T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

MİKROARRAY GEN EKSPRESYON VERİ SETLERİNDE RANDOM FOREST VE NAIVE BAYES SINIFLAMA YÖNTEMLERİ YAKLAŞIMI

Ebru KORKEM

Biyoistatistik Programı
YÜKSEK LİSANS TEZİ

ANKARA
2013

T.C.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

MİKROARRAY GEN EKSPRESYON VERİ SETLERİNDE RANDOM FOREST VE NAIVE BAYES SINIFLAMA YÖNTEMLERİ YAKLAŞIMI

Ebru KORKEM

Biyoistatistik Programı
YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI
Doç. Dr. Erdem KARABULUT

ANKARA
2013

ONAY SAYFASI

Anabilim Dalı	**:Biyoistatistik**
Program	**:Biyoistatistik**
Tez Başlığı	**:Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naive Bayes Sınıflama Yöntemleri Yaklaşımı**
Öğrenci Adı-Soyadı	**:Ebru Korkem**
Savunma Sınavı Tarihi	**:11.02.2013**

Bu çalışma jürimiz tarafından yüksek lisans/doktora tezi olarak kabul edilmiştir.

Jüri Başkanı:	**Prof. Dr. Osman Saraçbaşı** (Hacettepe Üniversitesi Tıp Fakültesi)
Tez danışmanı:	**Doç. Dr. Erdem Karabulut** (Hacettepe Üniversitesi Tıp Fakültesi))
Üye:	**Prof. Dr. C. Reha Alpar** (Hacettepe Üniversitesi Tıp Fakültesi))
Üye:	**Doç. Dr. S. Kenan Köse** (Ankara Üniversitesi Tıp Fakültesi))
Üye:	**Doç. Dr. Pınar Özdemir** (Hacettepe Üniversitesi Tıp Fakültesi))

ONAY

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun görülmüş ve Sağlık Bilimleri Enstitüsü Yönetim Kurulu kararıyla kabul edilmiştir.

Prof.Dr. Ersin FADILLIOĞLU

Müdür

TEŞEKKÜR

Yüksek lisans eğitimim süresince bilgi paylaşımlarıyla bana ışık tutan ve üzerimde emekleri çok olan tüm Biyoistatistik Anabilim Dalı ailesine, özellikle bu süreçte her konuda bana destek veren, değerli fikirleri ile beni yönlendiren, motivasyonum düştüğü anlarda beni tekrar motive eden danışman hocam Sayın Doç. Dr. Erdem KARABULUT'a, tez konusu belirleme sürecinde beni yönlendirip fikir veren Sayın Arş. Gör. Erdal COŞGUN'a, hiçbir zaman sıkılmadan sorduğum her soruya cevap veren bu çalışmada da desteğini hep hissettiğim sevgili arkadaşım Araş. Gör. Eda KARAİSMAİLOĞLU'na teşekkür ediyorum.

Bu süreçte bana destek olan ÖZCAN ailesine ve özellikle bıkmadan usanmadan ilgisini ve manevi desteğini her zaman olduğu gibi benden esirgemeyen Sayın Furkan ÖZCAN'a,

Her zaman benimle üzüntümle üzülüp benim mutluluğumla mutlu olan, bana varlığıyla hayata karşı güven hissi veren canım ablam Duygu KORKEM'e, tüm nazımı usanmadan çeken beni sevgisinden hiçbir zaman mahrum etmeyen ve çok sevdiğim kıymetli annem Güleser KORKEM'e ve tüm hayatını çocuklarına adayan her zaman bana doğru yolu gösteren, okumanın önemini bana aşılayan, bir çok fedakarlık yapıp bundan her zaman mutluluk duyan, sevgisini her zaman hissettiren, hasretle sevdiğim biricik Babam'a emekleri için çok teşekkür ederim.

ÖZET

Ebru KORKEM Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naive Bayes Sınıflama Yöntemleri Yaklaşımı. Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Programı Yüksek Lisans Tezi, ANKARA, 2012. Bu tez çalışmasında çeşitli kanser türlerine ilişkin farklı büyüklükteki ve farklı sınıf sayılarındaki veri setleri üzerinde, Random Forest ve Naive Bayes sınıflama yöntemleri karşılaştırılmıştır. İki sınıflama yönteminin de her bir veri seti için ayrı ayrı doğru sınıflama oranları, iki sınıflı veri setlerinde (hasta, sağlam) yöntemlerin başarısını karşılaştırmak için ise duyarlılık, seçicilik, kesinlik, F-ölçütü gibi performans ölçüleri hesaplanmıştır. Ayrıca Random Forest yönteminde her veri seti için önemli genler "doğru sınıflama oranındaki ortalama azalma" değerlerine göre hesaplanmış ve ilk 10 önemli genin gerçekte de sınıflar arasında farklılık gösterip gösterilmediğine bakılmıştır. Random Forest yönteminde ağaçlardaki en uygun değişken sayısına karar vermek için 5 farklı m_{try} değeri üzerinde çalışılmış (0.5P, 0.1P, P, 1.2 $\sqrt{P}$, $\sqrt{P}$) ve en uygun m_{try} değerine göre sınıflama yapılmıştır. Random Forest ve Naive Bayes'in kanser veri setleri üzerindeki performansları bu değerlerle karşılaştırılmıştır. Her iki sınıflama yönteminin hem ikili hem de çoklu sınıflama da performansları incelenmiş olup sonuçlarda, iki yöntem arasında fark gözlenmemiştir.

Anahtar Kelimeler: Random Forest, Naive Bayes, DNA Mikroarray, Sınıflama

ABSTRACT

Ebru KORKEM Random Forest and Naïve Bayes Approach in Microarray Gene Expressions Data Sets. Hacettepe University Institute of Health Sciences, Biostatistic Program Master Degree Thesis, ANKARA, 2012. In this study, two classification methods, Random Forest and Naive Bayes, were compared about various types of cancer on the data sets which are different in size or in number of classes. Correct classification rates of two classification methods for each data set were separately calculated and additionally in two-class data sets (patient-healthy), performance measures like sensitivity, specificity, precision and F-measure were calculated for success comparison of the methods. Also in Random Forest method, for each data set important genes were calculated according to the values of "mean decrease in accuracy" and it was investigated if in reality first 10 important genes were differed between the classes. In Random Forest method, five different m values (0.5P, 0.1P, P, 1.2 $\sqrt{P}$, $\sqrt{P}$) were studied to determine the appropriate number of variables in the trees and classification procedure was performed according to that appropriate m value. Performances of Random Forest and Naive Bayes on cancer data sets were compared with these values. In conclusion, each classification method's performance was examined in binary and multi-class classification and it was observed that there weren't any differences between the results of those two methods.

Keywords: Random Forest, Naive Bayes, DNA Microarray, Classification

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

CART	Sınıflama ve Regresyon Ağacı
cDNA	Tamamlayıcı DNA
CNS	Merkezi Sinir Sistemi
Cy3	Siyanin 3
Cy5	Siyanin 5
DNA	Deoksiriboznükleik Asit
FGF	Bulanık Aen Filtreleme
GI	Bilgi Kazancı
GR	Kazanç Oranı
MDA	Doğruluktaki Ortalama Azalma
MDG	Gini İndeksindeki Ortalama Azalma
mRNA	Mesajcı Ribonükleik Asit
NB	Naive Bayes
NCI	Ulusal Kanser Programı
OOB	Out-Of-Bag
PNET	Primitif Nöroektodermal Tümör
RF	Random Forest
SRBCT	Küçük, Yuvarlak, Mavi Hücreli Tümör
SVM	Destek Vektör Makinesi
Argmax	Maksimum Argümanı

ŞEKİLLER

TABLOLAR

1. GİRİŞ

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

Sınıflama; veri tabanındaki gizli kalıpları ortaya çıkarmakta kullanılan bir yöntemdir. Sınıflama ile veri tabanı belli özelliklere göre küçük homojen gruplara ayrılır. Sınıflama, yeni gelen bir verinin hangi sınıfa ait olduğunu gösteren bir analiz tekniğidir ve bir öğrenme algoritmasına dayanmaktadır. Bu algoritmanın amacı; bir sınıflama modeli oluşturarak, hangi sınıfa ait olduğu bilinmeyen bir veri için sınıf belirlemektir. Burada iyi belirlenmiş değişkenler kilit rolü oynamaktadır. Çeşitli sınıflama yöntemleri bulunmaktadır. Bunlardan ikisi son zamanlarda veri madenciliğinde sıkça kullanılmaya başlanan Random Forest ve Naive Bayes sınıflama yöntemleridir.

Leo Breiman 1996 yılında varyansı düşürücü etkisi olan ve aşırı öğrenmeye karşı güçlü bir yapısı olan "Bagging Sınıflama Yöntemi"ni geliştirmiştir. 2003 yılında Leo Breiman, Adele Cutler ile bir araya gelerek Bagging yöntemini ilerletmiş ve Random Forest Sınıflama Yöntemi'ni geliştirmişlerdir. Yaklaşık 10 yıldır Random Forest sınıflama yöntemi yüksek sınıflama performansı nedeniyle oldukça tercih edilir duruma gelmiştir. Random Forest yöntemi sınıflama performansını arttırmak için oylama ile sınıflandırıcıları birleştirme yolunu kullanır. En yüksek oyu alan sınıfı ise kazanan sınıf olarak belirler.

Naive Bayes Sınıflama Yöntemi'nin kökleri 1760'lı yıllarda keşfedilen Bayes Teoremine dayanmaktadır. Naive Bayes; Bayes Algoritmasının yalınlaştırılmış halidir. Bu yöntemde değişkenlerin birbirinden bağımsız olduğu varsayılır. Yeni örnekler üzerinde sınıflama ise bayes kuralına göre sınıflandırılacak örneğe en

yüksek olasılıkla benzerlik gösteren sınıf seçilerek yapılır. Tüm değişkenlerin birbirinden bağımsız olması gerçek hayatta neredeyse imkansız olsa da Naive Bayes'in sınıflandırmadaki başarısı çeşitli çalışmalarla kanıtlanmıştır.

1.1 Amaç

Bu çalışmadaki amaç, ortak kullanıma açık veri tabanlarından elde edilen 9 adet kanser türüne ilişkin veri setleri üzerinde Random Forest ve Naive Bayes sınıflama yöntemlerinin performanslarının karşılaştırılmasıdır. Veri madenciliğinde ve biyoinformatikte sıklıkla kullanılan ve başarısı çeşitli çalışmalarda gösterilmiş olan Random Forest sınıflama yöntemine göre daha çok metin sınıflamada kullanılan Naive Bayes sınıflama algoritmasının kestirim başarılarının farklı örneklem büyüklüğü ve farklı sınıf sayılarında benzer olup olmadığı incelenecektir. Araştırmaya konu olan hipotez ise; Naive Bayes sınıflama yönteminin performansının Random Forest sınıflama yöntemine göre daha düşük olduğudur.

2. GENEL BİLGİLER

2.1 Karar Ağaçları:

Karar Ağaçları günümüzde çok çeşitli alanlarda kullanılmaktadır. Radar sinyal sınıflandırma, karakter tanıma, uzaktan algılama, tıbbi teşhis, uzman sistemleri bunlardan sadece bir kaçıdır (1).

Karar Ağaçları aşamalı karar vermede kullanılan olası yaklaşımlardan biridir. Ağaç, dallar ve yapraklardan oluşur. Eğer yaprak artık dallara ayrılmıyorsa o yaprağa "karar düğümü" denir. Tüm yapraklar karar düğümü olana kadar ya da o yaprağa ait veri kalmayana kadar dallara ayrılmaya devam eder. Karar ağaçları, yorumlamasının kolay olması, veri tabanı sistemleri ile birleştirilebilmesi açısından tercih edilmektedir (1,2).

2.2 Bagging Sınıflama Yöntemi:

Bagging sınıflama yöntemi 1996 yılında Leo Breiman tarafından bulmuştur. Genelde ağaç tabanlı sınıflandırıcılarda kullanılan bir yöntemdir. Bagging sınıflama algoritması, hem sınıflama istikrarını hem de doğruluğunu arttırmak amacıyla kullanılır. Çünkü bu sınıflama yöntemi tahmin gücünü arttıran genel bir tekniktir. Varyansı düşürücü etkisi vardır, böylece sınıflama hatasını azaltmaktadır. Aşırı öğrenmeye ve eksik gözlemlere karşı güçlü bir yapısı vardır (3,4).

Bagging yöntemi, N çaplı eğitim setinden bootstrap örnekleme tekniği ile m adet n çaplı yeni eğitim setleri oluşturur ($n \leq N$). Bir bootstrap örneklemi, eğitim setinden yerine koyarak örnekleme yöntemiyle x adet örneğin seçilmesinden oluşur. m adet bootstrap örneklemi $B_1,\ldots B_m$ üretilir ve her bir bootstrap örneklemi için C_i gibi m (ağaç sayısı) adet sınıflayıcı oluşturulur. Final sınıfı ise m adet sınıflayıcı içerisinde en çok oyu alan sınıftır. Her bir örnekleme yerine koyarak örnekleme yöntemiyle yapıldığı için bazı örnekler, birden fazla kez aynı örneklemde bulunabilmektedir. Eğer n=N ise bu örneklemde orijinal eğitim setinin %63.2'sinin olması (farklı örnekler), geri kalanın ise bazı örneklerin tekrarından oluşması beklenir (3,4).

Ağaç tabanlı sınıflandırıcılar için Bagging yönteminde veri, eğitim ve test veri seti olarak ayrılır. Eğitim setinden bootstrap örnekleme yöntemiyle m adet ağaç

oluşturulur. Bu ağaçlarda dallara ayırıcı özellikteki değişken, tüm değişkenler arasından rastgele seçilir ve oylamaya tabi tutularak en yüksek oyu alan sınıf nihai (en son) sınıf olarak belirlenir (5).

Bagging yönteminin algoritması şöyledir;

Girdi: S=eğitim seti, I=indüktör, m=bootstrap örneklem sayısı

Her bir i=1 den m {S^l=Eğitim setinden bootstrap örnekleme C_i=I (S^l)}

$$C*(x) = \text{argmax} \sum_{i:C_i(x)=y} 1 \ (\text{en sık tahmin etiketi y}) \qquad (2.1)$$

Çıktı: C* sınıfı

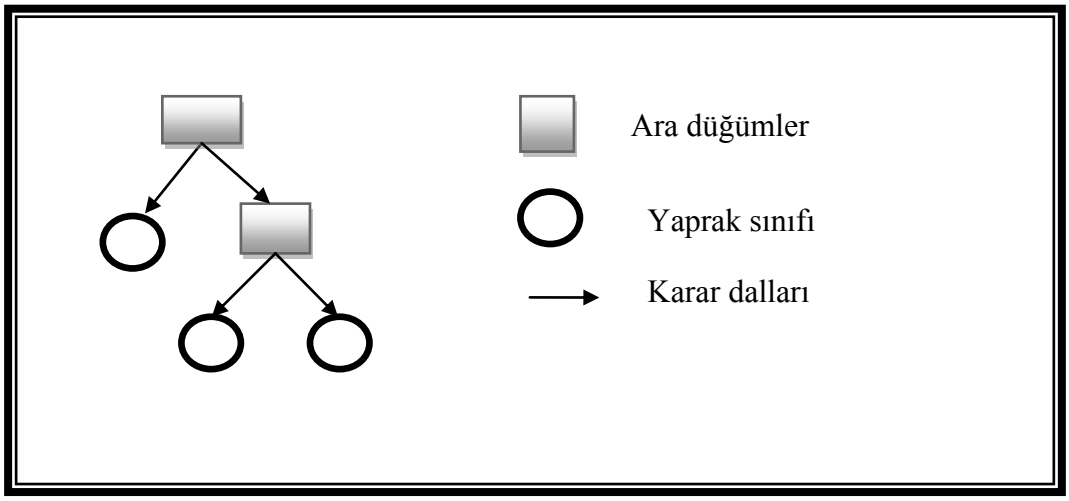
2.3 Random Forest Sınıflama Yöntemi:

Farklı veri grupları üzerinde sınıflandırıcıların başarısının değişkenlik göstermesinden dolayı bazen tek bir sınıflandırıcıdan alınan sonuçlar verimli olmayabilir. Bu durumda başarı oranını ve doğruluğunu artırmak amacıyla sınıflandırıcılar birleştirilir. Oylama yöntemi de bunlardan biridir. Oylama; sayısal tahminleri ya da olasılık değerlerini kullanarak sınıflandırıcıları birleştirme yöntemidir. Her sınıf için olasılık değerleri veya sayısal tahminler toplanır. En yüksek toplama sahip olan sınıf kazanan sınıftır (6).

Random Forest (RF); Leo Breiman ve Adele Cutler tarafından geliştirilen ve içerisinde oylama metodunu barındıran bir sınıflama yöntemidir. Birçok karar ağacının biraya gelmesiyle oluşur ve bireysel ağaçlar tarafından oylanarak kazanan sınıf belirlenir. Karar ağaçları, birbirinden bağımsızdır ve veri setinden bootstrap tekniği ile çekilen örneklerden oluşturulur. "Bagging" yönteminden farklı olarak, Random Forest'ta dallara ayırıcı özellikteki değişkenin, tüm değişkenler arasından rastgele olarak belirlenen m tane değişken içinden seçilmesidir. Her ağaç için m sayısı sabittir ve genelde $\sqrt{p}$ (p değişken sayısını ifade etmektedir) olarak alınması öngörülmektedir (7).

2.3.1 Ağacın Yapısı:

Ağaç; dallar ve yapraklardan oluşmaktadır. Her bir nitelik bir düğüm tarafından temsil edilir. En son yapı "yaprak" (terminal), en üst yapı "kök" ve bunların arasında kalan yapılar ise "dal" olarak adlandırılır. Random Forest yönteminde, ağaç bütün verinin oluşturduğu tek bir düğümle başlamakta, eğer örneklerin hepsi aynı sınıfa ait ise düğüm, yaprak olarak sonlanmakta ve sınıf etiketi verilmektedir. Eğer örnekler aynı sınıfa dahil değilse, örnekleri sınıflara en iyi bölecek olan nitelik seçilmektedir (6,8).



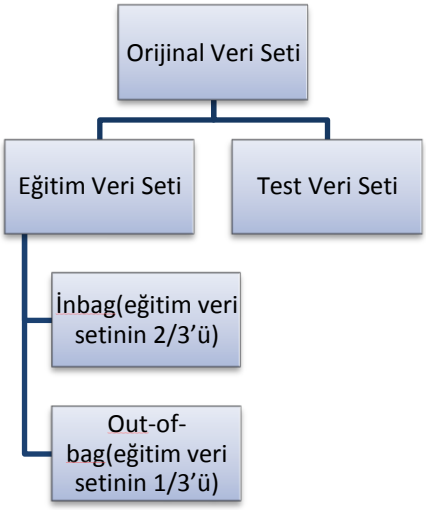
Şekil 2.1. Random Forest'ı Oluşturan Ağaçların Yapısı

2.3.2 Random Forest Sınıflama Algoritmasının Aşamaları:

1) a- Orijinal veri setinin kendine ait bir test seti yoksa; orijinal veri setinden bootstrap yöntemiyle n tane örnekleme seçilir. Her bir örneklemenin 2/3'ü ağaç oluşturmak için kullanılır ve bu verilere eğitim veri seti (inbag) adı verilir. Geriye kalan 1/3'ü ise hata oranını hesaplamak için kullanılır (Tekrarlı Holdout yöntemi) ve bu verilere test veri seti (out-of-bag) adı verilir.

b- Orijinal veri setinin kendine ait bir test seti varsa; hata oranı bu test setiyle de hesaplanabilir (Holdout hata oranı tahmin yöntemi).

Her iki yolla da elde edilen hata oranı birbirine yakın çıkmaktadır (9).



Şekil 2.2. Orijinal Veri Setinin Eğitim ve Test Veri Seti Olarak Ayrılış Şeması

2) Eğitim setinden rastgele olarak seçilen değişkenler arasında en iyi bilgiyi veren değişken dallara ayırıcı değişken olarak kullanılır.

Tüm değişkenler içinden rastgele m tanesi (m= $\sqrt{p}$) seçilerek entropiye dayalı bilgi kazancına bakılır ve dallara ayırıcı özellikteki değişken seçilir. Bu yöntem CART (classification and regression tree) algoritması olarak bilinir (11).

$$\text{Bilgi kazancı}(D,X) = E(D) - \sum_{i=}^{n} p(D_i)E(D_i)$$

X = Bölünecek olan değişken

D = Bölünecek olan özelliğin alt kümeleri (2.2)

E(D) = X değişkenine göre bölünme olmadan önceki entropi

$E(D_i)$ = X değişkenine göre bölünme olduktan sonraki i. alt bölünme entropisi

Dallara ayırma kriteri (cut-off değeri) ise gini indeksi ile belirlenir. Örneğin; dallara ayırıcı değişken olarak hastanın hemoglobin düzeyi seçilmiş ise ayırıcı kriter olarakta, hemoglobin değeri 10 mg/dL'nin altı ve üstü olark ayrılır (≤10 ve >10) (10).

3) Terminal (yaprak) düğüm elde edilene kadar bu işlem tekrar eder.

Terminal düğüm için şartlar:

- Bir düğümde bulunan bütün örnekler aynı sınıfa ait ise

- Bölünmenin yapılacağı değişken kalmamış ise, yani yaprak düğümüne gelene kadar bütün değişkenler kullanılmış ise düğüm terminal düğüm olarak adlandırılır.

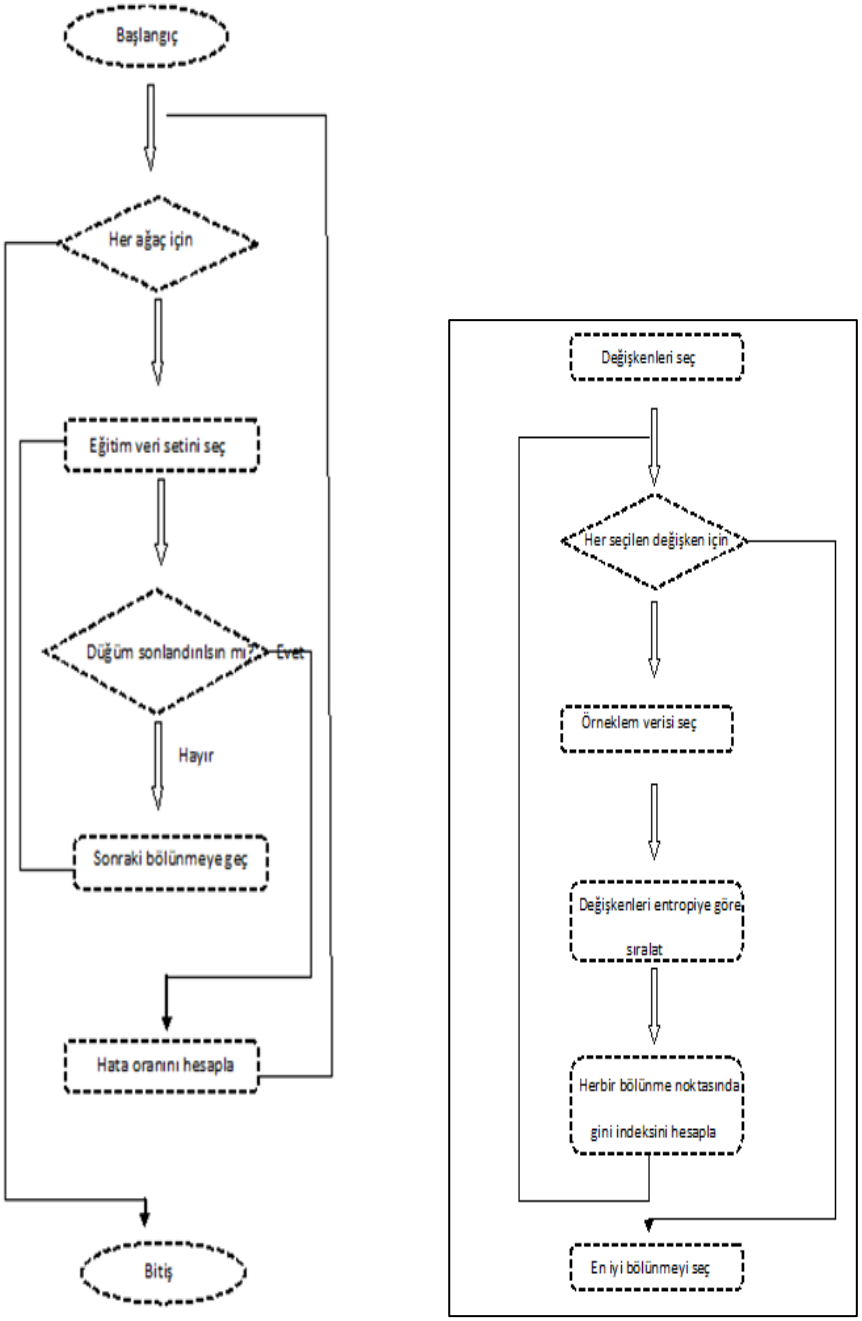
Yeni bir verinin sınıf tahmini için, n örneklemeden oluşturulan n adet ağacın tahminleri birleştirilir ve oylama ile en çok oyu alan sınıf yeni veri için final sınıfı olur. Bu oy, Out of Bag (OOB) hata oranı ile belirlenir (8,9).

4) Eğitim veri setiyle oluşturulan ağaç üzerinde, test veri seti uygulanarak OOB hata oranı hesaplanır. Bu hata; ağaçların her birinin ormandaki bireysel güçlerine (her ağacın kendi hata oranına) ve aralarındaki korelasyona (korelasyon arttıkça hata oranı da artar) bağlıdır.

Her ağaçta verilerin ve değişkenlerin farklı olmasından dolayı ağaçlar birbirinden bağımsız ve böylece hata oranı düşük olur ve doğru kestirilebilir. Hata oranı düştükçe sınıflama algoritmasının performansı artar. Ayrıca, her ağaçta verilerin ve değişkenlerin farklı olması Random Forest yöntemini aşırı öğrenmeye karşı güçlü kılar.

"m" sayısını azaltmak, değişkenler arasındaki korelasyonu azaltır ancak sınıflama gücünde de azalmaya sebep olur. Bu nedenle Breiman tarafından en uygun m sayısı $\sqrt{p}$ olarak belirlenmiştir.

OOB hata oranı ile her ağaca, hata oranı ile ters orantılı olarak ağırlık verilir ve bu ağırlıklara oy denir. Yeni veri için tüm ağaç tahmin değerleri (oyları) birleştirilir ve en çok oyu alan sınıf seçilir (11).



[12]

Şekil 2.3. Random Forest Sınıflama Yöntemi'nin Akış Şeması

Random Forest yöntemiyle sınıflamaya ek olarak değişkenlerin önemlilik derecesi de belirlenebilir. Karar ormanındaki her bir ağaç oluşturulduktan sonra OOB verileri ağaçta kullanılır. Rastgele seçilen i. değişken için OOB'den veriler seçilir ve ağaçta denenerek bir oy değeri elde edilir. Tüm ağaçlarda i. değişken için elde edilen ortalama oy değeri, i. değişkenin önemliliğini (importance) ifade eder. Özellikle

mikroarray çalışmalarında değişkenlerin önemlilik düzeyleri oldukça önemlidir. Değişkenlerin önemlilik derecesi sonucunda gereksiz değişkenler modelden çıkartılarak model indirgemesi yapılabilir (12).

2.3.3 Entropi

Entropi; (ID3, C4.5) karar ağaçları gibi ağaç tabanlı sınıflama algoritmalarında kullanılan bir belirsizlik ölçüsüdür. Entropi, belirsizliğin ölçüsünü verdiği için, ağaç tabanlı sınıflandırıcılarda, dallara ayırıcı özelliği belirlemede kullanılır (13).

Shannon, Metric, Renyi gibi çeşitli entropi hesaplama teknikleri vardır. Literatürde genellikle, klasik entropi hesabı olarak Shannon'nın formülü kullanılmaktadır (14).

$$\text{Entropi (S)} = \sum_{i=1}^{n} -p_i \log_2 p_i \qquad (2.3)$$

Burada logaritmanın 2 tabanında kullanılmasının sebebi; mümkün olan en yüksek belirsizliğin en iyi şekilde hesaplanabilir olmasıdır (13).

$P=(p_1,p_2,p_3,\ldots\ldots,p_n)$ olasılık dağılımları olmak üzere; $p_i \geq 0$ (i=1,2,…,n) ve $\sum_{i=1}^{n} p_i = 1$'dir. Dağılımın belirsizlik miktarı (entropi); her bir denemenin sonucuna ilişkin belirsizlik miktarıdır (15).

Entropi 0 ile 1 arasında değişmektedir. Eğer tüm örnekler aynı sınıfa ait ise entropi "0" olur. Eğer örnekler eşit sayıda ise (hastalar=sağlamlar) entropi "1" değerini alır ki bu da eşitsizliğin en çok olduğu durumdur. Örnekler eşit sayıda değil ise entropi 0 ile 1 arasında bir değer almaktadır (13).

Mikroarray gen ifade verilerinde entropinin yüksek olması, başka bir deyişle belirsizliğin çok olması, ilgili genin ifade düzeylerinin daha rastgele bir dağılıma sahip olduğunun göstergesidir (16).

2.3.4 Dallara Ayırıcı Değişken Optimizasyonu

Random Forest sınıflama yönteminde dallara ayırıcı değişken, tüm değişkenler arasından rastgele seçilen m adet değişken içerisinden seçilir. Genellikle bu m sayısı, eğitim setindeki değişken sayısı (p) kullanılarak hesaplanır. Leo Breiman tarafından m sayısı; regresyon ağaçları için p/3, sınıflama için ise $\sqrt{p}$ olarak önerilmiştir. En uygun m sayısının hesabı için birçok araştırma yapılmıştır. Çünkü m sayısı hem ağaçlar arasındaki korelasyonu hem de ağaçların bireysel güçlerini etkilemektedir. Başka bir ifadeyle; sınıflama performansı arttırılmak isteniyorsa, en uygun m sayısı seçilmelidir. "m" sayısı gereğinden küçük alınırsa, her bir bölünmede yeterli sayıda değişken dikkate alınamayacağı için sınıflama performansı düşer. "m" sayısının gereğinden fazla alınması durumunda ise ağaçlar arasındaki korelasyon artar ki bu da sınıflama performansını düşürür. En uygun m değeri hesabı için, farklı m değerleri kullanarak ağaçlar oluşturulur ve bu ağaçların hata oranlarına bakılır. En küçük hata oranına sahip olan ağacın m değeri en uygun sayı olarak belirlenir. Bir çok araştırmada denenecek olan farklı m değerleri değişkenlik göstermektedir. Bazı çalışmalarda; Breimanın önerisinin yanında m=p (bagging yöntemi), m=p/2 seçenekleri de önerilmiştir. Bazı çalışmalarda ise; 1.2*$\sqrt{p}$, 0.1p ve p seçenekleri önerilmektedir. Kısaca araştırmacı, farklı m değerleri belirlemeli ve her m değeri için OOB veri setinden yararlanarak hata oranı hesaplamalı ve en küçük hatayı sağlayan m değeri en uygun değer olarak seçilmelidir (17,18,19).

2.3.5 Akrabalık (Proximities)

Random Forest yönteminde en kullanışlı araçlardan biride akrabalık (proximities) dır. Akrabalık sayesinde veri seti içindeki aşırı gözlemler (outliers) belirlenir ve eksik değerler tahmin edilir. Akrabalık, N×N boyutlu matristen oluşmaktadır (N: veri setindeki toplam denek sayısı). Ağaçlar oluşturulduktan sonra tüm veriler, yukarıdan aşağıya doğru ağaca uygulanır. Eğer k. ve n. denek aynı terminal (yaprak) düğümde yer alıyorsa akrabalıkları 1 arttırılır. Aynı terminal düğümde yer alan denek sayıları toplanarak, ağaç sayısına bölünür ve akrabalık matrisi (simetrik bir matristir) oluşturulur (20).

Tablo 2.1. Random Forest Akrabalık Matrisi

	D1	**D2**	**D3**
D1	1	a	b
D2	a	1	c
D3	b	c	1

- D2 ve D1= a × T kadar aynı ağaçta yer almıştır.
- T= ağaç sayısı
- Yukarıdaki matris, aynı ağaçta yer alma sayılarının normalleştirilmiş halidir.

Büyük veri setlerinde N×N boyutlu matrisi oluşturmak hızı yavaşlatabilir, bu nedenle zorluk yaratabilir. Çok sayıda girdi olduğu durumlarda N×N matrisine alternatif olarak N×T matrisi kullanılabilir (20).

2.3.6 Kayıp Gözlem Kestirimi

Eğitim setindeki ve test setindeki kayıp gözlemlerin hesaplanması farklılık göstermektedir.

1) Eğitim veri setindeki kayıp gözlemlerin kestirimi:

Random Forest sınıflama yönteminde kayıp gözlem iki farklı şekilde kestirilebilir.

a)1. Yöntem: Eğer i. gözlem eksik ise ve sürekli ya da kesikli sayısal veri tipindeyse, ilgili sınıftaki değerlerin ortancası hesaplanır ve kayıp olan gözleme ortanca değeri atanır. Eğer i. kayıp gözlem kategorik veri tipinde ise; ilgili sınıfın en çok görülen (mod), yani frekansı en yüksek olan değer kayıp gözleme atanır.

b)2. Yöntem: Bu yöntem ilk yönteme göre hesaplaması daha uzun olan bir yöntemdir. Ancak kayıp gözlemi çok olan verilerde bile ilk yönteme göre daha iyi performans göstermektedir. Sadece eğitim setindeki kayıp gözlemlerin telafisi için kullanılır. Kabaca ve doğru olmayan bir şekilde kayıp gözlemin doldurulmasıyla başlanır. Bir orman üzerinde uygulanır ve akrabalıklar hesaplanır. Eğer y(a,b) kayıp bir gözlem ve sürekli sayısal veri tipindeyse, b. değerle kayıp olmayan gözlemler arasındaki akrabalık ile ağırlıklandırılan a. değerin kayıp olmayan gözlemleri

üzerinden ortalaması hesaplanır. Sonra bu değer kayıp gözleme atanır. Eğer y(a,b) kayıp gözlem kategorik veri tipindeyse, kayıp olmayan gözlemler arasında akrabalıkla ağırlıklandırılmış değerler içerisinde frekansı en yüksek olan değer kayıp gözlemimize atanır (8,17,21).

2) Test veri setindeki kayıp gözlemlerin kestirimi:

Random Forest sınıflama yönteminde test veri setindeki kayıp gözlemler; sınıf etiketlerinin belli olup olmama durumuna göre farklı kestirilmektedir.

a- Eğer sınıf etiketi belli değil ise; eğitim setinden değişim yapılarak kayıp gözlemler doldurulur.

b- Eğer sınıf etiketi belli değil ise; test setindeki her değer n adet sınıf kadar çoğaltılır. İlk çoğaltmanın 1. sınıfta olduğu varsayılır ve 1. sınıfta kayıp gözlem hesabı yapılarak yenisiyle değiştirilir. n adet sınıf için aynı işlem yapılır ve n adet sınıf içerisinde en yüksek oyu hangi sınıf aldıysa, kayıp gözlem o sınıfa atanır ve hesaplanır (8,17,21).

2.3.7 Random Forest'ın Kullanım Amaçları

Random Forest sınıflama yönteminde test seti ile algoritmanın hata oranı, değişkenlerin önemlilik düzeyleri, denekler arası akrabalık durumu ve aşırı değerler belirlenebilir. Ayrıca akrabalığa dayalı ölçek koordinatları belirlenebilir (kümeleme yapmak için kullanılır) (12).

2.3.8 Random Forest Yönteminin Avantajları

Random Forest sınıflama yönteminde her ağaçta verilerin ve değişkenlerin farklı olmasından dolayı aşırı uyum sorunu oluşmaz. Eksik verinin olduğu durumlarda ve mikroarray verileri gibi çok büyük veri setlerinde rahatlıkla kullanılabilir, yüksek başarımlı sonuçlar üretebilir. Ağaç sayısında herhangi bir kısıt yoktur tamamıyla araştırmacının isteğine bırakılmıştır. Bağımsız değişkenlere ilişkin sınırlama yapmaz (kesikli, sayısal, kategorik). İkili (binary) ya da çoklu (multiple) sınıflamada kullanılabilir. Her ağaçta verilerin ve değişkenlerin farklı olmasından dolayı RF, karar ağaçlarındaki gibi budamaya ihtiyaç duymaz. Değişkenleri önem sırasına göre otomatik olarak algoritma içinde sıralayabilir ki bu da mikroarray çalışmaları için oldukça önemlidir. RF sınıflama yönteminde analistin sadece 2

parametreyi belirlemeye ihtiyacı vardır (1.ağaç sayısı, 2.her düğümde rastgele seçilecek olan değişken sayısı).

2.3.9 Random Forest Yönteminin Dezavantajları

Random Forest yöntemi, sınıflamanın doğruluğuna ilişkin bir güven aralığı vermez, çünkü yapısında çok fazla ağaç bulundurmaktadır. Bu sebeple ağaçlar hem şekilsel olarak görülemez hem de belli bir güven aralığı hesaplanmasını engeller. Ancak Random Forest algoritmasında var olan "bootstrap" tekniği ile zaten yapılan sınıflama genellenmektedir. Bu nedenle güven aralığına ihtiyaç duyulmaz.

2.4 Naive Bayes Sınıflama Algoritması:

Naive Bayes sınıflama algoritması; Bayes teorisine dayanan, kolay anlaşılan ve hızlı çalışan bir yöntemdir. Naive Bayes; tüm değişkenlerin birbirinden bağımsız ve hepsinin aynı öneme sahip olduğu varsayımlarına dayanan bir algoritmadır (22,23,24).

Naive Bayes; hem tahmin edici hem de tanımlayıcı bir sınıflama tekniğidir. Makine öğrenmesi ve veri madenciliği için en etkili tümevarımsal öğrenme algoritmalarından biridir. Bağımsızlık varsayımı gerçekte çok nadir görülse de, yani bu varsayım çoğu zaman gerçek dışı olsa da, Naive Bayes'in sınıflamadaki başarısı ve diğer bazı sınıflama algoritmalarından üstünlüğü çeşitli çalışmalarda ortaya konmuştur. Bağımsızlık varsayımı her bir değişkenin tek tek öğrenilmesine olanak vermektedir. Böylece, çok değişkene sahip olan verilerde bile sınıflama işleminin hızlı olmasına olanak sağlamaktadır (24,25).

Naive Bayes sınıflama algoritması; bayes kuralına göre sınıflandırılacak örneğe, en yüksek olasılıkla benzerlik gösteren sınıf seçimi ile yapılır. Bu seçim hesaplanırken önsel olasılıklardan yararlanılır.

Naive Bayes sınıflamasında genel olarak; verinin %80'i eğitim seti olarak ve %20'si test seti olarak bölünür.

Naive Bayes sınıflaması, döküman sınıflaması, medikal teşhis gibi konularda sıklıkla kullanılmaktadır.

2.4.1 Değişkenlerin Kategorik Veri Tipinde Olduğu Durumlarda Naive Bayes Sınıflama Algoritması

Herhangi bir x örneği $<a_1,a_2,\ldots,a_n>$ gibi bir vektörle ifade edilsin. Buradaki a_i verinin değişkenlerini göstermektedir. Yeni örneğin ait olduğu sınıf, örneğin değişkenlerini gösteren $<a_1,a_2,\ldots,a_n>$ vektörüne göre MAP (maximum a posteriori) karar kuralı ile belirlenir (26).

$<a_1,a_2,\ldots,a_n> \Longrightarrow$ A örneğinin değişken vektörü

$k_1,k_2,\ldots,k_m \Longrightarrow$ m adet sınıf

MAP kuralına göre;

$$K_{MAP} = \operatorname{argmax} P(k_j \mid a_1, a_2, a_3, \ldots, a_n)$$

$$K_{MAP} = \operatorname{argmax} \frac{P(a_1, a_2, a_3, \ldots, a_n \mid k_j)P(k_j)}{P(a_1, a_2, a_3, \ldots, a_n)}$$

Yani;

$$K_{MAP} = \operatorname{argmax} P(K_j \mid A) \tag{2.4}$$

$$K_{MAP} = \operatorname{argmax} \frac{P(A \mid k_j)\, P(k_j)}{P(A)}$$

P(A) her örnekte sabit ise;

$$K_{MAP} = \operatorname{argmax} P(A \mid k_j)\, P(k_j)$$

Naive Bayes bağımsızlık varsayımı;

$$P(a_1, a_2, a_3, \ldots, a_n \mid k_j) = \prod_{i}^{n} P(a_i \mid k_j)$$

Naive Bayes Sınıflaması;

$$K_{MAP} = \operatorname{argmax} P(k_j) \prod_{i}^{n} P(a_i \mid k_j)$$

Örnek: Kişilerin yaşı, işinin stresi, ailede tansiyon hastası, halsizlik-baş dönmesi verilerine bakılarak ilgili kişinin tansiyon hastası olma durumu aşağıdaki tabloda verilmiştir.

Tablo2.2. Tansiyon Hastalığına İlişkin Veri Seti

Yaş	**İş Stresi**	**Ailede Tansiyon Hastası**	**Halsizlik-Baş Dönmesi**	**Tansiyon Hastası mı?**
≤30	Evet	Yok	Yok	Hayır
≤30	Evet	Yok	Var	Evet
1-40	Evet	Yok	Yok	Hayır
>40	Orta	Yok	Var	Hayır
>40	Hayır	Var	Yok	Hayır
>40	Hayır	Var	Yok	Evet
31-40	Hayır	Var	Var	Evet
≤30	Orta	Yok	Yok	Hayır
≤30	Orta	Var	Var	Evet
>40	Orta	Var	Yok	Hayır
>40	Orta	Var	Yok	Evet
31-40	Orta	Yok	Var	Evet
31-40	Evet	Var	Yok	Hayır
≤30	Orta	Yok	Var	Hayır

X= (yaş<=30, iş stresi=orta, ailede tansiyon hastası=var, halsizlik-başdönmesi=yok), yeni gelen X kişisinin tansiyon hastası olup olmadığının belirlenmesi;

$P(k_i)$: P(tansiyon hastası= “evet”) = 6/14 = 0.429

P(tansiyon hastası = “hayır”) = 8/14= 0.571

Veri seti dikkate alındığında; bu kişinin tansiyon rahatsızlığına sahip olma olasılığı 0.429 iken olmama olasılığı 0.571’dir.

Her sınıf için $P(x_i|k_j)$ olasılığının hesabı;

P(yaş = “<=30” | tansiyon hastası = “evet”) = 2/6 = 0.333

P(yaş = “<= 30” | tansiyon hastası = “hayır”) = 3/8 = 0.375

Tansiyon hastası olan birinin 30 yaşına eşit ve daha küçük yaşta olma olasılığı 0.33 iken, tansiyon hastası olmayan birinin 30 yaşına eşit ve daha küçük yaşta olma olasılığı 0.375’tir.

P(iş stresi = “orta” | tansiyon hastası = “evet”) = 3/6 = 0.50

P(iş stresi = “orta” | tansiyon hastası = “hayır”) = 4/8 = 0.50

Tansiyon hastası olan birinin işinin orta derecede stresli olma olasılığı 0.5 iken, tansiyon hastası olmayan birinin işinin orta derecede stresli olma olasılığı 0.5’tir.

P(ailede tan. hastası = “var” | tansiyon hastası = “evet) = 4/6 = 0.667

P(ailede tan. hastası = “var” | tansiyon hastası = “hayır”) = 38=0.375

Tansiyon hastası olan birinin ailesinde de tansiyon hastasının bulunma olasılığı %68 iken, tansiyon hastası olmayan birinin ailesinde tansiyon hastasının bulunma olasılığı 0.38’dir

P(Halsizlik-başdönmesi= “yok” | tansiyon hastası = “evet”) = 2/6 = 0.333

P(Halsizlik-başdönmesi = “yok” | tansiyon hastası = “hayır”) = 6/8 = 0.75

Tansiyon hastası olan birinin halsizlik ve baş dönmesi şikayetlerinin olmama olasılığı 0.33 iken, tansiyon hastası olmayan birinin halsizlik ve baş dönmesi şikayetlerinin olmama olasılığı 075’tir.

$P(X|k_i)$:

P(X| tansiyon hastası = “evet”) $= 0.333 \times 0.50 \times 0.667 \times 0.333 = 0.037$

P(X| tansiyon hastası = “hayır”) $= 0.375 \times 0.50 \times 0.375 \times 0.75 = 0.053$

Tansiyon hastası olan birinin X verisinin özelliklerine sahip olma olasılığı 0.037 iken, tansiyon hastası olmayan birinin X verisinin özelliklerine sahip olma olasılığı 0.053'tür.

$P(X|K_i) \times P(k_j)$: P(X| tansiyon hastası = "evet") × P(tansiyon hastası = "evet") = 0.016

P(X| tansiyon hastası = "hayır") × P(tansiyon hastası = "hayır") = 0.030

Böylece, X verisi ("tansiyon hastası = "hayır") sınıfına aittir.

2.4.2 Değişkenlerin Sürekli veya Kesikli Sayısal Veri Tipinde Olduğu Durumlarda Naive Bayes Sınıflama Algoritması

Başlangıçta Naive Bayes yöntemi kategorik veriler için üretilmişti. Ancak gerçekte sınıflamaların çoğunda ve veri madenciliğinde genellikle sürekli veya kesikli sayısal veriler ile uğraşılmaktadır. Sürekli ve kesikli veriler Naive Bayes sınıflama algoritmasında iki şekilde hesaba katılabilir (23,25).

2.4.2.1 Olasılık Yoğunluk Fonksiyon Hesabı ile Önsel Olasılıkların Bulunması

Naive Bayes sınıflama algoritmasının sürekli ve kesikli sayısal verilere uygulanabilmesi için değişkenlerin belli bir dağılım göstermesi gerekmektedir.

Değişkenlere ait değerler sürekli ve kesikli sayısal veri tipinde ise yine kategorik veri tipindeki formüller kullanılmaktadır. Ancak, burada $P(a_i / k_j)$ önsel olasılığı farklı hesaplanmaktadır (26).

Her bir k_j için genel varsayım; hedef değişkenin tahmin edicilerinin dağılımlarının birbirinden bağımsız olmasıdır. Her bir a_i değişkenin dağılımı genelde normal dağılım varsayılmakla birlikte ayrıca, lognormal, gamma, poisson gibi dağılım yapıları da göstermektedir. Gaussian (normal dağılım) dağılımı üzerinden gidilirse; her bir a_i sürekli veya kesikli değişkenin (özelliğin) dağılımı, ortalama ve standart sapma ile belirlenir (26,27).

$$\mu_{ij} = E\,(a_i / K = k_j)$$

$$\sigma^2_{ij} = E\,((a_i - \mu_{ij})^2 / K = k_j) \qquad (2.5)$$

Bu paremetrelerin tahmini için ya maximum likelihood estimates (MLE) ya da maximum a posteriori (MAP) tahminleri kullanılır. Ayrıca K'nın önsel olasılığı da $P(K=k_j)$ hesaplanmalıdır.

Dağılımlara ilişkin $P(a_i \mid k_j)$ önsel olasılıkların hesaplanması;

$P(a_i \mid k_j)$:

$$\frac{1}{\sigma_{ij}\sqrt{2\bar{a}}}\exp\left(\frac{-(a-\mu_{ij})^2}{2\sigma_{ij}}\right), -\infty < a < \infty, -\infty < \mu_{ij}, \sigma_{ij} > 0 \quad Normal$$

$$\frac{1}{a\sigma_{ij}(2\bar{a})^{1/2}}\exp\left\{-\left[\log\left(\frac{a}{m_{ij}}\right)\right]^2\right\}, 0 < a < \infty, m_{ij} > 0, \sigma_{ij} > 0 \quad Lognormal$$

Burada m_{ij}: Ölçek parametresi σ_{ij}: Şekil parametresidir.

(2.6)

$$\frac{\left(\frac{a}{b_{ij}}\right)^{k_{ij}-1}}{b_{ij}\Gamma(k_{ij})} exp\left(\frac{-a}{b_{ij}}\right), 0 \le a < \infty, b_{ij} > 0, c_{ij} > 0 \qquad Gamma$$

Burada b_{ij}: Ölçek parametresi k_{ij}: Şekil parametresidir.

$$\frac{\lambda_{ij}\exp(-\lambda_{ij})}{a!}, 0 \le a < \infty, \lambda_{ij} > 0, a = 0,1,2,\ldots \qquad Poisson$$

Burada $\lambda_{ij} =$ ortalamadır.

2.4.2.2 Verileri Kategorilere Bölerek Önsel Olasılıkların Hesaplanması

Burada, veriler, belli aralıklara bölünerek kategorik hale dönüştürülebilir. Böylece, daha önce bahsedildiği gibi; kategorik veriler için Naive Bayes sınıflama yönteminin tüm kural ve formülleri kullanılarak önsel olasılıklar hesaplanabilir. Domingos ve Pazzani 1997 yılında bu bölünme konusunu ele almışlar ve her bir kesikli veya sürekli sayısal veri içeren özelliğin eşit aralıklı 10 parçaya bölünmesi gerektiğini genellemişlerdir. Ancak, bu yöntem özellikle veri madenciliği için

kullanılmaya çok uygun değildir. Çünkü Naive Bayes sınıflama yönteminde verileri uygunsuz bir şekilde bölmek, bilgi kaybına neden olabilir ki bu da bazen veri madenciliği için felaket demektir **(**27,28).

Örnek; kişilere ilişkin yaş, cinsiyet ve beden kitle endeksi düzeyine bakarak kişilerin obezite hastası olup olmadığına ilişkin veriler aşağıdaki tabloda verilmiştir.

Tablo 2.3 Obezite Hastalığına İlişkin Veri

Yaş	**Cinsiyet**	**Beden Kitle Endeksi**	**Obezite Hastalığı**
≤20	E	26	-
21-30	K	25	+
21-30	K	28	-
>30	K	24	+
>30	E	29	-
≤20	K	23	+
21-30	E	24	+
>30	E	30	-
≤20	K	26	-
21-30	K	27	+

X=(yaş=21-30, cinsiyet=erkek, beden kitle indeksi= 27) kişisinin obezite hastası olup olmadığının belirlenmesi;

Obezite Hastası olduğunda	Obezite Hastası olmadığında;
ortalama= 24	ortalama= 27.667
varyans= 0.667	varyans= 2.667

p(obezitei hastası olan) = 5/10 = 0.5

p(obezite hastası olmayan) = 5/10 = 0.5

Veri seti dikkate alındığında; bir kişinin obezite olma olasılığı 0.50 iken olmama olasılığı yine 0.50'dir.

P(yaş=21-30 | Obezite yok) = 2/5 = 0.4
P(yaş=21-30 | Obezite var) = 2/5 = 0.4

Obezite hastası olmayan bir kişinin yaşının 21-30 arasında olma olasılığı 0.4 iken, obezite hastası olan bir kişinin yaşının 21-30 arasında olma olasılığı yine 0.4'tür.

P(cinsiyet=erkek | Obezite var) = 1/5 = 0.2
P(cinsiyet=erkek | Obezite yok) = 3/5 = 0.6

Obezite hastası olmayan bir kişinin cinsiyetinin erkek olma olasılığı 0.2 iken, obezite hastası olan bir kişinin cinsiyetinin erkek olma olasılığı 0.6'dır.

$$p(\text{beden kitle endeks düzeyi} = 27|\text{Obezite yok}) = 1/(\sqrt{2\pi}\,1.633)\,(e^{\frac{-(27-27.667)^2}{2(2.667)}}) = 0.13$$

$$p(\text{beden kitle endeks düzeyi} = 27|\text{Obezite var}) = 1/\left(\sqrt{2\pi}\,0.817\right)\left(e^{\frac{-(27-24)^2}{2(0.667)}}\right) = 0.20$$

Obezite hastası olmayan bir kişinin beden kitle endeksinin 25 olma olasılığı 0.13 iken, Obezite hastası olan bir kişinin beden kitle indeksinin 25 olma olasılığı 0.20'dur.

P(X|Sınıf= "Obezite yok") = 0.4 × 0.2 × 0.127=0.01

P(X|Sınıf="Obezite var") =0.4 × 0.6 × 0.199= 0.05

Sonuç :

P(X|(Obezite yok)) x P(Obezite yok) = 0.01 × 0.5=0.01

P(X|(Obezite var)) x P(Obezite var) = 0.05 × 0.5=0.03

P((Obezite var)|X) > P((Obezite yok)|X) olduğu için bu verilere sahip bir kişinin obezite hastası olduğu söylenebilir.

2.4.3 Sıfır Olasılık Sorunu

Naive Bayes sınıflama yönteminde karşılaşılan sorunlardan biri "sıfır olasılık" sorunudur. Yani gelen bir verinin herhangi bir değişkeni eğitim setindeki değişkenlerin içerisinde yoksa, $P(a_i \mid k_j)$ olasılığı 0 olmaktadır.

Sıfır olasılık, k_j sınıfı üzerindeki diğer tüm sonsal olasılıkların etkisini ortadan kaldırır. Bu sorun şöyle çözümlenmektedir; eğitim setinin oldukça büyük olduğu varsayımı altında; olasılığını hesaplamaya ihtiyaç olan değere "1" eklemek bu sorun üzerinde, tahmini olasılıkta ihmal edilebilecek bir farklılık yaratmakla birlikte sorunu ortadan kaldıracaktır. Bu teknik "Laplacion düzeltmesi (ya da Laplace tahmin edicisi) olarak adlandırılır. Laplace düzeltme metodu en basit düzeltme yöntemidir (27,29).

Gerçek olasılık $P(a_i|K) = \frac{N_{iK}}{N_K}$

Laplace $P(a_i|K) = \frac{N_{iK}+1}{N_K+C}$

C= nitelik sayısı

p= önsel (2.7)

Örnek: Bir veri setinde kişilerin gelirine ilişkin 1000 tane gözlem bulunmaktadır. Bunların; 990 tanesinin geliri orta, 10 tanesinin geliri yüksek olsun ve düşük gelire sahip olan kimse olmasın (gelir=düşük (0)). Düşük, orta ve yüksek gelirli kişilerin önsel olasılıkları hesaplanmak istendiğinde; düşük gelirli kişilerin önsel olasılığı "0" çıkmaktadır. Bu değişkenlerin Laplace düzeltme yöntemi ile hesaplanmış hali aşağıdaki eşitliklerde görülmektedir.

$$P(\text{gelir=düşük}) = \frac{0+1}{1000+3} = 1/1003$$

$$P(\text{gelir=orta}) = 991/1003$$

$$P(\text{gelir=yüksek}) = 11/1003$$

Yukarıdaki hesaplamada dikkat çeken nokta, olasılığı "0" olan düşük gelirli kişinin Laplace düzeltmesi sonucunda olasılığı neredeyse "0"a yakın çıkmıştır.

Ancak diğer değişkenlerin sonsal olasılıkları üzerinde herhangi bir etki yaratmamıştır.

2.4.4 Naive Bayes Sınıflama Yönteminde Kayıp Gözlem Durumu:

Eksik değerler, araştırmada sorun oluşturmaktadır. Naive Bayes eksik değer sorunuyla kolayca başa çıkan bir sınıflama yöntemidir. Verilerin eğitimi sırasında, eksik değer ya da değerler varsa, eğitim veri setine dahil edilmezler. Sınıflama sırasında ise, eksik olan değer ya da değerler hesaplamaya dahil edilmez. Eğer bir değişkenin çoğu değeri eksik ise o değişken, veri setinden çıkartılmalıdır (30).

Tablo 2.1. Hava, sıcaklık, nem ve rüzgar durumuna göre dışarıda oyun oynayıp oynanmayacağına ilişkin veri seti

	Hava		**Sıcaklık**			**Nem**			**Rüzgar**			**Oyun**	
	E	**H**		**E**	**H**		**E**	**H**		**E**	**H**	**E**	**H**
Güneşli	2	3	Sıcak	2	2	Yüksek	3	4	Yok	6	2	9	5
Bulutlu	4	0	Orta	4	2	Normal	6	1	Var	3	3		
Yağmurlu	3	2	Soğuk	3	1								
Güneşli	2/9	3/5	Sıcak	2/9	2/5	Yüksek	3/9	4/5	Yok	6/9	2/5	9/14	5/14
Bulutlu	4/9	0/5	Orta	4/9	2/5	Normal	6/9	1/5	Var	3/9	3/5		
Yağmurlu	3/9	2/5	Soğuk	3/9	1/5								

X= (sıcaklık = soğuk, nem = yüksek, rüzgar = var, hava = ?) ise X kişisi dışarıda oyun oynayabilir mi?

- Olabilirlik "evet" = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238
- Olabilirlik "hayır" = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343
- P("evet") = 0.0238 / (0.0238 + 0.0343) = 41%
- P("hayır") = 0.0343 / (0.0238 + 0.0343) = 59%

Bu kişi %59 olasılıkla oyun oynayamaz. Örnekte olduğu gibi yeni gelen verinin, sınıflanması sırasında gözlemlenen eksik değer (hava=?) hesaba

katılmamıştır. R gibi çeşitli istatistiksel programlardaki komutlar sayesinde Naive Bayes sınıflama yöntemindeki eksik değer sorunu kolayca ortadan kaldırılabilir.

2.4.5 Naive Bayes Sınıflama Yönteminin Avantajları

Naive Bayes sınıflama algoritması; anlaşılması kolay, kullanımı ve eğitimi hızlı olan bir sınıflayıcıdır. İkili ya da çoklu sınıflama için kullanılabilir. Bağımsızlık varsayımı genellikle gerçek dışı bir varsayım olsa da Naive Bayes çoğunlukla iyi bir sınıflama performansı göstermektedir**.** Olasılık tahmin hesaplamaları sırasında örnekten vazgeçerek kayıp değerle mücadele eder.

2.4.6 Naive Bayes Sınıflama Yönteminin Dezavantajları

Naive Bayes sınıflama algoritmasının en zayıf yönü; değişkenler arası bağımlılıkların modele alınmamasıdır. Ancak bağımsızlık varsayımı her ne kadar dezavantaj olarak sayılsa da, Naive Bayes sınıflama algoritması genellikle değişkenler arası bağımlılık durumunda da iyi performans göstermektedir.

2.5 Mikroarray

Bilgisayar teknolojisinin ve moleküler biyolojinin hızla gelişmesi sonucunda gen teknolojileri ortaya çıkmıştır. Bunlar arasında en dikkat çekicisi mikroarray teknolojisidir (31).

İnsan vücudunda birkaç istisna dışında her hücre kromozom ve benzer (özdeş) genler içerir. Bu genler üzerindeki bir parça, her bir hücre tipi için özgün özelliklerle "ifade" edilmeye açıktır ve "gen ifadesi" olarak adlandırılır. Gen ifadesi; DNA içindeki bilgilerin kopyalanmasını tanımlayan bir terimdir. Bu DNA'lar genetik bilgi deposudur (32).

DNA mikroarray teknolojisi; binlerce genin ifade düzeylerini (expression levels) ve DNA değişimlerini eşzamanlı olarak inceleme yöntemidir (31).

Miescher tarafından 1989'da DNA izolasyonunun gerçekleşmesi, 1975'de Southern Blotting ve hibridaizasyon (melezleştirme) ve son olarak 1985'de PCR teknolojisinin keşfinden sonra moleküler biyolojik çalışmalar önem kazanmıştır.

Önceleri araştırmacılar tek bir gen sekans analizi konusunda yoğunlaşmışken şimdilerde yayınlar tüm bir sekans analizi üzerinde durmaktadır (32).

DNA mikroarrayi cam, plastik veya silikon çip gibi katı bir yüzeye sıralı şekilde tutturulmuş mikroskopik DNA spotlarıdır. Yüzeye tutturulan bu parçalara "probe" adı verilir. Mikroarray teknolojisi DNA'nın bir yüzeye bağlanıp bilinen bir gen ya da fragmenti (parçası) ile probe hazırlanması şeklinde tanımlanabilecek "Southern Blotting" teknolojisinden türetilmiştir (33).

2.5.1 Mikroarray'in Aşamaları:

DNA mikroarray aşamaları; yüzey (substrat) seçimi, çip üretimi, etiketleme, melezleme, tarama ve veri normalizasyonundan oluşur (34).

2.5.2 Yüzey Seçimi

"Substrat" olarak adlandırılan yüzey, enzimlerin üzerinde etkili olduğu ve etki ettiği maddedir. Substrat ve enzim birbirine anahtar-kilit uyumu ile bağlıdır. Bu sebeple yüzey seçimi mikroarray çalışmalarında oldukça önemlidir. Mikroarraylerde yüzey olarak cam, plastik, silikon çip ya da naylon membranlar özellikle makro-dizilimler için tercih edilirken, cam yüzeyler mikro-dizilimler için daha idealdir. Her ne kadar standart mikroskop lamları DNA arraylerinde kullanılsa da DNA hedeflerinin bağlanmasını kolaylaştırmak için spotlama öncesi cam slaytların hazırlığa tabi tutulması daha iyi sonuçlar ortaya çıkartır. Cam slaytların nükleik asitle elektrostatik etkileşiminin artırılması için silil, aminosilin, lizin ve poli-L-lizin ile kaplanır. Ayrıca; bağlanma yeteneğini artırmak içinse "oncyte" adı verilen nitroselüloz kaplı lamellerde üretilmiştir. Böylece pürüzsüz bir yapı elde edilerek cDNA'ların bağlanması kolaylaştırılmış ve sağlamlaştırılmış olur. Bu sebeple son zamanlardaki mikroarray çalışmalarında cam slaytlar tercih edilmektedir (34,35).

2.5.3 Çip Üretimi

Çip üretimi; cDNA çipleri ve oligonükleotid çipleri olmak üzere iki kategoride incelenir.

2.5.4 Tamamlayıcı DNA (cDNA) Çipleri:

Tamamlayıcı DNA lar Standford Üniversitesinde Pat Brown ve arkadaşları tarafından geliştirilmiştir ve "geleneksel mikroarray metodu" olarak bilinmektedir.

cDNA'lar; ifade seviyeleri karşılaştırılmak istenen mRNA'lardan RT yöntemi ile sentezlenir. İki farklı kaynaktan alınmış mRNA'lardan (hasta ve sağlam bireyler gibi) sentezlenen cDNA'lardan hasta olanlar kırmızı sağlam olanlar ise yeşil renkte işaretlenir ve yüksek hızlı robotlar sayesinde destek ortamına aktarılır. Böylece cDNA çipi elde edilmiş olur. Bu aktarılma işlemine "hibridizasyon (melezleme)" denir (34,36).

cDNA'ların avantajı; gen polimorfizminden kaynaklanan değişikliklere daha az duyarlı oluşudur. Bu sebeple cDNA'lar hibridizasyon işleminde daha güçlüdürler (37).

2.5.5 Oligonükleotid Çipleri

Oligonükleotid arraylerde çipler öncede hazırlanmış oligonükleotid probları üzerine in situ sentezi ile sentezlenerek oluşturulurlar (34).

2.5.6 Etiketleme

Bir mRNA örneğinin içinden belirlenmiş anahtar bölgeler etiketlenmektedir. mRNA'lardan elde edilen cDNA'lar boyalar ile etiketlenir. Çünkü; mikroarray'a bağlanan cDNA'nın görülebilmesi, varlığını belli eden bir "reporter" moleküle işaretlenmesine bağlıdır. İncelenecek örnek iki farklı boya ile boyanır. Bu boyalar genelde Cy3 ve Cy5'tir. Örneğin; iki farklı kaynağımızdan biri hasta bireyler diğeri sağlam bireyler olsun. Hasta bireylerin cDNA etiketleri kırmızı boya ile, sağlam bireylerin cDNA'ları ise yeşil boya ile etiketlenir. Etiketlenen bu cDNA'lar yüksek hızlı robotlarla yüzey ortamına aktarılır ve hibridizasyon yani melezleme işlemi gerçekleştirilir (31,38).

2.5.7 Hibridizasyon(Melezleme)

Hibridizasyon; 1975'te Sibley ve Ahlquist tarafından geliştirilmiştir.Yeşil ve kırmızı olarak işaretlenmiş ve laboratuar ortamında hazırlanmış cDNA'lar yüksek hızlı robotlarla yüzeye tutturulurlar. cDNA tamamlayıcısını (komplimenterini) bulduğu spotta hibridize olur ve boyanın içindeki florsan sayesinde bu spot gözükür. Array üzerinde her spot ayrı bir deney olarak yorumlanır. Burada florsan yoğunluğu mRNA yoğunluğunu belirtir. Bir spota bağlı olan cDNA'nın miktarı, doğrudan başlangıçtaki RNA moleküllerinin sayısıyla orantılıdır (38).

2.5.8 Tarama

Melezleme işlemi sonrasındaki adım; melezlenmiş spotların taranmasıdır. Tarama; okuyucu denilen sistem ile yapılır. Okuyucu; bilgisayar ile kontrol edilen "inverted scaning flurenscent" mikroskoptur ve lazer ile çalışır. Mikroarraydeki spotlar lazer tarafından uygun dalga boylarında uyarılırlar ve taranırlar. Böylece kırmızı ve yeşil boyaları belirlerler. Uyarılma sonucunda yayılan floresan miktarı, bağlı nükleik asitlerin miktarına karşılık gelir ve buna göre kırmızı ya da yeşil rengi alır. Eğer bir gende her iki durum eşit ise sarı rengi ya da gen ifade edilmemiş ise her iki durumda da siyah rengi alır (39,40).

2.5.9 Veri Normalizasyonu

İki farklı boyayla etiketlenmiş cDNA'lar RNA ile melezleştirilir. Melezleme işlemi sonrasında görüntü üretimi için tarama işlemi yapılır ve bu görüntünün netleşmesi için görüntü analiz programı kullanılır. Bu programlar her spottaki arka ve ön plandaki kırmızı ve yeşil renk yoğunluklarını gösterirler. Ancak sahte sonuçlardan kaçınmak için dikkatli ve kaliteli bir şekilde veri seti kontrol edilmelidir. Örneğin; belli bir tip boya daha flüoresanlı olabilir, arka zeminin yoğunluğu beklide boyanın katılaşmasından ya da camın doğal floresan olması yüzünden spottan spota göre çeşitlilik gösterebilir. Bu sebeple gen ifade profillerinin analizi ve yorumu yapılmadan önce kırmızı ve yeşil renk yoğunluklarının birbirlerine göre normalleştirilmesi gerekir. Boya yanlılığı, yoğunluğun bir fonksiyonudur. Normalizasyon; gen ekspresyon analizinin ilk adımıdır ve melezlemedeki hataları ortadan kaldırmak için kullanılan bir yöntemdir (41,42).

Normalizasyonun pek çok yolu vardır bunlardan biri;

$$N_{Toplam} = \frac{\sum_{i=1}^{N_{array}} R_i}{\sum_{i=1}^{N_{array}} G_i} \qquad (2.8)$$

G_i : i. arraydeki yeşil renklerin yoğunluğu

R_i : i. arraydeki kırmızı renklerin yoğunluğu

Her bir spot için normalizasyon;

$$T_i = \frac{R_i}{G_i}$$

Bir diğer normalizasyon işlemi ise logaritma 2 tabanına göre yapılır;

$$M = log_2(R) - log_2(G)$$

Burada M=0 olursa; her iki ifade yoğunluğunda eşit olduğu, M=1 olursa RNA örnekleri arasında 2 kat değişim olduğu, M=2 olursa RNA örnekleri arasında 4 kat değişim olduğu anlamına gelir ve böyle devam eder (41,42).

2.5.10 Tekrarlama

Mikroarray deneyleri masraflı ve zaman alıcıdır. Bu sebeple mikroarray üreticileri çalışmalarda tekrara olan ihtiyacı vurgulamamışlardır. Sonuç olarak günümüzde mikroarray teknolojik çalışmaların bazılarında tekrar kullanılmamıştır. Ancak tekrar, mikroarray deneylerinde farklı gen ifadelerinin belirlenmesinde güvenilirliği arttırmaktadır. Bir çalışmada kaç tekrarın olacağı ise; gen ifade değişimlerinin büyüklüğüne, belirlenen istatistiksel güce, 1. tip hataya ve değişimi belirleyecek istatistiksel yönteme bağlıdır (43).

3. LİTERATÜR TARAMASI

Literatürde Random Forest ve Naive Bayes sınıflama yöntemlerinin mikroarray verilerine uygulanmasına ilişkin çalışmalar incelenmiştir. Literatürde Hu, Li, Wang ve Daggard'ın 2006 yılında "Mikroarray Veri Analizlerinde Sınıflama Yöntemlerinin Karşılaştırılması" başlıklı makalelerinde 7 farklı kanser verisi üzerinde (meme, akciğer, lösemi, lenf, barsak, yumurtalık ve prostat kanserleri) 5 farklı sınıflama yöntemini karşılaştırmışlardır. Bu yöntemler; C4.5, Random Forest, AdaBoost C4.5, Bagging C4.5 ve LibSVMs'dir. Hu ve arkadaşlarının veri ön işlemesi yaptıkları önceki çalışmalarına dayanarak, seçilen gen sayılarının sınıflamanın performansına etki ettiğini söylemişlerdir. Genel olarak sınıflama performansının 50 ile 100 gen içeren veri setlerinde daha başarılı olduğunu saptamışlardır. Bu sebeple bu çalışmalarında veri ön işleme sonrasında her veri seti için kesikli değere sahip 50 gen seçmişlerdir. Ayrıca verilere 10-kat çapraz geçerlilik uygulamışlardır. Sonuç olarak 10-kat çapraz geçerlilik uygulanan 7 kanser verisine ilişkin 5 farklı sınıflama yönteminin ortalama doğruluk oranları karşılaştırıldığında;

en yüksek doğru sınıflama ortalamasına sahip olan yöntem %94,8 ile Random Forest, ardından %94,1 ile AdaBoost C4.5, %93,2 ile Bagging C4.5, %89,6 ile C4.5 ve en sonda %88,3 ile LibSVMs gelmektedir.

Alexander Statnikov ve arkadaşları 2005 yılında "Mikroarray Tabanlı Kanser Sınıflaması İçin Random Forest ve Destek Vektör Makineleri (SVMs) Yöntemlerinin Kapsamlı Karşılaştırılması" başlıklı makalede hem veri setinin tamamını kullanarak hem de veri setinden gen seçimi yapılarak bu 2 sınıflama yönteminin performansları 22 veri seti üzerinde incelenmiştir. Gen veri setlerinin tamamı kullanıldığında, SVMs'nin 15 veri setinde RF'a göre daha üstün performans gösterdiği, 4 veri setinde ise RF'ın daha üstün performans gösterdiği ve 3 veri setinde ise 2 sınıflama yönteminin de aynı performansı gösterdiği gözlenmiştir. SVMs'nin daha üstün performans gösterdiği 15 veri setinden 7 tanesi istatistiksel olarak anlamlı iken RF'ın daha üstün performans gösterdiği 4 veri setinden hiçbiri istatistiksel olarak anlamlı değildir. Veri setinden gen seçimi yapıldığında, 17 veri setinde SVMs'nin RF'a göre daha üstün performans, 3 veri setinde RF'ın SVMs'ne göre daha üstün performans ve 2 veri setinde ise aynı performansı gösterdikleri gözlenmiştir. SVMs'nin daha üstün performans gösterdiği 17 veri setinden yalnızca 1'i istatistiksel olarak anlamlı iken RF'ın SVMs'ne göre daha üstün performans gösterdiği hiçbir veri setinde istatistiksel olarak anlamlılık gözlenmemektedir.

Ng Ee Ling ve Yahya Abu Hasan'nın 2008 yılında Matematiksel Bilimler Dergisinde yayınlanan "Mikroarray Verileri Üzerinde Random Forest Sınıflama Yöntemini Değerlendirme Yöntemi" başlıklı makalelerinde 4 farklı kanser verisi üzerinde (beyin kanseri, dağılmış büyük b-hücreli lenf kanseri, lösemi ve akciğer kanseri) Random Forest'ı 3 farklı yöntemle değerlendirmişlerdir. Bu yöntemler 10-kat çapraz geçerlilik, biri dışarıda bırakılmış çapraz geçerlilik (leave-one out cross validation LOOCV) ve bootstrap tahminidir. En iyi değerlendirme yöntemini en düşük hata oranını veren olarak belirlemişlerdir. Bu çalışmada Random Forest'ı en iyi değerlendirme yöntemi 10-kat çapraz geçerlilik olduğu sonucuna ulaşmışlardır.

Ng Ee Ling ve Yahya Abu Hasan'nın 2006 yılında yaptığı diğer bir çalışma ise mikroarray verilerinin sınıflanmasına ilişkindir ve bu konuyla ilgili makalede yazmışlardır. Bu çalışmada mikroarray verileri üzerinde çeşitli sınıflama tekniklerinin

karşılaştırılması yapmışlardır. Yapılan pek çok çalışmada iki sınıflı sınıflama üzerinde durulurken bu çalışmada çoklu sınıflama üzerinde durmuşlardır. 3 mikroarray verisine (beyin tümörü, lösemi ve akciğer kanseri) 5 farklı sınıflama yöntemi uygulamışlardır. Bu yöntemler; karar ağacı (J48), Bayes Teoremi (Naive Bayes), örnek tabanlı öğrenme (K-En Yakın Komşuluk), destek vektör makinesi (Sıralı Minimal Optimizasyon, SMO) ve sinir ağı (Çok Katmanlı Algılayıcılar, LP) dır. Bu 5 sınıflama yöntemlerini doğru sınıflama oranları ile karşılaştırılarak performansları ölçmüşlerdir. Veri setlerine özelik seçimi (feature selection) yaparak boyut indirgemesi yapmışlardır. Sonrasında 10-kat çapraz geçerlilik uygulamışlar ve bu verilerin %90'nını eğitim için %10'nu ise test için kullanmışlardır. Elde edilen verilere de parametre düzeltmesi (parametre tuning) yapmışlardır. J48 için her yapraktaki örnek sayısını 2, 5, 10, 15, 0, 25, 30 olarak, k-en yakın komşu için komşu sayısını 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 olarak MLP'de gizli katman sayısını 1,3,5,10 olarak almışlardır. SMO'da polinomal ve radyal çekirdek türlerinde uygulamasını yapmışlar ve Naive Bayes'te düzeltme yapmamışlardır. Tüm parametre düzeltmelerinin ayrı ayrı doğru sınıflama oranlarına bakmışlardır. Sonuç olarak tüm veri setlerinde k-en yakın komşu yöntemi en iyi performansı verdiğini, ancak komşu sayısı arttıkça performansın düştüğünü gözlemişlerdir. K-en yakın komşudan sonra ise en iyi performansı gösteren Naive Bayes yöntemidir.

Meir Peraz ve Tshilidzi Marwala'nın 2011 yılında yayınladıkları "Bulanık Gen Filtreleme: Sınıflandırıcı Performans Değerlendirmesi" başlıklı makalede 2 veri seti (prostat, diffüz büyük b-hücreli lenf kanseri) kullanmışlardır. Bu çalışmada mikroarray çalışmalarının en önemli aşaması olan gen seçiminde gen sıralaması yapmak için yaygın olan 3 yöntem (student-t testi, wilcoxon testi ve ROC eğrisi analizi) ve makalenin de temelini oluşturan bulanık gen filtreleme (FGF) yöntemleri kullanmışlardır. Yani; 2 veri seti üzerinde 4 farklı gen seçimi yöntemi ve 4 farklı sınıflama yöntemi uygulanmıştır. Bunlar; k-en yakın komşu, destek vektör makineleri, Naive Bayes ve yapay sinir ağlarıdır. Sonuç olarak prostat verisinde FGF yöntemi ile seçilen genlerin sınıflama sonucunda daha iyi performans gösterdiğini gözlemlemişlerdir. FGF yöntemiyle oluşturulan verilerin sınıflamasında ise Naive Bayes 0.94 sınıflama başarısıyla sonuncu sırada yer almıştır. Lenf veri setinde ise yine FGF yöntemiyle seçilen genleri sınıflama performansları daha yüksektir. FGF ile

oluşturulan veri setlerinin sınıflandırılmasında Naive Bayes 0.97 ile sonuncu sırada yer almaktadır.

4. GEREÇ ve YÖNTEM

4.1. Uygulama

RF ve NB sınıflama yöntemlerinin performanslarını karşılaştırmak için dokuz farklı kanser türüne ilişkin mikroarray gen ekspresyon verileri kullanılmıştır. Veri setlerine "http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html" adresinden erişilmiştir. Veri setlerine ilişkin Ramón Díaz-Uriart ve Sara Alvarez de Andrés'ın beraber yazmış olduğu "Gene Selection and Classification of Microarray Data Using Random Forest" makalesi 06 Haziran 2006'da BMC Bioinformatics dergisinde yer almış olup ayrıntılı bilgiye bu makaleden ulaşılmıştır. Veri setlerinin özellikleri aşağıdaki tabloda yer almaktadır.

Tablo 4.1 Uygulama Veri Setlerinin Başlıca Özellikleri

Veri Seti	Gen Sayısı	Hasta Sayısı	Sınıf Sayısı	Orijinal Referans
Adenokarsinom	9868	76	2	Ramaswamy et al.(2003)
Beyin Kanseri	5597	42	5	Pomeroy et al. (2002)
Meme Kanseri	4869	78	2	Van't Veer et al.(2002)
Barsak Kanseri	2000	62	2	Alon et al. (1999)
Lösemi	3051	38	2	Golub et al. (1999)
Lenf Kanseri	4026	62	3	Alizadeh et al. (2000)
NCI 60	5244	60	8	Ross et al. (2000)
SRBCT	2308	63	4	Khan et al. (2001)
Prostat Kanseri	6033	102	2	Singh et al. (2002)

Tüm veri setlerinin orijinal referans makaleleri incelenmiştir. Bu makalelere göre adenokarsinoma bir vücut organının duvarında veya iç yüzey hücrelerinde oluşan bir kanser türüdür. Bu çalışmada ilgili kanser türüne ilişkin metastaz olarak bilinen yayılmanın moleküler yapısı üzerine çalışılmıştır. Bunun için 12 metastatik adenokarsinoma nodülleri ile 64 primer (birincil) adenokarsinoma karşılaştırılmıştır.

Beyin kanseri veri setinde, “Gen ifadesine dayanan embriyonel tümör çıktısının merkezi sinir sisteminde (CNS) tahmini” başlıklı makalede 42 hastanın gen profilleri üzerinde çalışılmıştır. Bu veri setinde 10 örnek medulloblastomas denilen habis beyin tümörüne, 10 örnek böbrek ve böbrek dışı rabdoid tümöre, 5 örnek malignant gliome denilen embriyonel olmayan beyin tümörüne, 8 örnek supratentorial Primitif nöroektodermal tümörler (PNETs) ve 4 kişi ise normal insan serabraline sahip olan kişilerin gen ifadeleri üzerinde çalışılmıştır. Meme kanseri veri setinde “Gen ifade profilleri kullanılarak meme kanserinin klinik çıktısının tahmini” konulu makalede primer meme kanseri olan 78 hastanın gen profilleri incelenmiştir. Bunların 34’ü 5 yıl içinde uzak metastaz gelişimi yapan, 44’ü ise en az 5 yıldır herhangi bir belirti göstermeyen yani hastalığı yenmiş örnekler üzerinde çalışılmıştır. Bağırsak kanseri veri setinde, bağırsak kanserine ilişkin gen ifade profilleri üzerinde kümeleme analizleri yapılmıştır. Bunun için 22 normal bağırsak dokusuna sahip, 40 tümörlü bağırsak dokusuna sahip örneklerin gen ifadeleri üzerinde çalışılmıştır. Lenf kanseri veri setinde “Gen ifade profili ile belirlenen büyük B hücreli lenfomanın türleri” başlıklı makalede 3 tür lenf kanseri incelenmiştir. NCI 60 denilen veri setinde Ulusal Kanser Programı (NCI), çeşitli dokular ve organlardaki tümörlerden elde edilen 60 kanser türüne ilişkin hücre dizinleri üzerinde çalışılmış ve “Kanserli hücre dizininde gen ifade desenindeki sistematik varyasyonlar” başlıklı makale yazılmıştır. Bu makalede 8 varyasyon incelenmiştir. SRBCT veri setinde “Gen ifade profili ve yapay sinir ağları kullanılarak kanserlerin sınıflanması ve tanı tahmini” başlıklı makalede küçük, yuvarlak mavi hücreli tümörler (SRBCTs) model olarak kullanılmıştır. 63 örneğin üzerinde çalışılmış olup 4 sınıfta incelenmiştir. Prostat veri setinde “Klinik prostat kanser davranışının gen ifade verileri ile ilişkisi” başlıklı makalede 50 normal ve 52 tümörlü prostat dokuları üzerinde çalışılmıştır. Lösemi veri setinde “Gen ifade profili ile pediatrik akut lenfoblastik löseminin sınıflandırılması, alt türlerinin keşfi ve sonuç tahmini” başlıklı makalede sınıflandırma 102 örnek üzerinde çalışılmıştır.

4.2. Özellik Seçimi (Feature Selection)

Verilere değişken seçimi Clementine 12.0 programında yapılmıştır. Clementine programında değişken seçimi 4 ana başlık içerisinde hesaplanmaktadır.

1) Tüm değişkenlerin ve hedef değişkenin kategorik veri tipinde olduğu durumlarda;

a- Pearson ki-kare: Hedef ve tahmin edici arasında var olan ilişkinin yönüne ve gücüne bakılmaksızın yapılan bağımsızlığı test eder.

b- Olabilirlik oranı ki-kare: olabilirlik oranı pearson ki-kare'ye benzerdir. Ayrıca hedef ve tahmin edicinin bağımsızlığı test eder.

c- Cramer's V: Pearson ki-kare istatistiğine dayanan bir ilişki ölçüsüdür. Değer aralığı 0-1 arasında değişmektedir. 0 ilişkinin olmadığını 1 ise mükemmel bir ilişki olduğunu gösterir.

d- Lambda: Değişkenin, hedef değeri tahmin etmek için kullanıldığı zaman ortaya çıkan hatada oransal azalmayı yansıtan bir ilişki ölçüsüdür. Değer aralığı 0-1 arasında değişmektedir. 1 hedefin mükemmel tahmin edildiğini 0 ise tahmin edicinin hedef hakkında faydalı hiçbir bilgi sağlamadığını gösterir.

2) Bazı değişkenlerin ve hedef değişkenin kategorik veri tipinde olduğu durumlarda; değişken önemlilikleri ya Pearson ki-kare ya da olabilirlik oranı ki-kare yöntemleriyle sıralanabilir. Cramer V ve lambda tüm değişkenler kategorik olmadığı sürece kullanılamaz.

3) Değişkenler sürekli sayısal ve hedef değişkenin kategorik veri tipinde (ya da tam tersi) olduğu durumlarda; F istatistiği kullanılır. Bu çalışmada tüm değişkenlerimiz sayısal ve hedef değişkeni ise kategorik veri tipinde olduğu için değişken seçiminde F istatistiği kullanılmıştır.

4) Değişkenlerin ve hedef değişkenin sürekli sayısal veri tipinde olduğu durumlarda; korelasyon katsayısına dayanan t istatistiği kullanılır.

Değişken seçimi yapılıp boyutu indirgenmiş veriler üzerinde Random Forest ve Naive Bayes sınıflama yöntemleri uygulanmıştır. Önemli genlerin belirlenmesinde ise R programında yapılan Random Forest sınıflama yönteminde "Doğruluktaki Ortalama Azalma (MDA)" ve "Gini Değerindeki Ortalama Azalma (MDG)" değerleri kullanılmıştır. Çıktılar yorumlanırken ilk n kolon sınıflara ait olan doğruluktaki ortalama azalma miktarını (MDA), sonraki 2 kolonda ise sırasıyla MDA ve MDG değerleri verilmektedir. MDA değişkenlerin modele olan katkılarını ölçer. MDA her bir değişkenin önemini belirlemek için OOB verilerini tüm

ağaçlarda kullanarak ve doğru pozitif sonuçları sayarak kullanır. Sonra OOB deki değişkenler arasından biri rastgele değiştirilir ve bu örnek tekrar çalıştırılır. Bu 2'si arasındaki doğru pozitif oranlarının farkı ortalama olarak ormandaki tüm ağaçlar üzerinde MDA değerine eşittir (44).

Naive Bayes sınıflama yönteminin varsayımlardan biri olan değişkenlerin birbirinden bağımsız olma durumu gerçekte çok az rastlanan bir varsayımdır. Ancak Naive Bayes bu varsayımına rağmen oldukça iyi performans göstermektedir. Fakat bazı durumlarda da bu varsayım oldukça düşük performans göstermesine neden olmaktadır. Bunun için değişik yöntemler önerilmektedir. Yukarıda bahsedilen değişken seçimlerinden farklı olarak sınıflama yapmadan önce değişkenlere çeşitli sıralama teknikleri uygulanabilir. Yaygın olarak kullanılan 4 sıralama tekniği bulunmaktadır. Bunlar ki-kare, bilgi kazancı (IG), kazanç oranı (GR), relifF ve simetrik belirsizliktir.

Bilgi kazancı bilgi teorisi ve makine öğrenmesinde sıklıkla kullanılan bir ölçüdür. Bilgi kazancı; belirlenen bir değişkenin sınıf tahmini hakkında kazanılan bilginin ölçüsüdür. Her değişken için bir skor elde edilir. Bu skor; ilgili değişken kullanılarak sınıf hakkında ne kadar fazla bilgi edilebileceğine bakılarak elde edilir. Bunun için rastgeleliği, belirsizliği ve beklenmeyen durumun ortaya çıkma olasılığını gösteren entropiyi kullanır. Kazanç oranı bilgi kazancının hassaslaştırılmış halidir. Bilgi kazancı, çok sayıda değişkene sahip olan değişkenleri desteklerken, kazanç oranı yaklaşımı değişkendeki değerlerin sayısını minimize ederken bilgi kazancını maksimize eder. ReliefF her farklı sınıf için yerine koyarak örnekleme ile bir özelliğin kalitesini, o özelliğin tahmine katkısını öngörür ve ortalama her sınıfın önsel olasılığı ile ağırlıklandırılır. Simetrik belirsizlik değişkenler ve sınıflar arasındaki korelasyonun bir ölçüsüdür.

4.3. Yöntem

Literatürde Random Forest ve Naive Bayes veri madenciliğinde oldukça sık kullanılan iki sınıflama yöntemidir. Bu tezde iki sınıflama yönteminin, günümüz tıp biliminin üzerinde oldukça yoğun durduğu mikroarray verilerinde ki performansları incelenmiştir. Bunun için farklı boyut ve farklı sınıf sayılarında veri setleriyle çalışılmıştır.

Mikroarray verileri çok yüksek boyutlu veriler olduğu için öncelikle bu verilere değişken seçimi (feature selection) uygulanmıştır. Değişken seçimi Clementine 12.0 programında yapılmıştır. Bu programda veriler F istatistiği ile hesaplanan p değerlerine göre önemlilik sırasına dizilmiştir. Böylece verinin boyutu indirgenmiş ve veri gürültüden arındırılmıştır. Sonuç olarak ise boyut indirgeme sayesinde yapılacak işlemlerin verimliliği artırılmıştır. Değişken seçimi yapılan veriler sonrasında "R 2.13.2" programında analiz edilmiştir. R programı http://cran.r-project.org/ sitesinden ücretsiz olarak indirilebilmektedir. İndirildikten sonra herhangi bir ülkenin 'server'i üzerinden bağlantı kurulabilmektedir. R programının tercih nedenleri ise; yüksek boyutlu verilerde hızlı cevap verebilen bir yapıya sahip olması, kullanım açısından kolay olması (kullanıcı ara yüzüne ve elle komut girişi imkanı sağlar) ve ileriki safhalarda bu tez konusunun geliştirilebilmesi için sürekli geliştirilen bir program olmasıdır.

Random Forest sınıflama yöntemi için R programında "randomForest", Naive Bayes sınıflama yöntemi için ise "klaR" paketleri kullanılmıştır. randomForest (Breiman and Cutler's random forests for classification and regression) paketi Leo Breiman ve Adele Cutler tarafından Fortran programında bir ara yüz olarak önerilmiştir Random Forest'ı R programına ise Andy Liaw ve Matthew Wiener uyarlamıştır. "klaR" (Classification and visualization) paketini ise Christian Roever, Nils Raabe, Karsten Luebke, Uwe Ligges, Gero Szepannek, Marc Zentgraf oluşturmuştur. Her bir veri seti için iki sınıflama yönteminin doğru sınıflama oranları verilmiştir. Ayrıca iki sınıflı veri setlerinde ise doğru sınıflama oranına ek olarak gerçekte hasta olan bireyler arasında testin pozitif sonuç verme yani hasta olduğunu saptama oranı olan duyarlılık, gerçekte sağlam bireyler arasında testin negatif sonuç verme yani sağlam olduğunu saptama oranı olan seçicilik, sınıfı 1 olarak tahmin edilmiş doğru pozitif (true positive) örnek sayısının sınıfı 1 olarak tahminlenmiş tüm örnek sayısına oranı olan kesinlik (precision) değerleri hesaplanmıştır. Ancak kesinlik ve duyarlılık ölçütleri tek başına anlamlı bir karşılaştırma sonucu çıkarmamıza yeterli değildir. Bu sebeple her iki ölçütü beraber değerlendirmek daha doğrudur. İki sınıf sayılı veri setlerinde duyarlılık, seçicilik, doğru sınıflama oranı, kesinlik ölçütlerinin yanında ayrıca duyarlılık ve kesinliğin harmonik ortalaması olan F-ölçütü değerleri de verilmiştir

5. BULGULAR

Random Forest ve Naive Bayes sınıflama yöntemleri için sırasıyla doğru sınıflama oranları, duyarlılık, seçicilik, kesinlik ve F-ölçütü bulguları aşağıdaki tablolarda verilmiştir. Duyarlılık, seçicilik, kesinlik ve F-ölçütü değerleri sadece hasta ve sağlam olarak sınıflanan adenokarsinoma, meme, bağırsak ve prostat kanserine ilişkin veri setlerinde verilmiştir (Tablo 6.1 – Tablo 6.4). Burada "*" işareti daha yüksek performansın bir göstergesidir.

Tablo 6.1. Veri Setlerine İlişkin Doğru Sınıflama Oranları

	Doğru Sınıflama Oranı	
Veri Seti	**RF**	**NB**
Meme kanseri	0.69	0.73*
Bağırsak kanseri	0.87*	0.87*
Lösemi	0.95	1.00*
SRBCT	0.98*	0.90
Adenokarsinoma	0.83	0.84*
Beyin kanseri	0.81*	0.78
Lenf kanseri	1.00*	1.00*
NCI	0.68	0.70*
Prostat kanseri	0.92*	0.70

Tablo 6.2. Veri Setlerine İlişkin Duyarlılık, Seçicilik Oranları

	RF		**NB**	
Veri Seti	**Duyarlılık**	**Seçicilik**	**Duyarlılık**	**Seçicilik**
Meme kanseri	0,56	0,80*	0,65*	0,80*
Bağırsak kanseri	0,90*	0,82	0,88	0,86*
Adenokarsinoma	0,97*	0,08	0,88	0,67*
Prostat kanseri	0,89*	0,96*	0,71*	0,68

Tablo 6.3. Veri Setlerine İlişkin Kesinlik Oranları

	Kesinlik	
Veri Seti	**RF**	**NB**
Meme kanseri	0.68	0.71*
Bağırsak kanseri	0.90	0.92*
Adenokarsinoma	0.85	0.93*
Prostat kanseri	0.96*	0.70

Tablo 6.4. Veri Setlerine İlişkin F-Ölçüt Oranları

	F-ölçütü	
Veri Seti	**RF**	**NB**
Meme kanseri	0,61	0,68*
Bağırsak kanseri	0,90*	0,90*
Adenokarsinoma	0,91*	0,90
Prostat kanseri	0,92*	0,71

Random Forest Sınıflama Yönteminde MDA ile belirlenen ilk 10 önemli genin, gerçekte de sınıflar arasında farklılık gösterip göstermediğine Student t testi ve ANOVA testi ile bakılmıştır. Bulgular aşağıdaki gibidir.

Tablo 6.5. Uygulama Veri Setleri için Student t Testi ve ANOVA Testi Sonuçları

Adenokarsinoma			
Gen	**t**	**df**	**Düzeltilmiş P Değeri**
G642	2.82	12	0.015
G360	2.77	11.2	0.018
G1128	1.5	11.1	0.162*
G3	2.05	12.8	0.061
G132	1.46	11.3	0.170*
G79	2.47	11.7	0.030
G1027	2.19	11.3	0.050
G630	2.09	11.4	0.060*
G1042	2.63	11.7	0.022
G571	3.08	13.4	0.008
Meme Kanseri			
Gen	**t**	**df**	**P**
G758	-3.94	50.1	0.001
G640	3.33	75	0.001
G834	-2.35	52.8	0.023
G382	-5.5	75	0.001
G699	-2.15	48.6	0.036
G231	4.48	75	0.001
G683	3.7	75	0.001
G369	2.95	75	0.004
G234	-3	75	0.004
G249	2.16	75	0.034

Tablo 6.5. (Devamı)

Bağırsak Kanseri			
Gen	**t**	**df**	**P**
G490	7.2	60	0.001
G155	8.07	60	0.001
G131	7.92	60	0.001
G235	5.13	29.4	0.001
G92	7.38	60	0.001
G99	5.16	30.3	0.001
G475	-5.54	60	0.001
G399	3.36	59.3	0.001
G462	-4.14	60	0.001
G533	-5.9	60	0.001
Prostat Kanseri			
Gen	**t**	**df**	**P**
G925	14.03	100	0.001
G1971	-10.18	71.8	0.001
G1639	7.12	100	0.001
G634	5.58	100	0.001
G1931	-3.93	100	0.001
G615	9.99	100	0.001
G1338	-2.32	100	0.022
G1611	8.12	100	0.001
G597	6.16	71.5	0.001
G1186	2.84	100	0.006
Lösemi			
Gen	**t**	**df**	**P**
G325	-5.76	36	0.001
G748	-10.58	33.9	0.001
G296	-10.26	36	0.001
G134	-6.45	12.1	0.001
G833	6.24	36	0.001
G140	7.84	36	0.001
G279	-5.99	35.5	0.001
G98	3.24	22.6	0.001
G715	5.55	36	0.001
G908	-6.61	36	0.001

Tablo 6.5. (Devamı)

Beyin Kanseri			
Gen	**F**	**df**	**P**
G596	24.05	4 37	0.001
G607	17.09	4 37	0.001
G556	9.82	4 37	0.001
G263	21.35	4 37	0.001
G489	11.65	4 37	0.001
G414	14.84	4 37	0.001
G716	17.92	4 37	0.001
G364	9.75	4 37	0.004
G748	13.22	4 37	0.001
G333	12.35	4 37	0.001
Lenf Kanseri			
Gen	**F**	**df**	**P**
G482	98.7	2 59	0.001
G485	68.06	2 59	0.001
G483	86.8	2 59	0.001
G486	60.75	2 59	0.001
G505	57.22	2 59	0.001
G459	43.89	2 59	0.001
G626	60.31	2 59	0.001
G2425	52.36	2 59	0.001
G495	52.46	2 59	0.001
G2423	125.81	2 59	0.001

Tablo 6.5. (Devamı)

NCI-60			
Gen	**F**	**df**	**P**
G706	3.58	7 53	0.003
G745	5.85	7 53	0.001
G16	7.53	7 53	0.001
G854	8.55	7 53	0.001
G767	3.91	7 53	0.002
G577	4.39	7 53	0.001
G499	5.32	7 53	0.001
G876	7.16	7 53	0.001
G641	7	7 53	0.001
G491	4.1	7 53	0.001
SRBCT			
	F	**df**	**P**
G682	39.68	3 59	0.001
G480	79.98	3 59	0.001
G193	49.25	3 59	0.001
G83	52.32	3 59	0.001
G711	18.11	3 59	0.001
G458	21.42	3 59	0.001
G568	19.04	3 59	0.001
G588	13.22	3 59	0.001
G369	22.07	3 59	0.001
G629	13.94	3 59	0.001

Sınıflama performansının en üst seviyede olabilmesi için Random Forest'da çeşitli m_{try} değerlerine ilişkin OOB hata oranları aşağıdaki tabloda yer almaktadır.

Tablo 6.6. Random Forest Sınıflama Yöntemi İçin Dallara Ayırıcı Değişken Optimizasyonu Sonuçları

Veri Seti	m_{try}	OOB Hata Oranı
Adeno karsinoma	**0.5P**	0.17
	0.1P	0.15*
	1.2$\sqrt{P}$	0.16
	P	0.18
	$\sqrt{P}$	0.17
Beyin Kanseri	**0.5P**	0.19
	0.1P	0.14*
	1.2$\sqrt{P}$	0.17
	P	0.21
	$\sqrt{P}$	0.19
Meme Kanseri	**0.5P**	0.30
	0.1P	0.30
	1.2$\sqrt{P}$	0.26*
	P	0.30
	$\sqrt{P}$	0.29
Bağırsak Kanseri	**0.5P**	0.13
	0.1P	0.10*
	1.2$\sqrt{P}$	0.13
	P	0.16
	$\sqrt{P}$	0.13
Lösemi	**0.5P**	0.00*
	0.1P	0.00*
	1.2$\sqrt{P}$	0.03
	P	0.03
	$\sqrt{P}$	0.05
Lenf Kanseri	**0.5P**	0.02
	0.1P	0.00*
	1.2$\sqrt{P}$	0.00*
	P	0.03
	$\sqrt{P}$	0.00*
NCI 60	**0.5P**	0.36
	0.1P	0.36
	1.2$\sqrt{P}$	0.33*
	P	0.38
	$\sqrt{P}$	0.34

Tablo 6.6. (Devamı)

Veri Seti	**mtry**	OOB Hata Oranı
Prostat Kanseri	**0.5P**	0.07*
	0.1P	0.08
	1.2$\sqrt{P}$	0.07*
	P	0.09
	$\sqrt{P}$	0.08
SRBCT	**0.5P**	0.02*
	0.1P	0.02*
	1.2$\sqrt{P}$	0.02*
	P	0.03
	$\sqrt{P}$	0.02*

6. TARTIŞMA

Bu çalışmadaki amaç; çok boyutlu veriler üzerinde Random Forest ve Naive Bayes sınıflama yöntemlerinin sonuçlarını karşılaştırmaktır. Sonuçlarda Random Forest'in ve Naive Bayes'in 9 veri seti üzerindeki doğru sınıflama oranlarına, sınıfı hasta ve sağlam olan 4 veri setinde gerçekte hasta olan kişiler arasında doğru pozitifi belirleme oranı olan duyarlılığına, gerçekte sağlam kişiler arasında doğru negatifi belirleme oranı olan seçiciliğine, yöntemlerin aynı şartlar altında tekrarlanan ölçümlerinin aynı sonucu verme derecesi olan kesinliğine ve model başarımının değerlendirme oranlarından biri olan F-ölçütüne bakıldığında 2 yöntem arasında beklenin aksine farklılıklar gözlenmemiştir. Veri setleri üzerinde Random Forest ve Naive Bayes sınıflama yöntemlerinin başarısını ölçen bu oranlar hesaplanmış ve bu oranlar değerlendirilerek Random Forest ve Naive Bayes'in sınıflama performanslarının iyi olduğu sonucuna ulaşılmıştır. Bu yöntemlerden Random Forest'ın gerçekte de veri setlerinin sınıfları arasında da istatistiksel açıdan farklılık gösterip göstermediğine bakılmıştır. Bunun için ise R programında Doğruluktaki Ortalama Azalma (MDA) değerleri kullanılarak değişkenlerin modele olan katkısı belirlenmiş ve modele katkısı çok olan ilk 10 değişken üzerinden değerlendirme yapılmıştır. Genel olarak hasta ve sağlam bireyler arasında istatistiksel açıdan anlamlı bir farklılık gözlenmiştir. Random Forest'ın başarısı böylece bir kez daha gösterilmiştir. Naive Bayes'te ise her değişkenin aynı öneme sahip olma varsayımından dolayı önemli genler belirlenememiş ve böyle bir yorum yapılamamıştır.

7. SONUÇLAR ve ÖNERİLER

Bu tez çalışmasında mikroarray gen ekspresyon verileri üzerinde Random Forest ve Naive Bayes sınıflama yöntemleri karşılaştırılmak istenmiştir. Sonuçlarda, beklenildiği gibi büyük farklılıklar gözlenmemiştir.

Random Forest sınıflama yöntemi için en uygun m_{try} (dallara ayırıcı özelliğin seçilebilmesi için tüm değişkenler arasından rastgele belirlenen m adet değişken) değeri için 5 farklı formül kullanılmıştır. Bunlar 0.5P, 0.1P, $1.2\sqrt{P}$, P ve $\sqrt{P}$'dir. Sonuçlar incelendiğinde, 5 farklı formülün OOB hata oranları arasında büyük farklılıklar gözlenmemiştir. Genel olarak veri setlerinde m_{try} değeri 0.1P olarak alındığında OOB hata oranları daha düşük çıkmaktadır. 5 formülün OOB hata oranları arasında büyük farklılıklar gözlenmediği için bu çalışmada, Random Forest'ın geliştiricisi olan Leo Breiman'nın önerdiği $\sqrt{P}$ formülü kullanılmıştır. Random Forest sınıflama yöntemini kullanırken performans başarısının arttırılması için çeşitli m_{try} formülleri kullanılarak en uygun değerin OOB hata oranına göre belirlenmesi önerilmektedir.

9 adet çeşitli kanser türlerine ilişkin mikroarray gen ekspresyon verileri üzerinde Random Forest ve Naive Bayes sınıflama yöntemleri uygulanmış ve bu yöntemlere ilişkin doğru sınıflama oranları verilmiştir. Sonuçlar incelendiğinde Random Forest, Naive Bayes'e göre 3 veri setinde, Naive Bayes ise Random Forest'a göre 4 veri setinde daha yüksek bir performans göstermiştir. 2 veri setinde ise her iki yöntemde doğru sınıflama oranları eşittir. Genel olarak bakıldığında ise 2 sınıflama yöntemi arasında büyük farklılıklar gözlenmemektedir.

Çeşitli kanser türlerine ilişkin 9 veri setinden 4'ü (meme, barsak, adenokarsinoma ve prostat) hasta ve sağlam olarak sınıflandırılmıştır. Bu 4 veri setine ilişkin duyarlılık, seçicilik, kesinlik ve F-ölçütü oranları incelenmiştir. Sonuçlara bakıldığında;

- Random Forest 3 veri setinde, Naive Bayes ise 1 veri setinde daha duyarlı çıkmıştır. Kalan 1 veri setinde de 2 sınıflama yönteminin de duyarlılık oranları eşit çıkmıştır.

- Naive Bayes 2 veri setinde, Random Forest ise 1 veri setinde daha seçici çıkmıştır. Kalan 1 veri setinde de seçicilik oranları 2 sınıflama yönteminde de aynı çıkmıştır.
- Naive Bayes'in 3 veri setinde, Random Forest'ın ise 1 veri setinde kesinlik (precision) oranları daha yüksektir.
- Seçicilik ve kesinlik oranlarının birleştirilmesi ile elde edilen F-ölçütünde ise Random Forest'ın 2 veri setinde, Naive Bayes'in 1 veri setinde F-ölçütü oranı daha yüksektir. Kalan 1 veri setinde de 2 sınıflama yönteminin kesinlik oranı aynıdır.

Genel olarak duyarlılık, seçicilik, kesinlik ve F-ölçütü değerlerine bakıldığında her iki sınıflama yönteminde de büyük farklılık gözlenmemiştir. Araştırmalarda daha gerçekçi sonuçlar elde edilebilmesi için kesinlik ve seçicilik oranlarının birleşimi olan F-ölçütünün kullanılması önerilmektedir.

Random Forest'a ilişkin doğru sınıflama oranı, duyarlılık, seçicilik, kesinlik ve F-ölçütü gibi değerleri incelendikten sonra değişkenlere ilişkin MDA değerleri ile önemli ilk 10 gen belirlenmiştir. Bu genlerin gerçekte de sınıfları arasında farklılık gösterip göstermediğine student t testi ve ANOVA testi ile bakılmıştır. Sonuçlar incelendiğinde; adenokarsinomada bulunan G1128, G132 ve G630 genlerinde sınıflar arasında istatistiksel açıdan anlamlı bir farklılık gözlenmezken diğer tüm kanser veri setlerindeki ilk önemli 10 gende sınıflar arası farklılık istatistiksel açıdan anlamlı bulunmuştur. Yapılacak olan çalışmalarda sınıflama yöntemlerinin performansları hesaplandıktan sonra gerçek sınıflar arasında istatistiksel açıdan anlamlı fark olup olmadığına bakılması da önerilmektedir.

KAYNAKLAR

1. Safavian, R., Landgrebe, D. (1991), A Survey of Decision Tree Classifier Methodology [Elektronik Sürüm]. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 21, No. 3, pp 660-674.

2. Aydoğan, Ü. (2011). Destek Vektör Makinalarında Kullanılan Çekirdek Fonksiyonların Sınıflama Performanslarının Karşılaştırılması. Yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

3. Bauer, E., Kohavi, R. (1998). An Empirical Comparison of Voting ClassificationAlgorithms: Bagging, Boosting, and Variants [Elektronik Sürüm]. Machine Learning, Vol. 36, ıssue 1-2, 105-139.

4. Sutton, C. (2005). Classification and Regression Trees,Bagging, and Boosting [Elektronik Sürüm]. Handbook of Statistics, Vol. 24ISSN: 0169-7161, Elsevier B.V. All rights reserved. DOI 10.1016/S0169-7161(04)24011-1.

5. Breiman, L. (1996). Bagging Predictors [Elektronik Sürüm], Machine Learning, Vol. 24, 123–140.

6. Coşgun, E., Karabulut, E., Karaağaoğlu, E. (2009). Random Forest ve Destek Vektör Makinası Yöntemleri ile Gen Seçimi ve Sınıflaması.VI. Ulusal İstatistik Kongresi: 29 Mayıs 2009- Antalya, Türkiye.

7. Breiman, L. (2001). Random Forest [Eloktronik Sürüm]. Machine Learning, 45, 5-32.

8. Breiman, L., Cutlar, A. Random Forest. Erişim: 20 Mart 2010, University of Stat Barkly Ağ Sitesi. **http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm**

9. Pater, N. (2005). Enhancing Random Forest Implementation in weka. Comferance of Machine Learning: 28 Kasım 2005.

10. Gall, J., Razavi, N., Gool, L.V. (2011), An İntroduction to Random Forest for Multi-class Object Dedection [Elektronik Sürüm]. Outdoor and Large-Scale Real-World Scene Analysis, Vol. 7474, 243-263.

11. Montillo, A. Random Forest. Erişim: 25 Mart 2010, University of Temple Ağ Sitesi **http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForest_4-2-2009.pdf**

12. Breiman, L. Manual-Setting Up, Using, And Understanding Random Forests. Erişim: 20 mart 2010, University of California, Berkeley. **http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf**

13. Silahtaroğlu, G. (2009), An Attribute-Centre Based Decision Tree Classification Algorithm [Elektronik Sürüm]. World Academy of Science, Engineering and Technology, 56, 302-306.

14. Wolf, F. (1999). Calculation of information and complexity in time series - Concepts, Requirements and Limits. Doktora tezi, Bayreuth Üniversitesi, Almanya.

15. Reny, A., On measure of Entropy and Information (1961). Proccedings of the Berkely Symposium on Mathematical Statistics and Probablity: 20-30 Temmuz, 1961. Volume1:547-561.

16. Butte, A., Kohane, S. (2000). Mutual İnformation Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements. Pacific Symposium on Biocomputing: 20-30 Temmuz, 1961, Hawaii. Vol.5:415-426.

17. Svetnik, V., Liaw, A., Tong, C. (2004). Variable Selection in Random Forest with Application to Quantitative Structure-Activity Relationship. Erişim: 25.04.2010, National Taiwan University Ağ Sitesi.

18. Palmer D., O'Boyle N., Glen R., Mitchell J. (2006). Random Forest Models To Predict Aqueous Solubility [Elektronik Sürüm]. Journal of Chemical Information and Modeling 47(1):150-8.

19. Goldstein, B., Hubbard, A., Cutle, A., Barcellos L. (2010). AnrApplication of Random Forests to Agenome-Wide Association Dataset: Methodological Considerations & New Findings, Goldstein et al. BMC Genetics, 11:49.
http://www.csie.ntu.edu.tw/~b88052/tmp/vietri.pdf

20. Liaw, A. (2002). Classification and Regression by Random Forest [Elektronik Sürüm]. R News, Vol. 2/3, 18-22.

21. Qi, Y., Klein-Seetharaman, J.,Bar-Joseph, Z. (2005). Random Forest Similarity for Protein-Protein Interaction Prediction from Multiple Sources. Erişim: 30 Mart 2010, University of Stanford Ağ Sitesi.
http://psb.stanford.edu/psb-online/proceedings/psb05/qi.pdf

22. Ron, K. (2011), Scaling Up the Accuracy of Naive Bayes Classifiers: a Decision-Tree Hybrid. Erişim: 24.04.2010, Associatıon For The Advancement Of Artificial Intelligence Ağ Sitesi.
http://www.aaai.org/Papers/KDD/1996/KDD96-033.pdf

23. Rısh, I. (2001). An Emprical Study of the Naive Bayes. IBM Research Report: 2 Kasım, 2001.

24. Zhang H. (2004), The Optimality of Naive Bayes. In FLAIRS Conference : 2004, Miami Beach, Florida, USA.

25. Aydoğan, E. (2008), Veri Madenciliğinde Sınıflandırma Problemleri İçin Evrimsel Algoritma Tabanlı Yeni Bir Yaklasım: Rough-Mep Algoritması. Doktora tezi, Gazi Üniversitesi, Ankara.

26. Mitchell, T., Hill, M. (2005). Generative and Discriminative Classifiers: Naive bayes and logistic regression. Erişim: 02.05.2010, Carnegie Mellon University Ağ Sitesi.
http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

27. Kohavi, R., Sahami, M. (1996), Error – Based and Entropy – Based Discretization of Continuous Features. Erişim: 05.05.2010, University of Stanford Ağ Sitesi.
http://robotics.stanford.edu/users/sahami/papers-dir/kdd96-disc.pdf

28. Maimon, O., Rokach, L. (2005), Decomposition Methodology for Knowledge Discovery and Data Minig: Theory and Aplications. Singapur: World Scientific Yayınevi.

29. Sarkar, J., Lee, K. ve Lee, S. (2010). A Smoothed Naive Bayes Classifier for Activity Recognition [Elektronik Sürüm]. IETE Journals. Vol. 27, 107-119.

30. Naive Bayes. (2009). Erişim: 15.03.2010,
http://code.google.com/p/ourmine/wiki/LectureNaiveBayes#Bayes'_rule

31. Shakya, K., Ruskin, H. J., Kerr, G., Crane, M., Becker, J. (2010). Comparison of Microarray Pre-Processing Methods [Elektronik Sürüm]. National Center for Biotechnology Information (NCBI), 680:139-47, doi: 10.1007/978-1-4419-5913-3_16.

32. Microarrays: Chipping Away At The Mysteries of Science and Medicine. (t.y). Erişim: 17.03.2010,
http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html

33. Babu, M. (2004). Introduction to Microarray Data Analysis [Elektronik Sürüm], Computational Genomics (s. 225-249). İngiltere: Horizon Yayıncılık.

34. Yoltaş, A., Karaboz, İ. (2010), DNA Mikroarray Teknolojisi ve Uygulama Alanları [Elektronik Sürüm]. Elektronik Mikrobiyoloji Dergisi TR, Cilt: 08, sayı 1, 01-19.

35. Savlı, H. (2004). Dizilim Teknolojisi: Çip'lerden Sonsuzluğa [Elektronik Sürüm]. Türkiye Klinikleri J Med Sci, 24:534-540.

36. Zahurak, M., Parmigiani, G., Yu, W., Scharpf, R., Berman, D., Schaeffer, E., Shabbeer, S., Cope, L. (2007). Pre-processing Agilent Microarray Data [Elektronik Sürüm]. BMC Bioinformatics, 8: I 42.

37. Watson, A., Mazumder, A., Stewart, M., Balasubramanian, S. (1998). Technology for Microarray Analysis of Gene Expression [Elektronik Sürüm]. Current Opinion in Biotechnology, 9:609–614.

38. Peixoto, B., Vêncio, R., Egidio, C., Mota-Vieira, L., Verjovski-Almeida, S., Reis, E. (2006). Evaluation of Reference-Based Two-Color Methods for Measurement of Gene Expression Ratios Using Spotted cDNA Microarrays [Elektronik Sürüm]. BMC Genomics, **7**:35.

39. Ramdas, L., Zhang, W. (2006). Microarray Image Scanning [Elektronik Sürüm]. Springer Link,Vol. 319, 261-273.

40. Butte, A. (2002). The Use and Analysis of Microarray Data [Elektronik Sürüm], Nature Reviews / Durug Discovery, Vol. 1, 951-960.

41. Bilban, M., Buehler, L., Head, S., Desoye, G., Quaranta, V. (2002). Normalizing DNA Microarray Data [Elektronik Sürüm]. National Center for Biotechnology Information (NCBI), 4:57-64.

42. Ting, M., Kuo, F., Sklar, J. (2000). Importance of Replication in Microarray Gene Expression Studies: Statistical Methods and Evidence From Repetitive cDNA Hybridizations [Elektronik Sürüm]. PNAS, Vol. 97, 9834-9839.

43. Pan, W., Lin, J., Le, C. (2002). How Many Replicates of Arrays Are Required to Detect Gene Expression Changes in Microarray Experiments: A Mixture Model Approach [Elektronik Sürüm]. Genom Biology, Vol.3, No:5.

44. Armitag, D., Ober H. (2010). A comparison of supervised learning techniques in the classification of batecholocation calls [Elektronik Sürüm]. Ecological Informatics, 5, 465-473.