri madenciliği, çok büyük veri tabanlarındaki ya da veri ambarlarındaki veriler arasında bulunan ilişkiler, örüntüler, değişiklikler, sapma ve eğilimler, belirli yapılar gibi ilginç bilgilerin ortaya çıkarılması ve keşfi işlemidir. Daha basit bir tanım yapmak gerekirse, büyük ölçekli veriler arasından "değerli bilgiyi" elde etmektir. Bu sayede, veriler arasındaki ilişkileri ortaya koymak ve gerektiğinde ileriye yönelik tahminlerde de bulunmak mümkündür. Veri madenciliğinde sınıflandırma yöntemleri ise kurulan modele göre verilerin daha önceden nitelikleri bilinen gruplara dağılımına olanak tanımaktadır.

Bu çalışmada, biyoinformatik alanında mikroarray verilerine sınıflandırma yöntemleri uygulanmıştır. Kanserli olma ihtimali olan hücrelerden alınan verilere uygulanan algoritmalar sonucunda, örneklerin doğru bir şekilde sınıflandırılması amaçlanmıştır. Burada farklı kanser çeşitleri için analizler yapılmıştır. Sınıflandırma yöntemleri olarak Destek Vektör Makineleri, Naive Bayes, K-En Yakın Komşuluk ve Doğrusal Ayırma Analizi sınıflandırıcı kullanılmıştır.

Bilim Kodu : **902.1.019**
Anahtar kelimeler :**Biyoinformatik, veri madenciliği, öznitelik seçimi, sınıflandırma**
Sayfa Adedi : **60**
Tez Yöneticisi : **Yrd.Doç.Dr. Hasan Şakir BİLGE**

INVESTIGATION OF MICROARRAY GEN EXPRESSION OF CANCERED CELLS AND CLASSIFICATION BY USING DATA MINING METHODS
(M. Sc. Thesis)

Murat AKIN

GAZİ UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY
July 2012

ABSTRACT

Bioinformatics is an inter-disciplinary science that resulting from intersection of computer science, mathematic and statistics, which aims and evaluates collecting of informations about biological system and events. The aim of bioinformatics is to obtain information supplied from biological events and to create databases for storing this information and to provide description of biological events at molecular level by means of informatics technologies.

Data mining is a find out and a discover processing of interesting information such as patterns, changings, deviations and tendecies, specified structures, relationships presented with between data in huge databases or data warehouses. While it is neccesary to make a simpler definition, "valuable information" is to obtain among the data of large-scale. By this way, it is possible to find out relationships among the data and if necessary, to predict forward projection. Classification methods in data mining allows the distribution of data into groups with previously known attributes according to the established model.

In this study, classification methods has been applied to microarray data in the bioinformatics field. As a result of applied algorithms, data taken from being

potentially cancered cells, has been aimed to classify samples correctly. Here, it has been carried out analyses for various cancer types. Support Vector Machines, Naive Bayes, K-Nearest Neighbour and Linear Discriminant Analysis classifiers had been used as classification methods.

Science Code : **902.1.019**
Key Words : **Bioinformatics, data mining, feature selection, classification.**
Page Number : **60**
Adviser : **Assist.Prof.Dr. Hasan Şakir BİLGE**

TEŞEKKÜR

Yapmış olduğum tüm çalışmalar esnasında ilgi ve desteğini hiçbir zaman esirgemeyip tecrübeleri ile bana yol gösteren hocam Yrd.Doç.Dr. Hasan Şakir BİLGE'ye, akademik hayatım ve çalışmalarımda bana yön veren, yaşamımda önemli bir yeri olan, kendime rehber edindiğim saygıdeğer büyüğüm Prof.Dr.Metin GÜRÜ'ye, ayrıca maddi ve manevi her zaman beni destekleyen, yanımda olan sevgili dostlarıma ve aileme teşekkürü bir borç bilirim.

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELERİN LİSTESİ

Çizelge **Sayfa**

Çizelge **Sayfa**

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Kısaltmalar	Açıklama
ALL	Acute lymphoblastic leukemia
AML	Acute myelogenous leukemia
AUC	Area under curve (Eğri altında kalan alan)
BLAST	Basic Local Alignment Search Tool (Temel yerel eşleme arama aracı)
BUPA	British United Provident Association (Birleşik Britanya tasarruf derneği)
DDBJ	DNA Databank of Japan (Japonya DNA veri bankası)
DVM	Destek vektör makinesi
DAA	Doğrusal ayırma analizi (Linear Discriminant Analysis)
EMBL	The European Molecular Biology Laboratory (Avrupa moleküler biyoloji laboratuarı)
FN	False negative (Yanlış negatif)
FP	False pozitive (Yanlış pozitif)
K-ENK	K en yakın komşuluk
KA	Karar ağaçları
miRNA	mikroRNA
NB	Naive bayes
NCBI	National Center for Biotechnology Information (Biyoteknoloji bilgisi için ulusal merkez)
NLM	National Library of Medicine (Ulusal tıp kütüphanesi)
PIR	Protein Identification Resource (Protein kaynak tanımlama)

PIMA	Pima Indians (Pima yerlileri)
RF	Random forest (Rastgele orman)
ROC	Receiver operating characteristic (Alıcı işlem karakteristikleri)
TN	True negative (Doğru negatif)
TP	True pozitive (Doğru pozitif)
WDBC	Wisconsin diagnostic breast cancer (Wisconsin tanısal göğüs kanseri)
YSA	Yapay sinir ağları

1.GİRİŞ

Biyoinformatik; biyolojik olaylar hakkında bilgi toplanması ve toplanan bilginin değerlendirilmesi için, biyoloji alanının yanında kimya ve tıp biliminin bilişim bilimleri, matematik ve istatistik bilim dallarıyla entegrasyonu sonucu ortaya çıkmıştır. Bilişim teknolojilerinin biyolojik problemlerin çözümünde kullanılması esasına dayanan biyoinformatik, biyolojik olayların moleküler düzeyde açıklanmasına yardımcı olmaktadır. Biyoinformatikte, biyolojik bilgiler sayısal olarak elde edilir ve veritabanlarında saklanır. Biyoinformatik medikal bilimlerde çok önemli bir rol oynamaktadır. Medikal bilimlerde son yıllardaki uygulamalar gen ifade analizleri üzerine yoğunlaşmıştır. Genellikle farklı hastalıklardan etkilenen hücrelerin ifadeleri derlenerek, sağlıklı hücrelerle kıyaslanmakta ve aradaki farklılıklardan hastalık teşhisi yapılmaktadır [1].

Temel hedefi biyolojik süreci anlamamızı sağlamak olan biyoinformatik, farklı bilim alanlarının içinde yer aldığı bir bilim dalıdır. Mikroarray gen ifade, genom dizileme ve görüntü analizi çalışmalarında büyük ölçekte ve farklı tipte veriler elde edilmektedir. Bu verilerin düzenlenmesi, depolanması, araştırmacıların kolayca erişebilmesinin sağlanması, verileri incelemek için analiz araçlarının geliştirilmesi ve istatistiksel çözümlemelerin yapılması farklı bilim dallarından araştırmacıların bir arada çalışmasını zorunlu kılmaktadır. Bu konunun giderek önem kazanması sonucunda, biyoistatistikçiler/istatistikçiler kendilerini biyoloji alanında geliştirerek biyoinformatikte yapılan çalışmalara daha fazla katılmakta ve büyük katkı sağlamaktadır [2].

Biyoinformatik alanında yapılan çalışmaların ana amacı, bir canlıya ait organizmanın genom bilgisi ile fonksiyonlarının anlaşılması olarak ifade edilebilir. Mikroarray analizi de bu alanda mevcut verilerin genlerini incelemeye olanak tanımıştır. İnceleme sonuçlarında genler hakkında bilgi toplandığı gibi insanoğlunun savaş verdiği kanser gibi hastalıklar için hasta- sağlam birey ayırımında olumlu sonuçlar da alınabilmiştir.

Veri madenciliği ise biyoinformatik bilim dalı için sınıflandırma yöntemleri ile katkı sağlamıştır. Veri madenciliğinde bilgiye ulaşmak için farklı yöntemler kullanılmaktadır. Bu yöntemler ait birçok algoritma vardır. Bu algoritmaların hangisinin daha iyi neticeler verdiği üzerine pek çok çalışma yapılmış olup çalışmalardan farklı sonuçlar alınmıştır. Farklı sonuçlar elde edilmesinin başlıca sebepleri veri üzerindeki önişlem, öznitelik seçimlerinin farklı metotlarla yapılması, kullanılan algoritmaların parametrelerinin seçimi ve sınıflandırma yöntemlerinin uygulandığı programların farklılık göstermesi şeklinde sıralanabilir.

Bu çalışmada hedeflenen biyoinformatik alanında kanserli hücrelerden elde edilen verileri doğru sınıflara ayırmaktır. Sınıflandırmanın doğru yapılması ile hastalık teşhislerinde önemli bulgular elde edilebilir. Metodoloji olarak sınıflandırma algoritmalarından Destek Vektör Makineleri, K-En Yakın Komşuluk, Naive Bayes, Doğrusal Ayırma Analizi sınıflandırma kullanılmıştır. Sınıflandırma yöntemlerinin uygulandığı veriler mikroarray gen ifadeleridir. Bu gen ifadeleri, kolon, lenf ve göğüs kanseri, lösemi, BUPA karaciğer ve PIMA diyabet hastalarının verileridir. Önceki çalışmalar incelenerek, deneysel çalışmalar sonucunda elde edilen doğruluk, duyarlılık ve özgüllük değerlerinden daha iyi neticeler alınması amaçlanmıştır.

Yapılan çalışmada ilk olarak literatür ve genel bilgi başlığı altında biyoinformatik, mikroarray teknolojisi, veri madenciliği ile daha önce biyoinformatik alanında yapılan çalışmalar hakkında bilgiler içermektedir. İlk bölümde biyoinformatiğe giriş yapılmakta olup, biyoinformatiğin önemi, amacı, sık kullanılan veri tabanları anlatılmaktadır. Ayrıca veri madenciliği süreci de bu bölümde incelenmiştir. İkinci bölümde ise biyoinformatik alanında kullanılan sınıflandırma algoritmalarının metodolojisi anlatılmaktadır. Sonraki bölüm olan deneysel çalışmalar kısmında, sınıflandırma algoritmaları kullanarak elde edilen sonuçlar paylaşılmakta, son bölümde ise bu sonuçlar hakkında değerlendirme yapılmaktadır.

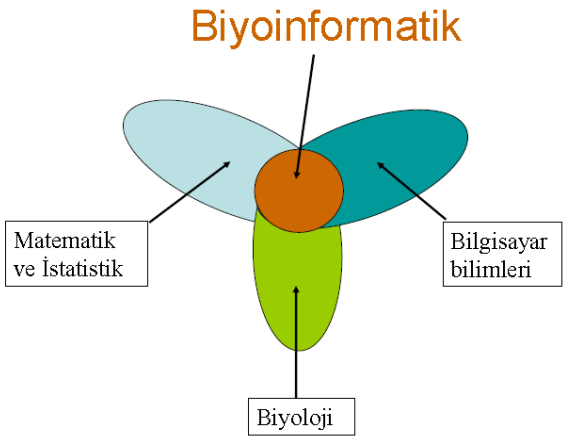
2. GENEL BİLGİ VE LİTERATÜR ÇALIŞMASI

2.1. Biyoinformatiğe Giriş

21. yy biyoloji sadece laboratuarda yapılan bir bilim olmaktan çıkmış, aynı zamanda bilgi teknolojisine de dayanan bir bilim dalı haline gelmiştir. Biyolojinin her alanında bilgisayarların ve bilgisayar yöntemlerinin kullanılmaya başlanması günümüzde vazgeçilmez olmuş, biyoinformatik olarak adlandırılan yeni bir bilim dalı ortaya çıkmıştır [3].

Biyolojik bilimlerin hızlı gelişimi ve bu gelişim sonucunda yapılan uygulamalar ve araştırmalar ile büyük miktarda veri birikimi oluşmuştur. Tüm dünyada, farklı ülkelerde pek çok laboratuarda yapılan çalışmalar sonucu çok büyük sayıda genomik bilgi elde edilmiştir. Bilimin bu hızlı ilerleyişi insanlar üzerine önemli bir bilgi yükü getirmiştir. Elde edilen bilgilerin bir araya getirilip kullanılır hale getirilmesi oldukça önemli ve zor bir iştir. Bilgilerin hızlı, güvenilir ve kolay erişebilir bir ortamda saklanması ihtiyacı bu nedenle ortaya çıkmıştır. Biyoinformatik biyolojik bilgilerin oluşturulması, saklanması için veritabanlarının oluşturulmasıdır.

Biyoinformatik, yaşam ve bilgisayar bilimleri arasında bir köprü kurmakla birlikte, verilerin son derecede etkin bir şekilde toplanması ve adımları hızlandırması nedeniyle son derecede önemlidir [4]. Şekil 2.1'de biyoinformatiğin diğer bilimler ile ilişkisini görülmektedir.



Şekil 2.1. Disiplinler arası bilim olarak biyoinformatik [5]

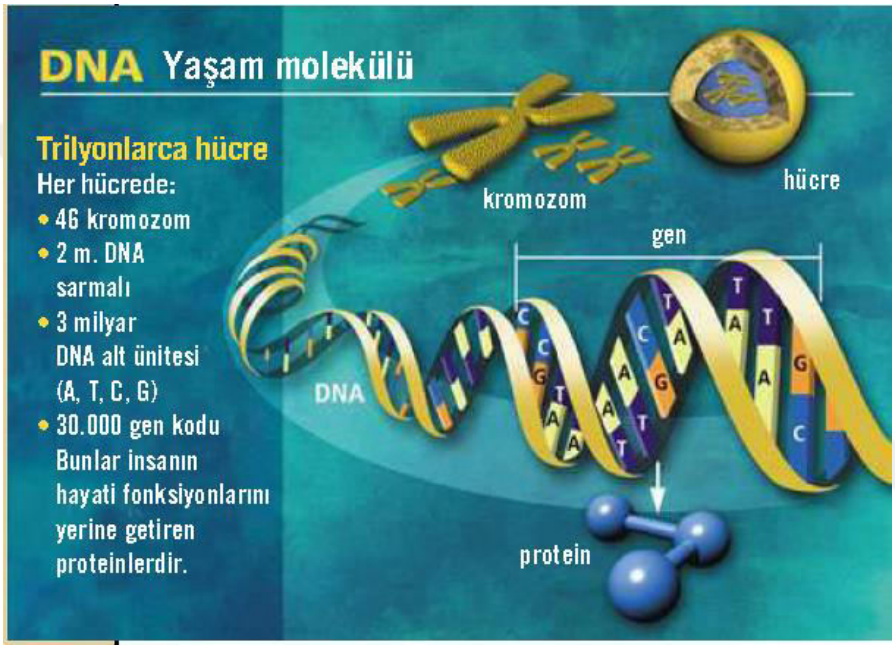
Biyoinformatiğin doğuşunda 2 ana sebep vardır.

- 20. yy ikinci yarısından itibaren biyolojik bilginin büyük bir artış göstermesi ve bu bilgiyi kendi içinde organize edebilmek için güçlü araçlara ihtiyaç duymasıdır.

- Biyolojik sistemler hakkında merak edilen sorular oldukça karmaşık olabilmekte ve bu sorulara aranan cevaplar insan beyninin kapasitesi ile araştırıldığı takdirde bulunamayacağı ortadadır [6]. Böyle karmaşık bilgilerin bütünleştirilmesi için kavram ve modeller geliştirmek ve anlaşılabilirliği için görselleştirilmesini sağlamak biyoinformatiğin en önemli uğraş alanıdır.

2001 yılının Şubat ayında Science ve Nature dergilerinde yayınlanan iki farklı makalede insan gen haritasının ilk taslağı kamuoyuna açıklanmıştır (Human Genome Project). Yayınlanan sonuçlar, modern biyoloji için yeni bir başlangıcın doğuşuna işaret etmektedir. Bundan sonra biyoloji ve tıpla ilgili araştırmalar protein ve DNA dizinlerine bağlı olarak yürütülecektir. Bu çalışmalar aslında bir asırlık birikimin sonucudur. 20. yy başında Mendel kanunlarının tekrar keşfedilmesi ile bu araştırmalar başlamış, yüzyılın ilk çeyreğinde biyologlar hücre içerisindeki bilgilerin nesilden nesile kromozomlar tarafından taşındığını bulmuşlardır. İkinci çeyrekte ise

DNA'nın üç boyutlu yapısı Watson ve Crick tarafından çözülmüştür. Şekil 2.2'de DNA yapısının sembolik yapısını da görülmektedir. Bu da sonraki çalışmalarda genetik özelliklerin taşınma mekanizmasının anlaşılmasını sağlamıştır. Bu gelişmelere paralel olarak laboratuar teknikleri ve deneysel yöntemlerde gelişmiştir. Bilim adamları bu teknikleri kullanarak laboratuar ortamında genetik deneyler yapmışlardır [3].

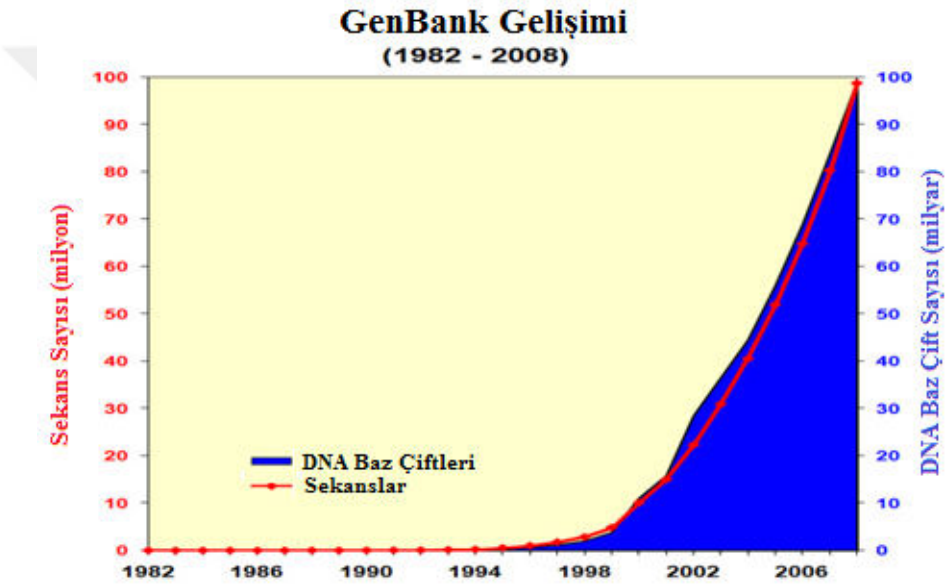


Şekil 2.2. DNA'nın yapısı [3]

2.2. Biyoinformatik Kaynak Bilgisi

Biyoinformatik uygulamalı bir bilimdir. Bilgisayar programları kullanımı ile genetik bilgi ve modern moleküler biyoloji arşivlerinden yararlanılarak geliştirilmiş sonuçlar çıkartılmakta ve bu sayede önemli tahminler yapılmaktadır. National Center for Biotechnology Information (NCBI), biyoinformatiği; "yaşam bilimleri(Biyoloji, Tıp, Biyokimya) bilişim teori ve teknolojileri ile matematik ve istatistiğe dayalı interdisipliner bir bilim dalıdır" şeklinde tanımlamıştır. Biyoinformatiğin ortaya çıkmasıyla yapılan çalışmalarından elde edilen veriler için bir veri bankası oluşturulmuş ve tüm dünyada biyoinformatik üzerine çalışan bilim adamlarının elde ettikleri verileri bu veri bankasına aktarabilmelerine olanak sağlamıştır [7]. Veri

bankasının oluşturulmasından sonra verilerde olağanüstü bir artış görülmüştür. Mevcut veriler 1986'da 3939 iken 1999 da 80 000 'e çıkmış ve 2004 yılında kayıtlara göre 160 000 veriye ulaşmıştır. Her geçen gün artış göstermekte olup günümüzde yaklaşık 300 000 civarında olduğu tahmin edilmektedir. Protein dizlim çalışmaları sonucu elde edilen verilen veri bankalarına kaydı ile ortaya çıkan veri patlamasıyla biyoinformatiğin gelişime önemli katkı yapmıştır. Şekil 2.3 ve Çizelge 2.1'de gen bankasında yıllara göre artış miktarları görülmektedir.



Şekil 2.3. Gen bankasında bilgi artışı [8]

Çizelge 2.1. Örnek yıllara göre gen bankasında bulunan veri miktarı [9]

GenBank Verileri		
Yıllar	Baz Çiftleri	Sekanslar
1982	680 338	606
1984	3 368 765	4 175
1990	49 179 285	39 533
1995	384 939 485	555 694
2000	11 101 066 288	10 106 023
2004	44 575 745 176	40 604 319
2007	83 874 179 730	80 388 382
2008	99 116 431 942	98 868 465

Biyoinformatik bu verileri hızlı bir şekilde değerlendirecek yeni algoritmalara dayalı yazılımların oluşturulmasını sağlamaktadır. Biyoinformatik elde edilmiş olan veriler üzerinden işlem yaptığından laboratuar çalışmalarına kıyasla önemli bir maliyete neden olmamaktadır. Yapılan çalışmalarda maliyetin azaltılması yanında ayrıca önemli sonuçlara da ulaşılabilmektedir. Nitekim biyoinformatik tabanlı algoritmaların geliştirilmesi ile bazı özel hastalıklara karşı teorik hastalık tedavilerinin yapılabileceği öngörülmüştür [10,11]. Dünya üzerinde birçok ülkede hastalık tedavileri için biyoinformatik tabanlı programlar kullanan, ilaç sanayi bağlantılı birçok şirket kurulmuş olup sayıları her geçen gün geçtikçe artmaktadır. Bu şirketlerin çoğunluğu ABD, İsviçre ve İngiltere'de bulunmakla birlikte yazılım alanında büyük atılım gerçekleştirmiş olan Hindistan'da da pek çok biyoinformatik şirketi kurulmuştur. Son yirmi yılda temel biyolojik araştırmaların klinik tıp uygulamaları ve klinik tıp bilgi sistemleri üzerindeki etkisi daha da belirleyici olmuş ve bu yeni kuşak tanı teşhis ve tedavi amaçlı modüllerin ortaya çıkmasına yol açmıştır. Biyoinformatik çalışmalar temel bilimsel araştırmalara yönelik görünmekle beraber önümüzdeki yıllar içinde klinik bilişim için vazgeçilmez olacaktır [1].

2.3. Biyoinformatiğin Tarihçesi

Biyoinformatiğin başlangıcını kesin olarak belirli değildir. 1951 yılında Pauling ve Corey'in proteinlerin sekonder yapılarının doğru anlaşılması için geliştirdikleri yaklaşım biyoinformatik için başlangıç bir çalışma kabul edilebilir. Çağdaş anlamda biyoinformatik bilimi, bilgisayara aşırı gereksinim duyduğundan, 1966 yılında Scientefic American dergisinde moleküler yapıların bilgisayarla çizimleri hakkındaki ilk makalenin yayınlamasını biyoinformatik başlangıç saymak daha gerçekçidir. Biyoinformatik terimi 1980li yılların ortalarından sonra kullanılmaya başlanmıştır. Biyoinformatik ile computional biology, biocomputing ve moleküler biyoinformatik terimleri de aynı anlamda kullanılmaktadır [6]. Temel moleküler ve genetik süreçlerin anlaşılmasında ve karmaşık verilerin analizi ve yorumlanması için yeni yöntemler geliştirilmesinde en etkin kurum olan NCBI 1988'de kurulmuştur. İnsan Genom Projesi (Human Genom Project-HGP) uluslar arası bir çabayı kapsamakta olup Ekim 1990 'da 30-35 bin insan geninin belirlenmesi ve biyolojik çalışmalarda

kullanılabilecek şekilde hizmete sunulmasını temel amaç edinmiştir. İnsan genom projesi biyoinformatiğin gelişiminde itici güç olmuştur [12]. Gen yapısının üç boyutlu modellenmiş hali Şekil 2.4'de görülmektedir.

Şekil 2.4. Gen yapısının 3b görünümü [13]

2.4. Biyoinformatiğin Önemi

Biyoinformatiğin en önemli görevi; insan dâhil tüm biyolojik türlerin genomlarına, protein sekanslarına ve proteinlerin üç boyutlu yapılarına metabolik yol veritabanlarına ve biyoçeşitliliğe bağlı bilgilerine ait niceleyici verilerin toplanmasıdır. Son yıllarda genom sekans projelerinde yaygın uygulamaları sayesinde biyoinformatik büyük önem kazanmıştır. Biyoinformatiğin insan genomu projesinin başarı ile tamamlanmasında çok büyük rolü olmuştur. Ayrıca biyoinformatik biyoteknolojiye dayalı üretim ve süreç geliştirmeye katkılar sağlamaktadır. Hastalık tespiti ve tespit sonucunda geliştirilen ilaç dizaynı, geliştirilmesi pahalı ve zaman alıcı bir süreçtir. Biyoinformatik bu süreçlerin hem maliyetini hem de gerçekleşme süresini çok kısaltmaktadır. Büyük ilaç biyoteknolojisi şirketlerinin hemen tamamında çok geniş biyoinformatik araştırma-geliştirme gurupları kurulmuştur. Son bilgilerle biyoteknolojinin en hızlı büyüyen üretim teknolojisi olduğunu göstermektedir. Bu teknoloji biyoinformatiğin önemli

uygulamalardaki başarıları ile ölçülebilir. Biyoinformatik, medikal tanı ve tedavi alanında da şimdiden önemli başarılar elde etmiştir [14]. Proteinlerin yapısını belirlemede kullanılan deneysel yöntemlerin çok pahalı olması ve deneylerinin uzun sürmesi sebebiyle istatistiksel ve bilişimsel yöntemlerin kullanılması zorunluluk haline gelmiştir. Protein veri bankasındaki bilgileri kullanarak, protein-protein ara yüzlerinin veri kümesini oluşturmak ve daha sonra bu veri kümelerini kullanarak ara yüzlerin istatistiksel olarak incelenmesi de araştırma konuları içindedir. Bu veriler proteinlerin birbiriyle etkileşim mekanizmalarının tayininde çok önemli bir rol alacaktır.

Biyoinformatik metotlar artık biyolojik araştırmaların vazgeçilmez unsuru haline gelmiştir. Aslında dizi analizleri için geliştirilen biyoinformatik günümüzde; yapısal biyoloji, genomik ve gen ifade çalışmaları gibi geniş bir konu aralığında hizmet vermektedir. Tüm biyoinformatik çalışmalarını prensip olarak destekleyen iki yaklaşım vardır. Birincisi biyolojik olarak anlamlı benzerliklere göre verilerin kıyaslanması ve sınıflandırılması, ikincisi ise bir veri türünün analiz edilerek diğer bir veri türünün anlaşılması ve değerlendirilmesidir. Bu yaklaşımlar biyoinformatiğin temel amaçları ile uyum içindedir [15].

2.5. Biyoinformatiğin Amacı

Biyoinformatiğin ana amacı genom bilgisi mevcut olan verilen bir organizma ile ilgili tüm fonksiyonların anlaşılması ve yaşam kalitesinin arttırılmasıdır. Biyoinformatiğin amaçları üç ana başlık altında toplanabilir. Bunlar veri organizasyonu, sistemlerin geliştirilmesi ve sistemlerin uygulanmasıdır. Veri organizasyonu; araştırmacıların biyolojik verilere kolayca ulaşabileceği ve elde ettikleri sonuçların kolayca aktarılabileceği bir şekilde organize edilmelidir. Bunun için, her bilim adamı tarafından sorunsuz kullanılabilecek basit veri tabanları kurulmalıdır (Örneğin: üç boyutlu makro moleküler yapı için kurulan Protein Data Bank-PDB). Sistemlerin geliştirilmesi; veritabanında bulunan bilgilerin analizini yapmalı ve bu analizleri basitleştirmelidir. Sistemlerin geliştirilmesi değişik alanlardan uzmanların işbirliğini gerekli kılmaktadır. Sistemlerin uygulanması;

geliştirilen sistemler ile biyolojik bakış açısıyla toplanılan veriler analiz edilmeli ve yorumlanmalıdır. Günümüzde protein yapıları ve nükleik asit dizilerinin saklandığı yüzden fazla bilgi bankası vardır. Bilgisayar destekli dizi analizleri ile bu değişimlerin sınıflandırması sayesinde akraba türlerinin belirlenmesi ve filogenetik ağaç denen soy ağacının ortaya çıkması ile mümkün olmaktadır [16].

2.6. Sık Kullanılan Veri Tabanları ve Programları

Milyonlarca nükleotidin depolanması ve organizasyonu için veritabanlarının oluşturulması, araştırıcıların bu bilgilere ulaşabilmesi ve yeni veriler girebilmeleri için ilk aşamadır. Nükleotid dizi bilgilerin organizasyonu ve depolanmasını sağlayan dünya üzerinde üç ana kuruluş vardır. Bunlar;

- GenBank (Gen Bankası, ABD-Maryland)
- EMBL (Avrupa Moleküler Biyoloji Laboratuarı, İngiltere-Hixon)
- DDBJ (DNA Japonya veritabanı, Japonya-Mishima) ‘dır.

Bu üç kuruluş araştırmacıların faydalanması için hizmet vermektedir. Bu kuruluşlar nükleotid dizi bilgilerinin toplanması ve dağıtılmasında işbirliği içinde çalışmaktadır. Protein dizi verileri ile ilgili başlıca hizmet sunanlar ise;

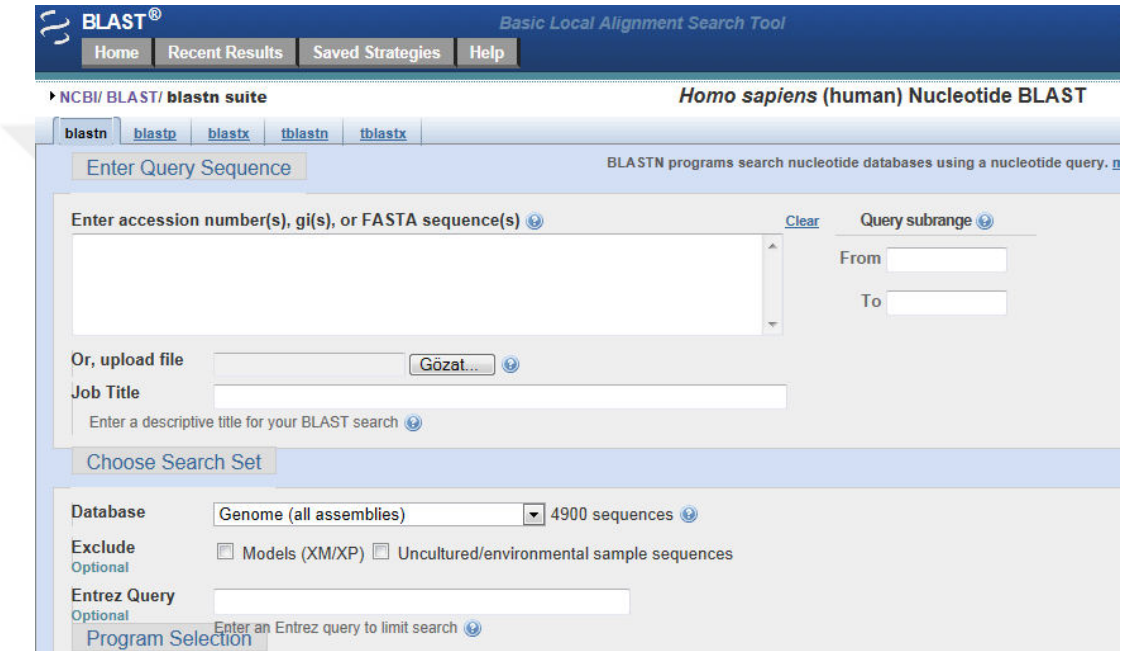
- EMBL
- GenBank
- Swiss-Prot
- PIR International (Protein Identification Resource)

Dizi bilgileri, veritabanlarında iki formatta bulunmaktadır. Birincisi; diziyi veritabanına ilk işleyenler, kaynak gösterimleri, biyolojik atıflar ve dizinin kendisiyle, intronlar, eksonlar, başlangıç ve bitiş kodonları gibi bilgiyi içeren bir tablodan oluşan tam bilgidir. İkincisi ise; FASTA formatı adı verilen ve hızlı benzerlik araştırmaları için kullanılan, sadece diziyi içeren formattır. Her bir diziyi belirleyen ve dizi veritabanına ilk kez girildiğinde verilen Accession (ulaşma)

numaraları özgün kimliklerdir. Patent ofisleri gibi çeşitli kaynaklar sayesinde dizi bilgileri içeren veritabanlarına ulaşılır. 1988'de ABD'nin en büyük tıp kitaplığı olan NLM'nin (National Library of Medicine) bir kolu olarak NCBI veritabanı kurulmuştur. Bu kuruluş halen web tabanlı en önemli biyolojik veritabanıdır. Birçok genom dizileri ile PubMed olarak bilinen biyomedikal ve biyoteknoloji ilişkili araştırma makalelerini kapsar. NBCI'nin sunduğu hizmetlerden biri ENTREZ servisidir. ENTREZ servisinin en önemli özelliği veritabanları arasında çapraz gezinme olanağı sunmasıdır. NCBI'nin bir başka hizmeti olan OMIM, genler ve genetik hastalıklarla ilgili ayrıntılı biyoteknolojik ve tıbbi bilgilerin bulunduğu servistir. Bu servis altında pek çok gende bugüne kadar tanımlanmış mutasyonlar ve ilgili klinik ilişkiler özetlenir [1]. NCBI web sitesi görünümü Şekil 2.5'te görülmektedir.

Şekil 2.5. NCBI web sitesinden bir görünüm [17]

Biyoinformatikte önemli yer tutan BLAST (Basic Local Alignment Search Tool) ile eldeki DNA dizisi ayrıntılı analiz edilmektedir. BLAST aranan dizi sırasını (nükleotid veya aminoasit) veritabanında bulunan mikroorganizmalara ait baz dizileri ile karşılaştırarak aynı veya en yakın olan dizi sırasının ait olduğu mikroorganizmayı yüzde benzerlik oranı ile veren bir bilgisayar programıdır [1]. BLAST dizi eşleştirme programı görünümü Şekil 2.6'da görülmektedir.



Şekil 2.6. BLAST kullanım sayfasından bir görünüm [18]

BLAST ile dizi analizi yapılarak aranan bölgenin baz dizisi belirlendikten sonra, bu dizi NCBI'nin web adresinde yer alan program kullanarak veritabanı ile karşılaştırılır. Tarama sonucu, aranan dizi sırasının hangi canlıya ait olabileceğini benzerlik yüzdesi ile verir. Hızlı sonuç almak için tasarlanan BLASTN bir nükleotid dizisi ile komplementer diziyi ele alarak nükleotid dizisi veritabanları ile karşılaştırır. Fakat bu uygulama yüksek duyarlılık aranan durumlar için uygun değildir. BLASTN ve BLASTX, EST verilerinin analizi, ekson yakalama yöntemi ve genomik dizi örneklemlerinin incelenmesinde kullanılır. Bir dizi için BLAST araştırması yapıldıktan sonra, ilgili gen ile ilgili literatür bilgileri MEDLINE'dan elde edilebilir. Daha sonra ilgili grafik programların yüklenmesi sonrasında protein yapısıyla ilgili veritabanları kullanılarak, proteinin iki veya üç boyutlu yapısı izlenebilir. Protein

dizilerindeki işlevsel motifleri araştırmak amacıyla kullanılan bazı veritabanları ise PROSITE ve BOLCKS'dur [1].

2.7. Mikroarray Teknolojisi

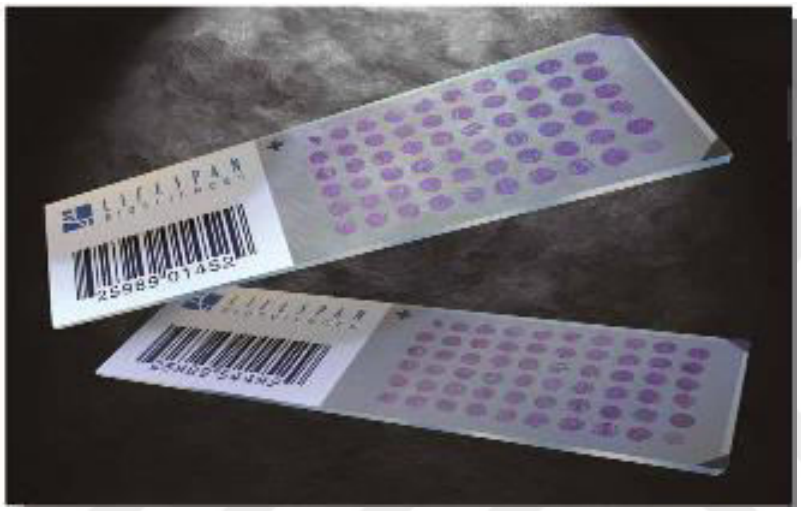
Moleküler biyolojiye paralel olarak bilgisayar teknolojisinin gelişimi, iki disiplini birbirine yaklaştırmıştır. Böylelikle, mevcut koşullar içerisinde biyoteknolojinin en son ulaşabileceği noktalardan biri olarak gen çip (mikroarray) teknolojisi ortaya çıkmıştır. Mikroarray tekniğinin ilk girişimleri Shalon ve Schena tarafından gerçekleştirilmiştir.

Moleküler biyolojideki geleneksel metotlara bakıldığında "bir deneyde bir gen" ilkesi geçerlidir. Bu kurama göre geleneksel metotlar ile gen fonksiyonlarının "bütün resmini" görmek 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.

Benzer uygulamaları ifade etmek için uygulanan yöntem ve amaç farklılıkları olmasına rağmen, "microchip", "DNA chip" ve "DNA array" gibi tanımlamalar da yapılmaktadır. Bununla birlikte, oligonükleotid sentezi tekniği ile üretilmiş olan yüksek yoğunluklu mikroarray platformları çip olarak ifade edilirken, "DNA array" terimi özellikle cam lamlar üzerinde hazırlanan daha küçük kapasiteli platformlar için tercih edilir. Genel olarak mikroarray teknolojisinin en önemli özelliği, küçük bir alanda çok sayıda genomik incelemenin yapılmasına olanak sağlamasıdır. Bir mikroorganizmanın tüm genleri "tırnak" boyutunda olan bir alanda mikroarraylenebilir ve binlerce genin ifade seviyeleri tek bir deneyde aynı anda çalışılabilir [19]. Cam bir mikroarray örneği Şekil 2.7'de görülmektedir.

Mikroarray yöntemi temel olarak iki hedefe yönelik olarak uygulanmaktadır.

1) Gen ifade düzeyinin ölçülmesi ve düzey farklılıklarının belirlemektir.
2) Genoma ait baz dizisinin araştırılması ve genomlar arası dizi farklılıklarını tespit etmektir.



Şekil 2.7. Cam bir mikroarray örneği [20]

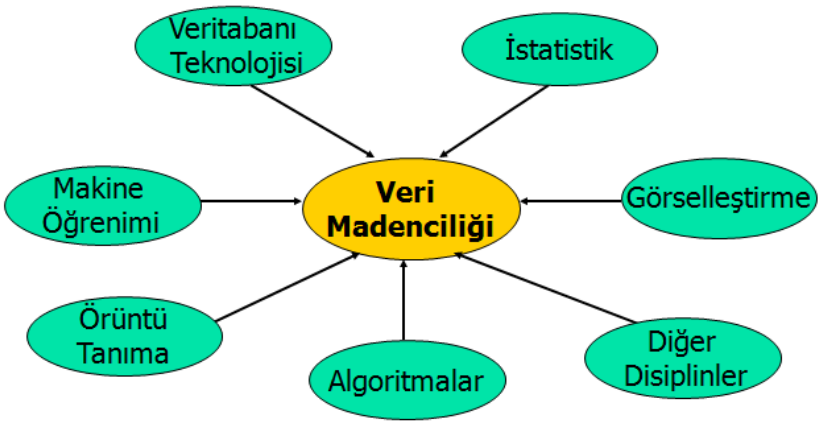
Bu tekniğin en heyecan verici yanı ise, yakın bir gelecekte küçük bir okuyucu cihaz aracılığı ile hasta başında çok sayıda hastalık/etken arasında ayırıcı tanı yapılmasına veya tedavi sonucunun değerlendirilebilmesine olanak sağlayacak olmasıdır.

2.8. Veri Madenciliği

Veri madenciliği, faydalı olabilecek anlamlı ve uygulanabilir bilgilerin veri yığınları içerisinden çıkarılmasına verilen addır. Çok büyük veri tabanlarından, önceden bilinemeyen, anlamlı ve kullanılabilir bilginin çıkarılma işlemi olarak ifade veri madenciliği tanımlanabilir. Diğer bir ifade ile çok büyük veri tabanlarındaki ya da veri ambarlarındaki veriler arasında bulunan ilişkilerin, örüntülerin, sapma ve eğilimlerin ortaya çıkarılması ve keşfi işlemidir. "Veri Tabanlarında Bilgi Keşfi" (Knowledge Discovery in Databases) uygulamaları ile birlikte faaliyet alanına yönelik karar destek mekanizmaları için gerekli ön bilgileri temin etmek için kullanılır [21].

Toplanan bilgilerin, istatistiksel ve sayısal yöntemlerle incelenip ilgili kurum ve yönetim destek dizgelerinde kullanılmak üzere değerlendirilmesi veri madenciliğinin amacı olarak ifade edilebilir. Veri madencisinin geleneksel yöntemlerde olduğunun aksine başlangıçta herhangi bir amacı ya da varmak istediği bir kavram yoktur [22].

Veri madenciliği konusunda çeşitli tanımlar mevcuttur. Bu tanımlar içerisinden basit bir tanım yapmak istenirse veri madenciliği, büyük ölçekli veriler arasından "değerli olan" bir bilgiyi elde etme işidir. Değerli bilginin elde edilmesinden sonra veriler arasındaki ilişkileri belirlemek ve gerektiğinde ileriye yönelik tahminlerde de bulunmak mümkün görülmektedir. Dolayısıyla, veri madenciliği üretilen tüm verilerin belirli yöntemler kullanılarak mevcut ya da ileride ortaya çıkabilecek önemli bilgileri ortaya çıkarma süreci olarak değerlendirilebilir. Bu açıdan bakıldığında, veri madenciliği işinin kurumların karar destek sistemleri için önemli bir yere sahip olabileceğini söyleyebiliriz. Veri madenciliğinin kullanım alanları Şekil 2.8'da görülmektedir.



Şekil 2.8. Veri madenciliği kullanım alanları [22]

Veri madenciliği aslında klasik istatistiksel uygulamalara çok benzemektedir. Ancak klasik istatistiksel uygulamaların çalışma alanı yeterince düzenlenmiş ve çoğunlukla özet verilerdir. Ayrıntılı bilgi olsa bile burada kayıtlar binlerce olabilir. Veri madenciliğinde ise milyonlarca hatta milyarlarca veri ve çok daha fazla değişken ile ilgilenir. Veri sayısı çok olmasından dolayı, bazı özel analiz algoritmaların

geliştirilmesi ve verinin saklandığı ortamların da örneğin veri ambarı biçiminde yeniden düzenlenmesini gerekmiştir [22].

2.9. Veri Madenciliği Uygulama Alanları

Veri madenciliği bankacılık, sigortacılık ve pazarlama gibi alanlarda kullanılmaktadır. Bu alanlarda çeşitli tespit ve değerlendirmeler yapılmaktadır [22]. Bunlardan kısaca bahsetmek gerekirse;

Bankacılık alanında;

- Gelen kredi taleplerinin değerlendirilmesi
- Farklı finansal değişkenler arasında bulunan gizli ilişkilerin tespit edilmesi
- Kredi kartı dolandırıcılıkların ve sahtekârlıkların belirlenmesi
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi

Sigortacılık alanında;

- Riskli müşteri gruplarının belirlenmesi
- Müşterilerin talep edeceği yeni poliçelerinin tahmin edilmesi
- Sektördeki dolandırıcılıkların tespiti

Pazarlama alanında;

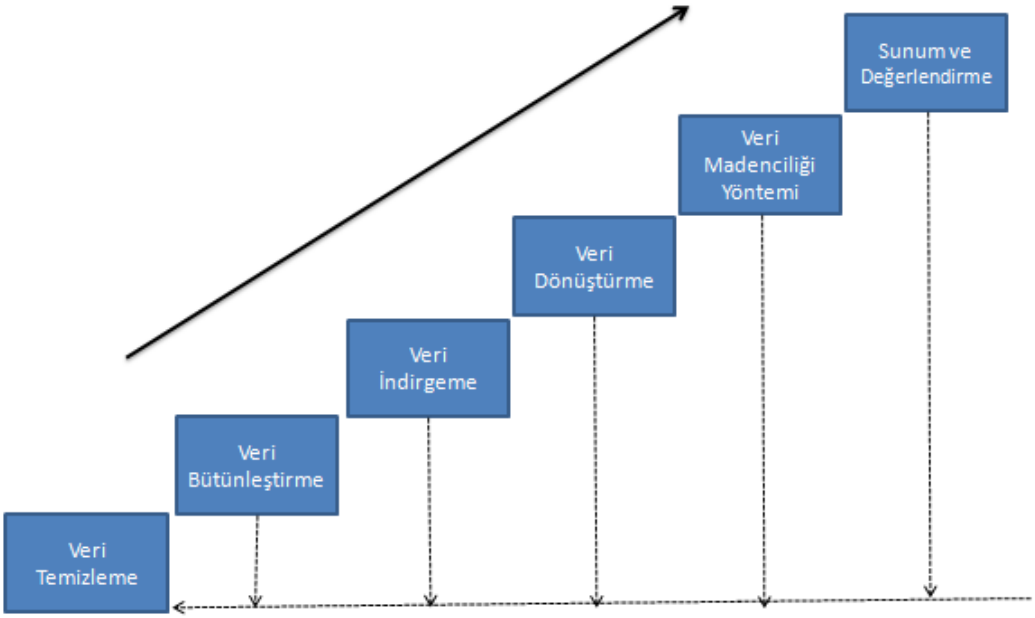
- Müşteri ilişkileri yönetimi
- Müşteri değerlendirme
- Müşterilerin satın alma alışkanlıklarının belirlenmesi
- Müşterilerin demografik özellikleri arasındaki bağlantıların ortaya konulması
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması
- Pazar sepeti analizi
- Satış tahmini yapılmaktadır ve veri madenciliği etkin olarak kullanılmaktadır.

2.10. Veri Madenciliği Süreci

Veri madenciliği bir süreç olarak tanımlanmaktadır ve bu süreç belirli adımlardan oluşmaktadır. Bu adımlar maddeler halinde yazılmak istenirse;

a. Veri temizleme
b. Veri bütünleştirme
c. Veri indirgeme
d. Veri dönüştürme
e. Veri madenciliği algoritmasını uygulama
f. Sonuçları sunum ve değerlendirme olarak yazılabilir [22].

Veri madenciliği süreci Şekil 2.9'da görülmektedir.



Şekil 2.9. Veri madenciliği süreci [22]

2.10.1 Veri temizleme

Bazı durumlarda, istenilen özelliklerin, üzerinde çalışılan verilerde bulunmadığı görülebilir. Eldeki mevcut verilerde eksik veya uygun olmayan örneklerle karşılaşılabilir. Bu istenmeyen anlamsız ve tutarsız veriler gürültü olarak adlandırılmıştır. Bu gibi durumlarda verinin söz konusu sorunlardan temizlenmesi gerekmektedir. Eksik verilerin yerine yeni veriler belirlenmeli ve yerine yeni veriler konulmalıdır. Bunun için aşağıda belirtilen yöntemlerden biri kullanılabilir.

a) Veri kümesinden eksik değer içeren kayıtlar atılabilir.
b) Genel bir sabit eksik değerlerin yerine kullanılabilir. Bütün kayıp değerler için aynı sabit kullanılmalıdır. Örneğin "bilinmiyor" değeri bu eksik veri yerine kullanılabilir. Ancak bütün değişkenlerde kayıp değerler yerine aynı sabit değerin kullanımı sorunlarda yaratmaktadır.
c) Başka bir yöntemde olarak değişkenin tüm verilerinin ortalaması hesaplanıp eksik veriler yerine kullanılabilir.
d) Değişkenin tüm verileri yerine, sadece bir sınıfa ait örneklerin değişken ortalaması hesaplanarak eksik değer yerine kullanılabilir.
e) Veriler arası ilişkiler belirlenerek veya uygun bir model kurularak eksik değer tahmin edilebilir ve eksik değer yerine kullanılabilir.

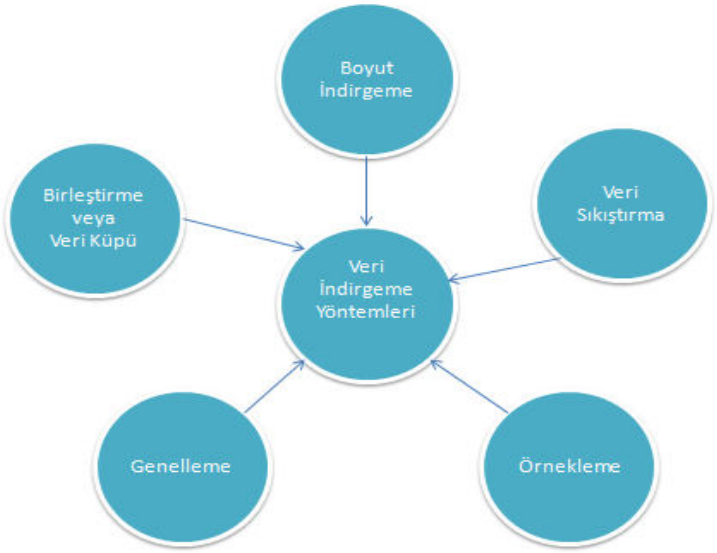
2.10.2. Veri bütünleştirme

Üzerinde çözümleme yapılacak veriler farklı veritabanlarından alınabilir. Bu verilerin birlikte değerlendirilmesi için farklı türdeki verilerin tek türe dönüşmesi gerekmektedir. Eğer veri madenciliği uygulaması için bir veri ambarı altyapısı hazırlanmış ise söz konusu veri bütünleştirme işleminin yapılmış olması gerekmektedir. Ancak böyle bir yapı yoksa söz konusu veri bütünleştirme işleminin doğrudan veri madenciliğine esas oluşturacak veriler üzerinde uygulanması gerekecektir.

2.10.3. Veri indirgeme

Veri madenciliği uygulamalarında çözümleme işleminin uzun zaman aldığı durumlarda ortaya çıkabilmektedir. Yapılacak olan çözümlemeden elde edilecek sonucun değişmeyeceği tahmin ediliyorsa, veri sayısı ya da değişkenlerin sayısı azaltılabilir. Bu azaltma işlemi çözümleme yapılacak verilerde;

a) Birleştirme veya veri küpleri oluşturma
b) Boyut indirgeme
c) Veri sıkıştırma
d) Örnekleme
e) Genelleme işlemleri yapılabilir. Veri indirgeme yöntemleri Şekil 2.10'de görülmektedir.



Şekil 2.10. Veri indirgeme yöntemleri [22]

Veriyi indirgeme aşamasında çok boyutlu veri küpleri oluşturarak veri indirgeme gerçekleştirilebilir. Böylece çözümlemeler sadece belirlenen boyutlara göre yapılır. Veriler arasında bir seçim yapılarak, anlamsız veriler veritabanından çıkarılır ve boyut azalması sağlanabilir. Veri sıkıştırma aşamasında, büyük veri kümelerinin sıkıştırılarak daha az çok kaplamasını önlemek mümkündür. Örnekleme aşamasında

ise büyük veri topluluğu yerine onu temsil eden daha küçük veri kümelerinin oluşturulması amaçlanır. Genelleme verilerin tek tek değil genel ifade edilmesi amaçlanır.

2.10.4. Veri dönüştürme

Bazı durumlarda eldeki mevcut verilerin veri madenciliği uygulamasına aynen katmak uygun olmamaktadır. Değişkenlerin ortalama ve varyansları önemli ölçüde farklı olduğu durumlarda büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerinde baskın olduğu görülmekte çözümleme sonucunu önemli bir şekilde etkilemektedir. Ayrıca değişkenlerin sahip olduğu çok büyük ve çok küçük değerler de çözümlemelerin sağlıklı biçimde yapılmasını engellemektedir. Bu nedenle bir dönüşüm yöntemi uygulayarak söz konusu değişkenlerin normalleştirilmesi veya standartlaştırılması uygun bir yol olacaktır.

2.10.5. Veri madenciliği algoritmasını uygulama

Veri madenciliği alanında birçok algoritma ve yöntem geliştirilmiştir. Bu yöntemlerin çoğu istatistiksel tabanlı olup, üç şekilde değerlendirilebilir.

- Sınıflandırma
- Kümeleme
- Birliktelik kuralı

Yapılan çalışmada bu yöntemlerden sınıflandırma yöntemi, sınıflandırma metotlarından DVM, K-ENK, NB ve DAA sınıflandırma algoritmaları kullanılacaktır.

2.10.6. Sonuçları sunum ve değerlendirme

Veri madenciliği algoritmaları veriler üzerinde uygulandıktan sonra sonuçlar değerlendirilerek sunulur. Sonuçlar çoğu kez grafiklerle desteklenir [22].

2.11. Literatür Araştırması

Biyoinformatik alanında geçmiş dönemlerde mikroarray analizi için birçok çalışmalar yapılmış olup, bunlardan biri de D.H Tran ve arkadaşlarının yapmış olduğu çalışmadır [23]. Tran ve arkadaşları mikroRNA (miRNA) ifade profillerinin tümör örneklerinde sınıflandırma ve analiz çalışmalarını yürütmüş olup mikroarray teknolojisi sayesinde binlerce miRNA aynı anda incelenebilmiştir. Yapılan çalışmalarda Gloub ve arkadaşlarının [24] yapmış olduğu çalışmalarda kullanılan 223 örnek ele alınmış, 151 miRNA öznitelik olarak kullanılmıştır. Sınıflandırma algoritması olarak DVM uygulanmış, ele alınan örnekler 2 sınıfa (tümörlü-normal) ayrılmıştır. Sonuç olarak Çizelge 2.2'deki değerler elde edilmiştir.

Çizelge 2.2. Tran ve ark. uyguladığı sınıflandırma algoritmaların doğruluk, duyarlılık ve AUC değerleri

DVM Kernel Fonksiyonu	Doğruluk	Duyarlılık	AUC
RBF	0,92	0,98	0,98
Linear	0,95	0,95	0,97
Polinominal	0,93	0,95	0,96

Yapılan bir başka çalışmada ise X.Fan ve arkadaşları [25] karaciğerde bulunan hepatatoks hücreleri ile kolon, lösemi, lenf kanser hücrelerinde mikroarray analizini uygulamışlardır. Hepatatoks veri kümesinin kullanıldığı çalışmada 318 örnek incelenmiş, 20500 gen öznitelik olarak kullanılarak sınıflandırma gerçekleştirilmiştir [26]. Kolon kanseri veri kümesi [27] 2000 gen ve 62 örnekten, lösemi veri kümesi [28] 7129 gen ve 72 örnekten, lenf kanseri veri kümesi [29] ise 4026 gen ve 96 örnekten oluşmaktadır. Çalışmanın sonuçları Çizelge 2.3'teki gibidir.

Çizelge 2.3. X.Fan ve ark. hepatatoks, kolon, lösemi ve lenf hücrelerinde uyguladığı sınıflandırma algoritmaların doğruluk değerleri

Uygulanan Sınıflandırma Algoritması	Hepatatoks	Kolon Kanseri	Lösemi	Lenf Kanseri
Karar Ağaçları	0,901	0,817	0,943	0,966
DVM	0,886	0,813	0,961	0,945
K-ENK	0,873	0,795	0,892	0,926

D. Liu ve arkadaşlarının [30] Wisconsin Diagnotistic Breast Cancer (WDBC) hücreleri veritabanında [31] yaptığı araştırmalar ile 699 örnek 9 öznitelik ile 2 sınıfa ayrılmıştır. Sınıflandırma algoritması olarak DVM kullanılmış, eğitim ve test verileri sırasıyla %50-50, %70-30, %80-20 şeklinde ayrılmıştır. Çalışmada alınan sonuçlar Çizelge 2.4'teki gibidir.

Çizelge 2.4. D.Liu ve ark. WDBC göğüs kanseri hücrelerinde uyguladığı DVM algoritmasının eğitim-test küme yüzdeleri ile doğruluk değerleri

Uygulanan Sınıflandırma Algoritması	Eğitim-Test Küme Yüzdeleri	Doğruluk Değerleri
DVM	%50-50	0,95
	%70-30	0,96
	%80-20	0,96

C. Chakraborty ve arkadaşlarının yapmış olduğu çalışmada [32] yine WDBC hücreleri veritabanında mikroarray analizi olmuştur. Çalışmanın amacı, göğüs kanserinde DVM sınıflandırıcı kullanarak yüksek doğruluk değerlerine ulaşmak olarak belirlenmiştir. Çalışmada iki çeşit veri kümesi kullanılmıştır. Bu veri kümelerinin ilki 699 örnekten oluşmakta olup 9 öznitelik ile 2 sınıf mevcuttur. Diğer veri kümesinde ise 569 örnek 10 öznitelik ile 2 sınıfa ayrılmaktadır. Bu veri kümelerine DVM algoritması uygulanırken DVM'nin polinominal ve Gauss fonksiyonları kullanılmıştır. Alınan sonuçları Çizelge 2.5'teki gibidir.

Çizelge 2.5. C. Chakraborty ve ark. WDBC göğüs kanseri hücreleri üzerinde polinominal ve Gauss kernelleri ile uyguladığı DVM algoritmasının duyarlılık(sensitivity) ve özgüllük(specificity) değerleri

Veri Kümesi - I	Polinominal	Gauss	Veri Kümesi - II	Polinominal	Gauss
Duyarlılık	0,9775	1	Duyarlılık	0,9269	0,945
Özgüllük	0,9762	0,9879	Özgüllük	0,9256	0,9298

Başka bir çalışmada M. Acı ve M. Avcı [33] K-ENK algoritmasını WDBC hücreleri veritabanında denemişlerdir. WDBC veri kümesi daha öncede belirtildiği üzere 699 örnek 9 öznitelik ve 2 sınıftan oluşmaktadır. K-ENK algoritmasında önemli olan uzaklık değerleri sırasıyla Manhattan, Öklid ve Minowski olarak alınmıştır. Çizelge 2.6'daki değerler çalışmanın sonuçlarını göstermektedir.

Çizelge 2.6. M. Acı ve M. Avcı'nın WDBC hücrelerinde uygulanan K-ENK algoritmasının farklı mesafelerde sınıflandırdığı yanlış örnek sayıları

	Manhattan	Öklid	Minowski
k=1	14	10	17
k=2	16	17	15
k=3	15	18	15
k=4	19	16	21
k=5	17	19	20
* Yukarıdaki değerler yanlış sınıflandırılan örnek sayıları göstermektedir.			

B. Han ve arkadaşları [34] mikroarray analiz çalışmalarını lösemi, beyin tümörü, kolon ve prostat kanser hücrelerinde denemişlerdir. Örnek veri kümeleri öznitelik seçim metotları ile sırasıyla 1, 5, 10, 20, 50, 100 gen olacak şekilde ayrılmış ve daha sonraki aşamada sınıflandırma algoritmaları kullanılmıştır. Sınıflandırma algoritmaları olarak DVM, K-ENK, RF, NB uygulanmıştır. Çizelge 2.7'deki sonuçlar kolon kanser hücresinin 5, 10 ve 20 öznitelik için sınıflandırma doğruluk değerlerini göstermektedir.

Çizelge 2.7. B. Han ve ark. 5, 10, 20 öznitelik kullanarak kolon kanser hücrelerinde uyguladığı farklı sınıflandırma algoritmaların doğruluk değerleri

Uygulanan Sınıflandırma Algoritması	Gen sayısı		
	5	10	20
DVM	0,840	0,863	0,851
K-ENK	0,853	0,903	0,886
RF	0,846	0,866	0,886
NB	0,887	0,911	0,866

D. Li ve arkadaşlarının yapmış olduğu çalışmada [35] Kaliforniya Üniversitesi'nin makine öğrenme veri ambarında bulunan veri kümeleri kullanılmıştır. Veri kümeleri olarak Echocardiogram, WDBC, BUPA karaciğer ve PIMA diyabet verileri kullanılmıştır [31]. Sınıflandırma algoritması olarak Gauss ve polinominal kernelleri ile DVM algoritması uygulanmıştır. Veri kümeleri sırasıyla 5, 10, 20, 30, 50, 100 eğitim örneği olacak şekilde ayrılmıştır. Çizelge 2.8 çalışmaların sonuçlarını göstermektedir.

Çizelge 2.8. D. Li ve ark. Echocardiogram, WDBC, BUPA ve PIMA veri kümelerinde Gauss ve polinominal kernelleri ile uygulanan DVM algoritmasının 5, 10, 20 ve 30 eğitim örneği seçilerek elde edilen doğruluk değerleri

Echocardiogram Veri Kümesi				
DVM Kernel Fonksiyonu	Eğitim Örnek Sayıları			
	5	10	20	30
DVM-Gauss	67,17	74,33	84,17	86,50
DVM-Polinominal	64,50	67,67	73,00	74,17
WDBC Veri Kümesi				
DVM Kernel Fonksiyonu	Eğitim Örnek Sayıları			
	5	10	20	30
DVM-Gauss	52,73	57,60	62,90	74,10
DVM-Polinominal	52,73	53,56	54,88	55,38
BUPA Karaciğer Veri Kümesi				
DVM Kernel Fonksiyonu	Eğitim Örnek Sayıları			
	5	10	20	30
DVM-Gauss	49,32	51,53	54,57	56,35
DVM-Polinominal	51,27	51,92	54,50	56,42
PIMA Diyabet Veri Kümesi				
DVM Kernel Fonksiyonu	Eğitim Örnek Sayıları			
	5	10	20	30
DVM-Gauss	55,00	59,22	60,87	60,03
DVM-Polinominal	55,03	60,05	61,68	61,68

Mikroarray tabanlı kanser sınıflandırma olarak geçen çalışmada ise X.Wang ve O.Gotoh [36] farklı sınıflandırma algoritmalarını sinir sistemi, kolon, akciğer, prostat, göğüs kanser hücreleri ile lösemi olan örnekler üzerinde denemiştirler. Sınıflandırma algoritmaları olarak DVM, NB, K-ENK ve KA uygulanmıştır. Çizelge 2.9'da kolon kanserinin 5, 10, 20, 50 ve 100 öznitelik ile uygulanan sınıflandırma sonuçları görülmektedir.

Çizelge 2.9. X.Wang ve O.Gotoh'un kolon kanseri veri kümesinde 5, 10, 20, 50 ve 100 öznitelik ile uyguladığı farklı sınıflandırma algoritmalarının doğruluk değerleri

Kolon Kanser Veri Kümesi					
Uygulanan Sınıflandırma Algoritması	Gen sayısı				
	5	10	20	50	100
DVM	59,68	82,26	88,71	83,87	87,10
NB	75,81	80,65	79,03	77,42	74,19
K-ENK (k=5)	67,74	82,26	85,48	85,48	88,71
KA	61,29	79,03	83,87	74,19	88,71

3. METODOLOJİ

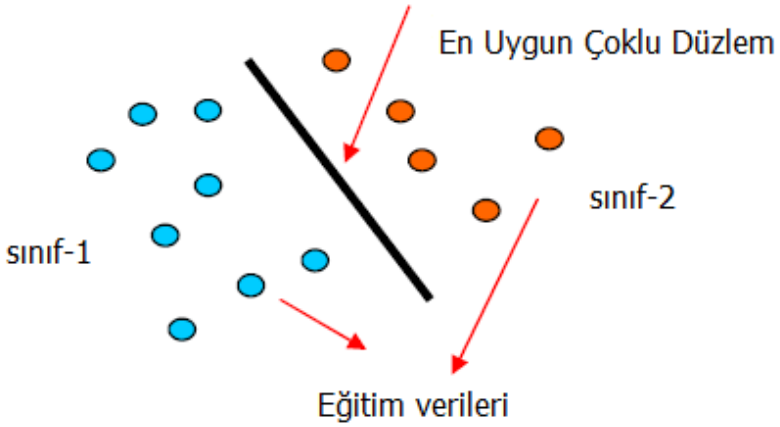
3.1. Sınıflandırma

Sınıflandırma makine öğrenimi ve istatistik alanlarında yeni bir gözlem örneğinin ait olduğu sınıfa ayırma problemi olarak tanımlanabilir. Bu ayırma işlemi yapılırken temel olarak sınıfı bilinen eğitim kümeleri baz olarak alınmaktadır. Her bir gözlem açıklanabilir değişken veya öznitelik olarak bilinen ölçülebilir özelliklerle analiz edilmektedir. Bu özellikler kategorik, sıralı, tamsayı değerleri veya reel sayılardan oluşan değerler olabilir. Somut bir uygulamada sınıflandırma yapan algoritma sınıflandırıcı olarak adlandırılmaktadır. Makine öğrenimi terminolojisinde sınıflandırma, doğru tanımlanmış eğitim kümelerinin mevcut olduğu bir danışmalı öğrenme örneği olarak kabul edilmektedir [37]. Yapılan çalışmalarda literatürde geçen sınıflandırma algoritmaları kullanılacaktır.

3.2. Sınıflandırma Yöntemleri

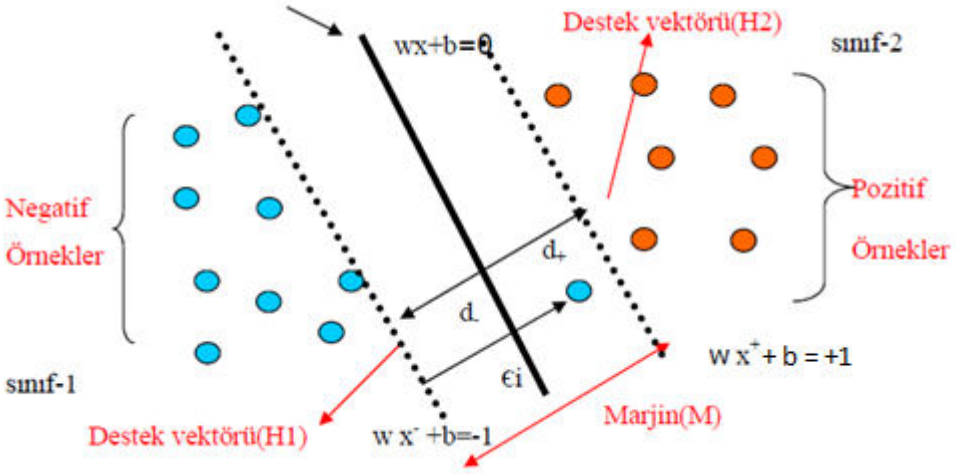
3.2.1. Destek vektör makineleri (DVM)

DVM, V.Vapnik tarafından geliştirilen bir makine öğrenme metodudur. Çekirdek tabanlı doğrusal olmayan sınıflandırıcıların sinyal işleme, yapay öğrenme ve veri madenciliği alanındaki pratik problemlerde iyi sonuçlar verdiği kanıtlanmıştır (Vapnik, 1999), (Weston, 1999). DVM temelde iki sınıflı problemlerle ilgilenir. Girişte olarak alınan veriler destek vektörler ile tanımlanabilen bir çoklu düzlem (hyperplane) tarafından Şekil 3.1'de gösterildiği gibi ikiye ayrılır. Amaç 2 sınıfı birbirinden ayırabilecek en uygun çoklu düzlemi bulabilmektir [38].



Şekil 3.1. Örnek bir çoklu düzlem gösterimi [38]

Burada amaç, en uygun çoklu düzlemi bulabilmektir. Her iki sınıfın en uygun çoklu düzlemine en yakın veri noktalarından geçen çoklu düzlemler çizilir ve bu iki düzlem birbirine paraleldir. En uygun çoklu düzleminin kalitesi bu düzlemler arasındaki mesafenin belirlenmesi ile ortaya çıkar. DVM iki sınıf arasındaki sınırı ayırt etme yüzeyini belirlemekte, yani eğitim kümesi ile ayırt etme yüzeyine en yakın noktaların arasındaki mesafeyi maksimumlaştırmaktadır. Destek vektörleri, çoklu düzlem ve marjin kavramları Şekil 3.2'de gösterilmiştir.



Şekil 3.2. Eğitim verileri ve çoklu düzlem [38]

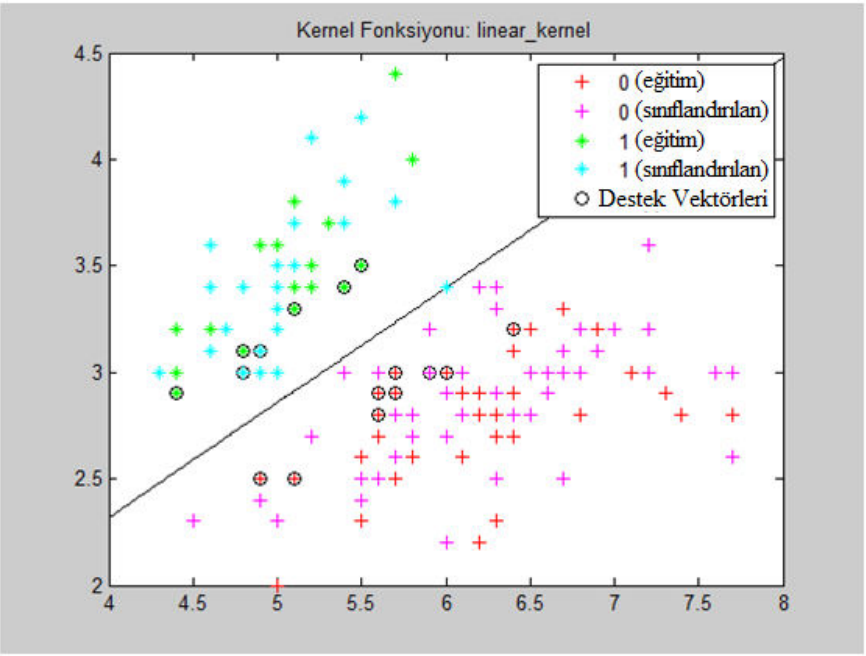
Bu iki grubu çok boyutlu bir düzlem üzerinde ayırdığımız zaman, düzlemi ve boyutları birer özellik olarak düşünmek mümkündür. Basit anlamda sisteme giren her girdinin (input) bir özellik çıkarımı (feature extraction) yapılmış ve sonuçta bu iki boyutlu düzlemde her girdiyi gösteren farklı bir nokta elde edilmiştir. Bu noktaların sınıflandırılması çıkarılmış olan özelliklere göre girdilerin sınıflandırılması ile mümkün olmaktadır.

Her iki sınıf arasında oluşan aralığa tolerans (offset) adı verilebilir. Bu düzlemdeki her bir noktanın tanımı Eş. 3.1'deki gibi gösterimi yapılabilir:

$$D = \left\{ \left(x_i, c_i \right) \middle| x_i \in R^P, c_i \in \{-1,1\} \right\}_{i=1}^{n} \tag{3.1}$$

Eş. 3.1'deki gösterim şu şekilde okumak mümkündür. Her x,c ikilisi için X vektör uzayındaki bir nokta ve c ise bu noktanın -1 veya +1 değerleri olduğunu gösteren değerdir. Bu noktalar kümesi i= 1'den n sayısına kadar gitmektedir.

Çoklu düzlemde her noktanın değeri "wx + b = 0" denklemi olarak ifade edilmektedir. Buradaki w çoklu düzleme dik olan normal vektörü ve x noktanın değişen parametresi ve b ise kayma oranıdır. Bu denklemi klasik "ax+b" doğru denklemine benzetmek mümkündür [39] ve Şekil 3.3'te DVM algoritmasının örnek bir uygulaması görülmektedir.



Şekil 3.3. Örnek bir DVM modellemesi

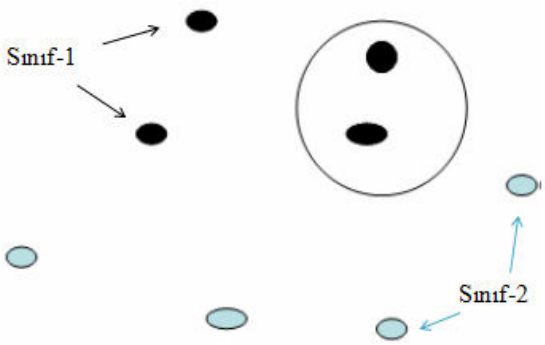
3.2.2 K-En Yakın Komşuluk (K-ENK)

K-ENK, sınıflandırılmak istenen örneğin en yakın örnekleri bulan danışmanlı bir metottur. Bu yöntem sınıflarda bulunan örnekler ile sınıflandırılmak istenen örnek arasındaki benzerliğe dayanmaktadır. Eğitim örnekleri n-boyutlu sayısal niteliklerle tanımlanmaktadır. Her bir örnek n-boyutlu uzayda bir noktayı gösterir. Bu sayede n-boyutlu örnek uzayda bütün eğitim örnekleri depolanır. Bu k-eğitim örnekleri, bilinmeyen örnek A'ya k en yakın komşudur. Yakınlık, Öklid (Euclidean Distance) uzaklığı hesaplanarak bulunur. $X=(X_1,X_2,\ldots,X_n)$ ve $Y=(Y_1,Y_2,\ldots,Y_n)$ arasındaki Öklid uzaklığı ile hesaplanır [38]. Eş 3.2'de bu uzaklık hesabını görebiliriz.

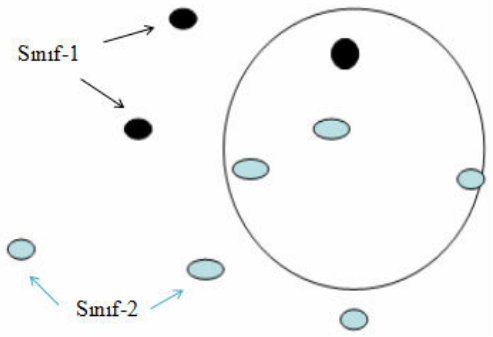
$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.2}$$

Bilinmeyen örnek, en yakın k-komşularının arasında en fazla yakınlığı bulunan sınıfa atanır. Şekil 3.4'te, k değerinin 1, Şekil 3.5'te k değerinin 3 olduğu duruma örnek

gösterilmektedir. K değerinin 1 olduğu durumda, bilinmeyen örnek, örnek uzayında en yakın eğitim örneğinin sınıfına aittir.



Şekil 3.4. k = 1 değeri için örnek bir sınıflandırma [38]



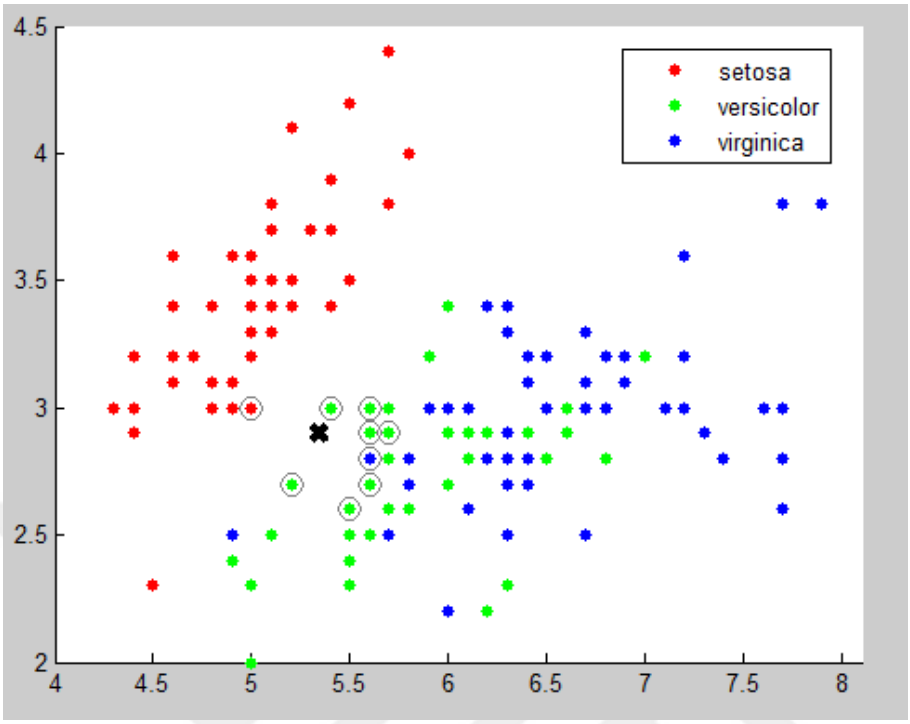
Şekil 3.5. k = 3 değeri için örnek bir sınıflandırma [38]

Bilinmeyen bir X örneğini, çizelge 3.1'deki eğitim verilerinden yararlanarak hangi sınıfa ait olduğunu K-En Yakın Komşuluk sınıflandırma yöntemi ile tahmin edilmiştir. Veri örneklerinin, araba sayısı, ev sayısı ve mağaza sayısı adında özellikleri mevcuttur. Elimizde AS, OS ve ZS olmak üzere 3 sınıf vardır. X = (araba sayısı= "3", ev sayısı = "2", mağaza sayısı = "1") örneği, sınıflandırmak istediğimiz bilinmeyen örnek olsun. Bilinmeyen örnek X'in, tüm sınıflardaki her elemanla arasındaki Öklid uzaklığı hesaplanır ve k değeri 5 seçildiği için en küçük ilk 5 değere sahip komşularının arasında en sık bulunan sınıfa atanır.

Çizelge 3.1. K-ENK algoritması için örnek bir uygulama [38]

Sınıf		Araba Sayısı	Ev Sayısı	Mağaza Sayısı	Öklid Uzaklığı
1	Zengin Sınıf(ZS)	6	8	3	$\sqrt{((6-3)^2+(8-2)^2+(3-1)^2} = 7.00$
2	Zengin Sınıf(ZS)	4	6	5	$\sqrt{((4-3)^2+(6-2)^2+(5-1)^2} = 5.74$
3	Zengin Sınıf(ZS)	4	7	8	$\sqrt{((4-3)^2+(7-2)^2+(8-1)^2} = 8.66$
4	Orta Sınıf (OS)	2	2	3	$\sqrt{((2-3)^2+(2-2)^2+(3-1)^2} = 2.24$
5	Orta Sınıf (OS)	2	2	2	$\sqrt{((2-3)^2+(2-2)^2+(2-1)^2} = 1.41$
6	Orta Sınıf (OS)	2	1	2	$\sqrt{((2-3)^2+(1-2)^2+(2-1)^2} = 1.73$
7	Alt Sınıf (AS)	1	1	0	$\sqrt{((1-3)^2+(1-2)^2+(0-1)^2} = 2.45$
8	Alt Sınıf (AS)	1	0	0	$\sqrt{((1-3)^2+(0-2)^2+(0-1)^2} = 3.00$
9	Alt Sınıf (AS)	0	1	0	$\sqrt{((0-3)^2+(1-2)^2+(0-1)^2} = 3.32$

Elde edilen ilk 5 değer sırasıyla; 1.41(OS), 1.73(OS), 2.24(OS), 2.45(AS), 3.00(AS) değerleridir. Bu değerler arasında en çok orta sınıfa ait elemanlar bulunduğundan X bilinmeyen örneği orta sınıfa atanır. Şekil 3.6'da da örnek bir K-ENK uygulaması görülmektedir.



Şekil 3.6. Örnek bir K-ENK uygulaması

3.2.3. Naive Bayes (NB)

NB yöntemi sık kullanılan bir sınıflandırma yöntemidir ve pratik, olasılığa dayanan bir sınıflayıcıdır. Diğer sınıflandırıcılarla kıyaslandıklarında birçok uygulamada hata oranının düşük olduğu görülmüştür. Bu yöntemde elimizde , S_1,S_2..S_n adında n adet sınıf olduğunu farz edelim. Herhangi bir sınıfa ait olmayan bir veri örneği X'in, hangi sınıfa ait olduğu NB sınıflandırıcı tarafından belirlenmektedir. Veri örneği X, verilen sınıflara ait olma olasılığı en yüksek değere sahip sınıfa atanmaktadır. Sonuç olarak, NB sınıflandırıcı bilinmeyen örnek X'i, S_i sınıfına atamaktadır. Her veri örneği, m boyutlu özellik vektörleri ile gösterilir, X = (X_1,X_2,..,X_m). Bütün özellikler aynı derecede önemli olup, birbirinden bağımsızdır. Bir özelliğin değeri başka bir özellik değeri hakkında bilgi içermez [38]. X = (X_1,X_2..X_m) örneğinin S_i sınıfında olma olasılığı Eş. 3.3'deki gibidir.

$$P\left(\frac{S_i}{X}\right)=\frac{P\left(\frac{X}{S_i}\right)P(S_i)}{P(X)} \tag{3.3}$$

Bilinmeyen bir X örneğinin çizelge 3,2'deki eğitim verilerinden yararlanarak hangi sınıfa ait olduğunu NB sınıflandırma yöntemi ile tahmini istenmektedir. Eğitim verileri Çizelge 3.2'deki veri olup, örnekleri vites, renk, yakıt ve kapı özellikleri ile tanımlanmaktadır. Elimizde A ve B olmak üzere 2 sınıf mevcuttur. Sınıflandırmak istenilen bilinmeyen örnek;

X = (vites = "otomatik", renk = "kırmızı", yakıt = "dizel", kapı = "4") olsun.

Çizelge 3.2. Naive Bayes için bilinmeyen örnek bir tablo [38]

Örnek	**Sınıf**	**Vites**	**Renk**	**Yakıt**	**Kapı**
1	A	Otomatik	Gri	Dizel	2
2	B	Normal	Gri	Benzinli	4
3	A	Otomatik	Kırmızı	Benzinli	2
4	A	Otomatik	Gri	Benzinli	4
5	A	Otomatik	Beyaz	Dizel	2
6	B	Otomatik	Kırmızı	Benzinli	4
7	A	Normal	Gri	Dizel	4
8	B	Otomatik	Gri	Benzinli	4
9	A	Normal	Beyaz	Benzinli	4
10	A	Otomatik	Kırmızı	Benzinli	4
11	B	Normal	Kırmızı	Dizel	4
12	A	Normal	Beyaz	Dizel	4
13	A	Normal	Gri	Benzinli	2
14	B	Otomatik	Beyaz	Benzinli	4
15	B	Otomatik	Beyaz	Benzinli	4

X bilinmeyen verisinin hangi sınıfa ait olduğunu bulabilmek için $P(X|S_i)P(S_i)$ değerini maksimumlaştırmak gerekmektedir. A sınıfı 9 elemandan, B sınıfı da 6 elemandan oluşmaktadır. $P(S_i)$ ve her sınıf için olasılık değerleri eğitim örneklerinden hesaplanmaktadır:

P(A) = 9/15 =0.600
P(B) = 6/15 =0.400

P(X|S_i), i =1,2 değerlerinin koşullu olasılıkları hesaplarsak,
P(vites = “otomatik” | A) = 5/9 = 0.556
P(renk = “kırmızı” | A) = 2/9 = 0.222
P(yakıt = “dizel” | A) = 4/9 = 0.444
P(kapı = “4” | A) = 5/9 = 0.556
P(vites = “otomatik” | B) = 4/6 = 0.667
P(renk = “kırmızı” | B) = 2/6 = 0.333
P(yakıt = “dizel” | B) = 1/6 = 0.167
P(kapı = “4” | B) = 6/6 = 1 elde edilir.

Hesaplanan bu olasılık değerleri kullanılarak;

P(X|A) = 0.556*0.222*0.444*0.556 = 0.030
P(X|B) = 0. 667*0. 333*0. 167*1 = 0.037
P(A) P(X|A) = 0.600*0.030 = 0.018
P(B) P(X|B) = 0.400*0.036 = 0.015 sonuçları elde edilir.

Elde edilen sonuca göre X arabası en yüksek olasılık değerine sahip olan A sınıfına aittir [38].

3.2.4. Doğrusal ayırma analizi (DAA)

DAA ilk olarak R.A Fisher tarafından öne sürülmüştür. Fisher iyi ayrılmış sınıflar elde etmek için doğrusal bir sınır bulmayı amaçlamıştır. Bu bağlamda DAA yöntemi, belirli bir veri kümesi içinde sınıf içi ve sınıflar arası varyans oranını maksimum tutarak nesneleri iki sınıfa ayırmaktadır [40,41]. Eş. 3.4’de bu maksimum oran ifadesi görülmektedir.

$$J(A) = \frac{A^t S_B A}{A^t S_w A} \tag{3.4}$$

Eş. 3.3'de hedef fonksiyon olarak J(A) olarak belirlenmiş, S_b ve S_w sınıflar arası ve sınıf içi saçılım matrislerini göstermektedir. Bu saçılım matrisleri Eş. 3.5 ve Eş. 3.6 ifadelerinde görülmektedir.

$$S_b = \sum_c (\mu_c - \bar{X})(\mu_c - \bar{X})^T \tag{3.5}$$

$$S_w = \sum_c \sum_{i \in c} (X_i - \mu_c)(X_i - \mu_c)^T \tag{3.6}$$

Eş 3.4 ve Eş. 3.5'de görülen ifadelerde μ_c değeri her bir sınıfın ortalaması, $\bar{X}$, tüm verileri ortalaması ve c, mevcut sınıf sayısıdır. Eşitliklere göre hedef fonksiyonun maksimum olduğu ve bunun yanında sınıf ortalamalarının maksimum, sınıf varyanslarının minimum değerlerini içeren doğrusal dönüşüm matrisi(A) elde edilmelidir. A matrisinin katsayıları v adında öz-vektör tarafından verilmektedir. Bu v öz-vektörü sıfırdan farklı değerleri olan $S_w^{-1}S_b$ matrisinden elde edilen λ değerine uygundur. Normalize edilen öz-vektör Eş 3.7'de gösterilen ifade ile sağlanır.

$$S_w^{-1}S_b v = \lambda v \tag{3.7}$$

DAA yöntemi direkt olarak bir sınıflandırıcı olarak kullanıldığı gibi daha karmaşık doğrusal olmayan sınıflandırıcılar için önişlem olarak da kullanılmaktadır.

3.2.5. Veri dönüştürme ve indirgeme yöntemleri

Z-score normalizasyonu

Z-score normalizasyon ile öznitelik değerleri ortalama ve standart sapma değerleri hesaplanarak belirli aralıklara çekilirler. Z-score normalizasyonu uygulanacak öznitelik değerlerinin alınan ortalaması Eş. 3.8'de görülmektedir [42].

$$\text{Ortalama(j)} = \frac{1}{n}\sum_{i=1}^{n} \text{öznitelik(j)} \tag{3.8}$$

Özniteliklerin ortalamadan sapmaları hesaplanır. Standart sapma hesabı Eş 3.9'da görülmektedir.

$$\text{standart sapma (j)} = \sqrt{\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\text{öznitelik değeri(j)- ortalama(j)}\right)^2} \quad (3.9)$$

Son olarak önceden hesaplanan ortalama ve standart sapma değerleri ile yeni öznitelik değeri hesaplanır ve Eş 3.10'da bu öznitelik değer hesabı görülmektedir.

$$\text{yeni öznitelik değeri (j)} = \frac{\text{öznitelik değeri(j)- ortalama(j)}}{\text{standart sapma(j)}} \quad (3.10)$$

Öznitelik seçimi

Sınıflandırma algoritmaları uygulanmadan eldeki veri kümesi birçok özniteliğe sahip olabilir. Bu öznitelikler gerekli olabileceği gibi sınıflandırmaya katkı sağlamayacak şekilde de olabilir. İşte bu yüzden öznitelikleri ayırmak ve sınırlamak gerekebilir. Bunun için öznitelik seçim yöntemleri kullanmak gereklidir.

Burada önemli olan özniteliklerin için nasıl bir gruplama yapılacağıdır. Gruplama işi bir ölçüt değeri ile yapılır. Gruplama yöntemi bir istatistik yöntemi olan t-test yöntemidir. T-test yönteminden kısaca bahsetmek gerekirse; hipotez testlerinde yaygın olarak kullanılan bir yöntemdir. T-testi ile iki grubun ortalamaları karşılaştırılarak, aradaki farkın rastlantısal mı, yoksa istatistiksel olarak anlamlı mı olduğuna karar verilir. Küçük örnekleme teorisi olarak da bilinen t dağımı, küçük örneklemlerle de çalışmaya imkân verdiğinden, araştırmacılar için büyük kolaylık sağlamaktadır [43].

3.2.6. Model başarım ölçütleri

Karışıklık matrisi (confusion matrix)

Doğruluk, hata oranı, kesinlik, duyarlılık, özgüllük ve F-ölçütü kavramları model başarım değerlendirmelerinde kullanılmaktadır. Doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atılan örnek sayısı nicelikleriyle model başarımı belirlenmektedir.

Test sonucunda ulaşılan sonuçların başarım bilgileri karışıklık matrisi ile ifade edilebilir. Karışıklık matrisinde satırlar test kümesindeki örneklere ait gerçek sayıları, kolonlar ise modelin doğru sınıflara ait örneklerini ifade etmektedir [44]. Karışıklık matrisi verileri Çizelge 3.3'te görülmektedir.

Çizelge 3.3. Karışıklık matrisi

		Doğru Sınıf	
		Sınıf=1	Sınıf=0
Test Sonucundaki Sınıf	Sınıf=1	TP	FP
	Sınıf=0	FN	TN
TP (True Pozitive) FN (False Negative)		FP (False Pozitive) TN (True Negative)	

Doğruluk–hata oranı

Karışıklık matrisinde kullanılan en popüler ve basit yöntem, modelin doğruluk oranıdır. Doğru sınıflandırılmış örnek sayısının (TP +TN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır ve Eş. 3.11'de görülmektedir. Hata oranı ise bu değerin 1'den çıkarılması ile elde edilir. Diğer bir ifadeyle yanlış sınıflandırılmış örnek sayısının (FP+FN), toplam örnek sayısına (TP+TN+FP+FN) oranıdır ve Eş. 3.12'de görülmektedir.

$$\text{Doğruluk} = \frac{(TP+TN)}{(TP+FP+FN+TN)} \tag{3.11}$$

$$\text{Hata Oranı} = \frac{(FP+FN)}{(TP+FP+FN+TN)} \tag{3.12}$$

Kesinlik (Precision)

Kesinlik, sınıfı 1 olarak tahmin edilen TP örnek sayısının, sınıfı 1 olarak tahmin edilmiş tüm örnek sayısına oranıdır ve Eş 3.13'deki gibi ifade edilmektedir.

$$\text{Kesinlik} = \frac{TP}{TP + FP} \tag{3.13}$$

Duyarlılık (Sensitivity)

Doğru sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranıdır ve Eş 3.14'deki gibi ifade edilmektedir.

$$\text{Duyarlılık} = \frac{TP}{TP + FN} \tag{3.14}$$

Özgüllük (Specificity)

Doğru sınıflandırılmış negatif örnek sayısının toplam negatif örnek sayısına oranıdır ve Eş 3.15'deki gibi ifade edilmektedir.

$$\text{Özgüllük} = \frac{TN}{TN + FP} \tag{3.15}$$

F-Ölçütü (F-Measure)

Elde edilen kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli olmayabilir. Her iki ölçütü beraber değerlendirmek daha doğru sonuçlar verebilir. Bunun için F-ölçütü tanımlanmıştır. F-ölçütü, kesinlik ve duyarlılığın harmonik ortalamasıdır ve Eş 3.16'daki gibi ifade edilmektedir.

$$\text{F-Ölçütü} = \frac{2\times\text{Duyarlılık}\times\text{Kesinlik}}{\text{Duyarlılık} + \text{Kesinlik}} \tag{3.16}$$

ROC eğrisi ve AUC

ROC eğrisi (Receiver Operating Characteristic)

ROC analizi, II. Dünya Savaşı sırasında Britanya'da radarda tespit edilen sinyallerin doğru tanımlanması, dost ve düşman ayrımının sağlanması için geliştirilmiştir. , ROC analizinin tıpta karar vermede kullanımını 1967 yılında Lusted önererek, 1969 yılında medikal görüntüleme cihazlarında kullanımını sağlamıştır. Sonraki yıllarda tıpta tanı testlerinin performansının değerlendirilmesinde kullanımı giderek yaygınlaşmıştır. ROC analizi ile ortaya çıkan gelişmeler, istatistiksel sonuçların değerlendirilmesi ve kıyaslanmasına duyulan gereksinimin doğal bir sonucudur [45,46].

Yapılan testinin kendi doğruluğunu belirlenmesi ve testler arasında güvenilir bir karşılaştırma yapmaya olanak sağlaması açısından ROC eğrisi sıklıkla kullanılmaktadır. Klinik çalışmalarda sürekli sayıların kullanıldığı ölçümlerde olguları ayırma (hasta/sağlam), çözümlemeyi karmaşık hale getirir ve hata olasılığını yükseltir. ROC analizi çeşitli durumlarda optimum eşik değerini ve yapısında var olan değerlendirme dışı bırakılacak olan değerleri (duyarlılık ve özgüllük arasında yer alan) belirler. ROC eğrisinin grafiksel yaklaşımı ölçümlerin duyarlılığı ile özgüllüğü arasındaki ilişkileri kavramayı kolay kılar.

ROC eğrisi, tanı koymak amacıyla kullanılan bir değişkenin değişim genişliği içinde aldığı tüm değerlerin sırasıyla kesim noktası kabul edilmesiyle hesaplanacak duyarlılık değerlerinin, testin yanlış pozitif oranına (1 - özgüllük) karşı noktalanması ile elde edilir. ROC eğrisinin oluşturulacağı koordinat sisteminde, Y ekseninde testin gerçek pozitif değeri (duyarlılık), X ekseninde ise yanlış pozitif değeri (1-özgüllük) yer alır. Her kesim noktasındaki doğru pozitif ve yanlış pozitife karşılık gelen

noktalar birleştirilerek ROC eğrisi çizilir [47]. ROC eğrisinin oluşumu için kullanılan değerler Şekil 3.7'de örnek olarak görülmektedir.

Şekil 3.7. ROC eğrisinin oluşumu için kullanılan değerlerin gösterimi [48]

AUC(Area under the curve)

AUC, testin hastalar ve sağlam bireyleri ayırmadaki doğruluk oranını belirler. AUC büyüklüğü üzerinde çalışılan tanı testinin ayırma yeteneğinin istatistiksel olarak önemini gösterir. Üzerinde çalışılan testin hiç ayırma yeteneği olmadığı durumda ROC eğrisi altındaki alanın beklenen değeri 0.50'dir. Mükemmel bir test ise sıfır yanlış pozitif ve sıfır yanlış negatif ile alanın değeri 1.00 olacaktır. Test, bu iki değerin arasında bir alana sahip olmalıdır.

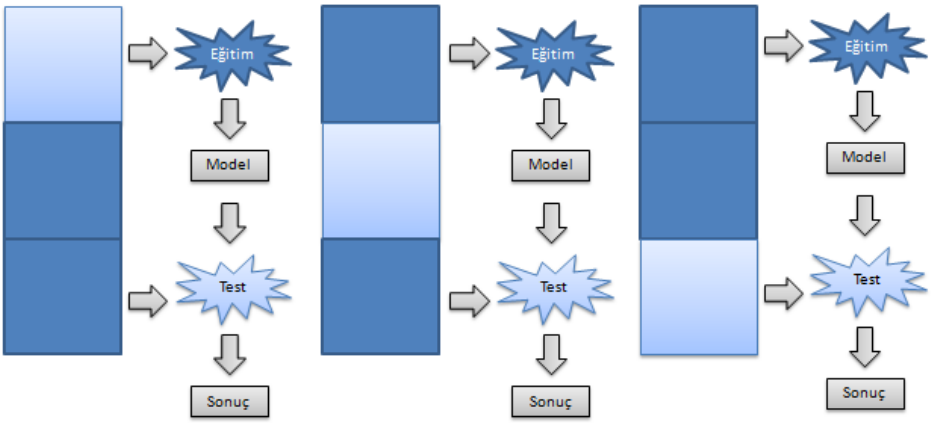
Eğri altındaki alanların yorumlanmasında aşağıda verilen derecelendirmeler kullanılabilir [49,50].

- 0,50-0,60 = başarısız
- 0,60-0,70 = zayıf
- 0,70-0,80 = orta
- 0,80-0,90 = iyi
- 0,90-1,00 = mükemmel

3.2.7. Çapraz doğrulama (Cross Validation)

Çapraz doğrulama, verileri, eğitim ve doğrulama olarak iki bölüme ayıran, öğrenme algoritmalarının kıyaslanmasında ve değerlendirilmesinde kullanılan bir istatistiksel bir metottur. Tipik bir çapraz doğrulama da, eğitim ve değerlendirme kümelerinin her birinin sırayla diğerlerine karşı bir doğrulanma şansı vardır. Çapraz doğrulamanın tipik formu k-kat çapraz doğrulamadır. Diğer doğrulama çeşitleri ya k-kat çapraz doğrulamanın özel bir durumu ya da k-kat çapraz doğrulamayı içeren tekrarlı formlarıdır [51].

K-kat çapraz doğrulamada veri öncelikle k bölüme veya yaklaşık olarak k bölüme ayrılır. Sonrasında k adımda her bir kısım test kümesi, k-1 bölüm ise eğitim kümesi olarak kullanılmaktadır. Şekil 3.8'de k=3 için çapraz doğrulamaya örnek bir modelleme görülmektedir.



Şekil 3.8. Çapraz doğrulama için örnek bir modelleme [51]

Eğer k=10 değeri alırsa 10-kat çapraz doğrulama adını alır ve bu yöntemde en sık kullanılan yöntemdir. Yapılan çalışmaların doğrulanmasında da 10-kat çapraz doğrulama yöntemi kullanılmıştır.

4. DENEYSEL ÇALIŞMALAR

4.1 Kullanılan veri kümeleri ve alınan sonuçlar

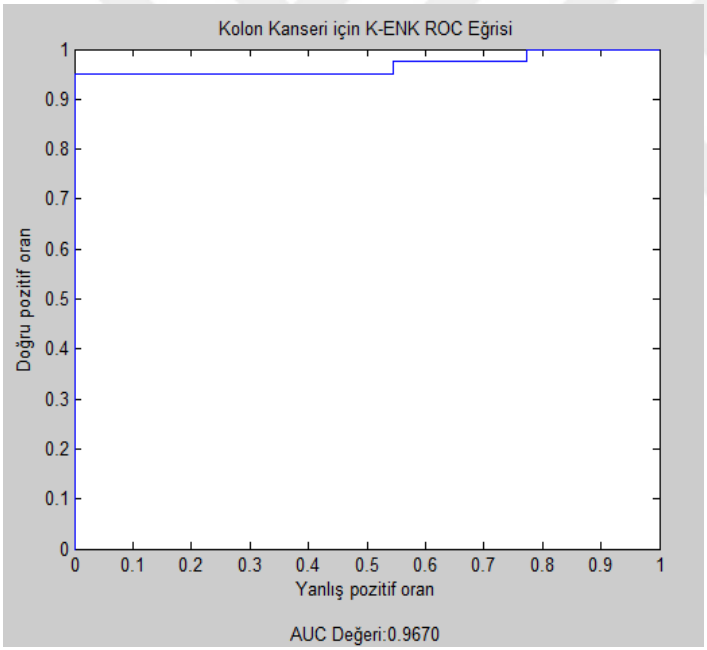
4.1.1 Kanserli hücre geni için mikroarray ifade profilleri

DNA mikroarray teknolojisi, gen ifade profillerinin izlenmesi ile eczacılık ve klinik araştırmalarda önemli rol oynamaktadır. Mikroarray teknolojisi ayrıca kanser olasılıklarının ve diğer hastalıklarının gen ifade seviyesinde sınıflandırılmasına da olanak sağlar. Farklı deneylerle elde edilmiş veri kümeleri üzerinde birçok sınıflandırma metodu uygulanmış ve bu konuda birçok çalışma yapılmıştır. Yapılan çalışmalarda çeşitli kanser hücrelerinden elde edilen mikroarray profilleri incelenmiştir. Bu çalışmada ilk olarak kolon ve lenf kanseri ile lösemi hücreleri ele alınmıştır. Kolon kanseri veri kümesi 62 örnekten ve 2000 genden meydana gelmektedir. Örneklerin 22 tanesi normal, 40 tanesi tümörlü hastadan oluşmaktadır. Lenf kanseri veri kümesi 96 örnek ve 4026 genden meydana gelmektedir. Örneklerin 42 tanesi DLBCL(Diffuse large B-cell lymphoma), 54 tanesi non-DLBCL hastadan oluşmaktadır. Lösemi veri kümesi ise 72 örnekten ve 7129 genden meydana gelmektedir. Örneklerin 47 tanesi ALL hastası, 25 tanesi AML hastasından oluşmaktadır. Sınıflandırma algoritmaları olarak K-ENK(K-En Yakın Komşuluk), DVM(Destek Vektör Makineleri), NB(Naive Bayes) ve DAA (Doğrusal Ayırma Analizi) kullanılmıştır. Kullanılan yöntemlerde sadece DAA için öznitelik seçimi uygulanmış olup, 30 gen seçilmiştir. Sınıflandırma sonucu doğruluk değerleri için karışıklık matrisleri hesaplanmış ayrıca çizdirilen ROC eğrisinin altında kalan AUC değeri hesaplanmıştır.

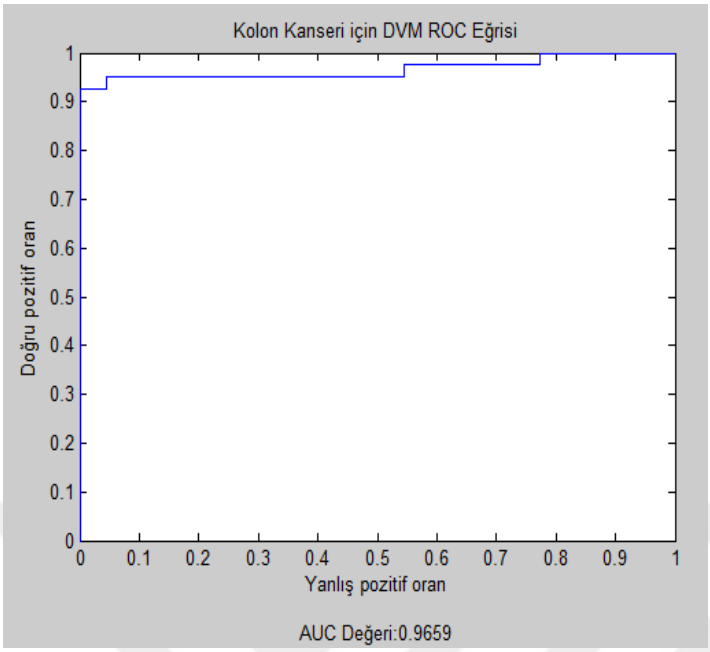
a) Kolon Kanseri veri kümesinden alınan sonuçlar Çizelge 4.1'de, uygulanan sınıflandırma algoritmalarının AUC sonuçları Şekil 4.1, Şekil 4.2, Şekil 4.3, Şekil 4.4'te gösterilmiştir ve doğruluk değerlerinin grafiksel gösterimi ise Şekil 4.5'teki gibi ifade edilmektedir.

Çizelge 4.1. Kolon kanseri veri kümesi için uygulanan sınıflandırma algoritmalarının karışıklık matrisleri

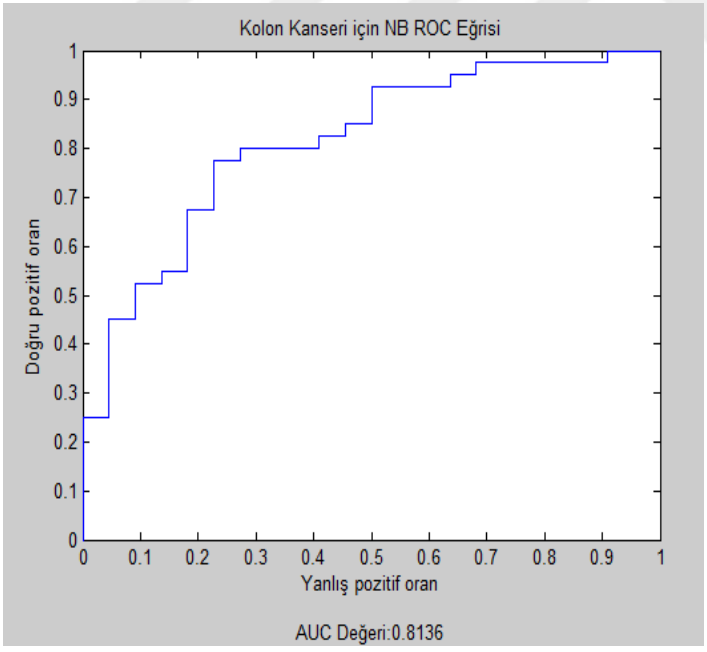
Kolon Kanser Veri Kümesi							
Karışıklık Matrisi K-ENK (k=3)		Karışıklık Matrisi DVM		Karışıklık Matrisi NB		Karışıklık Matrisi DAA (30 gen ile)	
22	2	21	2	18	5	18	7
0	38	1	38	4	35	4	33



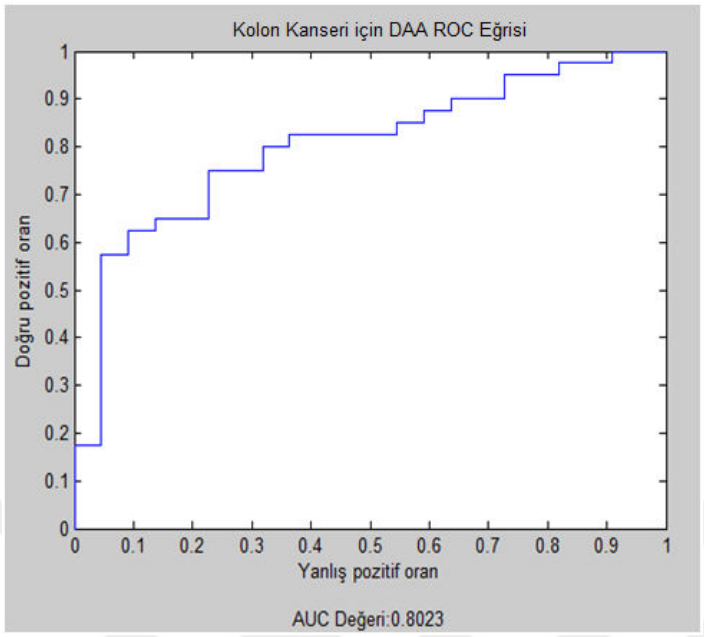
Şekil 4.1. K-ENK için ROC eğrisi ve AUC değeri



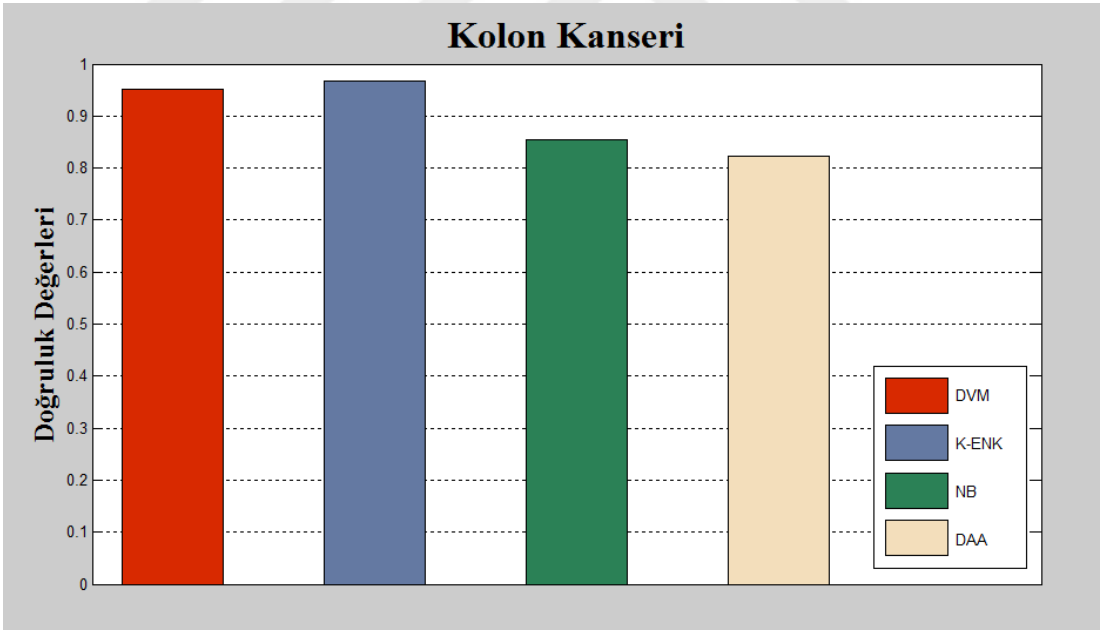
Şekil 4.2. DVM için ROC eğrisi ve AUC değeri



Şekil 4.3. NB için ROC eğrisi ve AUC değeri



Şekil 4.4. DAA için ROC eğrisi ve AUC değeri

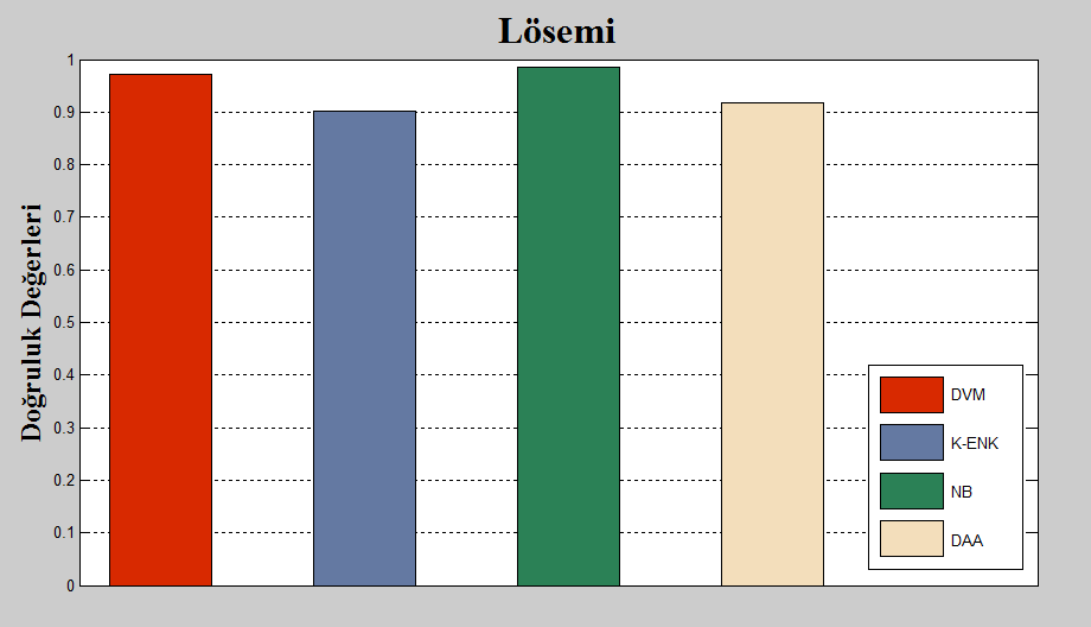


Şekil 4.5. Kolon kanseri veri kümesi için uygulanan sınıflandırma algoritmalarının doğruluk grafiği

b) Lösemi veri kümesinden alınan sonuçların karışıklık matrisleri Çizelge 4.2'de, doğruluk değerlerinin grafiksel ifadesi ise Şekil 4.6'da gösterilmektedir.

Çizelge 4.2. Lösemi veri kümesi için uygulanan sınıflandırma algoritmalarının karışıklık matrisleri

Lösemi veri kümesi							
Karışıklık Matrisi K-ENK (k=1)		Karışıklık Matrisi DVM		Karışıklık Matrisi NB		Karışıklık Matrisi DAA (30 gen ile)	
46	6	46	1	47	1	43	2
1	19	1	24	0	24	4	23

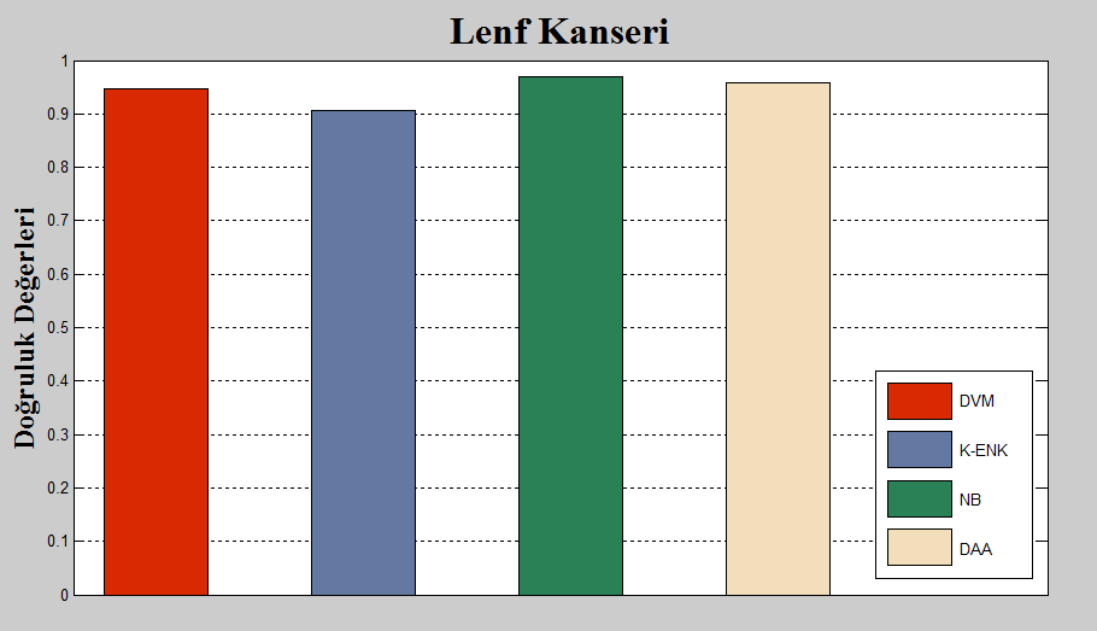


Şekil 4.6 Lösemi veri kümesi için uygulanan sınıflandırma algoritmalarının doğruluk grafiği

c) Lenf Kanseri veri kümesinden alınan sonuçların karışıklık matrisleri Çizelge 4.3'te, doğruluk değerlerinin grafiksel ifadesi ise Şekil 4.7'de gösterilmektedir.

Çizelge 4.3. Lenf kanseri veri kümesi için uygulanan sınıflandırma algoritmalarının karışıklık matrisleri

Lenf kanseri veri kümesi							
Karışıklık Matrisi K-ENK (k=3)		Karışıklık Matrisi DVM		Karışıklık Matrisi NB		Karışıklık Matrisi DAA (30 gen ile)	
46	6	52	3	52	1	51	1
8	41	2	39	2	41	3	41



Şekil 4.7. Lenf kanseri veri kümesi için uygulanan sınıflandırma algoritmalarının doğruluk grafiği

d) Uygulanan sınıflandırma algoritmalarının tüm kanser kümelerindeki doğruluk değerleri Çizelge 4.4'te gösterilmektedir.

Çizelge 4.4. Kolon ve lenf kanseri ile lösemi veri kümesinde uygulanan sınıflandırma algoritmalarının doğruluk değerleri

Uygulanan Sınıflandırma Algoritması	Kolon kanseri	Lösemi	Lenf kanseri
DVM	0,9516	0,9722	0,9479
K-ENK	0,9677	0,9028	0,9063
NB	0,8548	0,9861	0,9688
DAA	0,8226	0,9167	0,9583

4.1.2 WDBC göğüs kanserli hücre gen analizi ve DVM algoritmasının uygulanması

Tez kapsamında yapılan bu ikinci çalışmada WDBC göğüs kanser hücreleri incelenmiştir. WDBC veri kümesi 569 örneği içermekte olup 10 öznitelik ve 2 sınıftan oluşmuştur. Buradaki çalışmanın amacı WDBC kanser hücrelerinin doğru bir şekilde sınıflandırılmasını sağlamak, yüksek duyarlılık ve özgüllük değerlerine

ulaşmaktır. Sınıflandırma algoritması olarak DVM kullanılmış olup. DVM için iki farklı kernel yöntemi olan Gauss ve polinominal kernel yöntemleri uygulanmıştır. Yüksek değerler elde edilmesi için parametre değerleri Gauss kernel için sigma değeri 3, polinominal kernel için derecesi 2 alınmıştır. Karışıklık matrisi; doğruluk, duyarlılık ve özgüllük değerleri Çizelge 4.5'te görülmektedir.

Çizelge 4.5. WDBC veri kümesi için uygulanan DVM algoritmasının karışıklık matrisi ile doğruluk(accuracy), duyarlılık(sensitivity) ve özgüllük(specificity) değerleri

WDBC Veri Kümesi için DVM algoritması			
Karışıklık Matrisi (Gauss sigma = 3)		Karışıklık Matrisi (Polinominal derece = 2)	
191	9	196	13
21	348	16	344
Doğruluk	0,9473	Doğruluk	0,9490
Duyarlılık	0,9009	Duyarlılık	0,9245
Özgüllük	0,9748	Özgüllük	0,9636

4.1.3 WDBC göğüs kanserli hücre gen analizi ve K-ENK algoritmasının uygulanması

Yapılan bu çalışmada da WDBC göğüs kanser hücreleri incelenmiştir. Bu WDBC veri kümesi 699 örneği içermekte olup 9 öznitelik ve 2 sınıftan oluşmuştur. Çalışmada amaçlanan K-ENK algoritmasını Öklid mesafesi kullanarak, WDBC kanser hücrelerinin doğru bir şekilde sınıflandırılmasını sağlamaktır. Çizelge 4.6'da hatalı sınıflandırılan örnek sayıları görülmektedir.

Çizelge 4.6. Sınıflandırma sonucu hatalı sınıflandırılan örnek sayıları

Metot	Komşu sayısı	Hatalı Sınıflandırılan Örnek Sayısı
K-ENK	k = 1	4
	k = 2	9
	k = 3	13
	k = 4	18
	k = 5	26

4.1.4 Mikroarray verileri ile kolon kanseri veri kümesinin sınıflandırılması

Bu çalışmada ise kolon kanser hücreleri incelenmiştir. Daha önceki veri kümesi ile aynı olan kolon kanser hücreleri 62 örneği içermekte olup 2000 öznitelik ve 2 sınıftan oluşmuştur. Çalışmanın amacı farklı sınıflandırma algoritmaları kullanarak kolon kanser hücrelerinin doğru bir şekilde sınıflandırılmasını sağlamaktır. Algoritmalar uygulanmadan önce öznitelik seçimi yapılmıştır. T-test yöntemi ile öznitelik seçimi uygulanan veri kümesi, sırasıyla 5, 10, 20 öznitelik seçilerek işlem yapılmıştır. Uygulanan sınıflandırma algoritmalarının karışıklık matrisi sırasıyla Çizelge 4.7, Çizelge 4.8, Çizelge 4.9, Çizelge 4.10 ve doğruluk değerleri Çizelge 4.11'de görülmektedir.

Çizelge 4.7. Kolon kanseri veri kümesi için uygulanan K-ENK algoritmasının karışıklık matrisi

Kolon Kanseri Veri Kümesi için K-ENK algoritması					
Karışıklık Matrisi (5 Gen ile) k=3 için		Karışıklık Matrisi (10 Gen ile) k=3 için		Karışıklık Matrisi (20 Gen ile) k=3 için	
22	3	22	3	22	2
0	37	0	37	0	38

Çizelge 4.8. Kolon kanseri veri kümesi için uygulanan DVM algoritmasının karışıklık matrisi

Kolon Kanseri Veri Kümesi için DVM(lineer) algoritması					
Karışıklık Matrisi (5 Gen ile)		Karışıklık Matrisi (10 Gen ile)		Karışıklık Matrisi (20 Gen ile)	
21	1	22	1	22	1
1	39	0	39	0	39

Çizelge 4.9. Kolon kanseri veri kümesi için uygulanan NB algoritmasının karışıklık matrisi

Kolon Kanseri Veri Kümesi için NB algoritması					
Karışıklık Matrisi (5 Gen ile)		Karışıklık Matrisi (10 Gen ile)		Karışıklık Matrisi (20 Gen ile)	
21	3	22	2	22	2
1	37	0	38	0	38

Çizelge 4.10. Kolon kanseri veri kümesi için uygulanan DAA algoritmasının karışıklık matrisi

Kolon Kanseri Veri Kümesi için DAA algoritması					
Karışıklık Matrisi (5 Gen ile)		Karışıklık Matrisi (10 Gen ile)		Karışıklık Matrisi (20 Gen ile)	
22	3	21	3	22	2
0	37	1	37	1	38

Çizelge 4.11. Kolon kanseri veri kümesinde uygulanan sınıflandırma algoritmalarının doğruluk değerleri

Yöntem	5 Gen ile	10 Gen ile	20 Gen ile
K-ENK	0,9516	0,9516	0,9677
DVM (lineer)	0,9677	0,9839	0,9839
NB	0,9355	0,9516	0,9516
DAA	0,9516	0,9355	0,9516

4.1.5 PIMA diyabet hastaları ile BUPA karaciğer hastalarının sınıflandırılması

Son çalışmada; PIMA (Pima Indians) diyabet hastaları ile BUPA (British United Provident Association) karaciğer hastalarından alınan hücrelere ait veriler incelenmiştir. PIMA veri kümesi diyabet hastalarının hücrelerinden alınan verilerden oluşmakta olup 345 örneği içermekte 8 öznitelik ve 2 sınıftan oluşmaktadır. BUPA karaciğer veri kümesi ise 768 örnek 6 öznitelik ve 2 sınıftan oluşmaktadır. Çalışmanın amacı farklı kernel yapısı ile DVM algoritmasını kullanarak PIMA ve BUPA veri kümelerinin doğru bir şekilde sınıflandırılmasını sağlamaktır.

Algoritmalar uygulanmadan önce öznitelik seçimine tutulan veriler T-test yöntemi ile sırasıyla 5, 10, 20, 30 eğitim kümelerine ayrılmıştır. Her seferinde test edilen örnek sayısı 200 örnektir. BUPA ve PIMA veri kümeleri için uygulanan DVM algoritmasının karışıklık matrisleri ile doğruluk değerleri sırasıyla Çizelge 4.12, Çizelge 4.13, Çizelge 4.14 ve Çizelge 4.15'te görülmektedir.

Çizelge 4.12. BUPA veri kümesinde Gauss kernel ile uygulanan DVM algoritmasının karışıklık matrisi ve doğruluk değerleri

<table>
<tr><td colspan="8">Karışıklık Matrisi
DVM-Gauss-BUPA</td></tr>
<tr><td>4</td><td>86</td><td>42</td><td>50</td><td>34</td><td>61</td><td>37</td><td>53</td></tr>
<tr><td>5</td><td>105</td><td>38</td><td>70</td><td>25</td><td>80</td><td>27</td><td>83</td></tr>
<tr><td colspan="2">Doğruluk değeri:
0.55</td><td colspan="2">Doğruluk değeri:
0.56</td><td colspan="2">Doğruluk değeri:
0.57</td><td colspan="2">Doğruluk değeri:
0.60</td></tr>
<tr><td colspan="2">Test Verisi:
200 örnek
Eğitim Verisi:
5 Örnek</td><td colspan="2">Test Verisi:
200 örnek
Eğitim Verisi:
10 örnek</td><td colspan="2">Test Verisi:
200 örnek
Eğitim Verisi:
20 örnek</td><td colspan="2">Test Verisi:
200 örnek
Eğitim Verisi:
30 örnek</td></tr>
</table>

Çizelge 4.13. BUPA veri kümesinde polinominal kernel ile uygulanan DVM algoritmasının karışıklık matrisi ve doğruluk değerleri

<table>
<tr><td colspan="8">Karışıklık Matrisi
DVM-Polinominal-BUPA</td></tr>
<tr><td>58</td><td>32</td><td>38</td><td>54</td><td>68</td><td>27</td><td>60</td><td>30</td></tr>
<tr><td>57</td><td>53</td><td>29</td><td>79</td><td>54</td><td>51</td><td>48</td><td>62</td></tr>
<tr><td colspan="2">Doğruluk değeri:
0.56</td><td colspan="2">Doğruluk değeri:
0.59</td><td colspan="2">Doğruluk değeri:
0.60</td><td colspan="2">Doğruluk değeri:
0.61</td></tr>
<tr><td colspan="2">Test Verisi:
200 örnek
Eğitim Verisi:
5 Örnek</td><td colspan="2">Test Verisi:
200 örnek
Eğitim Verisi:
10 Örnek</td><td colspan="2">Test Verisi:
200 örnek
Eğitim Verisi:
20 Örnek</td><td colspan="2">Test Verisi:
200 örnek
Eğitim Verisi:
30 Örnek</td></tr>
</table>

Çizelge 4.14. PIMA veri kümesinde Gauss kernel ile uygulanan DVM algoritmasının karışıklık matrisi ve doğruluk değerleri

Karışıklık Matrisi DVM-Gauss-PIMA							
124	3	125	1	117	7	116	8
73	0	73	1	62	14	62	14
Doğruluk değeri: 0.62		Doğruluk değeri: 0.63		Doğruluk değeri: 0.66		Doğruluk değeri: 0.65	
Test Verisi: 200 örnek Eğitim Verisi: 5 Örnek		Test Verisi: 200 örnek Eğitim Verisi: 10 Örnek		Test Verisi: 200 örnek Eğitim Verisi: 20 Örnek		Test Verisi: 200 örnek Eğitim Verisi: 30 Örnek	

Çizelge 4.15. PIMA veri kümesinde polinominal kernel ile uygulanan DVM algoritmasının karışıklık matrisi ve doğruluk değerleri

Karışıklık Matrisi DVM-Polinominal-PIMA							
127	0	112	14	94	30	96	28
71	2	56	18	39	37	33	43
Doğruluk değeri: 0.65		Doğruluk değeri: 0.65		Doğruluk değeri: 0.66		Doğruluk değeri: 0.70	
Test Verisi: 200 örnek Eğitim Verisi: 5 Örnek		Test Verisi: 200 örnek Eğitim Verisi: 10 Örnek		Test Verisi: 200 örnek Eğitim Verisi: 20 Örnek		Test Verisi: 200 örnek Eğitim Verisi: 30 Örnek	

5. SONUÇ VE DEĞERLENDİRME

Bu tez kapsamında yapılan çalışmalarda ilk olarak kolon ve lenf kanseri ile lösemi hücrelerinin mikroarray ifadeleri incelenmiştir. Bu verilere DVM, K-ENK, NB ve DAA sınıflandırma yöntemleri uygulanmıştır. İlk olarak kolon kanser hücrelerinde yapılan çalışmada en yüksek doğruluk değerini %96,77 ile K-ENK algoritması vermiştir. K-ENK algoritmasını sırasıyla %95,16 ile DVM, %85,48 ile NB ve %82,26 ile DAA algoritmaları takip etmiştir. K-ENK ve DVM algoritmaları yakın ve yüksek doğruluk göstermiş, NB ve DAA ise kısmen düşük doğruluk değerlerinde kalmıştır. Lenf kanseri hücrelerinde ise en yüksek doğruluk değerine %96,88 ile NB algoritması ulaşmıştır. Diğer doğruluk değerleri sırasıyla DAA için %95,83, DVM için %94,79 K-ENK için ve %90,63'tür. Lösemi hücrelerinde %98,61 ile NB en yüksek doğruluk değerine ulaşırken, DVM %97,22, DAA %91,67, K-ENK %90,28 değerleri ile NB yöntemini takip etmiştir. Bu üç veri kümesine göre elde edilen sonuçlar için DAA yöntemi dışında herhangi birinde öznitelik seçim yöntemi uygulanmamıştır. Öznitelik sayısının az olduğu kolon kanserinde doğruluk değerleri düşük olan NB ve DAA'nin, öznitelik sayısının arttığı lösemi hücrelerinde doğruluk değerleri önemli oranda artış kaydedildiği açıkça görülebilir.

Bir diğer çalışmada farklı kerneller ile DVM algoritması 569 örneğe sahip WDBC veri kümesinde uygulanmıştır. Sınıflandırma sonucunda alınan DVM-Gauss doğruluk değeri %94,73, DVM-Polinominal doğruluk değeri %94,90 olarak bulunmuştur. DVM-Gauss için farklı sigma ve DVM-Polinominal için farklı polinom dereceleri çalışmada denenmiş olup en yüksek değerler sigma değeri için 3, polinom derecesi için 2 olarak belirlenmiştir. Duyarlılık ve özgüllük değerleri ise DVM-Gauss için %90,09 ve %97,48 DVM-Polinominal için ise %92,45 ve %96,36 olarak bulunmuştur. Duyarlılık ve özgüllük arasındaki farkların yanlış sınıflandırılan pozitif ve negatif örneklerden kaynaklandığını açıkça ifade edilebilir.

WDBC göğüs kanser hücrelerini kapsayan başka bir çalışmada sadece K-ENK algoritması uygulanmış olup yakın komşuluk için Öklid mesafesi dikkate alınmıştır. Çalışma sonucunda yanlış sınıflandırılan örnek sayıları k=1 için 4, k=2 için 9, k=3 için 13, k=4 için 18, k=5 için 26 olarak bulunmuştur. Öznitelik gen değerlerinin birbirine çok yakın olmasından dolayı ve K değerinin artması ile yanlış sınıflandırılan örnek sayısının arttığını söyleyebiliriz.

Kolon kanser hücrelerine uygulanan diğer bir çalışmada öznitelik seçim yöntemi uygulanmıştır. Öznitelik seçim yöntemleri ile gen sayıları 5, 10, 20 gen sayısı seçilmiştir. 5 öznitelik değeri ile DVM %96,77,K-ENK %95,16, NB %93,55, DAA %95,16 doğruluk değerine ulaşmıştır. 20 gen seçimi yapıldığı zaman doğruluk değerleri DVM için %98,39 K-ENK için %96,77, NB için %95,16 ve DAA için %95,16 olmuştur. Öznitelik seçimi ile ilgili genlerin sayısının artması doğruluk değerinin artmasını sağlamaktadır.

Son olarak BUPA ve PIMA veri kümeleri için DVM algoritması uygulanmıştır. DVM algoritması için Gauss ve polinominal kerneller kullanılmıştır. Test örnek sayısı her seferinde 200 örnek olarak belirlenmiş, eğitim kümeleri ise sırasıyla 5, 10, 20, 30 örnek olarak seçilmiştir. Eğitim kümesinin en az olduğu 5 örnek ile PIMA veri kümesinde DVM-Gauss için %62, DVM-Polinominal için %65 doğruluk değerleri elde edilmiştir. 20 eğitim kümesinin seçildiği uygulamada ise DVM-Gauss %65, DVM-Polinominal ise %70 doğru sonucu vermiştir. Eğitim kümesinin daha fazla seçimi doğruluk değerlerini arttırmış, az eğitim kümesi olmasının sınıflandırmanın öğrenme sürecini etkileyerek daha düşük doğruluk değeri vermesi sonucunu ortaya çıkarmıştır.

Sonuç olarak daha öncede ifade edildiği gibi veri kümelerinin önişleme tabi tutulması, sınıflandırma algoritmalarının ve kullandıkları parametre seçimi, ilgili özniteliklerin sayısının belirlenmesi gibi etkenler sınıflandırmanın doğruluk değerlerinde belirleyici unsurlar olmaktadırlar. Doğru seçimlerin yapılması biyoinformatik alanında daha iyi sonuçlar vereceği gibi ileriki dönemler için hastalık tespitinde klinik çalışmalara da ışık tutacaktır.

KAYNAKLAR

1. Polat, M., Karahan, A., "Multidisipliner yeni bir bilim dalı:biyoinformatik ve tıpta uygulamaları", ***S.D.Ü Tıp Fak. Derg.***,16(3): 41-50 (2009).

2. Karabulut, E., Karaağaoğlu E., "Biyoinformatik ve biyoistatistik", ***Hacettepe Tıp Dergisi***, 41:162-170 (2010).

3. İnternet: " Koç Üniversitesi – Kule" , http://home.ku.edu.tr/~ccbb/news/Bioinformatics_Kule.pdf (Kasım 2010).

4. Feagen L, Rohrer J, Garret A, Amthauer H, Komp E, Johnson D, Hock A, "Bioinformatics process management:information flow via a computional journal", ***Source Code for Biology and Medicine***, 2(9): 2-3 (2007).

5. İnternet: "Biyoinformatik", http://www.istanbul.edu.tr/fen/notlar/1260103637.ppt, (Mayıs 2012).

6. T Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S., "Bioconductor: open software development for computational biology and bioinformatics" ***Genome Biology***, 5(5), 80 (2004).

7. Gastgeiger E, Gattiker A, Hoogland C, Ivanyi I, Apel RD, Bairoch A. "ExPASy: The proteomics server for n-depth protein knowledge and analysis.", ***Nucleic Acids Res.***, 31:3784-3788 (2003).

8. İnternet: "GenBank Statistics (ca.2008)", http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/, (Mayıs 2012).

9. İnternet: "Growth of GenBank", ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt, (Haziran 2012).

10. Gatto ML, "The changing face of Bioinformatics", ***Drug Discov. Today***, 8: 375-376 (2003).

11. Doerry E, Douglas SA, Kirkpatrick AE and Westerfield M., "Participatory Design for Widely-Distributed Scientific Communities", ***Proceedings of the 3rd Conference on Human Factors and the Web,***USA (1997).

12. Collins FS, Morgan M And Patrinos A., "The Human Genome Project: Lessons from Large-Scale Biyology", ***Science,*** 300,286-290 (2003).

13. İnternet: “Gen İfadesi, http://tr.wikipedia.org/wiki/Gen_ifadesi”, (Haziran 2012).

14. Hogue CWV.,” Bioinformatics for Beginners.”, ***Trends Biochem. Sci.*** 27, 542-543 (2007).

15. Brooksbank C, Camon E, Midori AH, Magrane M, Martin MJ, Mulder N, O.Donovan C, Parkinson H, Tuli MA, Apweiler R, Birney E, Brazma A, Henrick K, Lopez R, Stoesser G, Stoehr P. and Cameron G, “The European Bioinformatics Institute’s Data Resources”***Nucleic Acids Research***, 31(1): 43-50 (2003).

16. Critchlow T, Musick R, and Slezak T., “Experiences applying meta-data bioinformatics”, ***Information Sciences***, 139:13-17 (2001).

17. İnternet: “National Center for Biotechnology Information”, http://www.ncbi.nlm.nih.gov/, (Mart 2012).

18. İnternet: “Nucleotide BLAST: Search nucleotide databases using a nucleotide query”, http://blast.ncbi.nlm.nih.gov/Blast.cgi, (Mart 2012).

19. Saraçlı, M.A, “DNA chip teknolojisi ve mikolojide uygulama alanları, ***İnfeksiyon Dergisi***, 21:181-184, (2007).

20. İnternet: “Mikraarray Yöntemi ve Virolojideki Uygulamaları”, http://www.belgeler.com/blg/2bjz/mikroarray-yntemi-ve-virolojideki-uygulamalari, (Mayıs 2012).

21. İnternet: “Veritabanı, Veri Madenciliği, Veri Ambarı, Veri Pazarı”, http://mail.baskent.edu.tr/~20394676/0302/bil483/HW2.pdf, (Kasım 2011).

22. Özkan, Y.,”Veri Madenciliği Yöntemleri 1. Basım”,Çölkesen R., Uğurkaya, C., ***Papatya Yayıncılık***, İstanbul, 37-44 (2008).

23. Tran, D.H., Ho, T.B., Pham, T.H., Satou, K., “MicroRNA Expression Profiles forClassification and Analysis of Tumor Samples”, ***IEICE Trans. Inf&Syst.***, 94:3, (2011).

24. Lu, J., Getz, G., Miska, A.E., Alvarez-Saavedra, J.,Lamb, J., Peck, D., Sweet-Cardero, A., Ebert, L.B., Mak, H.R., Fernando, A.A., Jacks, D.T., Horvitz, H.R., Golub, R.T., “MicroRNA expression profiles classify human cancers”, ***Nature***, 435:834-838 (2005).

25. Huang, J., Fang, H., Fan, X., “Decision forest for classification of gene ezpression data”, ***Computers in Biology and Medicine***, 40:698-704 (2010).

26. Lobenhofer, E., Auman, J., Blackshear, P., Boorman, G., Bushel, P., Cunningham, M., Fostel, J., Gerrish, K., Heinloth, A., Irwin, R.,"Gene expression response in target organ and whole blood varies as a function of target organ injury phenotype", ***Genome Biology***, 9 (6):100 (2008).

27. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", ***Proceedings of the National Academy of Sciences***, 96 (12) :6745–6750 (1999).

28. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M.,"Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", ***Science***, 286 (5439):531 (1999).

29. Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling",***Nature***, 403: 503–511 (2000).

30. Chen, H., Yang, B., Liu, J., Liu, D., "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis", ***Expert Systems with Applications***, 38:9014-9022 (2011).

31. İnternet: "Uci Machine Learning Reporsitory", http://archive.ics.uci.edu/ml/datasets/, (Nisan 2012)

32. Krishnan, M., Banerjee, S., Chakraborty, C., Chakraborty, C., Ajoy, K., "Statistical analysis of mammographic features and its classification using support vector machine", ***Expert Systems with Applications***, 37:470-478 (2010).

33. Aci, M., Avci, M., "K nearest neighbor reinforced expectation maximization method", ***Expert Systems with Applications***, 38:12585-12591 (2011).

34. Han, B., Li, L., Chen, Y., Zhu, L., Dai, Q., "A two step method to identify clinical outcome relevant genes with microarray data", ***Journal of Biomedical Informatics***, 44:229-238 (2011).

35. Li, D., Liu, C., "A class possibility based kernel to increase classification accuracy for small data sets using support vector machines, ***Expert Systems with Applications***, 37:3104-3110 (2010).

36. Wang, X., Gotoh, O., "A robust Gene Selection Method for Microarray-based Cancer Classification", ***Cancer Informatics***, 9: 15-30 (2010).

37. İnternet:"Statistical Classification", http://en.wikipedia.org/wiki/Statistical_classification, (Mayıs 2012).

38. Doğan, S.,"Türkçe dokümanlar için n-gram tabanlı sınıflandırma:Yazar, tür, cinsiyet", Yüksek Lisans Tezi, ***Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü***, İstanbul, 22-30 (2006).

39. İnternet : "Destek Vektör Makinesi(SVM)", http://www.yapay-zeka.org/modules/wiwimod/index.php?page=SVM, (Kasım 2011).

40. İnternet: "Fisher Linear Discriminant Analysis", http://www.ics.uci.edu/~welling/classnotes/papers_class/Fisher-LDA.pdf, (Mayıs 2012).

41. Kavas, C.K., Nasibov, E., "Classification of apoptosis proteins by discriminant analysis", Turkish Journal of Biochemistry, 37(1): 54-61 (2012).

42. Çalışkan, S.K.,"K'KKN:Kümeleme ve k en yakın komşu yöntemi ile ağlarda nüfuz tespiti", Yüksek Lisans Tezi, ***Gebze İleri teknoloji Enstitüsü Mühendislik ve Fen Bilimleri Enstitüsü***, Gebze, 34 (2008).

43. İnternet: "SPSS'de İstatistiksel Analizler", http://iys.inonu.edu.tr/webpanel/dosyalar/669/file/SPSS%20testleri.doc, (Mayıs 2012)

44. Baykal, A., Coşkun, C., "Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması", ***İnönü Üniv. Akademik Bilişim'11***, Malatya, 67 (2011).

45. Obuchowski, N.A.,"Receiver Operating Characteristic Curves and Their Use in Radiology" ,***Radiology***, 229 : 3-8 (2003).

46. Park, S.H., Goo, J.M., Jo, C.H.,"Receiver Operating Characteristic(ROC) Curve: Practical Review for Radiologists", ***Korean J Radiol***, 5:11-18 (2004).

47. İnternet:"Roc (Receiver Operating Characteristic) Eğrisi Yöntemi ile Tanı Testlerinin Performanslarının Değerlendirilmesi" http://tip.baskent.edu.tr/egitim/mezuniyetoncesi/calismagrp/ogrsmpzsnm12/10.2.pdf" (Mayıs 2012).

48. İnternet: "Infinity Teknoloji:ROC Eğrisi", http://infinityteknoloji2050.blogspot.com/2009/11/roc-egrisi.html, (Mayıs 2012).

49. Dirican, A.,"Evaluation of the diagnostic test's performance and their comparisons", ***Cerrahpaşa J Med***, 32:25-30 (2001).

50. Kanık, E.A., Erden, S.,"Tanı Testlerinin değerlendirilmesinde ROC (Receive Operating Characteristics) Eğrisinin Kullanımı", ***Mersin Üniversitesi Tıp Fakültesi Dergisi***, 3:260-264 (2003).

51. Refaeilzadeh, P., Tang, L.,"Cross-Validation", Encyclopedia of database systems, Özsu, M.T., Liu, L.(Editors), ***Springer***, New York, 532-538 (2009).

52. Wang, J..T.L,Zaki, M.J., Toivonen, H.T.T., Shasha, D., "Data Mining in Bioinformatics",***Springer***, USA, 9-24 (2005).

53. Keedwell, E., Narayanan, A., "Intelligent Bioinformatics", ***Wiley***, Cornwall, 103-132 (2005).

54. Claverie, J.M., Notredame, C., "Bioinformatics for Dummies 2nd Edition", ***Wiley***, New Jersey, 9-28 (2007).

55. Hu, H., Li, J., Plank, A., Wang, H., Daggard, G., "A Comparative Study of Classification Methods for Microarray Data Anlaysis", ***Proc.Fifth Australasian Data Mining Conference***, Toowoomba,33-37 (2006).

56. Castano, A., Fernandez-Navarro, F., Hervas-Matinez, C., Gutierrez, P.A., "Neuro-logistic Models Based on Evolutionary Generalized Radial Basis Function for the Microarray Gene Expression Classification Problem", ***Neural Process Lett***, 34:117-131 (2011).

57. Jung, K., Grade, M., Gaedcke, J., Jo, P., Opitz, L., Becker, H., Ghadimi, M., Beijbbarth, T., "A new sensitivity-preferred strategy to build prediction rules for therapy response of cancer patients using gene expression data", ***Computer Methods and Programs in Biomedicine***, 100:132-139 (2010).

58. Lorena, A.C, Costa, I.G., Spolaor, N., de Souto, M.C.P., "Analysis of complexity indices for classification problems:Cancer gene expression data", ***Neurocomputing***, 75: 33-42 (2012).

59. Obulkasim, A., Meijer, G.A., Van de Wiel, M.A., "Stepwise classification of cancer samples using clinical and molecular data", ***BMC Bioinformatics***, 12:422 (2011).

60. Lee, C., Mandoiu, I., Nelson, C.E., "Inferrring ethnicity from mitochondrial DNA sequence", ***BMC Proceedings***, 5:11 (2011).

ÖZGEÇMİŞ

Kişisel Bilgiler

Soyadı, adı : AKIN, Murat
Uyruğu : T.C.
Doğum tarihi ve yeri : 02.04.1982, Nazilli
Medeni hali : Bekar
Telefon : 0 (312) 838 68 00
Cep Tel : 0 505 664 38 10
E-mail : muratakin@gazi.edu.tr

Eğitim

Derece	Eğitim Birimi	Mezuniyet Tarihi
Yüksek lisans	Gazi Üniversitesi /Fen Bilimleri Enstitüsü/Bilgisayar Mühendisliği	-
Lisans	Kocaeli Üniversitesi/ Mühendislik Fakültesi/Bilgisayar Mühendisliği	2005
Lise	Nazilli Anadolu Lisesi	2000

İş Deneyimi

Yıl	Yer	Görev
2010	Gazi Üniversitesi ATAMYO	Öğretim Görevlisi
2007-2009	Erikoğlu Holding	Bilgi İşlem Uzmanı

Yabancı Dil

İngilizce

Hobiler

Futbol, yüzme, masa tenisi