T. C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
SAYISAL YÖNTEMLER ANABİLİM DALI

DOKTORA TEZİ

VERİ MADENCİLİĞİ YÖNTEMİ İLE PROSTAT KANSERİ İÇİN ERKEN UYARI PROTOKOLLERİNİN GELİŞTİRİLMESİ

ÖZİN KALEMCİ

2502110336

TEZ DANIŞMANI
PROF. DR. MEHPARE TİMOR

İSTANBUL 2018

T.C.
İSTANBUL ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ

DOKTORA
TEZ ONAYI

ÖĞRENCİNİN;

Adı ve Soyadı	: ÖZİN KALEMCİ	**Numarası**	:	2602110336
Anabilim Dalı / Anasanat Dalı / Programı	: SAYISAL YÖNTEMLER	**Danışmanı**	:	PROF.DR.MEHPARE TİMOR
Tez Savunma Tarihi	: 13.12.2018	**Saati**	:	14:00
Tez Başlığı	: VERİ MADENCİLİĞİ YÖNTEMİ İLE PROSTAT KANSERİ İÇİN ERKEN UYARI PROTOKOLLERİNİN GELİŞTİRİLMESİ.			

TEZ SAVUNMA SİNAVI, İÜ Lisansüstü Eğitim-Öğretim Yönetmeliği'nin **50. Maddesi** uyarınca yapılmış, sorulan sorulara alınan cevaplar sonunda adayın tezinin **KABULÜNE** OYBİRLİĞİ / ~~OYÇOKLUĞUYLA~~ karar verilmiştir.

JÜRİ ÜYESİ	İMZA	KANAATİ (KABUL / RED / DÜZELTME)
PROF.DR.MEHPARE TİMOR		Kabul
PROF.DR.TUĞBA GÜRSOY		Kabul
PROF.DR.ZUHAL TANRİKULU		Kabul
PROF.DR.ERGÜN EROĞLU		Kabul
PROF.DR.AHMET METE ÇİLİNGİRTÜRK		Kabul

YEDEK JÜRİ ÜYESİ	İMZA	KANAATİ (KABUL / RED / DÜZELTME)
DOÇ.DR.SEDA TOLUN TAYALI	—	—
DOÇ.DR.ERDAL DİNCER	—	—

ÖZ

VERİ MADENCİLİĞİ YÖNTEMİ İLE PROSTAT KANSERİ İÇİN ERKEN UYARI PROTOKOLLERİNİN GELİŞTİRİLMESİ

ÖZİN KALEMCİ

Veri madenciliği, karar vericilerin eldeki verilerden yola çıkarak doğru ve etkin kararlar almasına yardımcı olan yöntemler topluluğudur. Veri madenciliği yöntemleri, özellikle tıp alanında daha çok tahmin edici yönüyle kullanılmaktadır. Son yıllarda yapılan çalışmalar veri madenciliği yöntemlerinin kanser de dahil olmak üzere bir çok hastalığın teşhis edilmesinde umut verici sonuçlar ortaya koyduğunu göstermektedir. Bu çalışmada amaç, veri madenciliği yöntemleri kullanılarak, prostat kanserinin erken ve doğru teşhis edilebilmesi için anlamlı bir model oluşturmaktır. Çalışmanın diğer teşhis etme çalışmalarından farkı, PSA veya rektal tuşe vb. gibi hiçbir tıbbi sonuç değişkeni içermemesi, sadece genetik ve fiziksel değişkenler içermesidir. Çalışmada stacking topluluk metodu altında bayes sınıflandırıcılar, k en yakın komşuluk ve karar ağacı yöntemleri kullanılarak bir topluluk modeli oluşturulmuştur. Bu model ile prostat kanseri olan ve olmayanların en doğru şekilde sınıflandırılması amaçlanmaktadır. Yapılan çalışmada 989 kişiden oluşan, her kişiye ait 200 bin SNP ve 18 adet fenotip değişken içeren prostat verisi kullanılmıştır. Modelin performans sonuçlarına bakıldığında; doğruluk, kesinlik ve duyarlılık değerleri sırasıyla %84,13, %89,84 ve %74,23'dür. Bu sonuçlara bakıldığında modelin prostat kanserini tahmin etme yeteneği başarılıdır.

Anahtar Sözcükler: Sağlıkta Veri Madenciliği Uygulamaları, Karar Ağaçları, Stacking Yöntemi, K En Yakın Komşuluk Yöntemi, Naive Bayes Yöntemi, Bütünsel Genom İlişkilendirme, Tekli Nükleotid Polimorfizm

ABSTRACT

DEVELOPMENT OF EARLY WARNING PROTOCOLS FOR PROSTATE CANCER BY DATA MİNİNG METHOD

ÖZİN KALEMCİ

Data mining is a collection of methods that help decision makers to make accurate and effective decisions based on available data. Data mining methods are being used more often in the field of medicine, especially for predicting disease. Recent studies have shown that data mining methods have promising results in diagnosing many diseases, including cancer. The aim of this study is to establish a meaningful model for the early and accurate diagnosis of prostate cancer using data mining methods. The difference of the model from other diagnostic studies is that it does not involve any medical outcome variables, such as PSA or rectal key, but only genetic and physical variables. In the study, an ensemble model was constructed by using Bayesian classifiers, k nearest neighbor and decision tree methods under the stacking ensemble method. With this model, it is aimed to classify individuals with and without prostate cancer in the most accurate way. In the study, prostate data consisting of 989 individuals, 200 thousand SNPs per each person and 18 phenotype variables were used. When the performance results of the model are considered; accuracy, precision and sensitivity values are 84,13%, 89,84% and 74,23% respectively. Given these results, the model has a good ability to predict prostate cancer.

Keywords: Data mining applications in healthcare, Decision Trees, Stacking Method, K Nearest Neighbor Method, Naive Bayes Method, Genome Wide Association Studies, Single Nucleotide Polymorphism

ÖNSÖZ

Doktora eğitimim boyunca kendisinden çok şey öğrendiğim, tez dönemimde her daim manevi desteğini hissettiren çok değerli hocam Prof. Dr. Öner Esen'e ve tez çalışmam boyunca her zaman arkamda olan, bütün olumsuzluklara rağmen beni yüreklendiren, cesaretlendiren ve hep motive eden çok değerli hocam Prof. Dr. Mehpare Timor'a minnetlerimi sunarım.

Bilgi birikimi ile beni aydınlatan ve her zaman destek olan değerli hocam Prof. Dr. Umman Tuğba Şimşek'e ve zor zamanında bile bana vakit ayırarak her zaman hoş görüsüyle ve bilgi birikimiyle yanımda olan değerli hocam Prof. Dr. Zuhal Tanrıkulu'na çok teşekkür ederim.

En karamsar dönemimde desteğiyle beni motive eden ve zoru kolaylaştıran çok değerli arkadaşım Dr. Gürkan Üstünkar'a katkıları için çok teşekkür ederim. Sevgili arkadaşlarım Esra Baytören, Dr. Pakize Yiğit ve Dr. Yonca Erdem Demirtaş'a desteklerinden ve özverilerinden dolayı minnettarım.

Uygulamada kullandığım veriyi kullanıma sunan Amerika Birleşik Devletleri Ulusal Biyoteknoloji Bilgi Merkezi'ne (NCBI) teşekkürü bir borç bilirim.

Tez çalışmam boyunca sonsuz destekleri ve sabırları ile hep yanımda olan eşime, kızıma ve aileme sonsuz teşekkür ederim.

Bu tez çalışmam şu an hayatta olmayan sevgili anneme verdiğim söz olmasaydı belki de hiç yazılmayacaktı. Sevgili anneme ne kadar teşekkür etsem az.

İÇİNDEKİLER

İKİNCİ BÖLÜM
PROSTAT KANSERİ VE GENOM

ÜÇÜNCÜ BÖLÜM
UYGULAMADA KULLANILAN VERI MADENCILIĞI YÖNTEMLERININ AYRINTILI AÇIKLAMASI

DÖRDÜNCÜ BÖLÜM

UYGULAMA

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

KISALTMALARLİSTESİ

CRISP–DM	**The Cross Industry Standard Process for Data Mining**
VM	**Veri Madenciliği**
NB	**Naive Bayes**
K-NN	**K En Yakın Komşuluk**
NIH	**National Institute of Health**
NCBI	**National Center for Biotechnology Information**

GİRİŞ

Karmaşık problemlerle mücadele etmek ve rekabet gücünü korumak açısından şirketler, organizasyonlar, hükümetler, bilim insanları ve toplumlar için veri ve bunun en iyi şekilde kullanılması yeteneği giderek daha da önem kazanmaktadır. Veri madenciliği, tahmini analitik ve işletme analitiği, bu verileri kullanır, eşi benzeri olmayan bilgiler verir, daha iyi bilgilendirilmiş kararlar verir, tahminler sunar ve gittikçe artan karmaşık problemleri çözmeye yardımcı olur. Şirketler ve kuruluşlar her türlü iç ve dış kaynaktan artan miktarda veri toplar ve giderek daha fazla veri yönlendirir hale gelir. Veri analitiklerini kullanmak için güçlü araçlara hakim olmak ve nasıl kullanılacaklarını bilmek rekabetçi avantajlar kazanmak öngörüyü, etkinliği, verimliliği, büyümeyi ve verimliliği artırmak için büyük önem taşır.

Teknoloji, her zamankinden daha büyük miktarda veri yakalamamıza ve saklamamıza olanak tanısa da, verilerde altta yatan ilişkiler, trendler, anormallikler ve aşırıklar gibi ilgili bilgileri bulup bunları basit, anlaşılır ve sağlam nicel ve nitel modellerle özetlemek zor bir iştir. Veri madenciliği, verilerin altında yatan yapıların keşfedilmesine, verilerin bilgiye dönüştürülmesine yardımcı olur. Matematik, istatistik, mantık, bilgisayar bilimi ve bilgi teorisi, veri madenciliği ve makine öğrenimi ve istatistiksel öğrenme teorisi artık sağlam bir teorik temel ve bu zorluğu çözmek için güçlü yöntemler sunmaktadır.

Birçok alanda olduğu gibi sağlık verilerinde de altta yatan ilişkileri keşfetmek ve hastalıkların teşhisine destek olmak için veri madenciliği yöntemlerini etkili bir şekilde kullanmak çok önemlidir. Kronik hastalıkların teşhisi için erken uyarı sinyallerinin tesbit edilmesi tıbbın ilgilendiği önemli bir konudur ve bu erken uyarı sinyalleri veri madenciliği ile tesbit edilebilir.

Bu tezde de amaç veri madenciliği yöntemleri ile prostat kanseri için erken uyarı protokollerinin geliştirilmesidir. Çalışmada, stacking topluluk metodu altında bayes sınıflandırıcılar, K-NN ve karar ağaçları yöntemleri kullanılarak bir model kurulmuştur. Kurulan bu model, hem genetik hem de fiziksel değişkenler içeren çok etnikli prostat kanseri verisi üzerinde denenerek sağlıklı ve hasta bireyleri doğru bir şekilde sınıflandırmayı amaçlamaktadır. Bu şekilde prostat kanserinin doğru ve erken bir şekilde teşhis edilmesine katkı sağlanması amaçlanmaktadır.

Birinci bölümde, veri madenciliği kavramı, süreci, veri madenciliğinde karşılaşılabilecek önemli sorunlar, veri madenciliğinin ilişkide olduğu disiplinler ve tıpts veri madenciliği konularına değinilmiştir.

İkinci bölümde, prostat kanseri, prostat kanserinin tezde kullanılan fiziksel değişkenlerle ilişkisi, SNP'ler, SNP'lerin hastalıklarla ilişkisi ve prostat kanseri ile ilgili yapılmış SNP çalışmalarına değinilmiştir.

Üçüncü bölümde uygulamada kullanılan veri madenciliği yöntemleri ayrıntılı bir şekilde açıklanmıştır.

Dördüncü bölümde ise k-nn, naive bayes ve karar ağaçları yöntemleri kullanılarak stacking topluluk metodu altında bir model kurulmuştur.

BİRİNCİ BÖLÜM
VERİ MADENCİLİĞİ VE BİLGİ KEŞFI

Veri madenciliği, önceden bilinmeyen örtülü ve potansiyel olarak yararlı bilginin veriden çıkarılmasıdır. Otomatik olarak çıkartılan modeller, müşteri davranışını ve veri üreten süreçleri anlamanıza yardımcı olur, ayrıca nesneleri veya belgeleri veya resimleri otomatik olarak sınıflandırmak için uygulanabilir.

Ayrıca sayısal hedef değişkenlerini tahmin etmek, gözlemlenen zaman serisi verilerinin gelecekteki değerlerini tahmin etmek, müşteri kaybını önlemek, doğrudan pazarlama kampanyalarını optimize etmek, kredi riskini tahmin etmek ve azaltmak, makine arızalarının oluşmadan önce öngörmek ve önlemek için kullanılır. E-posta mesajlarının içeriğine dayalı olarak otomatik olarak yönlendirilmesi ve e-posta spaminin otomatik olarak algılanmasında da kullanılır. Bunun yanı sıra, verilerle daha iyi kararlar alınmasını sağlar ve hatta kararların ve süreçlerin otomatikleştirilmesine yardımcı olur.

Veri madenciliği yalnızca veritabanlarından yapılandırılmış verilere uygulanmaz, bunun yanında metin madenciliği (text mining), bu tekniklerin dokümanlar, haberler, müşteri geri bildirimleri, e-postalar, web sayfaları, İnternet tartışma grupları ve sosyal medya gibi yapılandırılmamış verilere uygulanabilirliğini genişletir. Görüntü madenciliği, ses madenciliği ve video madenciliği bu teknikleri daha başka türdeki verilere uygulanabilirliğini sağlar.

1.1 Veri Madenciliği Kavramı

En basit şekilde veri madenciliği, büyük miktarda veriden bilgiyi çıkarmak olarak ifade edilebilir. Terim aslında eksiktir. Kayadan veya kumdan yapılan altın madenciliği, kaya veya kum madenciliğinden ziyade altın madenciliği olarak adlandırılır. Benzer şekilde veri madenciliği “veriden bilgi madenciliği” olarak adlandırılmış olsa daha anlamlı bir ifade olurdu. “Bilgi madenciliği,” ifadesi tek başına büyük miktarda veriden yapılan madenciliği anlatmakta yeterli olmayabilir. Yine de madencilik, büyük miktarda ham maddeden oluşan küçük bir değerli seti tanımlayan süreci karakterize eden canlı bir terimdir (Şekil 1.3) (Hofmann ve Klinkenberg, 2014).

İstatistikçiler ve veri analistleri veri madenciliği terimini kullanmayı seçseler de bu terim yerine bilgi aktarımı, veri arkeolojisi, enformasyon keşfi, veri örüntü işleme ve

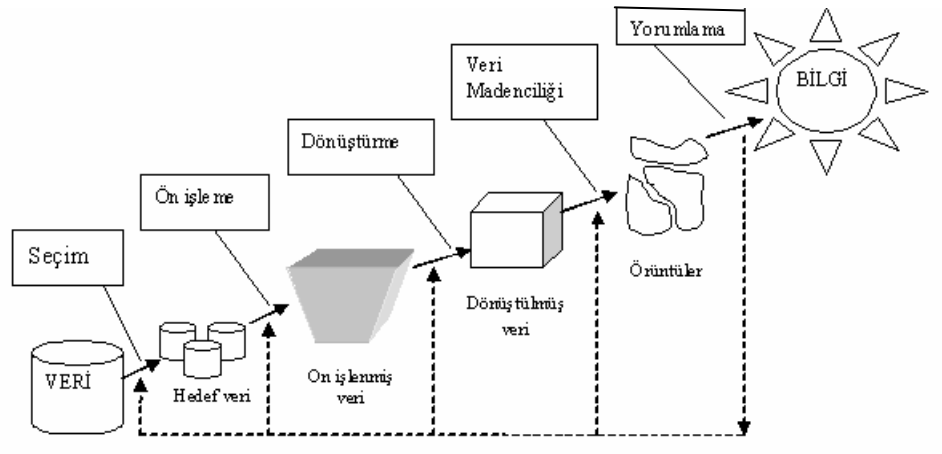
enformasyon hasadı gibi birçok farklı isim kullanılabilir (Koyuncugil ve Özgülbaş,2009).

Pek çok kişi veri madenciliğini popüler olarak kullanılan “veri tabanlarında bilgi keşfi” (VTBK) terimi ile eşanlamlı olarak kullanılır. Yaygın olarak VTKB, ingilizce terimi olan “Knowledge Discovery in Databases” in baş harflerinin kısaltması olan “KDD” şeklinde bilinir. Oysa veri madenciliğini VTBK sürecinin çok önemli temel bir adımıdır. Bu yüzden veri madenciliğini tam olarak tanımlamadan önce veri madenciliği sürecini de içine alan VTKB’yi tanımlamak gerekir.

1.1.1 Veri Tabanlarında Bilgi Keşfi Süreci

Koyuncugil ve Özgülbaş’a göre (2009) ‘Veri tabanlarında bilgi keşfi’ ifadesi ilk defa 1989 yılında VTKB toplantısında kullanılmıştır. Veri madenciliği VTKB sürecinin bir alt kümesidir, şöyle ki yararlı bilginin veriden çıkarılma sürecinin tamamı VTKB ile ifade edilirken bu sürecin tek bir adımı veri madenciliği ile ifade edilir. VTKB, istatistik, yapay zeka, makine öğrenimi, veri tabanları, örüntü tanıma, veri görselleştirme, uzman sistemler gibi alanları da içine alarak gelişmiştir. VTKB, veriden bilginin keşfinin tüm süreçlerine odaklanır, verinin depolanması, veriye erişim, algoritmaların büyük veri setlerine uyarlanması, algoritmaların çalıştırılması, sonuçların yorumlanması, sonuçların görselleştirilmesi VTKB’nin odaklandığı süreçlerdir. Veri madenciliği (VM) ise veriden örüntülerin aktarımı için özel algoritmaların uygulanmasıdır.

Bu süreçte istatistik, makine öğrenimi ve örüntü tanıma gibi teknikler kullanılır. VM, algoritmik anlamda veriden hangi bilgilerin aktarılıp, dikkate alınacağının ifadesi olarak da düşünülebilir (Koyuncugil ve Özgülbaş,2009).



Şekil 1.1. VTKB Sürecinin Adımları

Şekil 1. ‘de VTKB sürecinin adımları gösterilmektedir. Bu süreçler interaktif ve yenilemeli, karaların kullanıcı tarafında verilmesini gerektiren adımlardan oluşur.
VTKB’nin temel adımları aşağıda verilmiştir ((Hofmann ve Klinkenberg,2014).

1. Veri temizleme; Gürültülü ve tutarsız verileri temizleme amaçlı bir basamak
2. Veri birleştirme; Verilerin birleştirildiği basamak
3. Veri seçimi; Analizde kullanılacak verilerin veri tabanından alındığı basamak
4. Veri dönüşümü; Özet veya birleştirme işlemlerini gerçekleştirerek madencilik için verilerin uygun şekilde dönüştürüldüğü veya birleştirildiği basamak
5. Veri madenciliği; Verilerden anlamlı sonuçlar çıkarmak için akıllı yöntemlerin uygulandığı önemli bir basamak
6. Desen değerlendirmesi; Bazı ilginçlik ölçümlerine dayanarak, bilgiyi temsil eden gerçekten ilginç desenlerin tanımlandığı basamak
7. Bilgi sunumu; Görselleştirmenin ve bilgi temsili tekniklerinin, madencilik bilgisinin kullanıcıya sunulması için kullanıldığı basamak.

1'den 4'e kadar olan aşamalar, verilerin madencilik için hazırlandığı farklı veri ön işlem formlarıdır. Veri madenciliği adımı kullanıcı veya bilgi tabanı ile etkileşime girebilir. Desen değerlendirmesi basamağında, ilginç desenler araştırılıp bulunur ve kullanıma sunulur. Ayrıca yeni bilgi olarak saklanabilir. Bu adımlara bakıldığında, her ne kadar veri madenciliği önemli bir süreç olsa da bütün sürecin sadece bir adımıdır.

1.1.2 Veri Madenciliği Tanımı

Veri madenciliği, hedefler doğrultusunda, gelecek ile ilgili öngörüler yapmamızı kolaylaştıracak, büyük miktarda verinin depolandığı veri tabanlarından anlamlı olan veriye ulaşma ve veriyi kullanma işidir (Savaş, Topaloğlu ve Yılmaz, 2012)

Hand (1998), veri madenciliğini geniş veritabanlarında önceden öngörülemeyen bağlantıların ikincil analizi olarak tanımlarken, istatistik, veritabanı teknolojisi, makine öğrenme, örüntü tanıma gibi disiplinlerle etkileşimli yeni bir disiplin olduğunu belirtmiştir.

Alpaydın’a göre VM; gelecekle ilgili öngörülerde bulunmamıza yardımcı olacak bağlantı ve kuralların çeşitli bilgisayar programları kullanarak büyük veri kütleleri içinden, aranmasıdır (2000).

Veri Madenciliğini sade bir şekilde tanımlarsak, istatistikçilerin yıllardır yaptıkları işin otomatikleştirmektir; şöyle ki istatistikçiler istatistiksel açıdan önemli ilişkiler, örüntüler ve bağlantılar aramaktadır, veri madenciliğinin istatistikten farkı ise bu süreci otomatikleştirmesidir. **(**Koyuncugil ve Özgülbaş, 2009).

Veri madenciliği büyük veri yığınlarından ana bilginin çıkarılmasıdır, başka bir şekilde ifade etmek gerekirse, karmaşık veriler içerisinden aklagelmeyen değerli ve ilginç bağlantıların ortaya çıkarılması bilimi olarak da tanımlanabilir (Ganesh, 2002).

1.1.3 Veri Tabanı

Veri tabanı en genel şekilde, hedefe uygun olarak aranje edilmiş veriler topluluğu şeklinde tanımlanabilir. Şöyle ki veri tabanı büyük boyutlarda veriler içeren, ilişkili verilerin depolandığı, ihtiyaç halinde erişime izin veren alanlardır (Zhi-Hua, 2003).

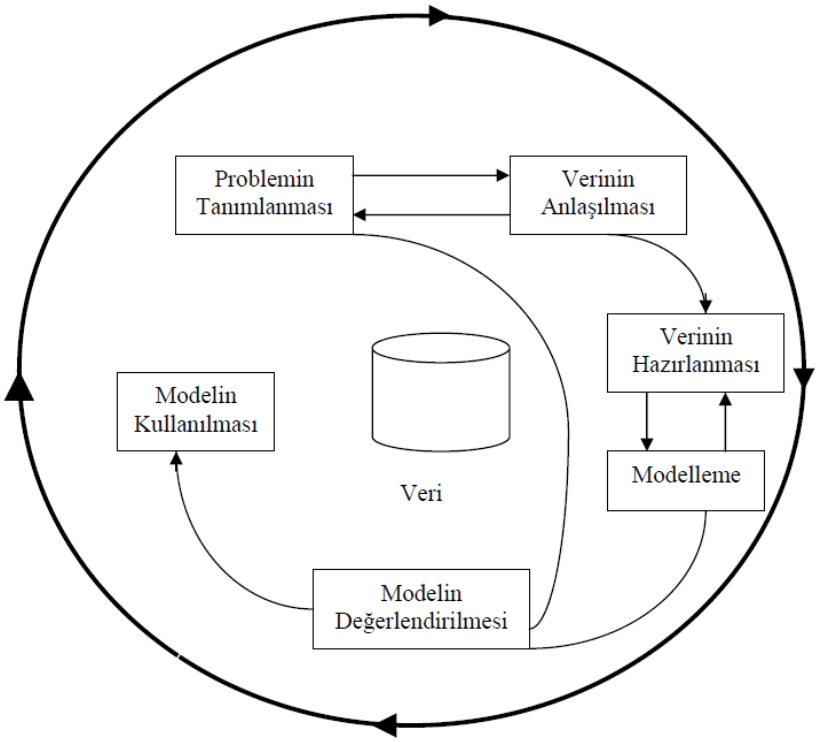
1.2 Veri Madenciliği Süreci

Verideki yararlı ilişkilerin ve örüntülerin metodolojik keşfi, veri madenciliği süreci olarak bilinen bir dizi yineleyici faaliyet tarafından sağlanır. Standart veri madenciliği süreci, (1) problemi anlamak, (2) veri örneklerini hazırlamak, (3) modeli geliştirmek, (4) modelin gerçek dünyada nasıl çalışabileceğini görmek için bir veri setine uygulanması ve (5) üretim dağıtımı aşamalarından oluşur. Veri madenciliği uygulamalarının geliştiği yıllar boyunca, çeşitli madencilik ve ticari kurumlar tarafından veri madenciliği süreci için farklı çerçeveler ortaya atılmıştır.

En popüler veri madenciliği süreci çerçevelerinden biri, Veri Madenciliği için Cross Endüstri Standart Süreci için bir kısaltma olan CRISP-DM'dir. Bu çerçeve, veri madenciliğine dahil olan üç şirketin bir konsorsiyumu tarafından geliştirilmiştir (Chapman ve diğ., 2000). Bu şirketler sırasıyla Daimler Chrysler, SPSS ve NCR'dir.

CRISP-DM süreci, veri madenciliği çözümleri geliştirmek için en yaygın olarak benimsenen çerçevedir. Şekil 1.2, CRISP-DM çerçevesinin görsel bir görünümünü sunmaktadır.

Tüm bu çerçeveler ortak özellikler sergiler ve dolayısıyla CRISP-DM sürecine benzeyen genel bir çerçeve kullanıyor olacağız. Herhangi bir süreç çerçevesinde olduğu gibi, bir veri madenciliği süreci en uygun çıktıyı elde etmek için belirli bir görev grubunun performansını önerir (Desphande ve Kotu,2014).



Şekil 1.2 CRISP-DM Veri Madenciliği Süreci

Verilerden bilgi çıkarma işlemi tekrarlanır. Veri madenciliği sürecindeki adımlar doğrusal değildir ve adımlar arasında geri ve ileriye giden birçok veri döngüsüne sahiptir ve bazen veri madenciliği problemini yeniden tanımlamak için ilk adıma dönmektedir.

Temel amaç veri madenciliği içeren herhangi bir sürecin analizi sorusunu ele almaktır.

Eldeki problem, hastalıkların sınıflandırılması veya bir hastalığın tahmin edilmesi olabilir.

Aşağıda veri madenciliği süreçleri ayrıntılı olarak incelenecektir.

1.2.1 Problemin Tanımlanması ve Verinin Anlaşılması

Problemin tanımlanması aşamasında analizin hangi hedef için yapılacağı net bir şekilde tanımlanır. Bu hedefin hangi problemi çözeceği ve ne işe yarayacağı net bir dille ifade edilmelidir (Tüzüntürk,2010). VM sürecinin en önemli aşaması bu aşamadır. VM uygulamalarının başarıya ulaşması için ilk önce, uygulamanın amacı açık ve net bir şekilde tanımlanmalıdır.

Bu aşamada önemli konulardan bir tanesi de elde edilecek sonuçların başarı seviyelerinin hangi yolla ölçülebileceğinin de iyi bir şekilde tanımlanmasıdır (Şimşek, 2006).

Sağlık ve tedavi verileri artık elektronik ortamlarda kolayca saklanabilmektedir. Tıpta ve sağlıkta VM uygulamlarının başlıcaları aşağıdaki gibi sıralanabilir; (Koyuncugil ve Özgülbaş, 2009) .

- Doğru sağlık politikalarının izlenmesi için sağlık hizmetlerinin kalitesinin arttırılması.
- Salgın hastalıkların tespit edilmesi ve gerekli önlemlerin alınması.
- Hastalıklara erken teşhis koyulmasının amaçlanması, böylece koruyucu hekimliğin geliştirilmesi ve doğru tedavi yöntemlerinin seçilmesi.
- Hastalıkları etkileyen faktörlerin ortaya çıkartılması (kanser vs).
- Koruyucu hekimliğin yaygınlaştırılması ve sağlık harcamalarının düşürülmesi.
- Sağlık harcamalarında yapılan sahtekarlıkların ortaya çıkarılmasının sağlanması ve maliyetin düşürülmesi
- Bir kadın hastanın, genetik mirası, alışkanlıkları, sağlık durumu vb değişkenlere bağlı olarak meme kanserine yakalanma riskinin tahmini.
- Sağlık veri tabanlarının ilaç geliştirici firmalar tarafından etkin ve doğru kullanılarak, doğru ilaçların geliştirmesinin sağlanması.

Bu aşamada veri madenciliği uygulamasının ne için yapılacağının iyi tanımlanması gerekir çünkü hedefin doğru konması uygulama sonuçlarının başarısını olumlu yönde etkiler (Şimşek,2006). Bir veri madenciliği projesinin başarılı olması için projenin gerçekleştirilebilir bir hedefinin olması, projenin dikkatle planlanması ve ölçülebilir bir hedefinin olması gerekir (Şimşek.2006).

1.2.2 Verinin Hazırlanması

Verinin hazırlanması aşaması veri madenciliği sürecinin en önemli aşamalarından biridir çünkü bu aşamada yapılan herhangi bir yanlışlık modelin kurulması aşamasında hataya sebep olarak tekrar bu aşamaya geri dönülmesine sebep olur (Akpınar, 2010).Her veri analizi süreci veri setinin toplanması, tanımlanması ve temizlenmesiyle başlar. Bu süreçten doğru bir şekilde yapıldıktan sonra ancak veriler analiz edilebilir ve doğru sonuçlara ulaşılabilir (Dasu ve Johnson, 2003).

Veri madenciliğinde kaliteli veri çok önemli bir konudur. Veri ön işleme süreci de veri madenciliğinde güvenilirliğin arttırılması için çok önemli bir süreçtir (Oğuzlar, 2003). Veriler aşağıdaki aşamalardan geçerek hazırlanır;verinin toplanması, verinin birleştirilmesi ve temizlenmesi ile verinin dönüştürülmesi.

1.2.2.1 Verinin Toplanması

Veri madenciliği modeli oluşturma sürecinde, verinin hazırlanması sürecindeki ilk adım verinin toplanmasıdır. Bu adımda problemin tanımlanması aşamasında tanımlanan problem için gerekli olduğuna karar verilen veriler toplanır ve verilerin toplanacağı veri kaynaklarına da karar verilir (Akpınar,2000). Bu adımda karar verilmersi gereken en önemli nokta toplanması gereken verilerin belirlenmesidir. Bu adımda da analizi yapıyor olmadaki ana amaç çerçevesinde hangi verinin toplanacağına karar vermek önemlidir.

Verinin toplanacağı kaynakların önceden belirlenmesi ve bu kaynakların güvenilirlikleri konusunda emin olunması, ileriki aşamalarda problemlerle karşılaşılma riskini önemli ölçüde azaltmaktadır.

1.2.2.2. Verilerin Temizlenmesi ve Birleştirilmesi

Verilerin temizlenmesi ve birleştirilmesi aşaması, verilerdeki tutarsızlıkların yok edilmesi, aykırı değerlerin ayırt edilmesi, eksik verilerin tamamlanması gibi işlemlerden oluşmaktadır. Eğer farklı veri tabanlarından veri alınacaksa ve birleştirilecekse bu aşamada gerçekleştirilir (Tüzüntürk, 2010).

Veri ambarı en temel anlamıyla, farklı kaynaklarda depolanmış verilerin ortak bir platformda birleştirilerek, tutarlı ve doğru bir şekilde, aynı ölçü ve zaman birimi boyutunda tutulduğu bir sistemdir (Şentürk, 2006). Veri uyuşmazlıklarının en büyük sebebi veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanmasıdır. Bu uyuşmazlıkların belli başlıları veriler arasında zaman tutarsızlıkları, ölçü birimi tutarsızlıkları ve kodlama farklılıkları şeklindedir. Farklı kaynaklardan toplanan verinin birleştirilmesi ve temizlenmesi gerekmektedir (Moss ve Atre,2003).

Tablo 1.1. Hastalık Durumunu Gösteren Olası Kodlama Biçimleri

Hasta	**Hasta Değil**
hasta	hasta değil
HASTA	HASTA DEĞİL
1	0
0	1
1	2

Örneğin bir kişinin kanser hastası olup olmama durumunun bilgisi gerekiyorsa ve bilgi farklı servisler tarafından toplanacaksa (dahiliye, genel cerrahi vs) bu servisler farklı kodlamalarla verilerini tutuyor olabilirler. Bu aşamada analizi uygulayan kişi gerekli dönüşümleri yapmalıdır.

Ayrıca gözden kaçırılmaması gereken bir başka nokta ise veride bulunan kayıp veya eksik değerlerdir. Örnek olarak, tablo 1.1.' de hastalık durumunu incelediğimiz bu kişilerin bazılarının veri tabanında yaş bilgisi mevcutken, bazı kayıtlarda bu yaş bilgisi hiç olmayabilir ya da eksik olabilir.

Bu örnekteki gibi eksik veriler "kayıp veriler" şeklinde tanımlanabilir. Bunun haricinde, bazı kayıtlarda aşırı uç değerler (outlier) veya yanlış girilmiş değerler olabilir. Bu tür bilgilere de gürültü (noise) adı verilmektedir. İdeal olan eksik verilerin olabildiğince tamamlanmasıdır.

Bir diğer alternatif ise eksik bilgilerin tahmin yöntemiyle tamamlanmasıdır. Veri analiz aşamasında yanlış girilen değerler önemsenmezse, analiz sonuçlarının doğruluğu ve güvenilirliği etkilenir. Aykırı değerlerin gözden geçirilip atıldıktan sonra yani verinin gürültüden arındıktan sonra analize sokulması çok önemlidir (Şimşek,2006).

Eğer bir veri tabanında kayıtlı kişilerin aynı şeyi ifade eden iki değişkene dair de bilgileri bulunuyorsa örneğin hem yaş hem doğum tarihi gibi aynı şeyi ifade eden değişkenler bulunuyorsa veride, bu değişkenlerden biri veriden çıkarılmalıdır. Böyle bir durumda iki değişkenin de veride tutulması saçmadır. Bir tanesinin işleme alınması yeterlidir.

1.2.2.3 Verilerin İndirgenmesi

Büyük hacimli veri kümesinden daha küçük hacimli veri kümesinin elde edilmesidir. Eğer veri kümesi çok fazla değişkenden oluşuyorsa ve bu değişkenlere oranla örnek sayısı azsa hiç indirgeme yapmadan yapılan analizin doğruluk değeri düşük olabilir.

Çünkü veri setindeki bazı değişkenlerin veri ile olan ilişkisi çok azdır; veriyi açıklama veya sınıflandırma oranı çok düşüktür.

1.2.3 Modelin Kurulması

Veri madenciliğinde bilgi kaynaklarından en fazla verimin alınabilmesi için, modelin kurulması aşaması çok önemlidir. Modelin iyi kurulması, analiz sonucunda elde edilecek sonuçların kalitesini de önemli ölçüde etkiler. İyi bir VM analisti, analiz bittikten sonra hangi ilişkilerin bulunabileceğini tahmin edebilmelidir. Veri hazırlama ve model kurma süreçleri en iyi modele ulaşıncaya kadar devam eder bu yüzden yenilenen bir süreçtir; en uygun model bulununcaya kadar çok sayıda modelin denenmesi gerekebilir (Savaş, Topaloğlu ve Yılmaz,2012)
Eğer model doğru kurulmazsa, veri seti içerisinde bulunabilecek kritik ilişkiler doğru bir şekilde sunulamaz ve önemli örüntüler tespit edilemez. Dolayısıyla modelden başarılı sonuç elde etme olasılığı da azalır.

1.2.4. Modelin Kullanılması

Kurulan modelin önce geçerliliğinin kabul edilmesi gerekir. Kurulan model ya başka bir uygulamanın alt parçası olarak kullanılabilir ya da kurulan bu model doğrudan uygulanır (Savaş, Topaloğlu ve Yılmaz,2012).

1.2.5. Modelin İzlenmesi

Kurulan model izlenmelidir. Zaman içerisinde kurulan model ürettikleri verilerde ortaya çıkan değişikliklere bağlı olarak yeni bir düzenlemeye ihtiyaç duyabilir (Tüzüntürk, 2010).

1.3 Veri Madenciliği ve İlişkide Olduğu Disiplinler

Veri madenciliği, Şekil 1.4'de görüldüğü gibi istatistik, makine öğrenimi, veri tabanı sistemleri, veri görselleştirme, yapay sinir ağları gibi konuları kapsayan disiplinlerarası bir alandır.

Şekil 1.3 Veri Madenciliği ve İlişkide Olduğu Disiplinler (Savaş, Topaloğlu and Yılmaz, 2012)

1.3.1 Veri Madenciliği ve İstatistik

İçinde bulunduğumuz bilgi çağında, teknolojinin gelişmesi ile hem boyut hem de karmaşıklık açısından artan istatistiksel problemlere çözüm bulunabilmektedir. Teknolojideki bu gelişmeler yeni bir alan olan veri madenciliğine işaret etmiştir.

Veri madenciliği ve İstatistiksel analizin hem benzer hem de farklı yönleri vardır; ikisinin farklılaştığı noktalar aşağıdaki tabloda özetlenmiştir (Moss ve Atre, 2003).

İstatistiksel teknikler özellikle verilerin indirgenmesi ve modellenmesi gibi temel veri ön işleme basamaklarında ve sonuçların değerlendirilmesi veya yorumlanması basamaklarında katkı sağlamaktadır.

İstatistik ile veri madenciliğinin birçok açıdan benzeştiği söylenebilir (Zhao ve Luan 2006). İkisinin de ortak yanı "verinin bilgiye dönüştürülmesi" (Kuonen, 2004) veya "veriden öğrenilmesi"dir (Ganesh, 2002) Her ikisinde de amaç verilerden doğru çıkarım yapmaktır. Her ikisi de belirsizliklerin üstesinden gelmek ve gelecekteki olaylar hakkında ön görüde bulunmak için bulunan araçlardır.

Her ikisi de bir olayı etkileyen önemli faktörleri belirler ve gelecekteki olayları daha iyi ön görmek için model üretirler (Tüzüntürk, 2010).

Tablo 1.2 Veri Madenciliği ve İstatistik Arasındaki Fark

İstatistiksel Analiz	**Veri Madenciliği**
İstatistikçiler genellikle bir hipotez ile başlarlar.	Veri Madenciliği hipoteze gerek duymaz.
Hipotezlerini eşleştirmek için kendi eşitliklerini geliştirmek zorundadırlar.	Veri Madenciliği algoritmaları eşitlikleri otomatik olarak geliştirir.
İstatistiksel analizler sadece sayısal verileri kullanır.	Veri Madenciliği farklı tiplerde data kullanır (örneğin metin, ses) sadece sayısal veriyi değil.
Kirli veriyi analizleri sırasında bulur ve filtre eder.	Veri Madenciliği temiz veriye dayanır.

1.3.2 Veri Tabanı Sistemleri

Veri tabanını "düzenli biçimde bilgisayar belleklerinde saklanan birbiriyle ilişkili veriler topluluğu" olarak tanımlamak mümkündür (Şimşek,2006). Kısaca veri tabanını en basit şekilde, belirli bir amaca yönelik düzenlenmiş veriler topluluğu şeklinde tanımlayabiliriz.

Biraz daha ayrıntıya inilirse veri tabanı (database), birbirleriyle ilişkisi olan verilerin depolandığı, kullanım amacına uygun olarak düzenlenmiş veriler topluluğunun saklandığı ve gerektiğinde tekrar bir erişime olanak sağlayan, büyük boyutlarda veriler içeren bilgi depoları şeklinde tanımlanabilir (Zhi-Hua, 2003).

Veri tabanındaki veriler uygulamada pek çok kuruluş (banka, şirket, fabrika, Genel Müdürlük, Bakanlık, özel veya kamu kuruluşları) tarafından kullanılırlar.

1.3.3 Yapay Sinir Ağları

Yapay sinir ağları (YSA) örneklerle ilgili bilgiler toplamakta, genellemeler yapmakta ve daha sonra hiç görmediği örnekler ile karşılaştırılınca öğrendiği bilgileri kullanarak o örnekler hakkında karar verebilmektedir. Yapay sinir ağları bu öğrenebilme ve genelleme özellikleri nedeniyle günümüzde birçok bilim alanında geniş uygulama olanağı bulmakta ve karmaşık problemleri başarı ile çözebilme yeteneğini ortaya koymaktadır (Ergezer vd, 2003).

YSA, insan beyninden esinlenerek geliştirilmiş, her biri kendi belleğine sahip işlem elemanlarından oluşan, biyolojik sinir ağlarını taklit eden bilgisayar programlarıdır (Elmas, 2003). Yapay sinir ağları veri madenciliği algoritmaları ile kıyaslandığında daha yavaş eğitilir.

1.3.4 Makine Öğrenmesi

Makine öğrenimi daha geniş olarak “Bilgisayarın bir olay ile ilgili bilgileri ve tecrübeleri öğrenerek, gelecekte oluşacak benzeri olaylar hakkında kararlar verebilmesi ve problemlere çözümler üretebilmesidir” şeklinde tanımlanabilir (Paul ve Watson, 1998) Uzman sistemler, genetik algoritmalar, bulanık mantık, yapay sinir ağları gibi makine öğrenmesi de, yapay zeka teknolojilerinden biridir.

“Makine öğrenmesi, bir problemi o probleme ait veriye göre modelleyen bilgisayar algoritmalarının genel adıdır” (Atalay ve Çelik, 2017). Mevcut veri seti ve kullanılan algoritma ile oluşturulan model, en yüksek performansı vermek üzere kurulmaktadır. Bu sebeple pek çok makine öğrenmesi yöntemi geliştirilmiş olup bunlardan bazıları; k-en yakın komşu algoritması, basit (naif) bayes sınıflandırıcı, karar ağaçları, lojistik regresyon analizi, k-ortalamalar algoritması, destek vektör makinaları ve yapay sinir ağlarıdır.

Bu yaklaşımların bir kısmı tahmin ve kestirim, bir kısmı kümeleme ve bir kısmı da sınıflandırma yapabilme yeteneğine sahiptir (Atalay ve Çelik, 2017).

1.3.5 Veri Görselleştirme

Veri görselleştirme teknikleri veriyi resme dökerek veri hakkında genel bir kanıya varılmasını sağlar (Carlis ve Kanstan,1998). Veri görselleştirmedeki ilk amaç kavramların, fikirlerin ve kuralların daha iyi anlaşılmasını sağlamaktır.

Veri görselleştirmenin ikinci amacı ise resimlerden ve grafiklerden yeni ilişkiler bulmak, yeni yapılar üretmek, bir hipotezin doğruluğunu test etmektir. İnsan 3 boyutu algılayabilir daha büyük boyuttaki verileri algılayamaz. Veri görselleştirme yöntemleri veriler arasındaki ilişkiyi koruyarak, çok boyutlu veriyi 2 veya 3 boyuta indirgeyerek görselleştirebilme yeteneğine sahiptir (Atalay ve Çelik, 2017). Veri görselleştirmede kısaca mantıksal problemlere çözüm bulmak için insanın görsel algılama sisteminden yardım alınır (Faloutsos, 1996).

1.3.6 Veri Ambarı

Veri madenciliği ile ilişkili diğer disiplinler olarak veri ambarı ve genetik algoritmalar sayılabilir.

Veri ambarları, özel bir veri tabanıdır; veri madenciliği sürecinin yürütüldüğü veriyi sağlarlar. Veri ambarları veri tabanlarının birleştirilmesiyle meydana gelir. Veri

ambarı, pek çok farklı yapıdaki verinin farklı kaynaktan depolandığı ve hepsinin aynı platformda kullanılmasının düşünüldüğü yapılardır (Koyuncugil ve Özgülbaş, 2009).
Verilerin toplandığı, işlendiği ve depolandığı her yerde veri kalitesini sağlamak önemli bir sorundur. Analizde kullanılacak olan verinin doğru olup olmadığını veya hatalı girişi olup olmadığını bilemeyiz. Verilerdeki bu hatalar, modelin temsil gücünü etkileyecektir.
Kuruluşlar, veri kalitesini iyileştirmek ve yönetmek için veri temizleme ve dönüştürme teknikleri kullanır ve bunları Veri Ambarı adı verilen şirket içi depolarda saklarlar.
İyi veri ambarlarından elde edilen veriler, uygun kontroller yapıldığı için kalitelidir, ayrıca yeni ve mevcut veriler için geçerli bir veri doğruluğu düzeyini sağlarlar. Ne olursa olsun, belirli bir veri kalitesi derecesini sağlamak için modeller oluşturmadan önce veri ve iş hakkında önceden bilgi edinilmesinin yanında, veri araştırma teknikleri kullanılarak verilerin kontrol edilmesi önemlidir (Desphande ve Kotu,2014).
Veri ambarları veri kümelerine VTKB aşaması için veri temizleme ve veri erişimi açısından yardımcı olurlar (Koyuncugil ve Özgülbaş, 2009).
Veri ambarları tipik olarak doğrudan veri madenciliğine izin vermeyi amaçlamamakla birlikte, veri madenciliği projeleri iyi tasarlanmış bir veri deposundan büyük ölçüde faydalanabilecektir, çünkü veri kalitesi veveri birleştirilmesi entegrasyonu ile ilgili birçok konu her ikisi için de çözülmelidir. Veri madencilerinin bakış açısına göre, veri ambarları, heterojen operasyonel verilerden, veri madenciliği için gerekli olan tek, birleştirilmiş bir analiz tablosuna giden yolda bir ara adım olarak görülebilir (Hoffman ve Klinkenberg,2014).
Veri ambarı uygulamaları genellikle OLAP (Online Analitik Prosesler)
olarak adlandırılmaktadır. OLAP araçları, birçok açıdan SQL'den üstündür. OLAP araçlarının odaklandığı şey, çok boyutlu veri analizi sağlamadır**.** Bu araçlar aynı zamanda interaktif veri analizi sağlama ve basitleştirmeyi hedeflemişlerdir. (Koyuncugil ve Özgülbaş, 2009).
Veri ambarı normalize edilmemiş veriyi saklar ve veri ambarı veri geçmişini korur. Veri ambarları iş anlayışı, geçmiş trendler, anlaşılabilirlik, verimin arttırılması ve gelişmiş veri tutatlılığı açısından önem arz eder.
Veri madenciliği, veri sorgusu ve OLAP'ın kullanım alanlarının farklılıklarına bakarsak;
Büyük veri tabanı ile çalışılmak istendiğinde ve ne arandığı bilindiğinde veri sorgusu kullanılmalıdır.
Büyük veri tabanlarında basit örüntüler bulunmak istendiğinde, OLAP kullanılmalıdır.
Veri madenciliği ise veri içerisinde açık bir şekilde göze çarpmayan, ilginç, gözle görülemeyen örüntü ve ilişkiler bulunmak isteniyorsa kullanılmalıdır. Fakat veri

madenciliği algoritmaları veri sorgusu ve OLAP'e göre yavaştır. Bu nedenle genellikle veri tabanının küçük olması gerekmektedir (Koyuncugil ve Özgülbaş, 2009).

1.4. Sağlıkta Veri Madenciliği Uygulama Alanları

Veri madenciliği sağlık endüstrisine birçok fayda sağlar. Veri Madenciliği, sağlık araştırmacılarının değerli kararlar almasına yardımcı olur. Aşağıda Veri Madenciliğinin sağlık hizmetlerinde uyguladığı çeşitli uygulamalar sıralanmıştır:

1.4.1.Hastane Kaynağının Etkin Yönetimi

Veri madenciliği, sağlık hizmetlerinde önemli bir görev olan hastane kaynaklarını yönetmeyi sağlamak için bir model oluşturulmasına destek sağlar.

Veri madenciliğini kullanarak, kronik hastalığı tespit etmek ve hastaların hastalığının komplikasyonuna dayanarak hastaları önceliklendirmek, böylece zamanında ve doğru bir şekilde etkili tedaviyi almalarını sağlamak mümkündür. Hastaların sağlık raporu ve demografik bilgileri de mevcut hastane kaynaklarını etkin bir şekilde kullanmak için yararlıdır. Veri madenciliği kullanan otomatik bir araç, Alapont ve arkadaşları tarafından, fiziksel ve insan kaynakları gibi hastane kaynaklarını yönetmek için önerilmiştir (Alapont v.d., 2005)**.**

Grup Sağlık Kooperatifi, veri madenciliği teknikleri kullanılarak daha düşük maliyetle çeşitli sağlık hizmetleri sunmaktadır (Koh ve Tan,2005).

Bu kooperatif, hastaların tıbbi bilgilerini çevrimiçi olarak kullanmalarını, çevrimiçi olarak reçete formunu doldurmalarını ve e-postanın sağlık hizmeti sağlayıcısıyla güvenli bir şekilde paylaşılmasını sağlayan kar amacı gütmeyen bir sağlık hizmetleri organizasyonudur. Seton Tıp merkezi ayrıca sağlık hizmeti kalitesini arttırmak, hastanın sağlığı ile ilgili çeşitli detaylar sağlamak ve hastaların muayne için kayıt sürelerini kısaltmak için hastanelerinde veri madenciliği kullanmışlardır (Dakins,2001).Veri madenciliği yardımıyla Blue Cross, hastalıkların etkin bir şekilde yönetilmesi, sonuçların iyileştirilmesi ve tedavi maliyetlerinin düşürülmesi için bir sistem sunmaktadır. Sierra Health Center tedavi için kılavuz bilgiler sağlar, tedavi maliyetini yönetir ve veri madenciliği kullanarak sağlık kalitesini iyileştirmek için alanları tespit eder (Schuerenberg,2003).

1.4.2 Hastane Sıralaması

Farklı veri madenciliği yaklaşımları, hastanelerin sıralamasını belirleyebilmek için gerekli olan çeşitli hastane detaylarını analiz etmek için kullanılmaktadır (Mary ve Mat,2004). Hastanelerin sıralaması, yüksek riskli hastaların üstesinden gelebilme yeteneklerine dayanarak yapılmaktadır. Daha yüksek sıradaki hastane, yüksek riskli hastayı öncelikli olarak ele alırken, alt sıradaki hastane risk faktörünü dikkate almamaktadır.

1.4.3 Daha İyi Müşteri İlişkisi

Veri Madenciliği, sağlık kurumunun müşterisi olan hastalarla daha iyi ilişki kurabilmek için onların ihtiyaçlarını, tercihlerini, davranışlarını ve kalitesini anlamasına yardımcı olur.
Veri Madenciliğini Kullanarak, Müşteri Potansiyel Yönetimi A.Ş., tüketici sağlık hizmetlerinin kullanımını temsil eden bir endeks geliştirmiştir . Bu endeks, müşterinin belirli sağlık hizmetlerine olan etkisini tespit etmeye yardımcı olur.

1.4.4 Hastane Enfeksiyon Kontrolü

Enfeksiyon kontrol verilerinde bilinmeyen veya düzensiz paternleri (örüntü, ilişki) bulmak için veri madenciliği teknikleri kullanılarak bir inceleme sistemi oluşturulmuştur. (Mary ve Mat,2004). Birliktelik kuralları kullanılarak halk incelemesi ve hastane kontrol verilerinden beklenmedik ve ilginç bilgiler elde edilmiştir. Hastanelerdeki enfeksiyonu kontrol etmek için, bu bilgiler bir uzman tarafından daha ayrıntılı olarak incelenir.

1.4.5 Daha Akıllı Muayene Teknikleri

Veri Madenciliği kullanılarak, hekimler ve hastalar farklı tedavi teknikleri açısından kolaylıkla karşılaştırılabilir. Hekimler bu yolla mevcut tedavilerin etkinliğini analiz edebilir ve hangi tekniğin daha iyi ve uygun maliyetli olduğunu keşfedebilirler. Veri Madenciliği ayrıca, hekimlerin belirli tedavilerin yan etkilerini belirlemelerine, tehlikeyi azaltmaya ve tedavi için akıllı metodolojiler geliştirmeye uygun karar vermelerine yardımcı olur.

1.4.6 Geliştirilmiş Hasta Bakımı

Elektronik sağlık kaydındaki ilerlemeyle büyük miktarda veri toplanabilir. Dijitalleştirilmiş formda mevcut olan hasta verileri, sağlık sistemi kalitesini iyileştirir. Bu büyük verileri analiz edebilmek için, bu büyük verilerden ilginç bilgileri keşfen ve sağlık hizmeti kalitesinin iyileştirilmesine yönelik kararlar alan veri madenciliği kullanılarak bir öngörü modeli oluşturulmuştur. Veri madenciliği, sağlık hizmeti sağlayıcılarının hastalarının mevcut ve gelecekteki ihtiyaçlarını ve onların memnuniyet seviyelerini artırma tercihlerini belirlemelerine yardımcı olur.

Milley (2000) ayrıca, veri madenciliğinin, sağlık kuruluşu tarafından sağlanan hizmetlerin iyileştirilmesi için belirli hastaların gereksinimlerini belirlemek için yararlı olduğunu önermiştir. Hallick (2001), veri madenciliği tekniklerinin hastaya çeşitli hastalıklar ve bunların önlenmesine ilişkin bilgileri sağlamasında yardımcı olduğunu ileri sürmüştür. Kolar, sağlık kuruluşunun hasta gruplaması için veri madenciliği tekniklerini kullandığını tespit etmiştir (Kolar,2001).

1.4.7 Sigorta Sahteciliğini Azaltma

Sağlık sigortası, veri madenciliği tekniklerini kullanarak tıbbi davalarda sahtekarlığı ve kötüye kullanımı tespit etmek için bir model geliştirir. Bu model, doktorlar, hastalar, hastaneler vb. Tarafından yapılan tıbbi davalarda uygunsuz reçetelerin, düzensiz veya sahte kalıpların tanımlanmasında faydalıdır. ABD vergi mükellefleri ayrıca 1997 yılında hastane faturalarındaki dolandırıcılık nedeniyle yüzlerce dolar kaybettiklerini bildirmiştir. ReliaStar finansal kuruluşu Dolandırıcılık ve suistimali tespit ederek yıllık tasarruflarını % 20 oranında artırmıştır. Doktor reçeteleri ve tedavi materyalleri çok miktarda veri üretmektedir.

Medicaid Dolandırıcılığından Utah Bürosu, sahtekarlığı tespit etmek için gizli ve yararlı bilgileri bulmak amacıyla bu verileri kullanmıştır (Milley,2000). Avustralya Sağlık Sigortası Komisyonu da bu büyük verileri analiz ederek ve milyonlarca dolarlık tasarruf sağladığını bildirmiştir (Christy, 1997). Texas Medicaid Sahtecilik ve Kötüye Kullanım Tespit Sistemi de dolandırıcılık ve kötüye kullanımı keşfetmek ve 1998 yılında milyon dolar tasarruf için veri madenciliği tekniklerini kullanmıştır (Anonymous,1999).

1.4.8 Yüksek Riskli Hastaları Tanımak

Amerikan Sağlık Sistemi, yüksek risk taşıyan hastaları tanımak için veri madenciliği kullanan bir tahmin modeli oluşturur. Bu sistemin temel amacı, diyabetik hastalarını tedavi etmek, bu hastaların sağlık kalitelerini yükseltmek ve hastaya maliyet tasarrufu hizmetleri sunmaktır. Tahminsel model kullanılarak, sağlık hizmeti sunan hasta, diğer hastalarla kıyaslandığında daha fazla bakım gerektiren hastayı tanır (Ridinger, 2002).

1.4.9 Sağlık Politikası Planlaması

Veri madenciliği, sağlık hizmetlerinin kalitesini düşürmenin yanı sıra sağlık hizmetlerini iyileştirmek için etkili bir sağlık hizmeti politikası oluşturulmasında önemli bir rol oynamaktadır. COREPLUS ve SAFS modelleri, hastaneler tarafından sağlanan tıbbi bakım hizmetlerinin sonuçlarını ve tedavi maliyetlerini analiz etmek için veri madenciliği teknikleri kullanılarak geliştirilmiştir (Tomar ve Agarwal, 2013).

1.5 Tıpta Veri Madenciliği Literatürü

Aşağıda sağlık alanında yapılmış bazı veri madenciliği çalışmalarına yer verilmiştir; MammaPrint TM kanser araştırmaları konusunda en başarılı örneklerden biri olarak gösterilebilir. Bu çalışmada, meme kanserine yakalanmış Amerikalı kadınlar üzerine araştırma yapılmıştır (Laura v.d., 2002) Bu çalışmada kadınlarda meme kanserinin metestaz yapma olasılığı üstünde durulmuştur.
(Marc v.d., 2002)
Ge ve Wong'un yaptığı çalışmada pankreas kanserinin teşhisi için kullandıkları C4.5 karar ağacı algoritmasının performansını, 6 farklı karar ağacı tabanlı sınıflandırıcı gruplama teknikleri (ensemble) ile karşılaştırmışlardır. Sonuç olarak sınıflandırıcı gruplama tekniklerinin pankreas veri seti için her zaman daha büyük doğruluğa (accuracy) ve daha küçük tahmin hatalarına sahip olduğunu ortaya çıkarmışlardır (2008).

1.6 Sağlık Hizmetlerinde Veri Madenciliği Zorlukları

Veri madenciliğinin sağlık hizmetlerinde en önemli zorluklarından biri, kalite ve ilgili tıbbi verileri elde etmektir. Kesin ve eksiksiz sağlık verileri elde etmek zordur. Sağlık

verileri doğası gereği karmaşık ve heterojen bir yapıya sahiptir, çünkü tıbbi laboratuvar raporları, hasta ile yapılan görüşmelerden veya hekimin muayenesi gibi çeşitli kaynaklardan toplanmıştır.

Sağlık hizmeti sağlayıcısı için, verilerin kalitesinin korunması önemlidir, çünkü bu veriler hastalara düşük maliyetli sağlık bakımı sağlamak için yararlıdır. Kaliteli veriler olmadan faydalı bir sonuç yoktur.

Başarılı veri madenciliği için, tıbbi verilerdeki karışıklık, tıbbi verilerin analiz edilmesinde önemli bir engeldir. Bu nedenle, veri madenciliğinde etkili bir karar verebilmek için kaliteli ve doğru verilerin elde edilmesi önemlidir. Sağlık hizmeti verisiyle ilgili başka bir zorluk da veri paylaşımıdır. Sağlık kuruluşları gizlilik nedeniyle verdikleri bilgileri paylaşmaya isteksizdir. Hastaların çoğu sağlık verilerini açıklamak istememektedir.

Böylece, Sağlık Bakım Organizasyonu ve Sağlık Sigortası Organizasyonu, hasta mahremiyetini korumak için verilerini dağıtmamaktadır.

Bu sağlık sigortasında dolandırıcılık tespit çalışmalarında engel teşkil etmektedir. Veri ambarının başlangıç maliyeti çok yüksektir. Veri madenciliği tekniklerini sağlık hizmeti verilerinde uygulamadan önce, farklı kaynaklardan gelen verileri, merkezi bir veri ambarına toplamak ve kaydetmek önemlidir. Bu da masraflı ve zaman alıcı bir süreçtir. Hatalı veri ambarı tasarımı etkili veri madenciliğinin önünde büyük bir engeldir.

İKİNCİ BÖLÜM
PROSTAT KANSERİ VE GENOM

2.1 Prostat Kanseri

Prostat**,** erkeklerde semenin (meni) yaklaşık olarak %30-40'lık kısmını oluşturan ve koyu kıvamlı salgı yapan bir aksesuar bezdir. Prostat bezi, ters dönmüş yuvarlak koni şeklinde olup ortalama ağırlığı yetişkin erkeklerde 18-20 gr arasında değişmektedir. Prostat bezi yerleşim olarak, mesanenin (idrar torbası) altında, rektumun (kalın barsağın son bölümü) ise önündedir. Mesaneden idrarın boşaltılmasını sağlayan idrar kanalının (üretra) ilk kısmını çepeçevre sarar.

Kanser, hücre büyümesi ve bölünmesi gibi normal hücre döngüsünü düzenleyen mekanizmalardaki oluşan bir sorun sonrası görülen, hücrelerin kontrolsüz veya normal olmayan bir şekilde büyümesi ve çoğalması ile karakterize patolojik bir durumdur. Normal şartlarda hücrelerin büyümesinde, çoğalma ve başkalaşım olarak isimlendirilen iki ana olay vardır. Hücrelerde bu iki ana olay iyi ayalarlanmış olup her zaman kontrol altındadır. Ancak bu iki ana olaydan bir veya her ikisi kontrol dışına çıkarsa, normal yapıdaki hücrelerin kanser hücresine dönüşme riski ortaya çıkar. Kanser hücrelerinde hem çoğalma esnasında, hem de yapısal ve işlevsel açıdan farklılıklar ortaya çıkabilmektedir. Temel olarak kanser hücrelerinin başlıca 3 önemli ana özelliği vardır. Bunlar; kontrolsüz çoğalma, çevre dokulara ve uzak dokulara yayılımdır.

Prostat kanseri ise prostat hücrelerinin kontrolsüz ve normal olmayan bir şekilde büyümesi ve çoğalmasıyla karakterize kötü huylu bir hastalıktır ve tüm dünyada erkekler arasında ikinci sıklıkta görülen bir kanser türüdür.

Son yıllardaki verilere göre prostat kanserinin insidansında ve öldürücülüğünde normalin üzerinde bir artış olduğu görülmüştür. Bu artış, prostat kanserinin tüm dünya için önemli bir sağlık sorunu olduğunu göstermektedir.

Prostat kanseri, diğer kanser türlerine göre nispeten daha yavaş ilerleyen bir kanser türü olmasına rağmen tüm dünyada kansere bağlı ölümler arasında ilk üç sırada yer almaktadır (Ekin ve Zorlu, 2013). Prostat kanseri, dünya üzerinde en sık Amerika Birleşik Devletleri'nde görülmektedir. Yaşam boyunca dünyada her altı erkekten birinde prostat kanseri gelişmesi beklenmektedir (Nahleh,2006). Amerikan Kanser Birliği'nin 2013 yılında yaptığı bir çalışmaya göre 238.590 yeni prostat kanseri vakası ve prostat kanseri sonucu ölüm

oranını yaklaşık 29.720 kişi olarak belirlemiştir (Siegel v.d., 2013). Benzer şekilde prostat kanseri, Avrupa birliği ülkelerinde de en sık tanı alan kanser türlerindendir

(Boyle ve Ferlay, 2004). Avrupa Birliği'nde erkeklerde görülen kanserlerinin % 11'ini prostat kanseri oluşturmakta ve erkekler arasındaki tüm kanser ölümlerinin % 9'undan sorumlu tutulmaktadır (Cansino ve Martinez,2006). Prostat kanseri, Türkiye'de yaşayan erkeklerde kanser tanısı ve kansere bağlı ölümlerde akciğer kanserinden sonra ikinci en sık nedendir ve Sağlık Bakanlığı verilerine göre Türkiye'deki prostat kanseri insidansının yüz binde 37.6 olduğu bildirilmiştir (Değirmenci, 2010).

Prostat kanseri, etyolojisinde hem çevresel hem de genetik faktörlerin rol oynadığı, çok yönlü bir hastalıktır. Prostat bezi hücrelerindeki bu genetik değişikliğin, yaşlanma ile nerdeyse kaçınılmaz olduğuna dair günümüzde güçlü kanıtlar vardır. İleri yaşa gelmiş hemen hemen tüm erkeklerde otopside ya da prostat ameliyatı sonrası çıkarılan örnekte prostat kanseri görülebilmektedir. Prostat bezindeki bu değişiklikleri etkileyen birçok faktör bulunmaktadır ve bu değişikliklerin ne zaman başladığı, hastalık evresinin hangi hızda ilerlediği hatta hayatı tehdit eden kanser haline dönüştüğü günümüzde halen araştırma konusudur. Prostat kanserinde de çoğu kanserde olduğu gibi açıkça ortaya konmuş hastalığın oluşumundan sorumlu tek bir ajan ya da süreç yoktur.

2.2 Prostat Kanserini Etkileyen Etmenler

Prostat kanserini etkileyen çeşitli etmenler vardır. Bu başlığın alt bölümlerinde prostat kanserini etkileyen, bizim de tezimizin değişkenleri olan bazı fiziksel değişkenler incelenecek, kullanılan bu değişkenlerin prostat kanseri ile ilişkisine dair yapılan çalışmalar değerlendirilecektir.

2.2.1 Diyetteki Yağ Miktarı

Prostat kanserinin ölüm oranları, tüm dünyada özellikle doymamış yağ asidinden zengin yağ tüketimi ile sıkı korelasyon göstermektedir (Bostwick v.d., 2004). Yüksek yağ tüketimi, prostat kanseri hücrelerinin çoğalmasını uyarabilmektedir. Bu konu ile ilgili olarak, Clinton ve arkadaşları (1998) yaptıkları çalışmada prostat kanserinde, yağdan fakir diyet alınmasının testosteron gibi ön maddesi kolestrol olan steroid kökenli hormonlara bağımlı olan tümör hücrelerinin büyümesini azaltabileceğini göstermiştir. Birden fazla etnik kökene sahip olan bir çalışmada ise kolesterolün, et kaynaklı yağların ve farklı yağ türlerinin prostat kanseri riskiyle net bir ilişkisinin olmadığı ortaya koyulmuştur.

Birbirinden farklı dört etnik grup (Afrika kökenli, Japon kökenli, Latin kökenli ve Beyaz Amerikalılar) üzerinde yapılan bir çalışmada ise yağ ve kırmızı et tüketimi ile prostat kanseri arasında anlamlı ve güvenilir bir ilişki bulunamamıştır (Park v.d., 2007). Yapılan çalışmalarda her ne kadar tutarsız sonuçlara ulaşılmış olunsa da omega-3 yağ asitlerinin kanser riskinin azalmasına yardımcı olduğu görülmüştür. Omega-3 yağ asitlerinin kanser riskini azaltabileceği ön görülerek yapılan prospektif epidemiyolojik çalışmaları içeren 38 makaleden oluşan bir analizde, omega-3 yağ asitlerinin diyette alımıyla kanser riski ilişkisi incelenmiştir. Ancak bununla ilgili de güçlü bir ilişkiye rastlanılmamıştır (MacLean v.d.,2006). Sonuç olarak yağ tüketimi ve prostat kanseri riski arasındaki ilişki ile ilgili yapılan gözlemsel çalışmaların geniş bir analizinde, yağın daha yüksek alımı ile prostat kanseri gelişmesi riski arasında sadece zayıf bir ilişki olduğu ve çalışmalar arasında da yüksek derecede heterojenite olduğu belirtilmiştir (Dennis v.d., 2004).

2.2.2 Vücut Kitle İndeksi

Her ne kadar prostat kanseri gelişiminde önemi çok anlaşılamamış olsa bile vücut kitle indeksinin de prostat kanserinde etkisi olduğu yönünde çalışmalar bildirilmiştir (Calle v.d. ,2003). Vücut kitle indeksi ile prostat kanseri arasındaki ilişki halen tartışmalı bir konu olup son zamanlarda yapılan epidemiyolojik çalışmalar yüksek vücut kitle indeksinin prostat kanseri ile ilişki gösterdiğini bildirmiştir (Ly, Reddy ve Klein, 2010).

Bu çalışmalarda prostat kanserine bağlı ölüm riski ile yüksek vücut kitle indeksi arasında yakın ilişki bulunmuştur. Buna göre vücut kitle indeksi 35-40 arasında olan erkeklerde, vücut kitle indeksi normal sınırlar içerisindeki erkeklere oranla %34 daha fazla oranda prostat kanserinden öldüğü saptanmıştır (Calle vd.,2003). Vücut kitle indeksi ile ölçülen obezitenin, prostat kanseri ile kolon kanserinin erkeklerde benzer tümör tipleri olduğundan, kolon kanseri ile görülme sıklığındaki artış nedeni ile prostat kanseri için de bir risk faktörü olabileceği öne sürülmüştür (Giovannucci v.d., 1995). Diyette alınan yağ miktarının azaltılması ve egzersizin artırılması ile obezitenin tedavi edilince, oksidatif stresin azaltılabilineceği ve bu şekilde yapılan hayat tarzı değişikliğinin prostat kanseri riskini azaltabileceği öne sürülmüştür (Roberts, Vasiri ve Barnard, 2002). Son zamanlarda yapılan üç büyük prospektif çalışmada, obezite ile prostat kanseri riski arasındaki ilişkinin hastalığın evre ve/veya tanı anındaki derecesine göre ayrıntılı incelenmesi sonucunda, obezitenin düşük dereceli hastalık için düşük, yüksek dereceli hastalık için ise yüksek risk

oluşturduğu öne sürülmüştür (Gong v.d., 2006); (Rodrigues v.d., 2007; Wright v.d.., 2007). Sonuç olarak bu konuda erişkin vücut kitle indeksi ile prostat kanseri riski arasındaki ilişkiyi inceleyen çalışmalar çelişkili sonuçlar vermiş olduğu söylenebilir.

2.2.3 Alkol Tüketimi

Yapılan epidemiyolojik çalışmaların büyük çoğunluğunda alkol tüketimi ile prostat kanseri gelişimi arasında bir ilişkinin olmadığı ileri sürülmüştür. Güncel olmamakla beraber 1998'den önce yapılan 35 çalışmanın meta-analizinde de bu sonuç çıkmıştır (Dennis,2000). Alkolün bir yandan östrojen ve testosteronu etkileyerek kanseri uyması, diğer bir yandan da kırmızı şarabın antioksidan etkisi ile kanseri önleyebileceğinin gösterilmesi nedeniyle, diğer kanserlerle olduğu gibi alkol tüketiminin ve prostat kanseri riski bilimsel açıdan ilgi çekici bulunmuştur.
Bu konuda yapılan çalışmalarda alkol alımı ve prostat kanseri gelişimi riski arasında pozitif bir ilişki (Sesso v.d.,2001; Velicer,2006) ve kırmızı şarabın ise prostat kanserine karşı pozitif yönde koruyucu etkisi olduğu (Schoonen v.d.,2005) gösterilmesine rağmen yeni yapılan geniş çalışmalarda alkol alımı ile prostat kanseri insidansı arasında net bir ilişki gösterilememiş ve alkolün bu hastalığın etiyolojisine önemli ölçüde katkıda bulunmadığı belirtilmiştir (Sutcliffe, 2007;Baglietto v.d., 2006);Calle v.d., 2003).

2.2.4 Sigara İçimi

Sigaranın çoğu kanser ile doğrudan ilişkisinin olduğu bilinmektedir ayrıca sigara kullanımı ile akciğer ve mesane gibi organlarda kanser gelişim riski normal popülasyona göre daha fazla oranda artmaktadır. Sigaranın içerdiği kadmiyum ve dolaşımdaki testosteron gibi hormonların düzeylerini artırmasından dolayı prostat kanseri için risk faktörü olabileceği belirtilmiştir. Sigaranın prostat kanseri riskini artırıp artırmadığı açısından ciddi kanıt bulunmamaktadır.
Islami ve ark, 1995 yılına kadar yapılan çalışmaları içeren sistematik derlemesinde sigara içimi ile prostat kanseri insidansı artışı arasında pozitif ilişki bulmuştur (2014). Bae ve ark, 14,450 hasta ile yaptığı çalışmada ise prostat kanseri riskinin eski sigara içicileri ve aktif sigara içenlerde hiç sigara içmeyenlere göre arttığını ancak bu farkın istatistiksel olarak anlamlı olmadığını göstermişlerdir (2015).
Bu konu ile ilgili genetik ve çevresel faktörlerin etkileşimini araştıran bir çalışmada, daha hassas ve kolay etkilenebilir genotipleri taşıyan ve sigara içen bir grup ile

sigara içmeyenler karşılaştırılmış ve sigara içen grupta prostat kanser riskinin anlamlı derecede yüksek olduğu vurgulanmıştır (Quinones v.d., 2006). Toplamda 21,579 hastanın ele alındığı 24 prospektif kohort çalışmasının meta-analizinde, halen aktif sigara içenlerde prostat kanseri gelişimi riskinin artmadığı ancak içtikleri sigara miktarı gibi diğer değişkenler ile katmanlanan verilerde prostat kanseri riskinin istatistiksel olarak arttığı gösterilmiştir. Hayatının belirli bir döneminde sigara içenlerde prostat kanseri riskinin hiç sigara içmeyenlere göre %9 daha fazla prostat kanseri gelişme riskinin olduğu gösterilmiştir (Michael v.d., 2010). Özetle çoğu çalışmada genel prostat kanseri insidansı ile sigara içilmesi arasında önemli bir ilişki gözlemlenememiştir (Hickey ve Green 2001; Huncharek v.d., 2010).

2.2.5 Likopen

Likopen, domateste yüksek oranda bulunan bir karotenoiddir ve ayrıca bilimsel olarak potansiyel antikanser özelliklerine sahip bir antioksidan ajan olarak tanımlanmıştır. Bugüne kadar likopen ile prostat kanseri riski arasındaki ilişkiyi araştıran çalışmalar tutarsız sonuçlar vermiştir.

Ping-Chen ve ark (2015), yirmi altı çalışmada 563,299 katılımcıdan bildirilen 17.517 prostat kanseri vakasını dahil ettiği meta-analizde tüm çalışmalarda likopen tüketimi ile prostat kanseri riski arasında ters bir ilişki bulunmamasına rağmen, daha yüksek likopen alımı ile prostat kanseri insidansının azaldığı yönünde bir eğilim olduğu görülmüştür.

Bu konu ile ilgili yapılan bir başka çalışmada, diyette yüksek miktarda likopen alımının, prostat kanseri gelişme riskini %21 oranında azalttığı bildiririlmiştir. Ayrıca prostat kanseri insidansının domates içerikli sosları çok tüketenlerde, az tüketenlere oranla %36 oranında daha az olduğu görülmüştür. Çalışmada enteresan bir bilgi olarak olarak, domates suyu tüketiminin prostat kanseri üzerinde koruyucu etkisinin olmadığını belirtimişlerdir ve bunun nedeninin de likopenin biyoyararlanımının, pişirilmesi ile arttığına bağlı olduğunu belirtmişlerdir (Reiter ve Dekernion, 2005).

2.2.6 Kalsiyum Alımı

Değiştirilebilir bir diyet faktörü olan kalsiyum alımı, son zamanlarda prostat kanseri için bir risk faktörü olduğu öne sürülmüştür. Diyette yüksek miktarda kalsiyum alınması ile prostat kanseri hücrelerinin gelişmesini ve bunların farklılaşmasını engelleyen D vitaminin sentezinin azalması sorumlu tutulmaktadır (Schwatz v.d.,

1995**).** Bu konu ile ilgili İsveç'te prostat kanseri olan ve olmayan erkekler ile yapılan bir olgu kontrol çalışmasında, diyetle alınan kalsiyumun prostat kanseri için bir risk faktörü olabileceği gösterilmiştir (Chan v.d., 1998b). Bu çalışmanın aksine, 2006 yılında Koh ve ark'ın (2006) yaptığı prospektif çalışmada ise ne günlük alınan kalsiyum miktarının ne de kalsiyum destek tedavisinin prostat kanseri riskini artırmadığı gösterilmiştir.

2.2.7 Ailesel Yatkınlık

Ailesel prostat kanseri, bir veya daha fazla akrabanın prostat kanseri tanısı alması olarak tanımlanır. Bol miktarda epidemiyolojik kanıt, prostat kanserinin hem ailesel hem de genetik bir bileşene sahip olduğunu düşündürmektedir. Ailesel kümelenmenin ilk raporları 20. yüzyılın ortalarında yayınlanmıştır ve prostat kanseri gelişme riskinin, bu hastalıktan etkilenen birinci derece akrabası olanlarda daha yüksek olduğu gösterilmiştir (Woolf, 1960). Daha sonraki zamanlarda yapılan vaka kontrolü ve kohort çalışmaları da bu gözlemi desteklemiştir (Eeles v.d, 1997).

Prostat kanserinde genetik geçişin önemli olduğu kabul edilmekle birlikte ailelerindeki diğer erkek bireylerde de prostat kanseri olan erkeklerde prostat kanseri riskinin arttığı, ayrıca prostat kanserinin görülme olgusunun daha düşük yaşlara inmesi sonucu ailenin diğer erkek üyelerinde prostat kanserine yakalanma olasılığının da arttığı gözlemlenmiştir (Carter v.d.,1993).

Birinci dereceden yakınlarında prostat kanseri olanlarda genel popülasyona kıyasla prostat kanseri tespit edilme olasılığı daha yüksektir.

Bu konuda yapılmış Zeegers ve arkadaşlarının (2003) yaptığı meta-analizde ailede etkilenen birey sayısı, akrabalık derecesi ve etkilendikleri yaşlar göz önünü alındığında rölatif riskin arttığı gösterilmiş. Bu çalışmada,yalnız babada prostat kanseri varsa riskin 2.17, yalnız erkek kardeşte varsa riskin 3.34, ailede 65 yaş altında tanı alan 1. derece akrabada varsa riskin 3.37, iki veya daha fazla birinci derece akraba etkilenmişse riskin 5.08, yalnız bir tane ikinci derece akraba etkilenmiş ise riskin 1.08 kat arttığı gösterilmiştir.

2.2.8 Fiziksel Aktivite

Ürogenital sistem kanserleri arasında, kanserin önlenmesi ve fiziksel aktivite arasındaki ilişkinin en çok araştırıldığı ve kesin olmasa da bazı sonuçların elde edildiği kanser türü, prostat kanseridir. Fiziksel aktivitenin artırılması ile prostat

kanserlerinin oluşmasının engellenebilineceği ön görülmektedir (Kushi v.d., 2006). Fiziksel aktivitenin androjen, insülin ve insülin benzeri büyüme faktörü (IGF) gibi kansere neden olabilecek bazı mekanizmalarda rol oynayan hormonları azaltma yolu ile bu sayede kanser riskini de azaltabileceği düşünülmüştür (Lehrer v.d., 2002). Bu mekanizmaya ek olarak yapılan fiziksel aktivitedeki artış, kansere neden olabilecek bazı durumları engelleyebilecek olan bağışıklık mekanizmaları ve antioksidan etkinliği düzenleyerek prostat kanseri ouşum riskini azaltabilmektedir (Chan v.d.,1998a). Toplamda 40 çalışmanın dahil edildiği bir derlemede, derlemelerin 22`sinde az miktarda da olsa yapılan fiziksel aktivitenin prostat kanseri riskini bir miktar azalttığı gösterilmiştir. Geri kalan 18 çalışmanın 14`ünde fiziksel aktivite ile prostat kanseri riski arasında herhangi bir ilişki saptanmamış iken 4 çalışmada ise tam tersi olarak fiziksel aktivite ile prostat kanseri riskinin bir miktar artış gösterdiği belirtilmiştir (McCaughan,2012).

Harvard Üniversitesinde 439 hasta ile yapılan çalışmada, ne total aktivite ne de aktivitenin kuvveti ile prostat kanseri riski ile arasında ilişki bulunamamıştır (Lee v.d., 2001). Liu ve arkadaşları (2011) ise toplamda 19 kohort ve 24 olgu-kontrol çalışmasını değerlendirdikleri bir analizde, fiziksel aktivitenin prostat kanseri riskini azalttığını ve özellikle 20-65 yaşları arasında yapılan fiziksel aktivitenin prostat kanseri riskini ciddi oranda azalttığını vurgulamıştır. Bu konu ile ilgili olarak, yapılan çalışmalarda genel olarak çok anlamlı ve kesin bir fark olmasa da fiziksel aktivitenin, prostat kanseri riski üzerine olumlu yönde etkisi bulunduğu söylenebilinir.

2.2.9 Etnik Köken (Irk)

Prostat kanserinin en dikkat çekici özelliklerinden birisi, onun ortaya çıkışında ve progresyonunda coğrafi varyasyonun derecesidir. Amerika Birleşik Devletleri'nde yaşayan erkekler arasında prostat kanseri, en sık tanı konan kanser türüdür ve kansere bağlı ölümlerin nedeni olarak da ikinci sırada yer alır (Jemal v.d., 2005). Irksal ve coğrafik orijin de prostat kanseri için risk oluşturmaktadır. Prostat kanserinin insidansı Amerika Birleşik Devletler' inde diğer ülkelere özellikle Asya ülkelerine göre oldukça yüksektir. Afrika kökenli Amerikalı erkekler dünyada en fazla risk oranına sahiplerdir ve ölüm oranları beyazlara göre daha yüksek olmaktadır. İskandinav ülkerindeki erkekler, güney Avrupa ülkelerinde yaşayan erkekler ile karşılaştırıldığında, prostat kanseri yüksek insidansına ve mortalitesine sahiptirler (Landis v.d.,1999).

Dünyada prostat kanseri insidansı en fazla oranda Amerikalı zencilerde görülür ve bu popülasyonda hayatları boyunca prostat kanserine yakalanma riski %9,8 dolayındadır. Amerikalı beyaz erkeklerde bu oran zenci popülasyona göre nispeten daha düşük olup %8'dir. Dünya genelinde ise prostat kanseri insidansının en düşük olduğu ülkeler ile Çin ve Japonya'dır.

Amerika'ya göç eden Japonlarda prostat kanser insidansı Japonya'da yaşayanlara göre daha yüksektir fakat Amerikalılara göre %50 daha düşüktür.

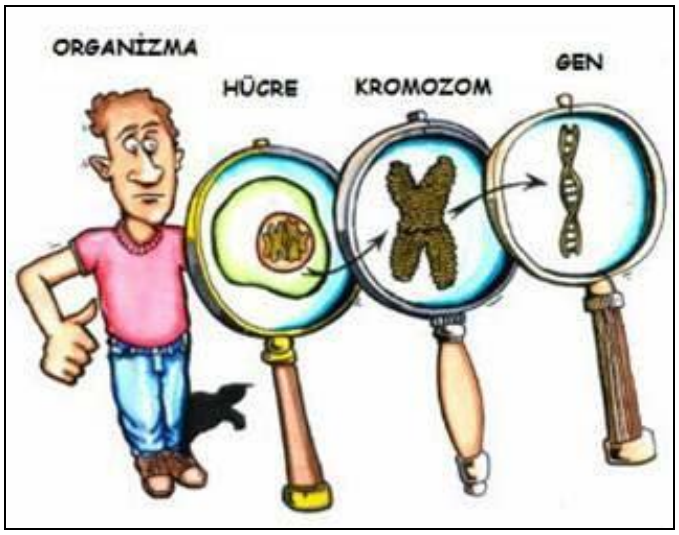
Etnik kökene göre görülen prostat kanser insidansı farklılıklarının sebebi kesin olarak açıklanamamıştır. Ancak genetik ve hormonal faktörler ile beslenme şeklinin, etnik kökenden kaynaklanan bu farkı açıklayabileceği düşünülmüştür (Platz v.d., 2000).

Prostat kanseri beyaz ırk ile karşılaştırıldığında Afrikalı Amerikalı erkeklerde 2-3 kat daha fazla görülmektedir. Ancak prostat kanseri Hispaniklerde beyaz ırk ile benzer oranda izlenmektedir. Afrikalı Amerikan erkeklerde prostat kanseri yaklaşık olarak her 100.000 erkekte 275,3 olarak saptanırken, beyaz ırka mensup erkekerde 100.00 erkekte 172,9 ve Hispaniklerde 100.000'de 127.olarak saptanmaktadır. Bununla birlikte Ssya kökenli erkeklerde oran 100.000 de 107.2 olduğu görülmüştür (American Cancer Society, 2012; Lunenfeld,2002).

Kanser epidemiyolojisinde coğrafyanın, ırkın ve yaşam tarzının birbiriyle olan yakın ilişkisini en iyi gösteren örnek göç olayıdır. Japonya'dan Hawaii adalarına göç edenler arasında, Japonya'da yaşamaya devam edenlere göre prostat kanserinin daha fazla oranda görülmesi, riskteki artışın yetişkin dönemde olduğu hipotezini geliştirmiştir (Haenszel ve Kurihara, 1968).

2.3.Tek Nükleotid Değişim (Tek Nükleotid Polimorfizm- SNP)

Tek nükleotid değişim (SNP- Single Nucleotide Polymorphism) en basit şekilde tek bir nükleotidde meydana gelen mutasyon (değişim) olarak tanımlanabilir. SNP'in ne olduğunu anlamak için bazı kavramları da bilmek gerekir. Aşağıda bu kavramlar açıklanacaktır.



Şekil 2.2 Kromozom ve Gen

2.3.1 Kromozom ve Genler

Eşeysel olarak üreyen türlerde, her organizma, anadan gelen yumurta ile babadan gelen spermin birleşmesiyle oluşur. Bu iki hücre, kalıtımsal özellikleri, ana-babadan döllerine geçirirler. Kromozomlar, bir bireyin her hücresinin çekirdeği içinde bulunurlar. Her hücrede, değişmez bir kromozom sayısı vardır. Bunların yarısı bireyin anasından, yarısı babasından gelir. Kalıtımsal maddeler, kromozomların üstünde yer alan genlerden oluşur.

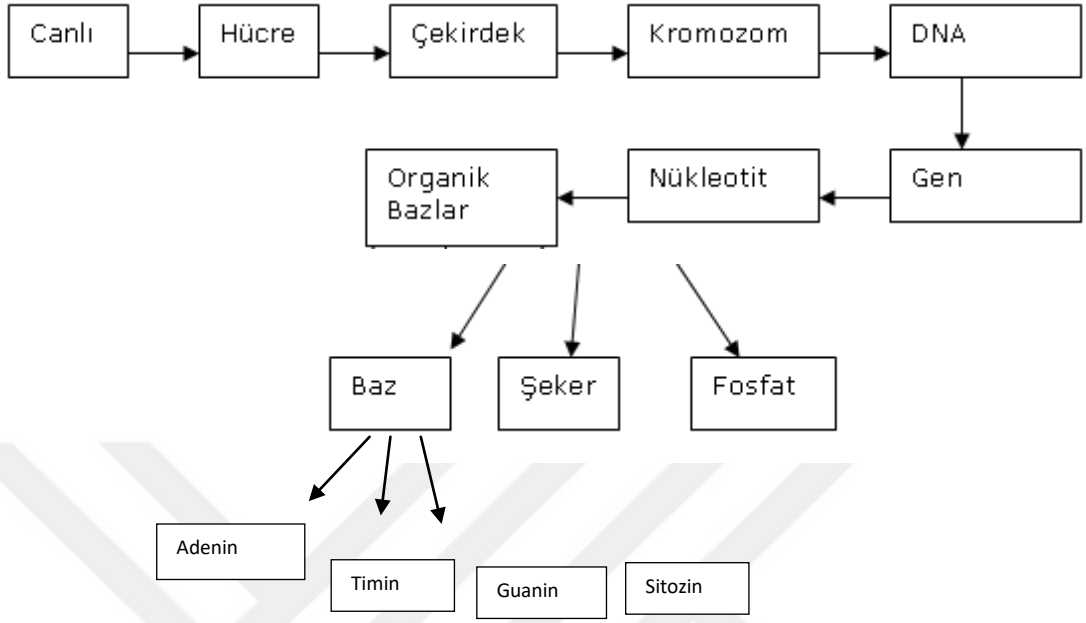
İnsanlarda 46 kromozom vardır: 22 çift otozom ve bir çift eşey kromozomu. Bu az sayıda kromozomun üstünde, binlerce kalıtımsal özelliğin bir kuşaktan ötekine taşınmasından sorumlu olan binlerce gen yer alır.

Şekil 2.1'de görüldüğü gibi bir organizma hücrelerden oluşur. Kromozomlar hücrelerin içindedir. Genler de kromozomların içindedir. Yani organizma hücrelerden, hücre kromozomlardan, kromozom da genlerden meydana gelir.

2.3.2 Genetik Şifre

Kromozomlardaki kimyasal bileşiklerin çoğu DNA ve RNA adı verilen nükleik asitler ve proteinlerdir. Bunlardan DNA, genetik şifreyi içermektedir. Bir organizmanın ne olduğu veya olacağı, o organizmanın hücrelerinde oluşan kimyasal tepkimelere göre belirlenir. Bu kimyasal tepkimelerin çoğu enzimlerle denetlendiğinden ve görev alan bu enzimler de protein oldukları için protein birleşimi büyük önem taşımaktadır. DNA moleküllerindeki genetik şifre, enzimlerin üretilmesini denetler ve bu sayede bireyin

birçok özelliğini etkiler. Örnek olarak saç ve göz renkleri gibi özellikler DNA tarafından denetlenmektedir.



Şekil 2.2 (Fenokulu, 2015) Canlıdan Organik Bazlara Doğru Sıralanışı

2.3.3 DNA'nın Yapısı

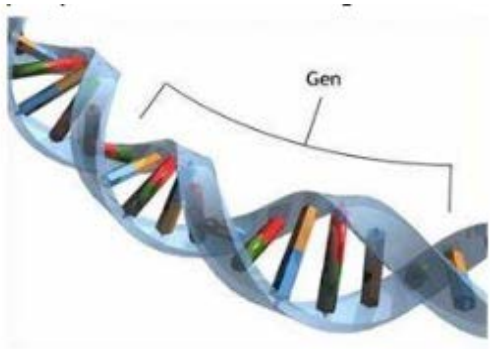
Şekil 2.3 (GenomTürkiye,2018) Kromozomun Yapısı

DNA, tüm organizmalar ve bazı virüslerin işlevleri ve biyolojik gelişmeleri için gerekli olan genetik bilgileri taşıyan bir çift sarmal yapıda bir nükleik asittir. DNA'lar hücre çekirdeğinde içerisindeki kromozomların içinde bulunurlar. DNA kabaca, 2-deoksiriboz adı verilen beş karbolu olan bir şeker, fosfat ve 4 adet bazdan oluşur. DNA'da bulunan 4 adet baz; Adenin, Timin, Guanin, Sitozin olarak adlandırılır. Bu

bazlar her zaman ikili olarak eşleşir; A-T (Adenin-Timin) ve G-C (Guanin-Sitozin). Bunlar baz çifti olarak adlandırılır. Yukarıdaki şekil 2.3'de görüldüğü gibi kromozomlar ortada uzun DNA zincirleri ve etrafında onu saran protein kılıfından oluşmaktadır.

Şekil 2.4 (GenomTürkiye,2018) DNA Zinciri

Gen bir kalıtım birimi olup, bir kromozomun belirli bir kısmını oluşturan nükleotid dizisidir olarak tanımlanabilinir. Genlerde, bir insanın saç rengi, parmak sayısı, boyu, göz hücreleri vb. gibi biyolojik özelliklerine ait bilgi kayıtlıdır. Genom ise organizmaya ait genlerin tamamına verilen isimdir. Basit bir bakteride yaklaşık olarak 600.000 civarında baz çifti (CG, TA gibi) bulunurken insanda bu sayı 3 milyardan fazladır.



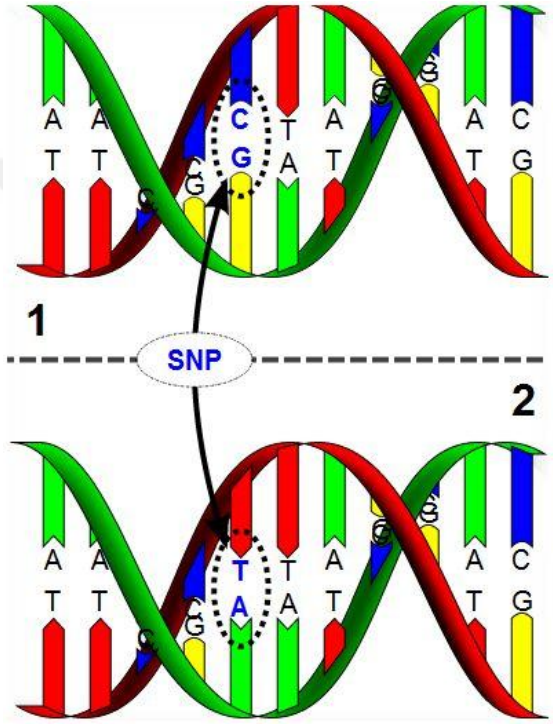
Şekil 2.5 Gen'in Yapısı (GenomTürkiye, 2018)

2.3.4 SNP

Genlerde coğrafi ve iklim gibi çevre şartlarına bağlı değişiklikler meydana gelebilir. Bu değişikliklere mutasyon adı verilmektedir. SNP (Single Nucleotide

Polymorphism) ise kısaca, tek bir nükleotidde meydana gelen mutasyon olarak tanımlanabilir. Yani SNP, DNA sekansında A, T, C veya G Gibi tek bir nükleotidin farklı olmasıdır. Popülasyonda belli bir nükleotid çifti farklılık gösteriyor ve farklı allel ve genotipler ortaya çıkartıyor ise buna tek nükleotid polimorfizmi denir. Mutasyonlar hastalık nedeni olabilirken, polimorfizmler hastalığa yatkınlık nedeni olabilirler. Verilere göre SNP'ler 1331 bazda 1 kez görülmektedir.

Aşağıdaki şekilde bir SNP gösterilmektedir. Şekle bakıldığında 1 ve 2 DNA zincirinde tüm baz çiftleri aynı iken 1'deki CG baz çifti 2'de TA ile yer değiştirmiştir.



Şekil 2.6 SNP'in Yapısı (GenomTürkiye, 2018)

SNP Single Nucleotide Polymorphism Türkçe olarak, Tek Nükleotid değişimleri anlamına gelmektedir. Şekilde de görüldüğü gibi SNP tek bir nükleotidde meydana gelen mutasyondur.

İnsan genomunda yaklaşık olarak 3 milyar kadar baz çifti yani 6 milyar nükleotid vardır. Yaklaşık 6 Milyon SNP vardır. Her 1000 Nükleotidde bir SNP gözlenmektedir. Aşağıdaki şekilde rastgele olarak seçilen iki kişinin kromozom 7'ye ait DNA dizisinin bir kısmının karşılaştırılması görülmektedir. Bu karşılaştırmaya bakıldığında iki birey arasında 2200 nükleotid içerisinde sadece 2 tane SNP görülmektedir.

Şekil 2.7 SNP

SNP'ler artık ilerleyen teknoloji ile beraber günümüzde birçok hastalıkla ilişkilendirilmiş olup tanı koymada gerekli ve kişiye özgü tedavilerin geliştirilmesinde kullanılabilen genetik değişimlerdir. İki farklı bireyin DNA dizisi karşılaştırıldığında her iki bireyin DNA dizisinin % 99.9 oranında aynı olduğu görülür. Farkı sağlayan % 0.1 oranındaki değişimin çok büyük bir kısmını da tek nükleotid değişimleri oluşturur. Bu % 0.1 oranındaki farkın çok az bir kısmı hastalıklar ile ilgilidir. Bunun büyük kısmı ise boy, ağırlık, yüz özellikleri, saç ve göz rengi gibi zararsız farklılıklardan sorumludur. Bazı SNP'ler görünüşümüzün farklı olmasından sorumlu iken bazıları ise ne tip hastalıklara sahip olacağımızı ya da herhangi bir ilaca nasıl tepkiler vereceğimizi etkileyebilmektedir. Fakat birçok SNP de görünürde hiçbir değişikliğe sebep olmamış gibidir. Sonuç olarak DNA dizilişlerimiz arasındaki farklılıklar ve benzerlikler o insanlarla ne kadar benzediğimizi veya farklı olduğumuzu göstermektedir.

SNP'ler genellikle genlerde bulunmalarına rağmen bazı durumlarda genler arası kodlanmayan bölgelerde ve genin intron adı verilen kodlanmayan kısımlarda da meydana gelebilir. Hastalıklar ile ilişkilendirilen SNP'ler genelikle gen içindeki ekson adlı bölgeyi kodlayan DNA'da meydana gelenlerdir. SNP'ler, proteinlere bağlı bazı mekanizmalarını bozarak genlerin fonksiyonlarını etkileyebilmektedir.

2.3.5 SNP’lerin Hastalıklar ile İlişkisi

Biyomedikal araştırmalarda, farklı popülasyonlardan seçilen hasta ve sağlıklı bireylerde SNP tayini ve bunların her iki gruptakiler ile karşılaştırılması yapılmaktadır. Araştırılan hastalıklar arasında Alzheimer, kalp-damar hastalıklar, bazı kanser türleri ve migren bulunmaktadır. Bu amaçla çeşitli hastalıklara ait SNP profilleri çıkarılmıştır. Bu SNP profilleri kullanılarak bireylerin hastalıklara olan yatkınlıklarını tahmin etmek mümkün hale gelmiştir. Günümüzde, 23andMe, FamilyTreeDNA, MyHeritage, Ancestry.com, SNP’leri kullanarak hastalık taraması yapan şirketlerden bazılarıdır (Snpedia,2018).

Eupedia’nın Genetics web sitesinde alerji, astım, tip1 ve tip2 diyabet gibi otoimmün hastalıklara, beyin, meme, kolon, mide, kan, akciğer, böbrek, yumurtalık, pankreas, prostat, cilt, lenfoma gibi çeşitli kanserlere, çeşitli kardiyovasküler hastalıklara ,çeşitli sindirim sistemi hastalıklarına, çeşitli kadınsal rahatsızlıklara, genetik hastalıklara, çeşitli nöropsikolojik bozukluklara ve bağımlılıklara karşılık gelen SNP’ler ve profilleri gösterilmiştir. Aşağıdaki tabloda daha önceki çalışmalardan derlenen, prostat kanseri ile ilişkili bulunan SNP’ler gösterilmiştir (Eupedia,2018).

Tablo 2.1 Prostat Kanseri ile İlişkili SNP'ler (Eupedia,2018)

PROSTAT KANSERİ İLE İLİŞKİLİ SNP'LER				
KROMOZOM	**GEN**	**SNP**	**RİSKLİ ALLELLER**	**TESTİ YAPAN KURUM**
B	EHBP1	rs2710646	AA, AC, CC	A, B
3	FYCO1	rs1545985	AA, AG, GG	A
		rs7652331	CC, CT, TT	A
4	KIAA1211	rs629242	CC, CT, TT	A
4	LOC152485	rs13149290	CC, CT, TT	A
5	FCHSD1	rs251177	CC, CT, TT	A
5	FGFR4	rs351855	CC, CT, TT	A
7	JAZF1	rs10486567		A
8	8q24 (1)	rs1447295	AA, AC, CC	A, B
8	8q24 (2)	rs6983267	GG, GT, TT	A, B
8	8q24 (3)	rs10505483	AA, AC, CC	A, B
8	8q24	rs16901979	AA, AC, CC	A
9	DAB2IP	rs1571801	GG, GT, TT	A
10	intergenic	rs10993994	CC, CT, TT	A, B
10	TCF7L2	rs12255372	GG, GT, TT	A
11	intergenic	rs10896449	AA, AG, GG	A, B
12	VDR	rs2107301	CC, CT, TT	A
		rs2238135	CC, CG, GG	A
13	FAM124A	rs10492519	AA, AG, GG	A
14	ESR2	rs2987983	CC, CT, TT	A
16	CDH1	rs16260	AA, AC, CC	A
17	TCF2	rs4430796	AA, AG, GG	A, B
17	17q24.3	rs1859962	GG, GT, TT	A, B
X	intergenic	rs5945572	A, G	A, B

Test Eden Şirketler:

A = 23andMe Şirketi

B = DeCODEme Şirketi

C = FTDNA Şirketi

tarafından test edildiğini ifade eder gösterir.

SNPedia (2018), insan genetiğini araştıran, Wikipedia'da modellenmiş bir sitedir. Bu sitede bilimsel yayınlardan alıntı yapılarak, DNA'daki varyasyonların etkileri hakkında bilgi paylaşılıyor. Ayrıca bu site Promethease tarafından DNA varyasyonlarını onlar hakkında yayınlanan bilgilere bağlayan kişisel raporlar oluşturmak için kullanılır. Şu anda SNPedia'da 108.198 SNP hakkında bilgi vardır. Ancestry.com, 23andMe, FamilyTreeDNA, MyHeritage, Genes for Goods, WeGene ve LivingDNA gibi kişisel DNA'ları inceleyerek yüzlerce hastalık açısından bu kişilerin risk faktörü taşıyıp taşımadığını belirleyen bu şirketler de SNPedia'dan destek almaktadır. SNPedia'da birçok hastalığa karşılık gelen SNP'lerle ilgili çalışmalar bulunmaktadır.

2.3.6 Prostat Kanseri İle İlgili Yapılmış SNP Çalışmaları

Daha önce de tanımladığımız gibi insan genomu içindeki yaygın varyasyonlara, tek nükleotid polimorfizmleri (SNPler) denir. SNP'lerin Prostat kanseri (PK) de dahil olmak üzere birçok karmaşık hastalığın gelişimine doğrudan katkıda bulunduğu düşünülür (Eeles v.d., 2013). Daha önce de belirtildiği gibi, yüksek verimli genotiplemedeki birçok gelişme, bütünsel genom ilişkilendirme (GWAS) performansını arttırmıştır. Bugüne kadar, bu genetik ilişkilendirme çalışmaları, artan PK duyarlılığı ile ilişkili yaklaşık 100 SNP'i belirlemiş ve PK'nin kalıtsal bileşeninin% 35'inden fazlasını açıkladığı düşünülmektedir (Eeles v.d., 2014)

Ayrıca, bu eşey (cinsiyetle alakalı olan) SNP'ler bir insanın yaşamı boyunca stabil olduğundan ve diğer hastalık süreçlerinden (örn., inflamasyon, enfeksiyon, iyi huylu prostat büyümesi) etkilenmediğinden, bu eşey SNP'lerin, biyolojik belirteçler olarak kullanımlarına büyük ilgi vardır.

2.3.6.1 Prostat Kanseri Riskinin Belirlenmesine Yönelik Yapılan SNP Çalışmaları

Bu bağlamda birçok çalışma PK risk ilişkili SNP topluluğunu değerlendirmiştir. Örneğin, 3500'den fazla PK vakası ve kontrol grubunu içeren bir çalışmada, PK geliştirme riskiyle ilişkili 5 SNP'den oluşan bir topluluk değerlendirilmiştir. Ailede PK öyküsü olan 5 risk alelinin tüm taşıyıcıları, aile öyküsü bulunmayan ve risk alelleri taşımayan erkeklerle karşılaştırıldığında, ilk grubun ikinci gruba oranla hasta olma riskinin 9,5 kat arttığı görülmüştür (Zheng v.d., 2008). Başka bir çalışmada, PK risk ilişkili 14 SNP'li küme incelendiğinde, PK ailesi öyküsü olan ve PK risk ilişkili 14 SNP'in taşıyıcısı olan 55 yaşındaki bir erkeğin, 20 yıllık dönemde hastalığa yakalanma riskinin % 50'den fazla olduğu görülmüştür. Buna karşılık, SNP genotipi ve aile öyküsü bilinmeden bu erkeklerin % 13'lük bir kısmının prostat kanseri riskine sahip olacağı tahmin edilmiştir (Xu v.d., 2009).

Kote-Jarai ve arkadaşları (2008), büyük PRACTICAL konsorsiyum'da (Genomdaki kanser ilişkili değişimleri araştıran prostat kanseri ilişkilendirme grubu) PK risk ilişkili 15 SNP'i değerlendirdi. Kanser Derneği Grubu, Genom'daki Kanserle İlgili Değişikliklerin Araştırılması için bu çalışmada 7370 vaka ve 5742 kontrol grubu kullandı.

Sonuç olarak bu SNP'lerin PK riski üzerinde güçlü bir kümülatif etkisi olduğu bulunmuştur. Bu 15 SNP'e dayanan risk dağılımının en üst % 10'luk risk kısmında bulunan erkekler genel nüfus oranlarına göre 2,1 kat fazla riske sahiptir.

Diğer araştırmalar son zamanlarda biyopside PK tanısı için doğrulanmış daha büyük ve farklı SNP panellerini değerlendirdi. Birinci derece aile öyküsü veya PSA gibi klinik özelliklerle kombine edildiğinde, PK riski ile ilişkili SNP'lerin, biyopsi sonuçlarının tahminini önemli ölçüde artırabildiği gösterilmiştir. Örneğin, yakın tarihli bir bütünsel genom ilişkilendirme çalışmasının sonuçları, prostat biyopsi sonuçlarının tahminini önemli ölçüde iyileştirmek için PSA ile kombine olarak, PSA'larla birlikte 23 PK risk ilişkili SNP'in kullanılabileceğini göstermiştir (Bansal v.d., 2000). Benzer şekilde,2005-2007 yılları arasında prostattan biyopsi yapılan İsveçli bir erkek kümesi kullanılmıştır. PK ile ilişkili SNP'leri ve mevcut klinik değişkenleri (yaş, serum PSA, total PSA oranı ve aile öyküsü) içeren genetik bir tahmin modelinin, Sadece klinik değişkenleri içeren tahmin modeline oranla önemli ölçüde daha iyi performans gösterdiği görülmüştür (Pilia v.d., 2006).

Spesifik olarak, aynı sayıda PK tümörünün saptanması için, PK riski ile ilişkili SNP'lerin modele eklenmesi, sadece klinik değişkenlerin kullanıldığı bir modele göre

gerekli biyopsi sayısını önemli ölçüde azaltmıştır. Başka bir çalışmada, Kader ve arkadaşları (2012), pozitif prostat biyopsisi tahmininde 33 PK riski ile ilişkili SNP'lerin performansı ile mevcut klinik parametrelerin performansını karşılaştırmışlardır. Denemedeki tüm erkekler başlangıçta negatif bir prostat biyopsisi geçirmişler ve 2-4 yıl içinde tekrar biyopsiye alınmışlardır (Andriole v.d., 2010) Bu çalışmada, PK riski en iyi klinik modele göre ve klinik ve genetik modelin birleştirilmesiyle oluşturulmuş kombine modele göre hesaplandı ve derecelendirildi (Kader v.d., 2012). Yazarlar, en iyi klinik modele SNP eklemenin, erkeklerin % 33'ünde PK riskini yeniden sınıflandırdığını ve yeniden sınıflandırılan riskin biyopsi sonuçları ile anlamlı derecede daha iyi korelasyona sahip olduğunu bulmuşlardır.

Ek olarak, prostat biyopsisini ayırt etmek için en iyi klinik model kullanıldığında eğri altında kalan alan (AUC) 0.62 iken, en iyi klinik model ve genetik model birleştirildiğinde bu alan 0.66'ya yükselmiştir (Akamatsu v.d., 2012).

Birlikte ele alındığında, bu sonuçlar, genetik belirteçlerin risk tahminini iyileştirmek için kullanılabileceğini düşündürmektedir. Rutin uygulamaya dahil edilirse, en yüksek SNP'leri taşıyan erkekleri hedef alan çalışmalarda gereksiz biyolojik biyopsi sayısını azaltmaya yardımcı olarak tarama uygulamalarını geliştirme olanağı sunarlar.

2.3.6.2 PSA Taramasının Performansını Arttırmaya Yönelik Yapılan SNP Çalışmaları

Genetik değişkenler aynı zamanda mevcut serum PSA taraması üzerinde bir gelişme fırsatı da sunmaktadır (Bansal v.d., 2000), (Guy v.d., 2009). Örneğin, önceden ölçülen serum PSA konsantrasyonlarındaki bireyler arasındaki farklılıkların % 40-45'inin genetik faktörler ile açıklanabileceği tahmin edilmiştir (Nam v.d., 2006), Son zamanlardaki çalışmalar, PSA'yı kodlayan genin içinde veya yakınında bulunan SNP'lerin (örn., Kallikrein ile ilişkili peptidaz 3 [KLK3]) serum PSA konsantrasyonlarını etkileyebileceğini ve daha sonra PK taraması ve tespit sıklığını etkileyebileceğini göstermiştir (Lose v.d., 2011), (Ishak ve Giri, 2011). Benzer şekilde, yeni bir bütünsel genom ilişkilendirme çalışmasının sonuçları PK risk-ilişkili SNP'lerin 6'sının serum PSA düzeyleri ile güçlü bir ilişki gösterdiğini ve bu SNP'lerin 4'ünün (PSA ilişkili 4 SNP) bağımsız olarak serum PSA konsantrasyonları ile ilişkili olduğunu göstermiştir.

Ayrıca, bir önceki çalışmayı yapan yazarlar İzlandalı toplulukta yaptıkları çalışma ile PSA ile güçlü ilişkili olan bu SNP'lerin ölçülen serum PSA'larını genetik olarak düzeltmek için kullanılabileceğini göstermiştir (Bansal v.d.,2000). Bu genetik olarak

düzeltilmiş PSA değerleri düzeltilmemiş değerler ile karşılaştırıldığında bir tarama aracı olarak PSA'nIn performansını önemli ölçüde artırmıştır. AUC değeri % 70.9 iken % 73.2 'ye çıkmıştır (Scandino,2013), (Hsu v.d., 2009). Son olarak, belgelendirilmiş PK'siz Avrupa soylarının erkeklerinden oluşan bir toplulukta, 4 PSA ilişkili SNP'in varlığına yönelik genetik düzeltmenin potansiyel olarak gereksiz biyopsi sayısında% 18 ila% 22'lik bir azalma ile sonuçlandığı gösterişlmiştir (gekersiz biyopsi, ölçülen serum PSA'sı, SNP'ler için düzeltme yapıldıktan sonra biyopsi eşiğinin altına düşmüş olan erkeklerde tanımlanmıştır).

Ek olarak, PSA-SNP'lerin yokluğu için genetik düzeltme potansiyel olarak gecikmiş biyopsilerde % 3'lük bir azalmayla sonuçlanabilir (ölçülen serum PSA'sı, SNP'ler için düzeltme yapıldıktan sonra biyopsi eşiğinin üzerine çıkmış olan erkeklerde tanımlanmıştır) (Guy v.d., 2009)

Genetik düzeltme yapıldığında Afrika kökenli Amerikalı erkeklerden oluşan bir topluluğa uygulandığında, aynı PSA-SNP'ler farklı sonuçlar vermişlerdir: genetik düzeltme gereksiz biyopsileri önlememiş, ancak hastaların %30'unda gerekli biyopsileri geciktirmek için kullanılabileceği ortaya çıkmıştır (Akamatsu v.d.,2012).Genetik düzeltmede ırk farklılıkları Afrika kökenli Amerikalı erkeklerin daha ileri evre hastalıklara yakalanma olasılığının daha yüksek olduğu ve Avrupalı erkeklere göre PK'e özgü ölümlerinin iki kat fazla olduğu bilinmektedir (Ren v.d., 2013). Hem Avrupalı hem de Afrikalı Amerikalıların genetik olarak düzeltilmiş PSA düzeyleri belki de Doktorların PK riskini daha doğru ölçmesine ve böylelikle gereksiz biyopsi ve / veya tanı gecikmelerini önlemesine izin verebilir (Aly v.d., 2011).

Her ne kadar PSA taramasının birçok yaşamı kurtardığı gösterilmiş olsa da, diğer taraftan agresif olmayan pek çok PK'nin aşırı tedavi edilmesine yol açmıştır (Carter v.d., 1992). Bu nedenle, agresif PK'ini agresif olmayanlardan ayırt etmeye yardımcı olabilecek biyobelirteçleri belirleme zorunluluğu vardır. Bugüne kadar tanımlanan en yaygın genetik varyasyonlar PK riski ile ilişkili olsa da, bu çalışmaların çoğu hastalığın agresifliğine odaklanmamıştır.

Bu nedenle, PK'e yatkınlık ile ilişkili bulunan yaklaşık 100 SNP'den herhangi birinin hastalığın agresifliği ile ilişkili olup olmadığı tartışmalıdır (Eeles v.d., 2013**)**. Son zamanlarda yapılan çalışmalar PK agresifliği konusundaki bu soruyu değerlendirmeye çalışmaktadır. Örneğin bir çalışmada, daha önce agresif veya ilerlemiş PK hastalığı ile ilişkili olan genler araştırılmıştır. Araştırmacılar, 150'den fazla farklı gende 1000'e yakın SNP'i araştırmışlar ve ölümcül PK tümörleri ile anlamlı bir şekilde ilişkili olan 5 farklı gende (LEPR, RNASEL, IL4, CRY1 ve ARVCF) SNP'ler bulmuşlardır (Akamatsu v.d., 2012). Bu çalışmalar, bir takım ilginç

genleri ortaya çıkardıysa da, bu bulguları doğrulamak ve daha agresif PK için kalıtsal riske katkılarını tam olarak anlamak için takip çalışmaları gereklidir.

2.3.6.3 Agresif Prostat Kanseri Alanında Yapılan SNP Çalışmaları

Son zamanlarda, bir takım bütünsel genom ilişkilendirme çalışmalarında PK agresifliği alanında çalışmalar yapıldı. Örneğin, Xu ve arkadaşları agresif hastalığı olan erkeklerde anlamlı olarak 17p12 ve 15q13 kromozomları üzerindeki SNP'lerin daha çok bulunduğunu rapor etmişlerdir (2011). Benzer şekilde, Bensen ve arkadaşları (2013), kromozom 3p12 ve 8q24 üzerindeki SNP'lerin PK agresifliği ile ilişkili olduğunu bildirmişlerdir. Lange ve arkadaşları, PK hastalığını değerlendiren bir çalışmada, 8q24 ve 11p15 üzerinde genom ilişkisinin anlamlı olduğunu bildirmiştir (2012). Birlikte ele alındığında, PK riski ile ilişkisi olan birçok SNP'in de agresif hastalığa karşı duyarlılığı artırdığı görülmektedir. Bir insanın metastatik hastalığı veya PK'inden ölme olasılığını artıran bu SNP'lerin bir topluluğunun belirlenmesi, PK için kişiselleştirilmiş tedavinin anahtarı olabilir.

Tümör içerisinde meydana gelen genetik değişiklikler (somatik düzeydeki mutasyonlar) kişiselleştirilmiş PK taramasına, saptanmasına ve tedavi edilmesine yardımcı olan potansiyel biyobelirteçler olarak karakterize edilebilir ve kullanılabilir. Aslında, prostat tümörlerinde somatik değişikliklerden yararlanan birçok farklı FDA onaylı ve Klinik Laboratuvar İyileştirme Değişikliği (CLIA) tabanlı klinik testler artık klinik karar vermede yol gösterici olmaktadır.

Prostat kanseri geni 3 (önceden DD3 olarak adlandırılan PCA3), prostatta üretilen kodlayıcı olmayan bir RNA'dır. Prostat tümörlerinin% 95'inde aşırı açıklayıcı olduğu ve komşu kanserli olmayan prostat dokusu ile karşılaştırıldığında medyanın 66 kat upregülasyon olduğu gösterilmiştir (Loeb ve Catalona,2014). Birçok çalışma, PCA3 skorlarının, biyopsi ile kanıtlanmış PK'li erkeklerin yüzdesi ile anlamlı derecede ilişkili olduğunu veduyarlılıklarının ve özgüllüklerinin sırasıyla% 76 ve% 90'a olduğunu raporlamıştır (De Kok v.d. 2002). Diğer araştırmalar, PCA3 skorlarının prostatektomi Gleason skoru ve tümör hacmi ile korelasyon gösterdiğini açıklamışlardır (Deros v.d., 2008).

Bununla birlikte, bu idrar temelli biyobelirteçin, biyopsi yapılması önerilen ve / veya agresif tümörü barındırabilen erkeklerin belirlenmesinde mevcut algoritmalara eklenme potansiyeli vardır. Genel olarak, PCA3 umut verici görünmektedir, ancak

bu sonuçların kesin olarak doğrulanması için daha geniş popülasyonlarda, daha ileri randomize prospektif çalışmalara ihtiyaç vardır.

2.3.6.4 Yapılan Çalışmalardan Çıkan Sonuç

Geçtiğimiz on yıl, genetik sıralama teknolojilerinde, PK'nin genetik temelini anlamamızı doğrudan ve orantılı olarak arttıran büyük ilerlemelere şahit oldu. Özellikle, artık PK'e yatkınlık ve agresif PK ile tekrar tekrar ilişkilendirilebilen neredeyse 100 farklı eşey SNP bulunmaktadır. Bu testlerin klinik uygulamaya dahil edilmesi;

1) PC taraması için hasta seçiminde
2) PSA yorumlamasında (örn. PSA SNP kullanarak genetik düzeltme kullanarak)
3) Biyopsi kararında (PC-risk SNP'leri kullanarak)
4) Tedavi kararında

destek olmakta ve gelişme kaydedilmesini sağlamaktadır.

Son olarak, PK'nin genetik temelinin belirlenmesinde mükemmel ilerleme kaydedilmiş olmasına rağmen, özellikle agresif tümörlerde, PK'e karşı genetik yatkınlığı daha iyi anlamak için daha fazla genetik araştırmaya ihtiyaç duyulmaktadır.

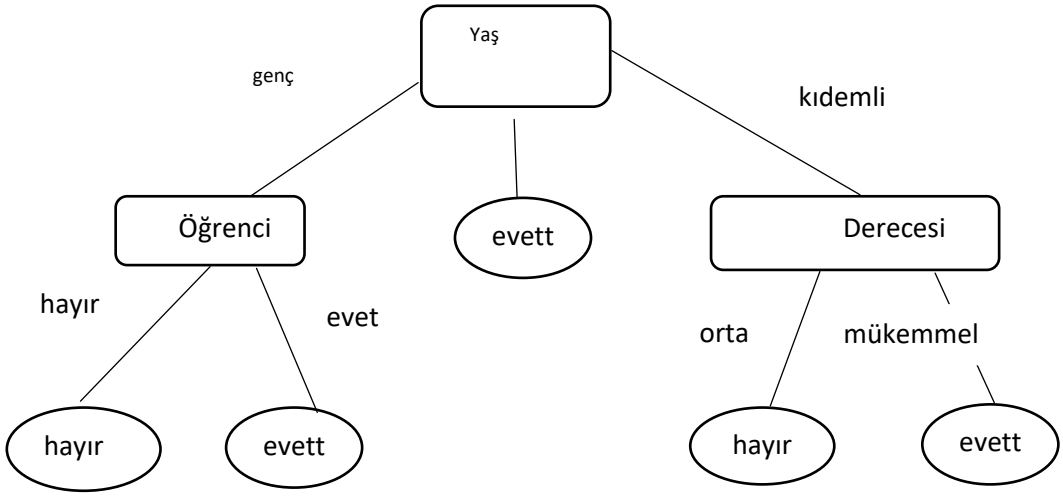
Bu tür genetik belirteçler, PSA taraması, aşırı tanı ve Pk'nin aşırı tedavisi hakkındaki güncel tartışmayı ele almak için son derece önemlidir. Bu genetik varyantların tanımlanması ve iyileştirilmesi, mevcut klinik uygulamaları iyileştirebilecek ek araçlara izin verecektir (Helfand, v.d., 2015).

ÜÇÜNCÜ BÖLÜM
UYGULAMADA KULLANILAN VERI MADENCILIĞI YÖNTEMLERININ AYRINTILI AÇIKLAMASI

3.1 Sınıflandırma

3.1.1. Karar Ağaçları

Karar ağaçları veri madenciliğinde en sık kullanılan tekniklerden birisidir. En büyük avantajlarından birisi, kullanımının kolay ve daha da önemlisi, uzman olmayanlar tarafından bile anlaşılmasının kolay olmasıdır. Karar ağaçlarının altında yatan temel fikir, böl ve al yaklaşımıdır. Her adımda, veri kümesi, farklı parçalara bölünürken, her parça, olası sınıflardan birini daha iyi temsil etmelidir. Karar ağacı süreç seması benzeri bir ağaç yapısıdır. Her dügüm (node) bir nitelikle ilgili testi gösterirken, her bir dal bir testin sonucunu, her yaprak (uç dügüm) ise bir sınıfı gösterir. Yani, karar düğümleri soruların sorulduğu, giriş verilerinin test edildiği ve hangi yöne yöneleceklerini belirler. Dal ise sorulan bu soruların cevaplarını temsil eder. Uç yapraklar da kategorilerin bulunduğu sınıf etiketlerini temsil eder.



Şekil 3.1 Karar Ağacı Örneği

Kök düğümden yaprak düğüme doğru giden herhangi bir yol bir karar kuralı oluşturmaktadır. Bir karar ağacında oluşan bütün kuralların toplamı da karar ağacının kendisini oluşturmaktadır (Akman,2010). Bu karar ağacında bir kişinin kredi alıp alamama durumu değerlendirilmektedir. En önemli kriter olarak yaş kök

düğümde yerini almaktadır. Şekil 3.1'de gösterilen karar ağacı için oluşturulan kurallar şu şekildedir.

Eğer kişinin yaşı genç ve öğrenci ise kredi alabilir.

Eğer kişinin yaşı genç ve öğrenci değil ise kredi alamaz.

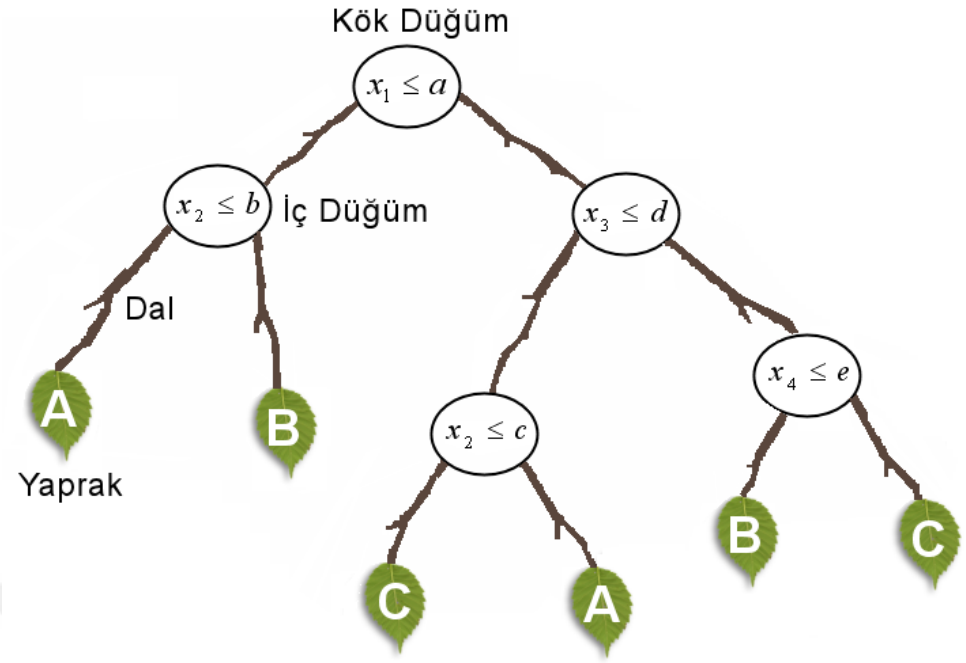
Eğer kişinin yaşı orta yaşlı ise kredi alabilir.

Eğer kişinin yaşı kıdemli ve kredi derecesi orta ise kredi alamaz.

Eğer kişinin yaşı kıdemli ve kredi notu mükemmel ise kredi alabilir

3.1.1.1. Karar Ağacının Temel Yapısı

Bir karar ağacı Şekil 3.2 'de görüldüğü üzere üç temel kısımdan oluşur;bunlar düğüm, dal ve yapraktır (Kavzoğlu ve Çölkesen, 2010). Bu ağaç yapısında her düğümde bir değişken vardır. Ağaç yapısının diğer elemanları ise dallar ve yapraklardır. Gerçek bitkilerde olduğu gibi bir karar ağacı yapısında da en üst kısım kök iken en son kısım yaprak adını alır . Dal ise kök ve yapraklar arasında kalan kısıma ise dal denir (Pal ve Mather,2003). Bir başka deyişle, bir ağaç yapısı; verilerin içinde bulunduğu bir kök düğümü, iç düğümler (dallar) ve uç düğümlerden (yapraklar) oluşur. Bir karar ağacı yapısı oluşturulmasındaki temel prensip, eğitim verilerine ait değişken bilgilerinden yararlanılarak verilere ilişkin sorulan bir dizi sorudan elde edilen cevaplar doğrultusunda hareket edilerek en kısa süre zarfında sonuca ulaşılması olarak ifade edilebilir. Karar ağacı, bu prensibi kullanarak sorduğu sorulara aldığı cevapları toplayarak karar kuralları oluşturur. Sorular ağacın ilk düğümü olan kök düğümünde sorulmaya başlanır. Böylece veriler sınıflandırılarak ağaç yapısı oluşturulur. Bu soru sorma işlemine dalları olmayan düğümler ya da yapraklar bulunana kadar devam edilir (Quinlan,1993). Test veri seti ise oluşturulan ağacın yeni bir veri seti için genellenip genellenemeyeceğinin belirlenmesi için kullanılır. Eğitim veri seti ile oluşturulan ağaç yapısına yeni gelen bir test verisi, ağacın kökünden girer. Test sonucuna göre, kökte test edilen bu yeni veri, bir alt düğüme gönderilir. Bu işleme ağacın belirli bir yaprağına gelene kadar devam edilir. Kökten her bir yaprağa tek bir yol veya tek bir karar kuralı ile gidilir



Şekil 3.2 Karar Ağacının Temel Yapısı

Yeni ve bilinmeyen bir durum daha sonra ağaçlardan biri yapraklardan birine ulaşıncaya kadar yönlendirilir. Her düğüm için, bu düğümde kullanılan değişken değerine bağlı olarak iki seçeneğimiz vardır. İç düğümlerden birinde bir test için bir nominal değişken kullanıyorsak, bu, veri kümesinin bu aşamada temel olarak bu değişkeninin farklı değerlerine göre bölünmüş olduğu anlamına gelir. Bu nedenle, bir nominal değişken bir kereden fazla test edilmeyecektir, çünkü daha sonraki tüm örnekler ağacın bu özel özellik için aynı değere sahip olacaktır. Bu, nümerik değişkenler için farklıdır: Burada, genellikle, değişken değeri, belirlenen bir sabitten daha büyük veya daha küçük mü diye test edilir. Değişken, farklı sabitler için birkaç kez test edilebilir (Tan, Steinbach ve Kumar, 2014).

3.1.1.2 Karar Ağacı Algoritması

Algoritma: Karar ağacı oluşturma. D veri bölümünün eğitim örneklerinden bir karar ağacı oluşturma (Han ve Kamber,2006)
Girdi: Veri bölümü
Bir takım eğitim örnekleri ve onlara karşılık gelen sınıf etiketlerinin kümesi olan D;
değişken listesi, aday değişkenlerin kümesi;
değişken seçim yöntemi, veri örneklerini sınıflara ayırmak için “en iyi” bölen ayrılma kriterlerini belirleme prosedürü

Bu kriter ya bir bölme noktası ya da bölme alt kümesi olan bir ayrılma özelliğinden oluşur

Çıktı: Bir karar ağacı.

Yöntem:

(1) bir N düğümü oluşturun;

(2) D'deki örneklerin hepsi aynı C sınıfındaysa,

(3) N düğümünü C sınıfı ile işaretlenmiş bir yaprak düğümü olarak döndürür;

(4) öznitelik listesi boşsa o zaman

(5) N'yi, D'de çoğunluk sınıfıyla etiketlenmiş bir yaprak düğümü olarak döndürür; // çoğunluk oyu

(6) “en iyi” ayırma kriterini bulmak için değişken seçim yöntemini (D, öznitelik listesi) uygula;

(7) bölme kriteri olan N etiket düğümü;

(8) ayırma değişkeni ayrıksa ve

çok yönlü bölünmeler izin veriliyorsa o zaman // ikili ağaçlarla sınırlı değil

(9) değişken listesi ok değişken listesi -bölme değişkeni; // bölme değişkenini kaldır

(10) bölme kriterinin her bir j sonucu için

// örnekleri bölümlere ayırın ve her bölüm için alt ağacı büyütün

(11) Dj D deki j sonucunu için yeterli veri örnekleri olmak üzere ; // bir bölüm

(12) eğer Dj boşsa o zaman

(13) D'de N düğümüne çoğunluk sınıfı olarak etiketlenmiş bir yaprak ekleyin;

(14)Dj boş değilse, N düğümüne karar ağacı oluşturma (Dj, öznitelik listesi) tarafından döndürülen bir düğüm ekleyin;

son

(15) N'i döndürün;

Algoritma

– Baslangıçta tüm ögrenme verisi agaç köküdedir

– Örnekler seçilen nitelige göre bölümlere ayrılır

– Nitelik seçimi için farklı matematiksel metrikler uygulanır (bilgi kazanımı – ID3, Gini Endeksi – CART, kazanım oranı – C4.5)

• Durma kriteri

– Bir dügüm için tüm örnekler aynı sınıfa aitse

– Bölmede kullanılacak nitelik kalmadıysa (tüm nitelikler agaca yerlestirildiyse)

– Hiç bir örnek kalmadıysa

3.1.1.3 Karar Ağaçlarında Kullanılan Algoritmalar

Karar ağaçları oluşturulurken hangi algoritmanın kullanıldığı önemlidir çünkü oluşturulan ağacın şekli kullanılan algoritmaya göre değişiklik gösterebilir. Farklı ağaç yapılarının oluşması demek farklı sınıflandırma sonuçları demektir. Bir ağaç yapısında kök düğümü oluşturan ilk düğümün farklı olması demek en uçtaki yaprağa ulaşırken izlenecek yolun farklı olması anlamında gelir. Bu da sınıflamayı değiştirecektir. (Silahtaroğlu,G.,2008).

3.1.1.3.1 CART Algoritması

CART Breiman, Friedman, Olshen ve Stone tarafından 1984 yılında geliştirilmiştir. CART algoritması ikili (binary) ağaç olarak büyüyen bir algoritmadır. Classification and Regression Trees kelimelerinin baş harfinden oluşmaktadır. Veri algoritma tarafından iki alt kümeye ayrılır. Böylece bir sonraki adımda oluşacak olan alt kümenin, bir önceki adımdaki kümeden daha homojen olması sağlanır.
Bu süreç kendini tekrarlayan bir süreçtir. Sonuç bulunana kadar devam eder. CART algoritması sınıflandırma ve regresyon analizi için kullanılabilen bir algoritmadır. Tek dezavantajı karmaşık bir algoritmadır. Büyük verilerle çalışıldığında sonuç bulması uzun sürmektedir (Şimşek,2006).

3.1.1.3.2 CHAID

CHAID (algoritması 1980 yılında Kaas tarafından geliştirilmiştir. CHAID algoritması ikili bir algoritma değildir. Bu da demektir ki bir ağaçta herhangi bir seviyede ikiden çok kategori üretebilir. Bu nedenle daha geniş ağaç yaratmaya eğilimlidir. Her tür değişken için kullanılabilen bir tekniktir. Chi-Squared Automatic Interaction Detector kelimelerinin baş harfinden oluşmaktadır. CHAID adından da anlaşılacağı gibi ayırma kriteri olarak Ki-kareyi kullanır. Oldukça başarılı bir karar ağacı tekniğidir.
CHAID algoritması, tahmin edici değişkenin tüm değerlerini dikkate alarak analiz yapar. Hedef değişkeni dikkate alarak istatistik olarak benzer olan değişkenleri birleştirir ve farklı olan değişkenlerle işlemi sürdürür. Daha sonra karar ağacının ilk dalını oluşturmak için en iyi tahmin edici değişkeni seçer. Her bir düğüm, seçilen değişkenin benzer değerlerinden oluşur. Bu süreç ağaç tamamıyla büyüyene kadar tekrarlanarak devam eder. Yapılacak testler hedef değişkenin türüne göre

değişmektedir. Eğer değişken kategorik (nominal/ordinal) bir değişken ise Ki- kare testi kullanılırken, sürekli bir değişken ise F testi kullanılır (Şimşek,2006).

3.1.1.3.3 C4.5

C4.5 1993 yılında Quinlan tarafından ortaya atılmıştır. Diğer algoritmalarla karşılaştırıldığında, C4.5 algoritması yeni bir karar ağacı algoritmasıdır. Bu algoritmada karar ağacı oluşturulurken kayıp veriler hesaba katılmaz (Dunham,2003). Kayıp veriler diğer veri ve değişkenler yardımı ile tahmin edilir. Böylece daha hassas ve daha anlamlı kurallar çıkartabilen bir ağaç üretilebilir (Doumpos,2002) C4.5 algoritması, kalitatif değişkenleri dikkate alır.

3.1.1.3.4 QUEST

QUEST (1997 yılında Loh ve Shih tarafından geliştirilmiş olan yeni bir tekniktir. Binary (ikili) büyüyen bir ağaç algoritmasıdır. Adını Quick, Unbiased, Efficient, Statistical Tree kelimelerinin baş harflerinden almıştır. Ayrı ayrı değişken seçimi ve ayırım noktası seçimi ile ilgilenir. QUEST'deki birim değişken ayıracı, tahmini olarak tarafsız değişken seçimini gerçekleştirir.
QUEST algoritmasının CART algoritmasına benzer avantajları olan bir algoritmadır, ancak ağaçlar yavaş büyüyebilir ve ikili olduğu için karar ağacı çok geniş olabilir (Şimşek, 2006).

3.1.1.4 Bilgi Kazanımı Yolu İle Nitelik Seçimi

Hangi kararın bilgi anlamında en yüksek kazancı doğurduğunu hesaplayabilmek için bilgiyi nasıl ölçebileceğimizi tartışacağız.
Bu bilginin bit olarak kolayca ölçülebileceğini ve bu miktarın hesaplanabilmesi için bir kavramın var olduğunu göreceğiz: entropi. Entropi, bir olayın olasılık dağılımı verilirse, bu olayın sonucunu tahmin etmek için, bitlerde gerekli bilgiyi (bitlerin kesirlerini de içerebilir) ölçmektedir.
Entropi, rastgeleliğin, belirsizliğin ve beklenmeyen durumun ortaya çıkma olasılığıdır.
Karar ağacı için sınıflandırmada bilgi kazancı ölçümü için entropi kullanılır**.**

- Entropi=0: Örneklerin aynı sınıfta olduğunu ifade eder.
- Entropi=1: Örneklerin sınıflar arasında eşit dağıldığını ifade eder.

- 0<entropi<1 ise örneklerin sınıflar arasında rastgele dağılmış olduğunu ifade eder.

N düğümü D bölümünün örneklerini temsil etsin. En yüksek bilgi kazancına sahip olan değişken node N için, ayrılma değişkeni olarak seçilmiştir.

Bu değişken, ortaya çıkan bölümlerdeki örnekleri sınıflandırmak için gereken bilgileri en aza indirir ve bu bölümlerdeki en az rastgeleliği veya safsızlığı yansıtır.

Böyle bir yaklaşım, verilen bir örneği sınıflandırmak için gerekli olan test sayısını minimuma indirir ve basit (ama en basit olmak zorunda değil) ağacın bulunduğunu garanti eder (Han ve Kamber,2006).

Formül şu şekildedir:

Entropi (D) = $-p_1\ log_2\ (p_1)$ - . . . - $p_m\ log_2\ (p_m)$

$$= -\sum_{i=1}^{m} p_i\ log_2(p_i) \qquad (3.1)$$

p_1, p_2,.., p_m toplamları 1 olan olasılıklardır.

Burada p_i, rastgele bir örneğin C_i sınıfına ait olma olasılığıdır ve I$C_{i,D}$I / IDI tarafından tahmin edilir.

2 tabanında bir log fonksiyonu kullanılır, çünkü bilgi bitlerde kodlanır

Entropi (D), D içinde bir örneğin sınıf etiketini tanımlamak için gereken ortalama bilgi miktarıdır.

Dikkat edersek, bu noktada sahip olduğumuz bilgi sadece her sınıfın örneklerinin oranına bağlıdır.

Entropi (D), Bilgi (D) olarak da bilinir.

Formülün küçük bir örnekle anlaşılması kolaydır. Tarafsız olarak bozuk bir parayı havaya atalım, her iki tarafın gelme olasılığı da eşit ve 0,5 dir.

Bu durumda,

entropi=-0.5 * ld (0.5) -0.5 * ld (0.5) = 1'dir.

Bozuk paranın hangi tarafının gerçekte geleceğine karar vermek için gerekli olan bilgiler bu nedenle 1 dir. Bu da eşit iki olası çıktısı olan bir ikili ile ilgili karara varabilmek için mükemmel bir değerdir.

Şimdi paranın taraflı olarak atıldığını varsayalım ve "% 75 olasılıkla tura gelirken % 25 olasılıkla yazı geldiğini var sayalım.

Bu durumda,

entropi -0.75 * ld (0.75) -0.25 * ld (0.25) = 0.81 olarak hesaplanacaktır.

Dolayısıyla karar vermek için daha az bilgiye ihtiyacımız vardır. Bu da doğal bir sonuçtur çünkü zaten tura gelmesini daha yüksek bir olasılıkla bekleriz. Burada rastgelelik ve belirsizlik ilk duruma göre daha azdır.

Şimdi,{a1, a2,..., av} gibi v ayrık değere sahip A değişkeninin D'deki örneklerini eğitim verilerinden gözlemlendiği böleceğimizi düşünelim

A, ayrık değerli ise, bu değerler, A'daki bir testin v çıktılarına doğrudan karşılık gelir. A değişkeni D'yi {D1, D2,...., Dv} şeklinde v parçaya bölmek için kullanılabilir. Bu bölümler, N düğümünden büyütülen dallara karşılık gelir. İdeal olarak her bölümün saf olmasını isteriz. Bununla birlikte, bölümlerin saf olmaması muhtemeldir (örneğin, bir bölüm, tek bir sınıftan ziyade farklı sınıflardan bir dizi örnek içerebilir). Tam bir sınıflandırmaya ulaşmak için (bölümlendirme sonrası) ne kadar daha bilgiye ihtiyacımız var sorusunun yanıtı

$$\text{Entropi}_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} \times \text{entropi}(D_j) \quad (3.2)$$

terimi ile ölçülür.

Buna herhangi bir A Niteliğinin bir örnek için entropisi denir.

$\frac{|D_j|}{|D|}$ j. bölümünün ağırlığı gibi davranır. $\text{Entropi}_A(D)$, A tarafından bölümlemeye dayanan D'de bir örneğin sınıflandırılması için gerekli beklenen bilgidir.

A bölümlendikten sonra hala gerekli olan beklenen bilgiler ne kadar küçükse, bölümlerin saflığı da o kadar büyük olur.

Bilgi kazancı, orijinal bilgi gerekliliği (yani, sadece sınıfların oranına bağlı olarak) ile yeni gereklilik (yani A yi bölümlendirdikten sonra elde edilen)
arasındaki fark olarak tanımlanır.

Bilgi Kazancı (A) = Entropi (D) - Entropi A (D) (3.3)

Kazancı en çok olan nitelik seçilir.

3.1.1.5 Bilgi Kazanım Yoluyla Nitelik Seçimi Örneği

Aşağıdaki örnekte hava durumu, sıcaklık, nem ve rüzgar değişkenleri ile birlikte tenis oynayıp oynamama kararı bilgileri yer almaktadır. Bu değişkenlere göre tenis oynama kararının karar ağacını çıkaracağız.

Tablo 3.1 Tenis Oynama Kararı İçin Veri Seti

GÜN	HAVA DURUMU	SICAKLIK	NEM	RÜZGAR	TENİS OYNAMA KARARI
1	Güneşli	Sıcak	Yüksek	Hafif	Hayır
2	Güneşli	Sıcak	Yüksek	Şiddetli	Hayır
3	Bulutlu	Sıcak	Yüksek	Hafif	Evet
4	Yağmurlu	Ilık	Yüksek	Hafif	Evet
5	Yağmurlu	Soğuk	Normal	Hafif	Evet
6	Yağmurlu	Soğuk	Normal	Şiddetli	Hayır
7	Bulutlu	Soğuk	Normal	Şiddetli	Evet
8	Güneşli	Ilık	Yüksek	Hafif	Hayır
9	Güneşli	Soğuk	Normal	Hafif	Evet
10	Yağmurlu	Ilık	Normal	Hafif	Evet
11	Güneşli	Ilık	Normal	Şiddetli	Evet
12	Bulutlu	Ilık	Yüksek	Şiddetli	Evet
13	Bulutlu	Sıcak	Normal	Hafif	Evet
14	Yağmurlu	ılık	Yüksek	Şiddetli	Hayır

1. Adım: Sistem entropisi hesaplanır.

9 tane evet, 5 tane hayır var.

$$\text{Entropi(Sistem)} = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right) = 0{,}94$$

2. Adım: Kök değişken belirlemek için her değişkenin kazancı hesaplanır.

Hava Durumu değişkeni için bilgi kazancı;

Tenis oynama kararı	Güneşli	Bulutlu	Yağmurlu
Evet	2	4	3
Hayır	3	0	2

Entropi (hava durumu) = 5/14* E (2,3) + 4/14 *E (4,0) + 5/14* E (3,2) = 0,694

$$= \frac{5}{14} \times \left(-\frac{2}{5}log_2\left(\frac{2}{5}\right) - \frac{3}{5}log_2\left(\frac{3}{5}\right)\right) + \frac{4}{14} \times \left(-\frac{4}{4}log_2\left(\frac{4}{4}\right) - \frac{0}{4}log_2\left(\frac{0}{4}\right)\right) + \frac{5}{14} \times \left(-\frac{3}{5}log_2\left(\frac{3}{5}\right) - \frac{2}{5}log_2\left(\frac{2}{5}\right)\right) = 0{,}694$$

Kazanç (hava durumu) = Entropi (Sistem)- Entropi (hava durumu)

= 0,940- 0,694

= 0, 246

Aynı şekilde entropi (sıcaklık), entropi (nem) ve entropi (rüzgar) hesaplanır.

Sıcaklık değişkeni için bilgi kazancı;

Tenis oynama kararı	Sıcak	Ilık	Soğuk
Evet	2	4	3
Hayır	2	2	1

Entropi (sıcaklık) =4/14 x E (2,2) +6/14 x E (4,2) + 4/14 x E (3,1) = 0,913
Kazanç (sıcaklık) = 0,027

Nem değişkeni için bilgi kazancı;

Tenis oynama kararı	Yüksek	Normal
Evet	3	6
Hayır	4	1

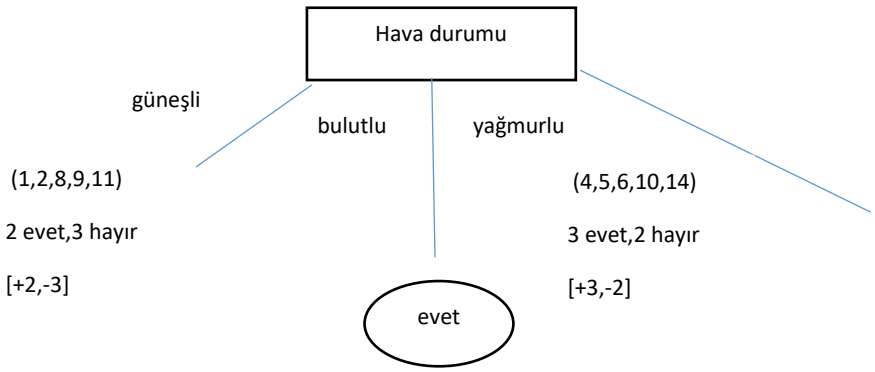
Entropi (nem) =7/14x E (3,4) +7/14 x E (6,1) = 0,790
Kazanç (nem) = 0,15

Rüzgar değişkeni için bilgi kazancı;

Tenis oynama kararı	Hafif	Şiddetli
Evet	6	3
Hayır	2	3

Entropi (rüzgar) =8/14x E (6,2) +6/14 x E (3,3) = 0,892
Kazanç (rüzgar) =0,048

Hava durumu değişkeni değişkenler arasında en yüksek bilgi kazancına sahip olduğu için bölme değişkeni olarak seçilir. Baştaki ana düğüm hava durumu değişkeni olarak etiketlenir. Dallar da her bir değişkenin değerine göre büyür. Örnekler daha sonra bu değişkenlere göre ayrılırlar. Şekil 3.3'de gösterilmiştir.
Benzer şekilde karar ağacının tamamı oluşturulur.



Şekil 3.3. Karar Ağacının İlk Basamağı

3.1.1.6 Rapidminerda Karar Ağacı Operatörü

Dikkat edilmesi gereken ana parametreler kriter menüsü ve minimum kazanç kutusudur. Kriterlerin her biri bir ayırma kriteridir ve bilgi kazancı, Gini endeksi ve kazanç oranından biri seçilir. Karar ağaçları en basit şekilde her bölünmeden sonra indirgenmiş veri kümesinde yer alan bilgileri arttırarak beş adımda inşa edilmiştir. Veriler doğası gereği belirsizlikler içerir. Belirsizlikleri sistematik olarak azaltabiliriz ve böylece sıralama veya sınıflandırma gibi aktivitelerle bilgiyi arttırabiliriz. Belirsizliği en aza indirgemek için sınıflandırdığımızda veya sıraladığımızda, bilgi anlamında büyük artış başarırız. Entropi iyi bir belirsizlik ölçüsüdür ve entropinin nasıl izleneceği, bilgiyi nicelleştirmemize olanak tanır. Bu da bizi RapidMiner içinde karar ağaçlarını bölmek için sunulan seçeneklere geri götürür. Aşağıda bu kriterler anlatılmıştır (Desphande ve Kotu, 2014)

1. Bilgi kazancı: Basitçe söylemek gerekirse, bilgi kazancı bölünmeden önceki bilgiden bölünmeden sonraki bilginin çıkartılması olarak hesaplanır. Çoğu durumda iyi çalışır yalnız çok sayıda değere sahip az değişken içeren kümeler için çok iyi çalışmaz. Bilgi kazancı çok sayıda değer içeren değişkenleri kök düğüm olarak seçme konusunda taraflıdır. Bu durum extrem durumlar dışında problem değildir. Örneğin, her müşteri kimliği benzersiz ve dolayısıyla değişken çok fazla değere sahiptir (her kimlik numarası farklıdır). Bu kimlik numaralarını ifade eden satırlar boyunca bölünmüş bir ağacın tahmini değeri yoktur.
2. Kazanç oranı (varsayılan): Bu bilgi kazanımının taraflı tutumunu azaltan bir modifikasyonudur. Genellikle en iyi seçenektir. Kazanç oranı bilgi kazancının bölünmeden önce elde edilecek dal sayısını dikkate alan yaklaşımındaki problemin üstesinden gelir. Bilgi kazancını bir bölünmenin gerçek bilgisini dikkate

alarak düzeltir. Gerçek bilgiden ne kast edildiği aşağıdaki örnek kullanılarak açıklanacaktır. 14 örneğin her birinin farklı bir kimlik numarasını (ID) ifade ettiğini varsayalım.

Öyleyse ID değişkeni için gerçek bilgi değeri

14 * (–1/14 * log (1/14)) = 3.807 şeklinde hesaplanır. Bir değişken için bilgi kazancının gerçek bilgisine bölünmesi ile kazanç oranı elde edilir.

Açıkça çok yüksek gerçek bilgiye sahip olan nitelikler (yüksek belirsizlik) bölünme üzerine düşük kazançlar sunma eğilimindedir ve dolayısıyla otomatik olarak seçilemez.

3. Gini Indeksi: Bu da bazen kullanılır, ancak kazanç oranı ile karşılaştırıldığında çok da bir avantajı yoktur.
4. Doğruluk: Bu aynı zamanda performansı artırmak için kullanılır. En iyi yol

Bu parametreler için en iyi yol optimizasyon operatörlerinin kullanılarak değerlerin seçilmesidir.

Diğer önemli parametre minimum kazanç değeridir. Teorik olarak bu

0'dan yukarı herhangi bir aralıkta olabilir. Pratikte, minimum 0,2 ila 0,3 kazanç

iyi sayılır. Varsayılan 0,1'dir.

Diğer parametreler (bölünme için minimum bütyüklük , minimum yaprak büyüklüğü, maksimum derinlik)

veri kümesinin büyüklüğüne göre belirlenir. B Bu parametreleri ayarlamanın en iyi yolu optimize ederek değerlerin seçilmesidir.

3.1.2 Naive Bayes

Bayes sınıflandırıcıları istatistiksel sınıflandırıcılardır. Bu sınıflandırıcılar, sınıf üyelik olasılıkları da denilen bir sınıfın belirli bir sınıfa ait olma olasılıklarını tahmin edebilirler.

Bayes sınıflandırması aşağıda tanımlanacak olan Bayes teoremine dayanmaktadır. Sınıflandırma algoritmalarını karşılaştıran araştırmalar, karar verme ağacı ve seçilmiş sinir ağı sınıflandırıcılarıyla performans açısından karşılaştırılabilir olan naif bayes sınıflayıcısı olarak bilinen basit bayes sınıflandırıcıyı bulmuştur (Han ve Kamber,2006).

Bayes sınıflandırıcıları, büyük veri tabanlarına uygulandığında yüksek doğruluk ve hız sergilemiştir.

Naive Bayes sınıflandırıcıları, belirli bir sınıfta bir özellik değerinin etkisinin diğer özellik değerlerinden bağımsız olduğunu varsayar.

Bu varsayıma, sınıf koşullu bağımsızlık denir. Bu varsayım ilgili hesaplamaları basitleştirmek için yapılır ve bu anlamda "naif" olarak kabul edilir. Bayes inanç ağları, naif bayes sınıflandırıcılarının aksine, özelliklerin alt kümeleri arasındaki bağımlılıkların temsil edilmesine izin veren grafik modellerdir. Sınıflandırma için Bayes inanç ağları da kullanılabilir.

3.1.2.1 Bayes Teoremi

Bayes teoremine göre:

$$P(H/X) = \frac{P(X/H).P(H)}{P(X)}$$ şeklindedir. (3.4)

Bayes teoremi ismini 18. yüzyılda olasılık ve karar teorisinde erken çalışmalar yapan, toplum kurallarına uymayan İngiliz bir papaz olan Thomas-Bayes'den almıştır.

X bir veri örneği olsun. Bayes terimleriyle X, "delil" olarak kabul edilir. Genel olarak X, n değişkenli bir küme üzerinde yapılan ölçümlerle tanımlanır.

H, "X veri takımı belirli bir C sınıfına aittir" şeklinde bir hipotez olsun. Sınıflandırma problemleri için, verilen delil için veya gözlemlenen X veri örneği için H hipotezinin gerçekleşme olasılığını ifade eden P (H / X) olasılığını belirlemek istiyoruz. Diğer bir deyişle, X örneğinin özelliğini bildiğimiz durumda X örneğinin C sınıfına ait olma olasılığını arıyoruz.

P (H/X), H'nin X üzerinde koşullu olasılığıdır.

Veri örneklerinin sırasıyla yaş ve gelir değişkenleriyle tanımlanan müşterilere ait olduğunu ve X'in 35 yaşında 40.000 TL'lik gelire sahip olan bir müşteri olduğunu varsayalım. H hipotezi ise "müşteri bilgisayar alacaktır" şeklinde olsun. O zaman P (H / X) olasılığı , 35 yaşında 40.000 TL'lik gelire sahip olan bir X müşterisinin bir bilgisayar satın alma olasılığını ifade eder (Hofmann ve Klinkenberg,2014).

Buna karşılık, P (H), H'nin ön olasılığı veya öncül olasılığıdır. Bir bilgisayar satın alma olasılığını ifade eder. Bu olasılık herhangi bir müşterinin yaşa, gelir ya da başka herhangi bir bilgiye bakılmaksızın bilgisayar satın alma olasılığını ifade eder. Ardıl olasılık, P (H / X), X'den bağımsız olan P (H) olasılığından daha çok bilgi içerir (ör. Müşteri bilgisi gibi).

Benzer şekilde, P (X / H), H üzerinde koşullandırılan X'in ardıl olasılığıdır, ya da X'in H üzerindeki koşullu olasılığıdır.

Yani bilgisayar alacak olan müşterinin 35 yaşında olma ve 40.000 TL. gelir kazanma olasılığını ifade eder. Bu olasılıkların nasıl tahmin edileceği sorusuna gelince; verilen

veriden P (H), P (X / H) ve P (X) tahmin edilebilir. Bayes teoremi P (H),P (X / H) ve P (X) 'den ardıl olasılık P (H / X)'i hesaplamanın bir yolunu sunduğu için faydalıdır.

3.1.2.2 Naive Bayes Algortiması

Hoffman ve Klinkenberg'e göre Naive Bayes algoritması, güçlü bağımsızlık varsayımlarıyla Bayes teoreminin uygulanmasına dayanan basit olasılıkçı bir sınıflandırmadır. Basitçe söylemek gerekirse, Naive Bayes algoritması, bir özelliğin belirli bir değerinin varlığının diğer herhangi bir özellik değerinin varlığıyla ilgisi olmadığını varsayar (2014). Örneğin, yeşil, yuvarlak ve çapı 6,5 cm olan bir top tenis topu olarak sınıflandırılabilir. Bu özellikler birbirine bağlı olsa bile, Naif Bayes algoritması bu özelliklerin tümünü, bu topun bir tenis topu olma olasılığına bağımsız olarak katkıda bulunmak için değerlendirir. X'i sınıflandırmak istediğimiz bir örnek olsun. X, bir dizi n niteliğinde yapılan ölçümlerle açıklanmaktadır. H hipotezi: " X örneği C sınıfına aittir" şeklinde olsun.

Sınıflama problemleri için P (H / X)' i belirlemek istiyoruz. Burada P (H/X) olasılığı verilen X örneği için H hipotezinin olasılığını belirtir. Diğer bir deyişle, X'in öznitelik tanımının bilindiği durumda X örneğinin C sınıfına ait olma ihtimalini arıyoruz.

P (H / X), H'nin X üzerinde koşullandırılmasının ardıl olasılığıdır. Örneğin, veri kümesi, sıcaklık ve hava durumu gibi verilen hava koşullarında golf oynanıp oynanmayacağı ile ilgili olsun. Böylece, sıcaklık ve hava durumu bizim özelliklerimizdir. Örnek X, 25 derece sıcaklık ve güneşli bir havayı ifade etsin. H hipotezi de "Verilen hava koşullarında golf oynanacaktır" şeklinde olsun. P (H / X), sıcaklığı ve hava durumunu X hava koşullarında golfun oynanma olasılığını ifade eder. Koşullu olasılık P (H / X) ,X'den bağımsız olan öncül olasılık P (H) 'den daha fazla bilgi, yani hava durumu bilgisi içerir. Benzer şekilde, P (X / H), H üzerinde koşullandırılan X'in ardıl olasılığıdır. Yani bu olasılık golfun oynanacağı durumlarda X havasının 25 derece sıcaklıkta ve güneşli olma olasılığını ifade eder. P (X), X'in öncül olasılığıdır. Örneğimizde, tüm veri setinde havanın 25 derece ve sıcak olma olasılığını ifade eder.

Örnek: Aşağıdaki örnekte tenis oynama kararı ve bu kararı almak için çeşitli değişkenler verilmiştir.

Bu değişkenler; hava durumu, sıcaklık, nem ve rüzgar durumudur. Bu verilere göre tenis oynanıp oynayıp oynamama kararını naif bayes yöntemi ile vereceğiz.

Tablo 3.2 Tenis Oynama Kararı Almak İçin Kullanılacak Veri Seti

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

P (evet)= 9/14

P (hayır)= 5/14

Test: H**ava durumu= güneşli, sıcaklık= soğuk, nem= yüksek, rüzgar=** şiddetli iken

Maks { 9/14*2/9*3/9*3/9*3/9, 5/14*3/5*1/5*4/5*3/5} = {0,0053, 0,0206}= {0,0206}

Sonuçta tenis oynamama kararı çıkıyor.

3.1.2.3 Naive Bayes Sınıflandırıcısı

Naive Bayes sınıflandırıcı veya basit Bayes sınıflandırıcısı aşağıdaki gibi çalışır (Han ve Kamber,2006):

1. D, örneklerin eğitim kümesi olsun ve bunlarla ilişkili sınıf etiketleri olsun. Genel olarak, her örnek n-boyutlu bir öznitelik vektörü, X = (x1, x2, ...,xn) ile temsil edilir, bu n değişken li örnek üzerinde yapılan ölçümler sırasıyla , A1, A2,…, An. Şeklindedir.

2. C1, C2,….., Cm'in m tane sınıf olduğunu varsayalım. X örneği verildiğinde, sınıflandırıcı X'in en yüksek ardıl olasılığa sahip olan sınıfa ait olduğunu tahmin edecektir .Yani, naif bayes sınıflandırıcısı, X örneğinin Sınıf Ci'ya ancak ve ancak P (Ci / X) >P (Cj / X) ; i ≤j≤ m olması durumunda ait olduğunu tahminleyecektir. Dolayısıyla, P (Ci / X) 'in maksimize edilmesi gerekiyor. P (Ci / X) 'in maksimize edildiği sınıf Ci'ye, Maksimum posterior (ardıl) hipotezi denir.

Bayes Teoremi'ne göre;

P (Ci /**X**) =P (**X** /Ci).P (Ci) / P (**X**)

3. P (X) tüm sınıflar için sabit olduğu için yalnızca P (X / Ci) P (Ci) maximize edilir. Sınıf öncül olasılıklar bilinmiyorsa, sınıfların eşit olduğu, yani P (C1) = P (C2) = _ _ = P (Cm) olduğu varsayılır ve bu nedenle P (X / Ci) maximize edilir.

Aksi takdirde, P (X / Ci). P (Ci) 'yi maksimize ederiz. Sınıf öncül olasılıkları P (Ci) = I$C_{i,D}$I / IDI, tarafından tahmin edilir; burada, I$C_{i,D}$I, D sınıfındaki eğitim örneklerinin sayısıdır.

4. Birçok özelliğe sahip veri kümeleri göz önüne alındığında, P (X / Ci) hesaplamak son derece pahalı olacaktır. Bu maliyeti azaltmak için sınıfsal koşullu bağımsızlığın naif varsayımı yapılır. Bu varsayım şu şekildedir: Örneğin sınıf etiketi verildiğinde özelliklerin değerleri koşullu olarak birbirinden bağımsızdır. Diğer bir değişle değişkenler arasında bağımlı bir ilişki yoktur. Böylece

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i). \quad (3.5)$$

Eğitim örneklerinden P (x1 / Ci), P (x2 / Ci),:::, P (xn / Ci) olasılıklarını çok kolay bir şekilde tahmin edebiliriz. Burada xk , X kümesi için Ak özniteliğinin değerini belirtir. Her öznitelik için, özelliğin kategorik veya sürekli değerli olup olmadığına bakılır. Örneğin, P (X / Ci) 'yi hesaplamak için aşağıdakileri göz önünde bulunduruyoruz:

(A) Eğer Ak kategorik ise, o zaman P (xk / Ci), D içindeki Ci sınıfındaki örnek sayısına karşılık gelir

(B) Ak sürekli değerli ise, şunu hesaplamak gerekir; Sürekli değerli bir değişkenin genellikle μ ortalama ve σ standart sapma ile bir Gauss dağılımına sahip olduğu varsayılır.

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (3.6)$$

Böylece

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}). \quad (3.7)$$

Örneğin, X = (35, 40,000TL) olsun, burada A1 ve A2 yaş ve gelir değişkenleri olsun Sınıf etiketi özelliği "bilgisayar almak" olsun. X'e karşılık gelen sınıf etiketi evet (yani, bilgisayar satın alıyor = evet). Diyelim ki yaş aralıklı değil de sürekli bir değişken olarak ifade edilmiş olsun. Eğitim setinden, bir bilgisayar satın alan D'deki

müşterilerin yaşlarının 38± 12 olduğunu bulduğumuzu varsayalım. Bir başka deyişle, bu sınıftaki yaş değişkeni için μ = 38 yaş ±12 yaş'tır.

5. X sınıf etiketini tahmin etmek için P (X / Ci) P (Ci) her Ci sınıfı için hesaplanır. Sınıflandırıcı, X kümesinin sınıf etiketinin ancak ve ancak P (X / Ci) P (Ci)> P (X / Cj) P (Cj) Ci 1≤ j ≤m i≠j olduğu durumda Ci sınıfı olduğunu öngörür. Başka bir deyişle, tahmin edilen sınıf etiketi, P (X/Ci) P (Ci)'yi maksimum yapan Ci sınıfıdır.

"Bayes sınıflandırıcıları ne kadar etkilidir?" Bu sınıflandırıcının ampirik çalışmaları Karar ağacı ve sinir ağı sınıflandırıcıları ile karşılaştırıldığında bazı alanlarda karşılaştırılabilir oldukları bulunmuştur. Teorik olarak, Bayes sınıflandırıcılarının diğer tüm sınıflandırıcılara kıyasla minimum hata oranı vardır. Bununla birlikte, pratikte, sınıf koşullu bağımsızlığı ve mevcut olasılık verilerinin olmaması gibi varsayımlarda kullanılan yanlışlıklar nedeniyle her zaman böyle değildir.

3.1.2.4 Naive Bayes Avantaj ve Dezavantajları

Avantajlar:

– Kolayca uygulanır

– Çogunlukla iyi sonuç verir

• Dezavantajlar

– Sınıflar ve nitelikler arası kosulsal bagımsızlık varsayımı vardır

– Niteliklerin bagıntılı oldugu durumlar modellenemez

3.1.2.5 Rapidminer'da Naive Bayes Operatörü

RapidMiner'deki Naive Bayes operatörünün yalnızca bir parametresi vardır, o da Laplace doğrulama (Laplace correction) 'dır. Bu uzman parametre, Laplace düzeltmesinin sıfır olasılığını önlemek için kullanılıp kullanılmayacağını belirtir. Sıfır olasılıklardan kaçınmak için, eğitim verilerinin çok büyük olduğu ve ihtiyacımız olan her sayıya bir tane eklenmesinin tahmini olasılıklarda önemsiz bir fark yaratacağı varsayımı kullanılabilir. Bu teknik Laplace düzeltme olarak bilinir (Hoffman ve Klinkenberg, 2014).

Bayesian sınıflama yöntemi kosullu olasılıkların sıfırdan farklı oldugunu varsayar (diger türlü tahmin olasılıgı sıfır olur)

• 1000 örnekten olusan bir veri düsünelim:

– Gelir düzeyi düsük – 0 kisi

– Gelir düzeyi orta – 990 kisi

Gelir düzeyi yüksek 10 kisi

•Test örneginde gelir düzeyi düsük ise ne olur?

• Bu problemi Laplace düzeltmesi ile çözeriz:

P olasılığı ifade etmek üzere;

P (gelir = düsük) = 1/1003

P (gelir = orta) = 991/1003

P (gelir = yüksek) = 11/1003

3.1.3 K En Yakın Komşuluk

K-en yakın komşu yöntemi (K-NN Yöntemi) ilk olarak 1950'lerin başında tanımlanmıştır. Bu yöntem, büyük eğitim setlerine uygulandığında çok uğraştırıcı olmuştur ve artan hesaplama gücünün elde edildiği 1960'lara kadar popülerlik kazanamamıştır. O zamandan beri model tanıma alanında yaygın bir şekilde kullanılmıştır (Han ve Kamber,2006) .

K-NN algoritması, en temel örnek tabanlı öğrenme algoritmaları arasındadır. Örnek tabanlı öğrenme algoritması demek, eğitim setinde tutulan verilere dayalı olarak öğrenme işleminin gerçekleşiyor olması demektir. Yeni bir örnek ile karşılaşıldığında bu örnek, eğitim setinde yer alan örnekler ile arasındaki benzerliğe bakılarak sınıflandırılmaktadır.

En yakın komşu sınıflandırıcılar benzetme yoluyla öğrenmeye dayanır; diğer bir deyişle, gözlemlenen test örnekleri, ona benzeyen eğitim örnekleriyle karşılaştırılır. K-NN algoritmasında, eğitim setindeki örnekler n boyutlu sayısal değişkenler ile gösterilir. Tüm örnekler n boyutlu bir örnek uzayında saklanır ve her örnek n boyutlu uzayda bir noktayı temsil eder. Bilinmeyen bir örnek ile karşılaşıldığında, yeni örneğin sınıf etiketi belirlenirken, eğitim setinden ilgili örneğe en yakın k tane örnek belirlenerek, belirlenen bu k tane en yakın komşusunun sınıf etiketlerinin çoğunluk oylamasına bakılır . Bu k eğitim seti bilinmeyen veri setinin k "en yakın komşuları" dir (Tan, Steinbach ve Kumar, 2014).

"Uzaklık", Öklid mesafesi veya başka bir mesafe metriği ile tanımlanır. Uzaklık ölçütlerine ayrıntılı bir şekilde ileride değinilecektir.

Genellikle, mesafe metriği hesaplanmadan önce her değişkenin değeri normalize edilir.

Bu normalize işlemi başlangıçta geniş aralıklara (gelir gibi) sahip değişkenlerin başlangıçta daha küçük aralıklara sahip (ikili öznitelikler gibi) değişkenlerden daha ağır basmasını önlemek içindir.

Örneğin, sayısal bir A değişkeninin Min-maks normalizasyonu [0, 1] aralığında v değerini v' değerine dönüştürmek için şunu hesaplayarak kullanılabilir (Han ve Kamber,2006):

$$V' = \frac{v - minA}{maxA - minA} \quad (2.8)$$

Burada minA ve maxA sırasıyla A değişkeninim minimum ve maksimum değerlerinin ifade eder.

3.1.3.1 K-NN'de Sayısal ve Kategorik Değişkenlerin Sınıflandırılması

K-en yakın komşu sınıflandırması için, bilinmeyen örnek, k en yakın komşuları arasında en genel sınıfa atanmaktadır. k = 1 olduğunda bilinmeyen örnek, model uzayında kendisine en yakın olan sınıfa tayin edilir.

En yakın komşu sınıflandırıcıları, tahmin için kullanılabilir, yani bilinmeyen bir örnek için gerçek değerli bir tahminde bulunmak için kullanılabilir. Bu durumda, sınıflayıcı, bilinmeyen örneğin en yakın k komşularına karşılık gelen gerçek değerli etiketlerin ortalama değerini döndürür.

"Ancak değişken, sayısal değil de kategorik ise, örneğin renk gibi değişkenler için nasıl hesaplanabilir?" Yukarıdaki tartışma, örnekleri tanımlamak için kullanılan değişkenlerin hepsinin sayısal olduğunu varsaymaktadır. Kategorik değişkenler için, basit bir yöntem, X1 kümesindeki değişken değerini, X2 kümesindeki değişken değeriyle karşılaştırmaktır. Eğer ikisi aynı ise (örn., X1 ve X2 küpleri her ikisi de maviye sahipse), ikisi arasındaki fark 0 olarak alınır. Eğer ikisi farklıysa (örn. X1 kümesi mavi, ancak X2 kümesi kırmızıdır), o zaman ikisi arasındaki fark 1 olarak alınır.

Diğer yöntemler, diferansiyel derecelendirme için daha karmaşık düzenleri içerebilir (örneğin, mavi ve beyaz için, mavi ve siyah için olduğundan daha büyük bir fark puanı tahsis edilir).

3.1.3.2 K-NN'de Eksik Değerlere Yaklaşım

"Eksik değerler ne olacak?" Genel olarak, belirli bir A niteliğinin değeri, X1 ve X2 kümelerinde eksikse, mümkün olan maksimum farkı alırız. Her özniteliğin [0, 1] aralığına eşlendiğini varsayalım. Kategorik nitelikler için, A'nın karşılık gelen

değerlerinden birinin veya her ikisinin eksik olması durumunda, fark değeri 1 olarak alınır. A, nümerik bir değişken ise ve her iki X1 ve X2 kümesinde de eksikse, fark 1 olarak alınır. Yalnızca bir değer eksikse ve diğeri (ki v' olarak adlandıracağız) mevcut ve normalleştirilmişse, o zaman farkı ya I1-v' I veya I0-v' I şeklinde alabiliriz. (Yani, 1-v' veya v'), hangi değer daha büyükse o değer kullanılır.

3.1.3.3 K-NN'de k Parametresi İçin En Uygun Değerin Bulunması

"K için iyi bir değer nasıl bulunabilir, komşuların sayısı nasıl bulunur?" Bu, deneysel olarak belirlenebilir. K = 1 ile başlayarak, sınıflandırıcının hata oranını tahmin etmek için bir test seti kullanılır.

Bu işlem her seferinde bir daha fazla komşuya izin vermek için k'yi artırarak tekrar edilebilir. Minimum hata oranını veren k değeri seçilir. Genel olarak, eğitim örneklerinin sayısı ne kadar büyük olursa, k'nin değeri de o kadar büyük olur (böylece sınıflandırma ve tahmini kararları depolanmış örneklerin daha büyük bir bölümüne dayanır).

k=1 iken eğitim örneklerinin sayısı sonsuza yaklaştığında, hata oranı bayes hata oranının iki katından daha kötü olamaz. Eğer k da sonsuzluğa yaklaşırsa, hata oranı bayes hata oranına yaklaşır.

3.1.3.4 K-NN Avantajları ve Dezavantajları

En yakın komşu sınıflandırıcılar, her değişkene eşit ağırlık atayan uzaklık temelli karşılaştırmaları kullanır (Liu ve Zhang,2012). Bu nedenle, gürültülü veya alakasız değişkenler söz konusu olduğunda doğruluk değeri düşebilir. Bununla birlikte, yöntem değişken ağırlıklandırması ve gürültülü veri örneklerinin budaması için modifiye edilmiştir. Bu da yöntemi gürültülü eğitim verilerine karşı dirençli kılar. K-NN algoritması basit bir yapıya sahiptir bu yüzden az sayıda parametresi vardır. K-nn algoritmasını gerçekleştirmek kolaydır, kolay uyarlanabilirdir. Eğitim kümesi büyük olduğunda, K-NN algoritması, gayet iyi sonuçlar vermektedir. Bu sebeplerden dolayı sınıflandırma uygulamalarında özellikle tercih edilmektedir Bhatia ve Vandana, 2010). Bu avantajlarının yanında bazı dezavantajları da beraberinde getirir; K-NN algoritmasının en büyük dezavantajı, hesaplama maliyetinin yüksek olmasıdır. Büyük veri setleri için, sınıf etiketi belirlenmek istenen örnek ile veri setinde yer alan örnekler ile arasındaki uzaklığın belirlenmesi, maliyetli bir iş olabilmektedir. K-NN

algoritmasının bu maliyetli yapısını ortadan kaldırmak için temel bileşenler analizi gibi boyut azaltma yöntemleri ile birlikte kullanılabilmektedir (Shmueli,Patel ve Bruce,2010). K-NN algoritmasının bir diğer dezavantajı ise çok boyutlu veri setlerinde etkin olmamasıdır. Çünkü k-nn algoritması yüksek belleğe ihtiyaç duyar ve uzaklık ölçütü ve komşu sayısı gibi parametrelere duyarlıdır (Duda,Hart ve Stark,2000).

O zaman kısaca bu dezavantajları değişken ve veri seti büyüklüğü arttıkça işlem maliyetin ve işlem süresinin önemli ölçüde artması, performansın k komşu sayısı, uzaklık ölçütü ve değişken sayısı gibi parametre ve değişkenlere bağlı olarak etkilenmesi, yüksek bellek alanına ihtiyaç duyulması, şeklinde sayabiliriz.

Ayrıca, en yakın komşu sınıflandırıcılar test örneklerini sınıflandırırken son derece yavaş olabilir (Onan ve Taşçı,2016).

3.1.3.5 K-NN Algoritması

K-En Yakın Komşu algoritması benzetme yoluyla öğrenmeye, diğer bir deyişle verilen bir test örneğini buna benzer eğitim örnekleriyle karşılaştırarak elde etmeye dayanmaktadır. Eğitim örnekleri n değişkenle tanımlanır. Her örnek, boşluktaki n-boyutlu bir noktayı ifade eder. Bu şekilde, tüm eğitim örnekleri n-boyutlu bir model uzayında saklanır. K-NN algoritması, bilinmeyen bir örnek verildiysebilinmeyen örneğe en yakın k eğitim örnekleri için model arar**. (**Taşçı ve Onan,2016) "Yakınlık", Öklid mesafesi gibi bir mesafe metriği ile tanımlanır.

Komşular, doğru sınıflamanın yapıldığı bir dizi örnekten alınmıştır veya regresyon durumunda, etiketin değeri bilinmektedir. Bu, algoritma için ayarlanmış eğitim kümesi olarak düşünülebilir, ancak açık bir eğitim basamağı gerekmemektedir.

Temel bir k-En Yakın Komşu algoritması iki adımdan oluşur (Hofman ve Klinkenberg,2014):

• Görünmeyen örneğe en yakın k eğitim örnekleri bulunur.

• Bu k örnek için en sık rastlanan sınıflandırma alınır ya da regresyon durumunda bu k etiket değerlerinin ortalaması alınır ve bu değer bu eğitim veri seti örneğinin etiketi olarak atanır.

Bu iki basamak, görülmeyen tüm örnekler, yani eğitim veri setinin tüm örnekleri için gerçekleştirilir.

Tablo 3.3 K-NN Algoritmasının Genel İşleyişi

Algoritma 1: K-NN algoritmasının genel işleyişi [8]

Eğitim Algoritması

♣ Eğitim setinde yer alan her bir örneği (x, f (x)) eğitim örnekleri listesine ekle.

Sınıflandırma Algoritması

♣ Sınıflandırılmak üzere verilen x_q örneğini aşağıdaki kurala göre sınıfla:

- Eğitim örnekleri arasında yer alan $x_1, x_2, \ldots\ldots, x_k, x_q$ örneğine en yakın k tane örneği temsil etmek üzere, x_q örneğinin sınıf etiketinin belirlenmesi:

$$f(x_q) \leftarrow argmax_{v \epsilon V} \sum_{i=1}^{k} \delta(v, f(x_i)$$

- Burada, a ve b eşit olduğu takdirde δ (a,b) =1 olarak, aksi takdirde δ

3.1.3.6 K-NN Parametreleri

Örnekler arası yakınlığın nasıl ölçümleneceği, K-NN algoritmasının performansı için kritik bir öneme sahiptir. Öklid uzaklığı ya da bir başka uzaklık ölçütü kullanılarak yakınlık hesaplanabilir. Temel K-NN algoritmasında, sınıf etiketi çoğunluk oylamasına dayalı olarak belirlenir. Uzaklık ölçütü ve komşu sayısı (k) K-NN algoritmasının performansına etki eden en önemli parametrelerdir. Uzaklık Ölçütleri Uzaklık ölçütleri olarak, Minkowski, Öklid, Manhattan, Chebyschev ve Dilca uzaklığı kullanılmaktadır.

3.1.3.6.1 Minkowski Uzaklığı

Minkowski uzaklığı, Öklid, manhattan gibi uzaklık ölçütlerinin en genel halidir. Öklid ve manhattan uzaklık ölçütleri veri madenciliği uygulamaları ve makine öğrenmesi uygulamalarında çok sık kullanılır.

$P = (x_1, x_2, \ldots\ldots, x_n)$ ve $Q = (y_1, y_2, \ldots\ldots, y_n)$ iki nokta olmak üzere,

İki nokta arasındaki Minkowski uzaklığı olmak üzere, aşağıdaki eşitliğe göre hesaplanır:

Minkowski uzaklığı formülünde p=2 olduğu durumda bu uzaklık öklid uzaklığını ifade eder. p=1 olduğu durumda ise Manhattan uzaklığını ifade eder. $n \rightarrow \infty$ olursa bu uzaklık Chebyschev uzaklığını ifade eder (Kresse ve Danko,2012).

3.1.3.6.2 Öklid Uzaklığı

Kümeleme ve sınıflandırma algoritmalarında en çok öklid uzaklığı kullanılır. $P = (x_1, x_2, \ldots\ldots, x_n)$ ve $Q = (y_1, y_2, \ldots\ldots, y_n)$ iki nokta olmak üzere olmak üzere, aşağıdaki eşitliğe göre öklid uzaklığı hesaplanır (Kresse ve Danko,2012) :
Öklid uzaklığı, iki nokta arasındaki doğrusal uzaklıktır. K-NN algoritması, K-Öklid uzaklığı sınıflandırma ve kümeleme algoritmalarında yakınlığın ölçülmesi için kullanılan temel uzaklık ölçütüdür.

3.1.3.6.3 Manhattan Uzaklığı

İki nokta arasındaki farkların mutlak değerlerinin toplamı olarak ifade edilen uzaklığa Manhattan uzaklığı denir. Bu noktalar n boyutludur.
$P = (x_1, x_2, \ldots\ldots, x_n)$ ve $Q = (y_1, y_2, \ldots\ldots, y_n)$ n boyutlu iki nokta olmak üzere, Manhattan uzaklığı aşağıdaki eşitliğe göre hesaplanır (Kresse ve Danko, 2012):

3.1.3.6.4 Chebyschev Uzaklığı

Chebyschev uzaklığı (maksimum değer uzaklığı), daha önce de değinildiği gibi Minkowski uzaklığının, n→∞ olduğu özel bir halidir durumudur (Xu, Zong ve Zang, 2013).
Chebyschev uzaklığı $P = (x_1, x_2, \ldots\ldots, x_n)$ ve $Q = (y_1, y_2, \ldots\ldots, y_n)$ n boyutlu iki nokta olmak üzere, olmak üzere, aşağıdaki eşitliğe göre hesaplanır, ayrıca bu uzaklık iki nokta arasındaki farkların mutlak değerlerinin maksimumu olarak tanımlanmaktadır.

3.1.3.6.5 Dilca Uzaklığı

Dilca “Distance Learning in Categorical Attribute” kelimelerinden türetilmiştir. Bu uzaklık, kategorik değişkenlerin değerleri arasındaki uzaklığı ölçümlemek için kullanılır (Ienco, Pensa ve Meo, 2012).

3.1.3.7 Rapidminer’da K-NN Operatörü

K-en yakın komşuluk algoritması, RapidMiner'deki k-NN operatörü tarafından uygulanmaktadır. Bu operatör, Operatörler Penceresinde "Modelleme" "Sınıflandırma ve Regresyon" Tembel Modelleme "olarak adlandırılmıştır. Bu operatör, bir ExampleSet girdi olarak beklemekte ve verilen ExampleSet'den en yakın Komşu modelini üretmektedir. Bu model bir sınıflandırma veya regresyon olabilir. Model, verilen Örnek kümeye bağlı olarak değişir. Örnek küme etiketinin türü polinominal veya binominal ise, bu operatör bir sınıflandırma modeli oluşturur.
Etiketin türü sayısal ise, bu operatör bir regresyon modeli üretir. Bu operatörün bazı önemli parametreleri şunlardır (Hoffman ve Klinkenberg,2014):

• k: Bu parametre, görünmeyen örnek için aranacak en yakın komşu sayısını belirtir. Bu parametre, k-En Yakın Komşu Algoritması'ndaki k değişkenine eşdeğerdir.
Parametre k = 1 olarak ayarlanırsa, örnek basitçe kendisinin en yakın komşu sınıfına eşlenir.

•ağırlıklandırılmış oy (weighted vote) : Bu parametre oyların benzerlikle ağırlıklandırılıp ağırlıklandırılmayacağını belirtir.
Bu parametre true olarak ayarlanırsa örneklerin ağırlığı da dikkate alınır. Bu komşuların katkılarını ağırlıklandırmak için yararlı olabilir, böylece daha yakın komşular uzak olanlardan daha fazla katkıda bulunurlar.

• ölçü tipleri: Bu parametre, en yakın komşuları bulmak için kullanılacak ölçü türünü seçmek için kullanılır. Başka bir deyişle, bu parametre, örneklerin yakınlığını ölçmek için kullanılacak ölçü türünü belirler. Aşağıdaki seçenekler mevcuttur: karışık ölçekler (mixed measures), nominal ölçekler, nümerik ölçekler ve Bregman ıraksamaları (Bregmen divergences).

Tablo 3.4 Sınıflandırma Yöntemleri Karşılaştırması

Methods	Avantajları	Dezavantajları
K-NN	1. Uygulaması kolaydır. 2.Eğitim hızlıdır Model hızlı bir şekilde eğitilir.	1. Büyük bir depolama alanı gerektirir. 2. Gürültüye duyarlıdır. 3. Test yavaştır.
Decision Tree	1. Karar ağacının inşasında alan bilgisi gerekmez. 2. Karmaşık kararların belirsizliğini en aza indirir ve çeşitli eylemlerin sonuçlarına kesin değerler atar. 3. Yüksek boyutlu verileri kolayca işleyebilir. 4. Yorumlanması kolaydır. 5. Karar ağacı hem sayısal hem de kategorik verileri işleyebilir.	1. Bir çıktı değişkeni ile sınırlandırılmıştır. 2. Kategorik bir çıktı üretir. 3. Kararsız bir sınıflandırıcıdır, yani sınıflandırıcının performansı, veri kümesinin türüne bağlıdır. 4. Veri kümesinin türü nümerik ise karmaşık bir karar ağacı oluşturur.
Bayesian Belief Network	1. Hesaplama sürecini kolaylaştırır. 2. Büyük veri kümeleri için daha iyi hız ve doğruluğa sahiptir.	Değişkenler arasında bağımlılığın olduğu bazı durumlarda doğru sonuçlar vermemektedir.

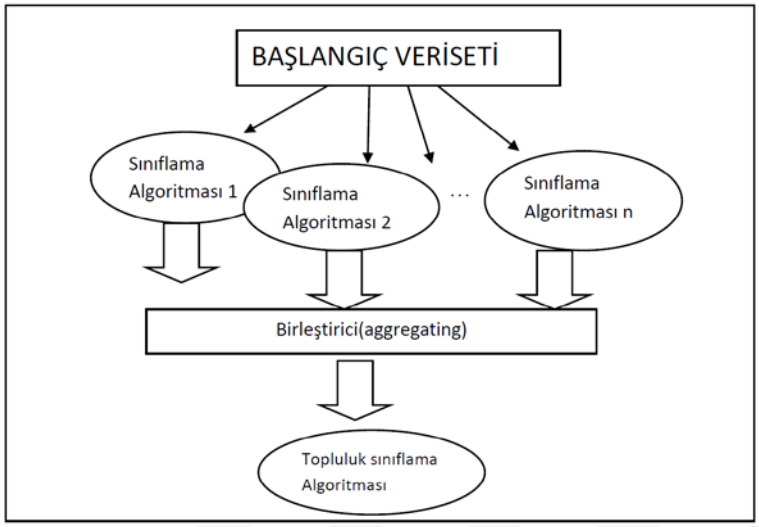
3.2 Topluluk Yöntemleri

3.2.1 Topluluk Yöntemlerine Genel Bakış

Veri madenciliği ve makine öğrenimi yöntemlerinden biri olan topluluk (ensemble) yöntemleri son 10 yılda çok büyük bir gelişme kaydetmiştir. Topluluk öğrenme yöntemlerinde (Ensemble Learning) birden çok sınıflayıcının ortaya koyduğu sonuçlar bir araya getirilerek, topluluk adına tek bir karar verilmektedir.

Başka bir deyişle, topluluk sınıflandırıcıları, yeni örnekleri sınıflandırmak için bireysel tahminleri değişik şekillerde birleştirilen (oylama, istifleme vs.) sınıflandırıcılardan oluşan bir gruptur (Dzeroski ve Zenko, 2004). Topluluk modelleri birden çok modeli, bileşenlerinin oluşturduğu modellerin en iyilerinden daha iyi bir biçimde birleştirir. Topluluk öğrenme yöntemleri, çok sayıdaki sınıflayıcının kararını birleştirdiği için daha güvenilir tahminler ortaya koymaktadır. Denetlenen öğrenmede en aktif araştırma alanlarından biri, sınıflandırıcıların iyi topluluklarının oluşturulması için yöntemler araştırmaktır (Dietterich, 1997). Bu konunun makine öğrenimi araştırmacılarına cazibeli gelen tarafı, toplulukların genellikle onları oluşturan sınıflandırıcılardan çok daha doğru sonuç vermesi fikrine dayanmaktadır. (Dzeroski

ve Zenko, 2004 içinde (Dietterich, 1997; Gams, Bohanec ve Cestnik, 1994)). Topluluk yöntemleri, yatırım zamanlamasından ilaç keşfine, sahtekarlık algılamadan öneri sistemlerine kadar endüstriyel zorluklara kritik destekler sağlayabilir. Kurulan modelin tahmini doğruluğu, modeli yorumlamaktan daha büyük önem taşır. Topluluk yöntemleri bütün modelleme algoritmaları ile oluşturulabilir. (Seni ve Elder, 2010).



Şekil 3.4 Topluluk Öğrenme Stratejisi (Akman,2010)

Topluluk öğrenme yöntemleri temel veya tekil öğrenme algoritmalarının ortaya koyduğu tahminlerin doğruluk oranını arttırmaktadır ve bu sebepten dolayı tekil öğrenme yöntemlerine göre daha başarılıdırlar.

Sınıflayıcı toplulukları üzerine yapılan araştırmaların çoğu, karar ağacı öğrenmesi veya sinir ağı eğitimi gibi tek bir öğrenme algoritması (Dietterich, 2000) kullanılarak topluluk üretmek ile ilgilidir (Dzeroski ve Zenko, 2004). Bu şekilde topluluk üreten yöntemlerden Bagging (Breiman 1996) ve Boosting en çok üzerinde çalışılan ve bilinen topluluk öğrenme algoritmalarıdır (Akman,2010). Oluşturulan sınıflandırıcılar, genellikle çoğunluk veya ağırlıklı oylama ile birleştirilir. Bir diğer yaklaşım ise, farklı öğrenme algoritmalarını (heterojen model sunumlarıyla) tek bir veri setine uygulayarak sınıflandırıcılar üretmektir. Modelleri birleştirmek için iki yaklaşım vardır; bunlar oylama (voting) ve istifleme (stacking)'dir.

Oylama yönteminde, birbirinden farklı çok sayıda sınıflayıcının yaptığı sınıf tahminleri oylamaya tabi tutulur ve oylama sonucunda en çok oyu alan sınıf, topluluğun sınıf tahmini olarak sunulur. En basit oylama şeması çoğul oylamadır. Bu oylama şemasına göre, her bir temel düzey sınıflandırıcı, kendi tahmini için bir oy kullanmaktadır. Örnek, en çok oy toplayan sınıfta sınıflandırılmıştır (Dzeroski ve Zenko,2004).

Yığınlanmış Genelleme (İstifleme)' de temel fikir; Eğitim verilerinin düzgün öğrenilip öğrenilmediğinin öğrenilmesidir. İki aşamalıdır; Birinci aşamada belli bir temel sınıflandırıcı özellik uzayının belirli bir bölgesini yanlış bir şekilde öğrendiyse, ikinci aşamadaki (meta) sınıflandırıcı bu istenmeyen durumu algılayabilir. Diğer sınıflandırıcıların öğrenmiş olduğu davranışlarla birlikte, meta sınıflandırıcı bu yanlış eğitimi düzeltebilir. İki aşamalı olması sebebiyle hata oranı düşer ve doğruluk oranı artar.

İstifleme ile oylama arasındaki fark;

- İstiflemenin aksine, sınıflandırıcıları bir oylama şemasıyla birleştirirken (çoklu, olasılıklı veya ağırlıklı oylama gibi), meta düzeyinde bir öğrenme gerçekleşmez. Oylama şeması, tüm farklı eğitim setleri ve öğrenme algoritmaları setleri (veya temel seviye sınıflandırıcılar) için aynı kalır.
- Oylama kullanılırken, en sık belirli bir örneğe atanan etiket doğru tahmin olarak seçilir.
- Stacking yönteminde (Wolpert, 1992) sıklıkla sınıflandırıcıların topluluğuna ek olarak birleştirme yöntemi öğrenmek için kullanılır. Daha sonra, oylama (voting) , öğrenilen birleştiricilerin karşılaştırıldığı sınıflandırıcıları birleştirmek için temel bir yöntem olarak kullanılır.
- Tipik olarak, oylama (voting) ile karşılaştırıldığında stacking ile daha iyi performans elde edilir ((Dzeroski ve Zenko, 2004)
- Verilerden modeller çıkarmak için çok çeşitli yöntemler mevcuttur ve her birinin kendisine göre güçlü yönleri vardır.

3.2.2 Regülarizasyon

İstatistiksel ve Makine Öğrenimi modelinin çıkarımı için yaygın olan en önemli prensip, doğruluk ve basitlik ilkesidir. Ancak, ikisi arasında bir zıtlık vardır: modelin karmaşıklığı arttıkça doğruluğu da artar ama basitlik ilkesinden uzaklaşmış olunur böylece modelin aşırı uyumsuz olma şansı (overfitting) artarak genelleme olasılığı düşer. Daha karmaşık bir modele daha yüksek doğruluğa sahiptir, ancak aşırı uyumsuz olma (overfitting) şansı artar ve esnek bir modeli genelleme olasılığı düşer. Regülarizasyon teknikleri, hata fonksiyonunu arttırarak model uyum prosedürünün esnekliğini azaltır. Artan hata kriterini minimize etmek için doğruluk payında belli bir artışa ihtiyaç vardır. Regülarizasyon bugün, modern topluluk algoritmalarının üstün performansının ana nedenlerinden biridir. Bu konuda iyi bir araştırma, Tibshirani'nin doğrusal modeller için Kement düzeltme tekniğini kullanmasıydı (Tibshirani, , 1996).

Kement, modeldeki katsayıların mutlak değerinin toplamını ceza fonksiyonu olarak kullanır ve Breotte tarafından Garotte (1996) olarak adlandırdığı katsayı sonrası işleme tekniğinde yaptığı köklü işi yapar. Bir diğer önemli gelişme Efron ve diğerleri tarafından LARS algoritması ile kaydedildi. (2004). LARS algoritması, Kement çözümünün verimli yinelemeli bir hesaplamasına olanak tanır. Kısa süre önce, Friedman, Lasso hatasını çeşitli kayıp (hata) işlevleriyle birleştiren (Friedman ve Popescu, 2004) Path Finder (PS) adında bir teknik yayımladı ve orijinal Kement belgesini En Küçük Kareler kaybıyla sınırlandırdı (Seni ve Elder,2010)

3.2.3 Topluluk Modelleri Oluşturmak

Bir topluluk oluşturmak iki adımdan oluşur: (1) çeşitli modeller kurmak ve (2) kurulan bu modellerin tahminlerini birleştirmek. Örneğin, değişken vaka ağırlıkları, veri değerleri, yol gösterici parametreler, değişken alt gruplar veya girdi alanının bölümleri vasıtasıyla bileşen modeller üretebilir (Seni ve Elder,2010).
Bileşen modeller “voting” (oylama) yöntemi ile gerçekleştirilebilir, diğer yöntemler bagging, boosting random forest, stacking olarak sayılabilir.

3.2.3.1 Bagging Yöntemi (Bootstrap Aggregating)

Daha önce sınıf katılımcılarını oylama ile birleştirmeyi tartıştık. Bagging (bootstrapped aggregating) oyu sağlayan sınıflandırıcıları oluşturmak için oylamayı bir yöntemle birleştiren bir topluluk sınıflama yöntemidir (Akman, 2010). Bagging metodu, her bir temel sınıflandırıcının çeşitlilik kazanmak amacıyla farklı rastgele alt kümeler tarafından eğitilmesine izin verir. İyi performans gösteren farklı (yani farklı hatalar yapan) temel sınıflandırıcılar üretmenin yollarını bulmak topluluk metodları araştırması içindeki en önemli alt konulardan biri olmuştur. Bagging, çeşitli temel sınıflayıcıların üretilmesi yönteminin basit bir örneğidir; daha da ayrıntıya girecek olursak, Bagging, çeşitli temel sınıflayıcıların üretilmesi yöntemin basit bir örneğidir; daha da ayrıntıya girecek olursak, bagging, orijinal eğitim setinden çoklu bootstrap eğitim setleri üretir ve her birini topluluğa dahil etmek üzere bir sınıflandırıcı oluşturmak için kullanır (Oza ve Tumer,2008).
Bagging sınıflama ve regresyon modelleri için uygulanmaktadır. Aşırı öğrenmeye karşı güçlü olan bu yöntem sınıflamada doğru sınıflama oranını arttıran ve varyansı düşüren bir yöntemdir. Veri setinde kayıp verilerin olduğu durumlarda da sınıflamada oldukça başarılıdır.

Bagging yöntemi, birçok sınıflama modeline uygulanabilmekle birlikte daha çok karar ağaçları için kullanılmaktadır. Bagging yöntemi veri setinden sınıf yapısını bozmayacak şekilde rastgele örnekler seçilerek (bootstrap) oluşturulan çok sayıdaki karar ağacının yaptığı sınıf tahminleri oylamaya tabi tutularak en çok oyu alan sınıfı nihai sınıf tahmini olarak belirleyen öğrenme yöntemidir. Bagging yönteminde art arda oluşturulan ağaçlar önceden oluşturulan ağaçlara bağımlı değildirler ve ağaçlar orijinal veri setinden bootstrap örnekleme yapılarak oluşturulmaktadır. Bagging ve bootstrap örneklemesi yapmak için algoritmalar (değiştirme ile örnekleme) Şekil 3.5'de gösterilmiştir (Oza ve Tumer,2008).

Bagging (T,M)
Her bir m= 1,2,…..M için,
T_m= Yerine Koyma ile Örnekleme (T, ITI)
$h_m = L_b (T_m)$
$h_{fin}(x) = argmax_{y \in Y} \sum_{m=1}^{M} I(h_m(x) = y)$ fonksiyonunu döndür

Yerine Koyma İle Örnekleme (T,N)
S=Ø
İ= 1,2,…..N e kadar,
r= rastgele tamsayı (1,N)
T[r] 'yi S'ye ekle
S'yi döndür

Şekil 3.5 Toplu Torbalama Algoritması ve Yerine Koyma ile Örnekleme
T: N örnek sayılı orjinal eğitim seti, M: öğrenilecek temel modellerin sayısı, Lb:Temel model öğrenme algoritması, hi'ler: Yeni örnekleri girdi olarak alan ve olası Y:sınıfları kümesi içinden tahmini sınıfları döndüren sınıflama fonksiyonları. Rastgele tamsayı (a; b), tamsayıların her birini a'dan b'ye eşit olasılıkla döndüren bir fonksiyon.
I (A) fonksiyonu, A olayı doğru ise 1, yanlış ise 0 döndüren bir işaret fonksiyonu.

N boyutlu bir eğitim kümesinden bir bootstrap eğitim seti oluşturmak için N tane çok terimli deneme gerçekleştiririz. Her denemede N örnekten biri oluşturulur. Her denemede N örnekten biri oluşturulur. Her denemede her örnek için oluşturulma olasılığı eşit ve 1/N 'dir. Şekil 3.2.2 deki ikinci algoritma da tam olarak bunu yapar; algoritma N kez, 1 den N'e kadar olan sayılar arasından rastgele bir r sayısı seçer ve r. eğitim örneğini bootstrap eğitim kümesi olan S'ye ekler. Açıkçası, orijinal eğitim örneklerinden bazıları bootstrap eğitim setine dahil edilmek üzere bir kez bile seçilmeyebilir ve diğerleri bir veya birden fazla kez seçilebilir. Ortalama olarak, oluşturulan her bootstrap eğitim kümesi gerçekte N eğitim örneği içerecek olsa bile 0.63N benzersiz eğitim örnekleri içerecektir. N gerçek eğitim örnekleri. Bagging

yönteminde bu gibi M tane bootstrap eğitim kümesi oluşturuyoruz ve her biri kullanılarak sınıflandırıcılar oluşturuyoruz.

Bagging h1, h2, ..., hM taban modellerinden en fazla oyu alan y sınıfını döndürerek yeni örnekleri sınıflandıran h (x) fonksiyonunu döndürür. Bagging yönteminde, yaratılan M bootstrap eğitim setlerinin bazılarının farklı olması muhtemeldir.

Bu farklılıklar M temel modelin performansı iyiyken bu modeller arasında fark edilebilir farklılıklar yaratmak için yeterliyse topluluk muhtemelen bireysel modellerden daha iyi performans gösterecektir. Genel olarak, bagging toplulukları, temel model öğrenme algoritmaları kararsız ise

temel modellerine daha iyi sonuç verirler. Bir modelin kararsız olmasından kasıt, farklı eğitim kümelerinin önemsenecek derecede farklı modeller oluşturmasıdır (Breiman,1996). Bunu anlatmanın başka bir yolu ise,bagging temel modellerde varyansı azaltmak için biasi azaltmaktan daha fazlasını yapar. Yani temel modeller yüksek varyansa ve düşük biase sahipken bagging temel modellerden daha iyi sonuç verir. Karar ağaçları kararsızdır, bu da bagging karar ağaçlarının neden bireysel karar ağaçlarından daha iyi performans gösterdiğini açıklar;

benzer şekilde, aynı sebepten bagging veya genel bootstrapping sinir ağı sınıflandırıcıları için iyi çalışır (Tumer ve Gosh,1996). Ancak, Naive Bayes sınıflandırıcıları veya k-en yakın komşu (kNN) sınıflandırıcıları stabildir, bu da bagging işleminin neden özellikle bu sınıflandırıcılarda etkili olmadığını açıklar (Oza ve Tumer,2008).

3.2.3.2 Boosting Yöntemi

Boosting, Kearns (1988) tarafından; “Çok sayıda zayıf öğrenicinin oluşturduğu sınıflayıcı grup bir araya gelerek güçlü bir öğrenici oluşturabilir mi?” sorusundan esinlenerek oluşturulmuş bir topluluk öğrenme yöntemidir. Zayıf öğrenici, doğru sınıflama ile çok az ilişkili bir sınıflayıcı iken güçlü öğrenici doğru sınıflama ile çok fazla ilişkili olan bir sınıflayıcıdır.

Boosting, belirli bir algoritmayla kısıtlı değildir, ancak çok sayıda boosting algoritması belirli bir dağılıma tabi olarak zayıf sınıflayıcıların sonuçlarını tekrarlı bir şekilde toplayıp en son güçlü sınıflayıcıyı oluşturmaktadır. Her yeni zayıf öğrenici bir sonuç ortaya koyduğunda veri yeniden ağırlıklandırılmaktadır. Burada toplama işlemi devam ederken, önceki oluşturulan zayıf öğrenicilerden yanlış sınıflama

yapanlara daha fazla odaklanılarak yeni zayıf öğreniciler oluşturulmakta ve sonuç olarak güçlü bir sınıflayıcı ortaya konulmaktadır.

Boosting yöntemi de çok sayıda oluşturulan karar ağaçlarının sonuçlarını ağırlıklı oylamaya tabi tutarak son sınıf tahminini yapmaktadır. Boosting yönteminde sonradan oluşturulan ağaçlar önceden oluşturulan ağaçlara bağımlıdırlar.

Daha önceki oluşturulan ağaçlardan yanlış tahminde bulunan tahmin edicilere daha fazla ağırlık verilerek ard arda yeni ağaçlar oluşturulmaktadır. Sonunda da final tahmin için ağırlıklı oylama yapılmaktadır. En çok bilinen boosting algoritması Yoav Freund ve Robert Schapire (1996) tarafından formüle edilen Adaboost"tur.

AdaBoost, Temel sınıflayıcılar tarafından seçilecek örneklerin sınıflandırılmasında bagging'den farklı olarak çalışmaktadır. Bagging algoritmasında bütün örneklerin ağırlıklandırması eşit olmaktadır. Buna karşın Adaboost'ta sınıflandırma işlemine müdahale etmeyip yanlış sınıflandırılan örneklerin ağırlıklarını arttırır. Böylelikle doğru olarak snıflandırılan verilerle tekrar işlemek yerine yanlış sınıflandırılan örneklere yoğunlaşıp sınıflandırma başarımını arttırmayı hedeflemektedir (Cingiz, Albayrak ve Amasyalı,2013).

Boosting Algoritmaları içinde en popüler olan algoritma AdaBoost algoritmasıdır. Adaboost algoritması, eğitim kümesi üzerinde farklı ağırlık dağılımları ile temel model dizilerini oluşturur. AdaBoost algoritması, Şekil 3.6' da gösterilmiştir (Oza ve Tumar,2008).

Adaboost ({ $(x_1,y_1),\ldots (x_n,y_n)$}, L_b,M)

Başlat D_1 (n) = 1/N her nє {1,2,…,N}.

m= 1,2,…..M için:

$h_m = L_b$ ({ $(x_1,y_1),\ldots, (x_N,y_N)$}, D_m).

H_m'nin hatasını hesaplayalım: $\epsilon_m = \sum_{n:h_m(x_n)\neq y_n} D_m(n).$

Eğer $\epsilon_m \geq ½$ ise,

M=m-1 yap ve bu döngüden çık.

D_m dağılımını güncelle:

$$D_{m+1}(n) = D_m(n) \times \begin{cases} \frac{1}{2(1-\varepsilon_m)} & eğer\ h_m(x_n) = y_n\ ise \\ \frac{1}{2\varepsilon_m} & diğer\ durumlarda \end{cases}$$

Çıktı: son model:

$$h_{fin}(x) = argmax_{y\in Y} \sum_{m:h_m(x)=y} log \frac{1-\varepsilon_m}{\varepsilon_m}$$

Şekil 3.6 AdaBoost Algoritması
f (x1; y1),…., (xN; yN). g eğitim örneklerinin kümesi,
Lb: temel model öğrenme algoritması
M oluşturulacak temel modellerin sayısı

Algoritmanın girdileri, bir N elemanlı eğitim örneği seti, Lb temel model öğrenme algoritması ve birleştirmek istediğimiz M tane temel model. AdaBoost aslında iki sınıflı sınıflandırma problemleri için tasarlanmıştır. Ancak, AdaBoost daha fazla sayıda sınıfla birlikte de kullanılabilir.

AdaBoost'un ilk adımı, eğitim seti üzerinde D1 ağırlıklarının başlangıç dağılımını oluşturmaktır. Bu dağılım, N eğitim örneklerinin tümüne eşit ağırlık verir. Şimdi döngüyü algoritmaya giriyoruz. İlk temel modeli oluşturmak için, eğitim seti üzerinde D1 dağılımı ile Lb 'yi çağırıyoruz.

h1 modeline geri döndükten sonra, eğitim kümesi üzerinde h1'in yanlış sınıflandırdığı eğitim örneklerinin ağırlıklarının toplamı anlamına gelen h1 modelinin $Є_1$ hatasını hesaplıyoruz. $Є_1$ <1 / 2 olmalı (bu zayıf öğrenme varsayımıdır; hata sınıfı rastgele tahmin ederek elde edeceğimizden daha küçük olmalıdır.

Bu koşul sağlanmasa, duruyoruz ve önceden oluşturulmuş temel modellerden oluşan topluluğa geri dönüyoruz.

Bu koşul sağlanırsa, aşağıdaki gibi eğitim örnekleri üzerinde yeni bir D2 dağılımı hesaplıyoruz; h1 modeli tarafından doğru şekilde sınıflandırılan örneklerin ağırlıklarının katsayıları 1/ 2 (1-$ϵ_1$) ile çarpılırken,h1 modeli ile yanlış sınıflandırılmış örneklerin ağırlıklarının katsayıları 1/2$ϵ_1$ ile çarpılıyor.

$Є_1$ <1 /2 koşulu nedeniyle doğru sınıflandırılmış örneklerin ağırlıklarının azaldığını ve yanlış sınıflandırılmış örneklerin ağırlıklarının arttığını unutmayın. Özellikle, h1 tarafından yanlış sınıflandırılmış örneklerin toplam ağırlığı D2 altında 1 /2'ye yükselirken, h1 tarafından doğru sınıflandırılmış örneklerin toplam ağırlığı D2 altında 1 /2'ye düşmüştür. Bu adımdan sonra eğitim kümesini ve yeni D2 dağılımını kullanarak h2 temel modelini oluşturmak için döngünün bir sonraki adımına geçiyoruz.

Bu şekilde M tane temel modeli oluşturuyoruz.

AdaBoost tarafından döndürülen topluluk modeli, girdi olarak yeni bir örnek alan ve her temel modelin ağırlığının log (1-$ϵ_m$/$ϵ_m$) olduğu M tane temel model üzerinde maksimum ağırlıklı oyu alan sınıfı döndüren bir fonksiyondur. Her bir temel modelin ağırlığı olan log (1-$ϵ_m$/$ϵ_m$), kendisine sunulan ağırlıklandırılmış eğitim kümesi üzerinde temel modelin doğruluğu ile orantılıdır. Açıkçası, AdaBoost algoritmasının kalbi dağılım güncelleme adımıdır.

Altında yatan fikir aşağıdaki gibidir. Algoritmadan, yanlış sınıflanmış örneklerin ağırlıklarının toplamının $ϵ_m$ olduğunu görebiliriz. Yanlış sınıflandırılmış örneklerin ağırlıkları 1/2$ϵ_m$ ile çarpıldı. Böylece ağırlıklarının toplamı $ϵ_m$ * 1/2$ϵ_m$ = ½ ' ye yükseltilir.

Doğru sınıflandırılmış örneklerin toplam ağırlığı 1- ϵ_m ile başlar, ancak bu ağırlıklar 1/2 (1- ϵ_m) ile çarpılır. Bu nedenle, ağırlıklarının toplamı1- ϵ_m * 1/2 (1- ϵ_m)= ½ 'ye düşer. Asıl nokta, bir sonraki temel modelin zayıf bir öğrenİci tarafından üretilmesidir (yani, temel model 1 /2'den küçük bir hataya sahip olacaktır); dolayısıyla,
önceki temel modelin yanlış sınıflandırdığı örneklerden en azından bazıları öğrenilecektir.
Boosting, sapmayı (bias) azaltmak için varyansı azaltmaya göre daha çok şey yapar. Bu sebepten dolayı, boosting en çok yüksek sapmaya ve düşük varyansa sahip olduklarında temel modellerini geliştirmeye eğilimlidir.
Bu modellerin örnekleri Nave Bayes sınıflandırıcıları ve karar ağaçlarıdır. Boostingin sapmayı azaltma yaklaşımı eğitim seti üzerindeki dağılımını ayarlayabilmesinden gelir.
Bir temel modelin yanlış sınıflandırdığı örneklerin ağırlıkları arttırılır ve bu da temel model öğrenme algoritmasının bu örneklere daha fazla odaklanmasını sağlar. Temel model öğrenme algoritması, belirli eğitim örneklerine karşı önyargılıysa, bu örnekler daha fazla ağırlık alır ve bu önyargıyı düzeltme olasılığını ortaya çıkarır. Ancak, bu eğitim seti dağılımının ayarlanması yöntemi eğitim verisi gürültülü olduğunda boostingin zor durumda kalmasına sebep olur (Dietterich,2000). Gürültülü örnekler normalde zor öğrenirler. Bu nedenle, gürültülü örneklere verilen ağırlıklar genellikle diğer örneklere verilen ağırlıklardan çok daha yüksek olur. Bu da boostingin gürültülü örneklere çok daha fazla odaklanmasına ve veriye fazla uyuma (overfitting) neden olur. Buna rağmen, boosting şu anda en iyi denetlenen sınıflayıcı öğrenme algoritmalarından biridir (Oza ve Tumer,2008).

3.2.3.3 Rastgele Orman (Random Forest)

Random Forests, yukarıda bahsedildiği gibi topluluk öğrenme yöntemidir. Bireysel olarak oluşturulan karar ağaçları bir araya gelerek karar ormanını oluşturmaktadır. Karar ormanındaki her karar ağacı, orijinal veri setinden bootstrap tekniği ile farklı örneklemler seçilerek oluşturulmaktadır. Orman, ağaçların yapmış olduğu sınıf tahminleri bir araya getirerek, nihai sınıf tahminini yapmaktadır.
Random Forests yöntemi, Breiman (1996) tarafından önerilen Bagging tekniği ile Ho (1998) tarafından önerilen Random Subspace tekniğini birleştirmiştir. Bagging yönteminde karar ağaçları veri setinden bootstrap tekniği ile örneklem seçilerek birbirinden bağımsız olarak oluşturulmaktadır. Random Subspace yöntemi ise, karar ağacının her düğümünde en iyi dallara ayırıcı değişkenin, tüm değişkenler

arasından rastgele (random) seçilen az sayıdaki değişken içinden seçilmesidir (Akman, 2010).

Random Forests yönteminde sonradan gelen veriye ait tahmin yapılmasının yanında, değişkenlerin önem derecesi de hesaplanmaktadır. Veri setinde çok sayıda değişken varsa, değişken önem derecesinin hesaplanması model indirgemesi açısından oldukça kullanışlıdır. Örneğin binlerce değişkenin bulunduğu veri setinde, Random Forests yöntemiyle elde edilen önem derecesine göre, kurulacak yeni modelde önem derecesi yüksek değişkenler kullanılarak daha etkin tahminlerin yapılması sağlanabilir (Akman,2010).

Random Forests yönteminde, model kurulurken, modeli test etmek için ayrı bir test veri seti varsa veya orijinal veri setinden test veri seti ayrılmışsa, modelin kurulması için ayrılan öğrenme veri seti, kendi içinde 2/3 oranında öğrenme veri seti, 1/3"ü ise test veri seti olarak ayrılmaktadır.

3.2.3.4 Stacking (İstifleme)

Stacking bir meta öğrenme yöntemi olarak düşünülmelidir. Kelimenin tam anlamıyla, meta öğrenme öğrenmeden öğrenmek demektir. Uygulamada, meta-öğrenme, ilk aşamada öğrenme yoluyla üretilen sonuçları girdi olarak alır ve bunlarla ilgili genellemeler yapar. Meta öğrenme konularından olan uygun bir öğreniciyi seçmeyi öğrenme, uygun bir önyargıyı seçmeyi öğrenme ve temel düzey sınıflandırıcıların tahminlerini birleştirmeyi öğrenme makine öğrenimi topluluğunda değerlendirilmiştir. Aşağıda her biri ayrıntılı biçimde açıklanmıştır (Dzeroski ve Zenko,2004).

- En uygun öğrencinin nasıl seçileceğinin öğrenilmesi. Bir öğrencinin belirli bir alan için uygunluğu, bazı kriterlere göre değerlendirilir. Bu ölçüt genellikle tahmini doğruluktur (predictive accuracy). Genel fikir, her alan adını, öğrenme algoritmalarının performansıyla alakalı bir dizi meta-özellik ile tanımlanmasıdır. Birden fazla alanın bu meta özellikler açısından tanımlanması, bu alanlardaki algoritma performanslarıyla birlikte, bir meta öğrenenin uygulandığı bir meta alanı oluşturmaktadır. Ortaya çıkan meta sınıflandırıcı, yeni bir alan adı için en uygun öğrenciyi önerebilmelidir.
- Öğrenme algoritması için önyargıyı dinamik olarak nasıl seçeceğinizi öğrenme. Burada amaç, eldeki alanın daha iyi kapsanması için hipotez alanını değiştirebilecek bir öğrenme algoritması oluşturmaktır.

- Temel düzey sınıflandırıcıların tahminlerini nasıl birleştireceğinizi öğrenmek. Taban seviyesinde sınıflandırıcıların (veya bazı özelliklerinin) doğru sınıf değerleri ile birlikte tahminleri bir meta düzey veri kümesi oluşturmaktadır.

3.2.3.5 Stacking Çerçevesi

Stacking, tek bir S veri seti üzerinde farklı öğrenme algoritmaları L1, ..., LN kullanılarak üretilen birden fazla sınıflandırıcıyı birleştirmekle ilgilidir. S veri kümesi si = (xi, yi) örneklerinden oluşur. Burada (xi)'ler ϵ R^n olan özellik vektörlerini, (yi)'ler de bu vektörlerin sınıflandırmalarını temsil eder. Birinci aşamada, Ci = Li (S) şeklinde olan bir dizi taban düzeyde sınıflandırıcı C1, C2, ..., CN üretilir. İkinci aşamada, taban sınıflandırıcılarının çıktılarını birleştiren bir meta düzey sınıflandırıcı öğrenilecektir. Meta düzey sınıflandırıcıyı öğrenme amaçlı bir eğitim seti oluşturmak için birini bırakma (leave-one- out) veya çapraz doğrulama (cross validation) prosedürü uygulanır .Birini bırakma (Leave -one-out) yönteminde test için bir örnek bırakarak kalan veri kümesine taban öğrenme algoritmalarının her biri uygulanır: ∀i = 1, ..., n: ∀k = 1, ..., N : C_k^i = L_k (S - s_i). Daha sonra, s_i leri tahmin etmek için öğrenilen sınıflandırıcılar kullanılır: $s_i = \hat{y}_i^k$ = C_k^i (x_i). Meta düzey veri kümesi ($\hat{y}_i^1, \ldots\ldots, \hat{y}_i^n$), y_i) örneklerinden oluşur. Burada özellikler taban düzeyinde sınıflandırıcıların tahminleridir ve sınıf eldeki örneklerin doğru sınıflarıdır. Örneğin 10 kat çapraz doğrulama gerçekleştirirken, her seferinde bir örneği çıkarmak yerine orijinal veri kümesinin onda biri çıkarılır ve test kümesi olarak o kullanılır. Yani öğrenilen taban düzeyde sınıflandırıcıların tahminleri bunlar üzerinden elde edilir (Dzeroski ve Zenko,2004).

Stacking Algoritması:

Girdi: Veri Seti: D= $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m); x_i \in R^n, y_i \epsilon Y\}$

İlk seviye öğrenme algoritmaları $\mathcal{L}_1, \ldots, \mathcal{L}_T$

İkinci seviye öğrenme algoritmaları $\mathcal{L}$

İşlem:

for t=1,.....T:

$h_t = \mathcal{L}_T(D)$ %orjinal veri seti D ye ilk seviye öğrenme algoritması $\mathcal{L}_T$ nin

%uygulanması ile ilk seviye öğrenci h_t nin eğitilmesi.

end;

D' = BOŞ KÜME %Yeni bir veri seti oluşturmak

For i= 1,.....m;

For t=1,.....,T:

$Z_{it} = h_t(x_i)$

End;

D'= D' U $\{((z_{i1}, z_{i2}, \ldots, z_{iT}), y_i)\}$

End;

h' = $\mathcal{L}$(D') %Yeni veri seti D' ye ikinci seviye öğrenme algoritması $\mathcal{L}$ nin

% uygulanması ile ikinci seviye h' nin eğitilmesi.

Çıktı: H(x) = h' $(h_1(x), \ldots, h_T(x)$

Şekil 3.7 Stacking Algoritması

Yukarıdaki algoritma aşağıdaki şekilde işler;

1. Eğitim kümesi ayrık iki kümeye bölünür.
2. İlk bölümde birçok taban öğrenici eğitilir.
3. İkinci bölümdeki taban öğrenciler test edilir.
4. 3'deki tahminler girdi, doğru cevaplar da çıktı olarak kullanılarak daha üst seviyedeki öğrenici eğitilir.

Dikkat edilirse 1'den 3'e kadar olan adımlar çapraz doğrulama ile aynıdır, ancak en iyi öğrenicinin seçilmesi yaklaşımının kullanılması yerine stackingde temel öğreniciler çoğu zaman doğrusal olmayan şekilde birleştirilir.

Aşağıdaki tablo 3.5 ve tablo 3.6'de topluluk metodlarının eğitim yaklaşımları, sınıflayıcıları, karar verme yöntemleri, sınıflayıcı önceliği, girdi parametresi, artıları ve eksileri karşılaştırılmıştır.

Tablo 3.5 Topluluk Metodlarının Karşılaştırılşması -1

METOD	EĞİTİM YAKLAŞIMI	SINIFLAYICILAR	KARAR VERME YÖNTEMİ
Bagging	Yeniden Örnekleme	Yeniden örneklenmiş setler üzerinde değişken öğrenciler çıktı olarak farklı modeller verirler	Çoğunluk oyu
Bossting	Yeniden Örnekleme	Zayıf öğrenici her yenilemede yeniden ağırlıklandırılır	Ağırlıklandırılmış Çoğunluk Oyu
Stacked Generalization	Yeniden Örnekleme ve katlama	Çeşitli taban sınıflandırıcıları	Meta sınıflandırıcı

Tablo 3.6 Topluluk Metodlarının Karşılaştırılması -2

METOD	SINIFLAYICI ÖNCELİĞİ	GİRDİ PARAMETRESİ	ARTILARI	EKSİLERİ
Bagging	YOK	Eğitim verileri, sınıf sayısı yineleme sayısı	Basit, anlaşılması ve uygulaması kolay	Doğruluk değeri diğer topluluk yaklaşımların dan düşüktür
Bossting	YOK	Eğitim verileri, sınıf sayısı yineleme sayısı	Zayıf öğrenicilerin performansı nı arttırır	Gürültüden Etkilenir
Stacked Generalization	YOK	Eğitim verileri, sınıf sayısı yineleme sayısı	Performansı diğerlerine kıyasla daha iyidir	Depolama alanı büyük ve analiz süresi uzun olabilir

DÖRDÜNCÜ BÖLÜM
UYGULAMA

Bu çalışmanın amacı prostat kanserinin erken ve doğru teşhis edilebilmesi için veri madenciliği yöntemleri kullanılarak anlamlı bir model oluşturmaktır. Çalışmada, Karar Ağaçları, Bayes Sınıflandırıcılar ve K-NN yöntemleri kullanılarak hibrit bir topluluk model (ensemble modeli) kurulmuştur. Kurulan bu model, hem genotip hem de fenotip değişkenler içeren çok etnikli prostat kanseri verisi üzerinde denenerek hasta ve hasta olmayan bireyleri doğru ve anlamlı bir şekilde sınıflandırmayı amaçlamaktadır. Böylece prostat kanserinin doğru ve erken bir şekilde teşhis edilmesine katkı sağlanması amaçlanmaktadır.

Bu çalışmada kullanılan veri fenotip değişkenler yanında SNP adı verilen genotip değişkenleri de içerir. SNP'lerin çeşitli kanser türleri (Easton ve Eeles, 2008), diyabet (Reddy v.d, 2011), kardiyovasküler hastalıklar (Lettre v.d, 2011) ve akıl hastalıkları (Ryten v.d, 2009) gibi karmaşık hastalıklarla ilişkisi araştırmalara konu olmaktadır (Hardy ve Singleton, 2009). SNP'lerin çeşitli hastalıklarla bağlantısının araştırıldığı çalışmalarda kullanılan veriler oldukça büyük miktardadır ayrıca çok da değişken içerir. Bu yüzden bu tarz verilerin incelenmesinde, istatistiksel yöntemler yetersiz kalır ve veri madenciliği yöntemleri tercih edilmektedir (Yücebaş,2016).

Modelde kullanılan her bir yöntem literatür titizlikle incelenerek seçilmiştir. Bilakis genom ilişkilendirme çalışmaları (GWAS) ve SNP'lerin karmaşık hastalıklarla ilişkilendirildiği çalışmalara bakıldığında karar ağaçları, K-nn ve naive bayes ilk sıralarda yer almaktadır. SNP'ler ile hastalıklar arasındaki bağlantıların araştırıldığı çalışmalarda Karar Ağaçları (Anunciacao v.d,2010) ve Bayes Ağları (Jiang v.d,2010) sıklıkla kullanılır. Ayrıca en iyi 10 veri madenciliği algoritmasına bakıldığında Karar Ağaçları, K En Yakın Komşuluk ve Bayes Sınıflandırıcılar en iyi yöntemler arasındadır (Wu v.d,2008).

İlgili yöntemlerin uygulaması için Java tabanlı bir veri madenciliği aracı olan ve literatürdeki uygulamalarda da tercih edilen Rapid Miner'ın 7.3 sürümü kullanılmıştır.

4.1 Verinin Yapısı

Çalışmada kullanılan veri kümesi NCBI'ın dbGaP veri tabanındaki phs000306.v4 numaralı çok etnikli prostat kanseri veri kümesinden oluşturulmuştur. Oluşturulan bu kümede 458 hasta, 531 sağlıklı birey yer almakta ve her bireye ait 200.000 SNP ve

18 adet fenotip bulunmaktadır. (status, ind, severity) Veri kümesindeki bireylerin 135'i Japon ve kalan 854'ü Latin etnik kökenindendir.

4.2 Problemin Tanımlanması

Veri madenciliği sürecinin en kritik basamağı problemin tanımlanması basamağıdır. Bir çalışmada problem net ve doğru bir şekilde tanımlanmadıysa o probleme çözüm de açık bir şekilde bulunamaz (Şimşek,2006). Bu çalışmada çeşitli genotip ve fenotip değişkenler kullanılarak kurulan en uygun topluluk modeli ile bir erkeğin prostat kanseri olup olmadığını doğru bir şekilde sınıflandırmak amaçlanmaktadır.

4.3 Verinin Hazırlanması

4.3.1 Verinin Toplanması

Veri Amerika Birleşik Devletleri Ulusal Biyoteknoloji Bilgi Merkezi'nden (NCBI) istenmiştir. Ulusal Biyoteknoloji Bilgi Merkezi (NCBI), Ulusal Sağlık Enstitüsü'nün (NIH) bir kolu olan Amerika Birleşik Devletleri Ulusal Tıp Kütüphanesi'nin (NLM) bir parçasıdır. ABD Sağlık ve İnsani Hizmetler Departmanının bir parçası olan Ulusal Sağlık Enstitüsü (NIH), ülkenin sağlık araştırma ajansı olup, sağlığı iyileştiren ve hayatları kurtarıcı önemli keşifler yapmaktadır.

NIH'ın bir kolu olan Ulusal Tıp Kütüphanesi (NLM),Ulusal Sağlık Enstitüsünün Bethesda, Maryland'daki kampüsünde bulunur .1836'da kurulduğu günden bu yana bir bilgi inovasyonu merkezi haline gelmiştir. Dünyanın en büyük biyomedikal kütüphanesi olan NLM, geniş bir baskı kolleksiyonuna sahiptir. Ayrıca, dünya çapında milyonlarca insan tarafından her yıl milyarlarca kez araştırılan çok çeşitli konularda elektronik bilgi kaynakları toplar ve üretir. Bunun yanında, biyomedikal bilişim ve sağlık bilişim teknolojisinde araştırma, geliştirme ve eğitimi destekler ve yürütür. Buna ek olarak, kütüphane, Amerika Birleşik Devletleri'ndeki topluluklardaki sağlık bilgilerine erişim sağlayan 6 bin 500 üyeli Ulusal Tıp Kütüphaneleri Ağını koordine eder. NCBI, Bethesda, Maryland'de yer almakta ve 1988 yılında Senatör Claude Pepper'ın destek olduğu yasalarla kurulmuştur.Ulusal Biyoteknoloji Merkezi, biyomedikal ve genomik bilgilere erişim sağlayarak bilim ve sağlığını geliştirir.

4.3.2 Verinin Birleştirilmesi ve Temizlenmesi

Çok boyutlu ve büyük miktarda veri içeren çalışmalarda özellikle SNP gibi çok değişkenli verilerde analizden anlamlı sonuçlar çıkarabilmek için boyut indirgeme yöntemleri kullanılır **(**Zhou ve Wang,2007).

İlgili çalışmada kullanılan veri kümesindeki SNP sayısının indirgenerek üç adımlı bir indirgeme yapılmıştır. Bu boyut indirgemeler yapılırken ilk adımda PLINK analizi ikinci adımda SPOT analizi kullanılmıştır. Daha sonra ise Rapidminerda değişken indirgeme yöntemi olan bilgi kazanımı oranı yoluyla ağırlıklandırma (weight by information gain ratio) operatörü kullanılmıştır.

İnsan Genetik Araştırma Merkezi'nde Shaun Purcell tarafından geliştirilen PLINK, GWAS için geniş bir fonksiyon seti sunan açık kaynak kodlu bir bütün genom ilişkilendirme analiz aracıdır (2007). Plink veri yönetimi, kalite kontrolü için özet istatistikler, populasyon tabakalaşma tespiti, temel ilişkilendirme testi, sayı değişkeni analizi, meta-analiz, sonuç açıklama ve raporlama gibi analizleri içerir. PLINK, genetik hastalıkların çeşitli özelliklerini inceleyen bilim dalının (epidemiyolojinin) İsviçre-ordu-bıçağı olarak kabul edilebilir, çünkü onunla yapılamayacak kadar az şey vardır. Ancak ne yazık ki, PLINK komut satırından çalıştırılması gereken bir program olarak sunulmaktadır. Bu da programı araştırmacı için kullanıcı dostu yapmaz. İşlevsellik, kullanıcı bilgisayar meraklısı değilse, en iyisi olmak için yavaş olan bir dizi operasyonda sunulur. Aslında, komut dizisi içinde kaybolmak oldukça kolaydır ve bazen gerekli işlevi yerine getirmek için doğru komutu bulmak zaman alır. SNP'lerle prostat kanseri arasındaki ilişkinin istatistiksel gücünü bulabilmek için ilk önce PLINK analizi yapıldı. Daha sonra SNP'leri istatistiksel ve biyolojik anlamlılıklarına göre önceliklendirmek için SPOT analizi yapıldı. SPOT analizi sonucunda 22.000 olan SNP sayısı 5000'e indirgendi.

Veri önişlemenin ilk adımında hastalıkla ilişkisi belirli bir kuvvet değerinin üzerinde olan SNP'lerin seçimi için genom ilişkilendirme çalışmalarında kullanılan bir açık kaynak kod aracı olan PLINK tercih edilmiş ve p değerleri bulunmuştur. Daha sonra bu p değerleri SPOT analizine sokulup istatistiksel anlamlılık yanında SNP'lerin fonksiyonel özelliklerini de kullanan bir sıralama yapıldı. Bu iki adımda veri kümesindeki SNP sayısı 5000'e indirgenmiştir. 3. adımda bilgi kazanım oranı yoluyla ağırlıklandırma operatörü kullanılarak SNP sayısı 400'e indirgenmiştir. Bilgi kazanım oranı yoluyla ağırlıklandırma operatörü bilgi kazancını kullanarak sınıf özniteliğine göre niteliklerin ağırlığını hesaplar. Bu operatöre göre, bir özelliğin ağırlığı ne kadar yüksekse o özellik o kadar önemlidir ve o özellik veri setinde kalmalıdır. Bilgi

kazancı, genellikle, bir niteliğin alaka düzeyinin belirlenmesinde iyi bir ölçüttür. Fakat çok sayıda farklı değer alabilen niteliklere bilgi kazancı uygulamadan önce karar vermek için bu değişkenlerin önemli olup olmadığını iyi düşünmek gerekir. Bu operatör sadece nominal değişkenli veri setlerine uygulanabilir. Böylece tüm model 418 değişken (400 SNP ve 18 fenotip değişken)'e indirgenmiştir. 1037 kişiden alınan veriler, 458 hasta, 531 kontrol grubu ve 48 kayıp değerden oluşmaktadır. 48 kayıp değer analizden çıkarılarak 989 kişi ile devam edilmiştir.

4.3.3 Veriyi Dönüştürme

Verideki tüm değişkenler kullanılacak yöntemler nominal değişkenleri kabul ettiği için nominal değişkenlere dönüştürülmüştür.

4.4 Veri Setindeki Değişkenler

Aşağıdaki tabloda veri setindeki değişkenler ayrıntılı bir şekilde açıklanmaktadır.

Tablo 4.1 Veri Setindeki Değişkenler

Değişken Adı	Açıklama	Değişken Tipi	Değer Kodlama
1-)Ethnicity	Etnik köken	string	J= Japon L= Latin
2-)Study ID	çalışmaya katılan her bir kişi için numara	kodlanmış tamsayı	
3-)Status	durum	kodlanmış tamsayı	Prostat kanseri tanısı alanlar (case)= 1 Prostat kanseri olmayanlar (control) = 2
4-)Age_cat	Gruptakilerin 5 yıllık aralıklarla ifade edilmiş yaşı. Hasta olan ve olmayanları eşleştirmek için kullanılmıştır.	kodlanmış tamsayı	1: 45-49 yaş aralığındakiler 2: 50-54 yaş aralığındakiler 3: 55-59 yaş aralığındakiler 4:60-64 yaş aralığındakiler 5: 65-69 yaş aralığındakiler 6:70-74 yaş aralığındakiler 7: 75 + (75 yaş ve üstü)
5-) Agedx_cat	Kanser hastaları için kanserin teşhis edildiği yaşı ifade eder.	kodlanmış tamsayı	1: 45-49 yaş aralığındakiler 2: 50-54 yaş aralığındakiler 3: 55-59 yaş aralığındakiler 4:60-64 yaş aralığındakiler 5: 65-69 yaş aralığındakiler 6:70-74 yaş aralığındakiler 7: 75 + (75 yaş ve üstü)
6-) Age_co cat	Kanser olmayan kontrol grubu için kanın alındığı yaş	kodlanmış tamsayı	1: 45-49 yaş aralığındakiler 2: 50-54 yaş aralığındakiler 3: 55-59 yaş aralığındakiler 4:60-64 yaş aralığındakiler 5: 65-69 yaş aralığındakiler 6:70-74 yaş aralığındakiler 7: 75 + (75 yaş ve üstü)
7-)BMI_index	BMI (Body Mass Index) Vücut Kitle İndexi	kodlanmış tamsayı	1: BMI <= 22,5 2: 22,5<= BMI <= 24,9 3: 24,9<= BMI<= 29,9 4: BMI >= 30
8-) currsmoke	Şu anda sigara içme durumu	kodlanmış tamsayı	0: hayır 1:evet . : kayıp değer
9-) eversmoke	Hayatının bir döneminde sigara içme durumu	kodlanmış tamsayı	1: hayır 2:evet . : kayıp değer

10-) Severity (hastalığın ilerleme durumu)	Prostat kanseri olanlar için hastalığın ilerleme durumunu gösterir. Aşağıdaki evre tablosu ve Gleason Derece tablosuna göre hesaplanmıştır. Evre Tablosu 1: lokalize 2: bölgesel 3: lenf düğümleri ile beraber bölgesel 4: hem 2 hem 3 beraber 5: bölgesel Nos ? 7: Uzak Metastaz 8: Soyutlanmış değil 9: Evresi belirlenememiş ya da bilinmiyor **SEER Ayrıştırması** 1: Gleason Derecesi= 2,3,4 2: Gleason Derecesi = 5,6,7 3: Gleason Derecesi: 8,9,10 4: Ayrıştırılamamış 9: Bilinmiyor SEVERITY KODLAMASI 1: 1. Evre ve Ayrıştırma 1 / 2 2: 2-7. Evre ve Ayrıştırma 1-9 VEYA 1. Evre ve Ayrıştırma 3 / 4 veya 9. Evre ve Ayrıştırma 3 / 4	kodlanmış tamsayı	1 : lokalize 2: gelişmiş 3: kayıp değer
11-)d_lyco cat	Likopen alım miktarı (bir günde 1000 kilocalori başına alınan mikrogram miktarını ifade ifade eder.)	Kodlanmış tamsayı	1: <= 752 mikro gram 2: 752- 1077 mikro gram 3: 1078-1436 mikro gram 4: 1437- 2022 mikro gram 5: >= 2023 mikro gram . : kayıp değer Not: yukarıdaki kategoriler cinsiyete özgü belirlenmiş kesim noktası değerlerine dayanmaktadır
12-)P_fat_cat	Vücuda alınan kalorinin yağ yüzdesi (%)	Kodlanmış tamsayı	1: <= 24,1 2: 24,2-28,6 3: 28,7-32,3 4: 32,3-36,2 5: >= 36,3 . : kayıp değer Not: yukarıdaki kategoriler cinsiyete özgü belirlenmiş kesim noktası değerlerine dayanmaktadır.
13-)d_-calc_cat	Kalsiyum alım miktarı (bir günde 1000 kilocalori başına alınan miligram miktarını ifade ifade eder.)	Kodlanmış tamsayı	1: <= 245 mg 2: 246-304 mg 3: 305-362 mg 4: 363-438 mg 5: >= 439 mg . : kayıp değer Not: yukarıdaki kategoriler cinsiyete özgü belirlenmiş kesim noktası değerlerine dayanmaktadır.
14-)fh_pca	Ailede (babada veya erkek kardeşte) prostat kanseri öyküsü	Kodlanmış tamsayı	0: yok 1: var 9: bilinmiyor
15-) Pa_cat	Günde ortalama kaç saat fiziksel aktivitede bulunulduğu	Kodlanmış tamsayı	1: <= 0,32 2: 0,33-0,70 3: 0,71-1,06 4: 1,07-2,03 5: >= 2,04 . : kayıp değer Not: yukarıdaki kategoriler cinsiyete özgü belirlenmiş kesim noktası değerlerine dayanmaktadır.
16-) Ethanol_ca	Günlük alkol tüketimi (12gr 1 içki olacak şekilde hesaplanmıştır)	Kodlanmış tamsayı	1: yok 2: <= 1 3: 1-2 4: 2+ .: kayıp değer

Çalışmaya katılmış 989 erkeğin 458'i prostat kanseri tanısına sahipken, 531'i kanser değildir.

4.5 Modelin Kurulması

Bu tezin amacı veri madenciliği yöntemleri kullanılarak hasta olup olmayanları doğru olarak sınıflandıran anlamlı bir model oluşturmaktır. Çalışmada, prostat kanserinin erken ve doğru teşhis edilebilmesine yönelik, karar ağaçları, bayes sınıflandırıcılar ve K-NN yöntemleri stacking metodu ile birleştirilerek bir topluluk modeli (ensemble modeli) kurulmuştur. Kurulan bu model, hem genotip hem de fenotip değişkenler içeren çok etnikli prostat kanseri verisi üzerinde denenerek hasta ve hasta olmayan bireyleri doğru ve anlamlı bir şekilde sınıflandırmayı amaçlamaktadır. Böylece prostat kanserinin doğru ve erken bir şekilde teşhis edilmesine katkı sağlanması amaçlanmaktadır.

Analiz RapidMiner 7.3 ile yapılmıştır.

Tablo 4.2' de modelde kullanılacak olan veri gösterilmektedir. Görüldüğü üzere veri, 989 erkeğin 414 özelliğini gösteren bir veri setidir. Bu özelliklerden 14'ü fenotip özelliklerdir.

Tablo 4.2 Modelde Kullanılan Veri

ExampleSet (989 examples, 2 special attributes, 412 regular attributes) Filter (989 / 989 examples): all

	agedx_cat	ageco_cat	bmi_cat	fh_prca	pa_cat	packyrs_ca	ethanol_ca	d_lyco_cat	p_fat_cat	d_calc_cat	currsmoke	eversmoke	severity
2	2	0	1	1	1	2	5	5	0	0	1	0
6	3	0	5	1	2	5	4	5	0	0	1	0
7	1	0	4	4	2	1	1	5	0	1	2	1
6	4	0	2	4	1	2	4	3	1	1	?	1
7	4	0	2	4	2	3	3	3	0	1	1	0
4	3	0	4	2	2	4	2	5	0	1	1	0
6	4	0	2	5	?	?	?	?	0	1	1	0
4	3	0	5	2	4	4	2	5	0	1	1	0
7	3	0	2	3	4	1	1	3	1	1	1	0
6	4	0	5	6	1	5	5	4	1	1	1	0
3	3	0	1	2	1	3	5	2	0	1	2	1
3	3	0	4	4	3	4	3	3	0	1	1	1
4	3	0	1	4	1	3	3	4	0	1	1	1
7	3	0	1	?	1	3	5	5	0	1	1	1
4	3	0	5	1	4	1	2	3	0	0	1	0
4	3	0	3	2	2	4	3	4	0	1	1	0
7	4	9	1	6	2	4	2	3	0	1	1	0
6	2	0	2	2	1	5	3	3	0	1	1	0
5	2	9	4	6	1	3	4	2	1	1	1	0

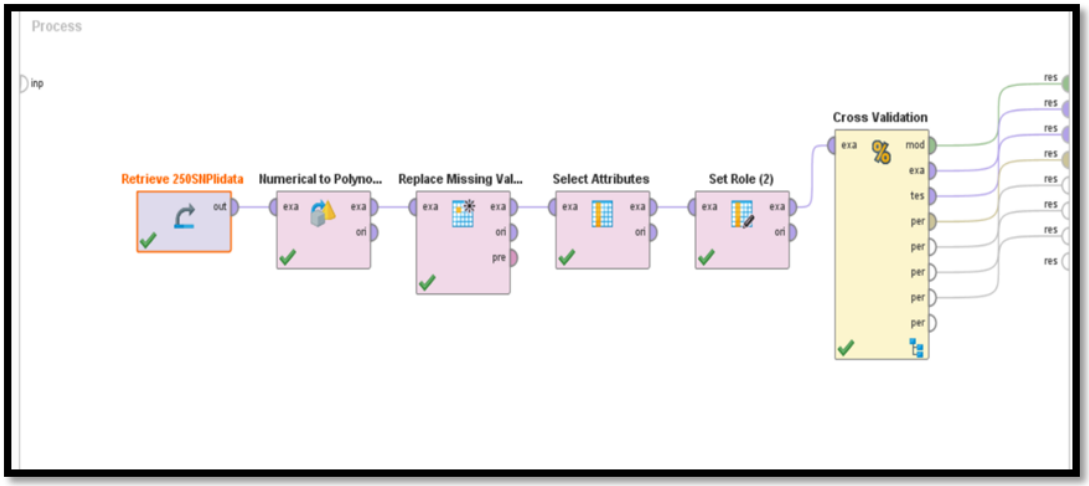
4.5.1 Kurulan Topluluk Modeli

Topluluk sınıflandırıcıları, yeni örnekleri sınıflandırmak için bireysel tahminleri değişik şekillerde birleştirilen sınıflandırıcılardan oluşan bir gruptur. Denetlenen öğrenmede en aktif araştırma alanlarından biri, sınıflandırıcıların iyi topluluklarının oluşturulması için yöntemler araştırmaktır (Dietterich, 1997). Bunun sebebi, toplulukların genellikle onları oluşturan sınıflandırıcılardan çok daha doğru sonuç

vermesidir (Dietterich, 1997; Gams, Bohanec ve Cestnik, 1994). Sınıflayıcı toplulukları üzerine yapılan araştırmaların çoğu, karar ağacı veya sinir ağı gibi tek bir algoritma (Dietterich, 2000) kullanarak topluluk üretmek ile ilgilidir. Oluşturulan sınıflandırıcılar, çoğunlukla çoğunluk veya ağırlıklı oylama ile birleştirilir (voting) . Oylama operatörü, alt süreçleri içinde çoklu temel modelleri barındıran topluluk öğrencisidir (Mierswa v.d, 2006). Oylama süreci model çıktısı herhangi bir sınıflandırma modeli gibi davranır ve karar ağacı yönteminin kullanılabildiği her senaryoya uygulanabilir (Dedspande ve Kotu,2014). Modeli uygulama aşamasında, tüm temel sınıflandırıcılar tarafından tahmin edilen sınıflar toplanır ve en çok oy alan sınıf, topluluk modeli için tahmini sınıf olarak seçilir. Bir diğer yaklaşım ise, farklı algoritmaları (heterojen model sunumlarıyla) tek bir veri setine uygulayarak sınıflandırıcılar üretmektir (Merz, 1999). İstifleme (Stacking) sıklıkla sınıflandırıcıların topluluğuna ek olarak birleştirme yöntemini öğrenmek için kullanılır (Wolpert, 1992). Yığınlanmış Genelleme (İstifleme)'de temel fikir; Eğitim verilerinin düzgün öğrenilip öğrenilmediğinin öğrenilmesidir.2 aşamalıdır;

Belli bir temel sınıflandırıcı özellik uzayının belirli bir bölgesini yanlış bir şekilde öğrendiyse, ikinci aşamadaki (meta) sınıflandırıcı bu istenmeyen durumu algılayabilir. Diğer sınıflandırıcıların öğrenmiş olduğu davranışlarla birlikte, meta sınıflandırıcı bu yanlış eğitimi düzeltebilir. Oysa oylamada daha önce bahsettiğimiz gibi sınıflandırma işlemi temel sınıflandırıcılar tarafından en çok oyu alan sınıfa göre yapılır. Oylama iki aşamalı değildir. İstifleme ile oylama arasındaki farka baktığımızda; İstiflemenin aksine, sınıflandırıcıları oylama şeması ile birleştirirken meta düzeyde herhangi bir öğrenme gerçekleşmez. Tipik olarak, oylama (voting) ile karşılaştırıldığında stacking ile daha iyi performans elde edilir (Dzeroski ve Zenko, 2004). Bu çalışmada istifleme kullanılmasının nedenleri daha iyi performans vermesi ve iki aşamalı bir sınıflandırma yapmasıdır.

Bu çalışmada stacking topluluk metodu (stacking ensemble method) ile karar ağaçları, K-NN ve naive bayes yöntemleri birleştirilerek bir topluluk sınıflandırıcı modeli kurulmuştur. Bu model fenotip ve genotip özellikler içeren prostat kanseri verisine uygulanarak erkeklerin hasta olup olmadığını doğru bir şekilde sınıflandırmayı amaçlamaktadır.



Şekil 4.1. Topluluk Modeli

RapidMiner'da retrieve operatörü ile depolama alanlarına erişilir. RapidMiner'ın kullanımını kolaylaştıran tam meta veri işleme sağladığı için tüm dosya erişimlerinde bu operatör kullanılır. Ham bir dosyaya erişmenin tersine, verinin tam meta verisini sağlar, böylece tüm meta veri dönüşümleri mümkündür.

Depodan bir nesneyi yüklemek için verilerin saklandığı bölmeden gerekli nesne sürükleyip bırakılırsa da kurulan modelin süreç sayfasına otomatik olarak istenen nesnenin doğru yolunu içeren bir veri alma operatörü (retrieve operatörü) eklenecektir.

Şekil 4.1'de görüldüğü gibi, ilk aşamada model kurmak için daha önce birleştirilen, temizlenen ve dönüştürülen verinin son hali süreç sayfasına retrieve operatörü ile alınmıştır.

Topluluk metodları içierisinde kullanılacak olan karar ağaçları, K-NN ve naive bayes sınıflandırıcı yöntemleri sınıflandırma işleminde nominal ve polynominal değişken tiplerini kabul ederler. Bu yüzden öncelikle yapılması gereken değişkenlerin tümünün nominal veya polynominal olduğunun programa tanıtılmasıdır.

Bu sebeple Şekil 4.2'de görüldüğü üzere "Nümerik Değişkenleri Polinom Değişkenlere Dönüştürme Operatörü" (Numerical to Polynominal) ile tüm değişkenler polinom olarak programa tanıtılmıştır.

Şekil 4.2 ümerik Değişkenleri Polinom Değişkenlere Dönüştürme Operatörünün Parametreleri

Daha sonraki aşamada Şekil 4.2'de görüldüğü üzere kayıp değerlerin analizi bozmasını engellemek için bütün kayıp değerler ortalama değerler ile değiştirilmiştir. Bunun için de "Kayıp Değerleri Değiştirme" (Replace Missing Values) operatörü kullanılmıştır.

Şekil 4.3 Kayıp Değerleri Ortalama Değerler ile Değiştirme Operatörünün Parametreleri

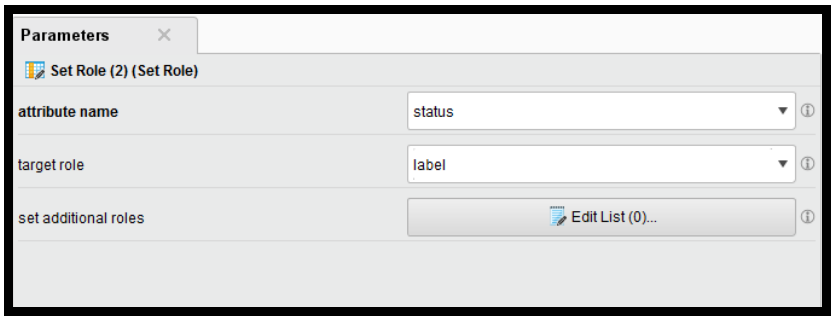
Uygulama kısmının başında da bahsedildiği gibi, "severity" (hastalığın ilerleme durumu) değişkeni modelin doğruluk değerinin %100 çıkmasına ve kurulan modeldeki karar ağacına göre bir bireyin hasta olup olmamasının sadece "hastalığın ilerleme durumu" değişkenine bağlı olduğunu göstermesine sebep olduğu için model kurulurken değişken analizden çıkarılmıştır. "Hastalığın ilerleme durumu" değişkeni prostat kanseri ile doğrudan ilişkilidir ve hasta olan grup için hastalığın şiddetini ifade eder. Fakat çalışmaya dahil edilince prostat kanserinin başka hangi değişkenlere

bağlı olduğunu görmemizi engellediği için modelden çıkarılması uygun görülmüştür. Ayrıca bu çalışmada prostat kanserinin var olup olmadığının tahlil ve tetkiklerden bağımsız eldeki bazı genotip ve fenotip değişkenler bilinirken sınıflandırılması amaçlanmıştır. Severity değişkeni ise hasta olanlar için hastalığın ilerleme durumunu gösteren ve belirli tahlil ve tetkiklerle anlaşılan bir değişkendir.

Şekil 4.4'de görüldüğü üzere "Değişken Seçme Operatörü " (Select Attributes) ile (severity) "hastalığın ilerleme durumu" değişkeni seçilmiş ve aşağıdaki "invert selection" seçeneği tıklanarak "seçilenlerin tersini seçmek" anlamına gelen severity değişkeni dışında kalan tüm diğer değişkenler filtrelenmiştir. Böylece "severity" değişkeni modelden çıkarılmıştır.

Şekil 4.4 Değişken Seçme Operatörünün Parametreleri

Daha sonraki adımda Şekil 4.5'deki "Rol Atama" operatörü ile ("Set Role") bir bireyin hasta olup olmadığını gösteren durum (status) değişkenine "label " yani etiket rolü verilmiştir. Böylece kurulan modelde sınıflandırma kriteri olarak durum (status) değişkeni anlaşılacaktır.



Şekil 4.5 Rol Atama Operatörü

En son adımda "Çapraz doğrulama" ("Cross Validation") operatörü ile modelin performansı tahmin edilecektir. Bu operatör, genellikle bilinmeyen veri kümeleri üzerinde öğrenen bir operatörün istatistiksel performansını tahmin etmek için çapraz

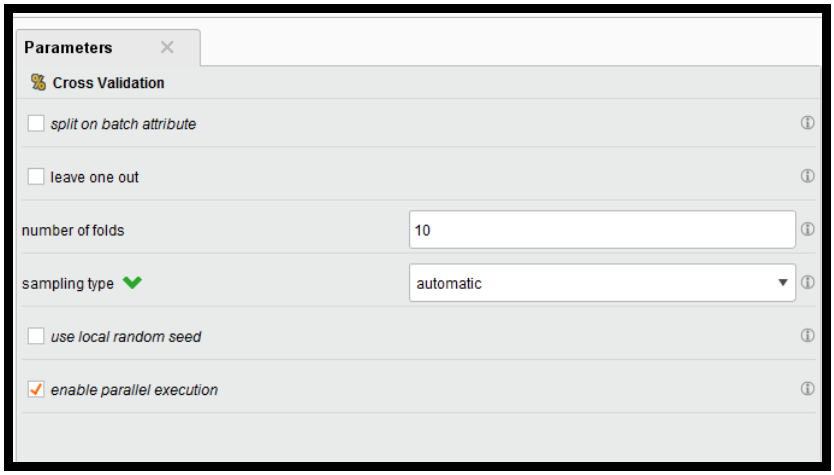
doğrulama gerçekleştirir. Esas olarak belirli bir öğrenme operatörü tarafından öğrenilen bir modelin uygulamada ne kadar doğru bir performans göstereceğini tahmin etmek için kullanılır. Girdi verisi, eşit büyüklükte k alt kümeye bölünür. K alt kümenin tek bir alt kümesi test veri kümesi olarak tutulur.

Yani k alt kümeden k-1 altküme eğitim veri seti olarak kullanılırken, bir küme de test verisi olarak kullanılır. Çapraz doğrulama işlemi daha sonra k kere tekrarlanır, k alt kümelerinden her biri tam olarak bir kez test verisi olarak kullanılır. Daha sonra k tekrarlarından elde edilen k tane sonucun tek bir performans değeri üretmek için ortalaması alınır. k değeri, doğrulama sayısı parametresi kullanılarak ayarlanabilir. K değeri olarak en yaygın olarak 10 kullanılır. Bu çalışmada da K değeri olarak 10 alınmıştır.

Çapraz doğrulama, en popüler iç içe geçmiş operatörlerden birisidir. Eğitim seti bir döngüde k parçaya bölünür ve k - 1 parça eğitim kümesi olarak ele alınırken kalan 1 parça test kümesi olarak ele alınır. Bu durum k kez tekrarlanarak k tane performans hesaplanır ve son performans hesaplanırken k tane performansın ortalaması alınır. Böylece aynı veri seti içinden farklı örnek küme dizilerinin performans ortalamaları alınarak nihai performansa ulaşılır.

Yapılan analizlerin çoğunda mantıklı performans hesaplamaları elde etmek için X-doğrulama veya diğer adıyla çapraz doğrulama operatörü kullanılır.

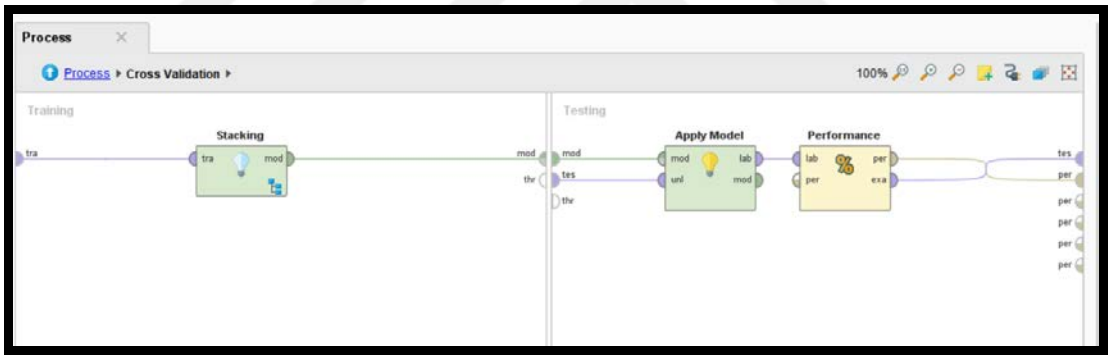
Örneğin 10 kat çapraz doğrulama gerçekleştirirken, her seferinde bir örneği çıkarmak yerine orijinal veri kümesinin onda biri çıkarılır ve test kümesi olarak kullanılır, geri kalan dokuzda biri ise eğitim kümesi olarak kullanılır. Bu işleme 10 kez devam edilir ve her seferinde hesaplanan performansların ortalaması alınarak son performans değeri hesaplanır.



Şekil 4.6 Çapraz Doğrulama Operatörü

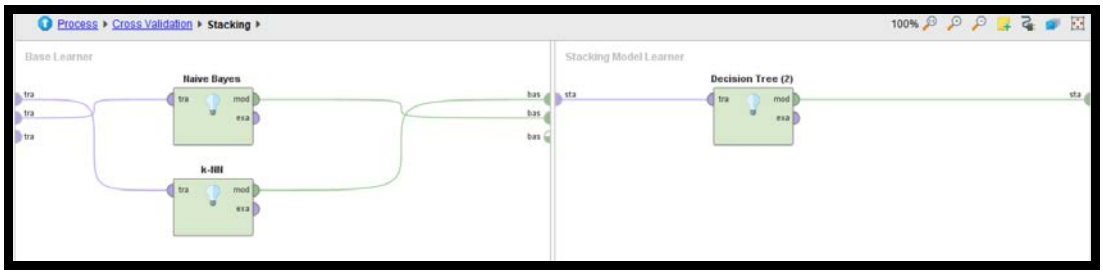
Çapraz Doğrulama operatörü iç içe geçmiş bir (nesting) operatördür. İki alt süreci vardır: biri eğitim alt süreci öteki ise test alt sürecidir. Şekil 4.6'da görüldüğü gibi süreç eğitim (training) ve test diye iki alt sürece bölünmüştür. Eğitim alt süreci bir modeli eğitmek için kullanılır. Eğitimli model daha sonra test alt sürecinde uygulanır. Modelin performansı da test aşamasında ölçülür.

Doğrulama operatörü bir model oluşturmamıza ve aynı adımda doğrulama verilerini uygulamamıza izin verir. Bu iki operasyon anlamına gelir - model oluşturma ve model değerlendirmesi - aynı operatör kullanılarak yapılandırılmalıdır. Bu, iç içe bir operatör olan doğrulama operatörü üzerine çift tıklanarak gerçekleştirilir. RapidMiner'deki tüm iç içe operatörler iki alt pencereye sahiptir. Bu operatör "açıldığında" İçeride iki bölüm olduğu görülür (bkz. Şekil 5.7. Sol kutucuk, stacking operatörünün yerleştirildiği ve eğitim veri örneklerinin% 90'ının kullanılarak modelin oluşturulduğu yerdir. Sağ kutucuk ise bu eğitimli modelin test verileri örneklerinin kalan% 10'unda "Model Uygula" (apply model) operatörünü kullanılarak uygulandığı ve "Performans Operatörü" (performance) kullanılarak modelin performansının değerlendirildiği yerdir (Desphande ve Kotu,2014).



Şekil 4.7 Çapraz Doğrulama Operatörü Alt Süreçleri

Bu çalışmada eğitim sürecinde daha önce de belirtildiği gibi stacking topluluk modeli kurulmuştur. İstifleme (stacking), bir meta öğrenme kavramını tanıtan çoklu modelleri birleştirmenin bir yoludur. Bagging ve boosting yöntemlerinin aksine, istifleme, farklı türdeki modelleri birleştirmek için kullanılır. İstifleme de iç içe geçmiş bir operatördür ve alt süreçlerinde K-NN, naive bayes ve karar ağacı modelleri birleştirilerek bir istifleme (stacking) modeli kurulmuştur. İstifleme (stacking) metodu stacking modeli kısmında ayrıntılı olarak açıklanacaktır.



Şekil 4.8 Stacking Operatörünün Alt Süreçleri

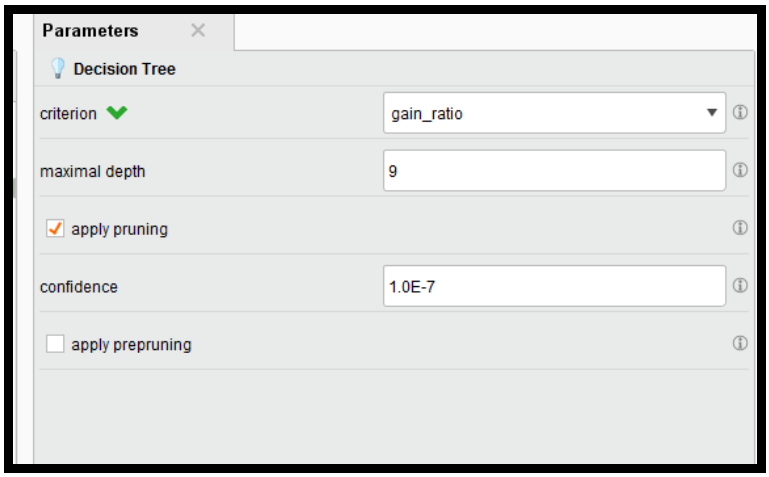
Kullanılan her bir yöntemin parametreleri optimize edilerek en iyi sonucu veren parametreler seçilmiştir. Buna göre, karar ağaçlarının önemli parametreleri tablo 4.3'deki şekilde optimize edilmiştir. Optimize edilen bu değerler de şekil 4.9'daki şekilde modelde kullanılmıştır.

Tablo 4.3 Karar Ağacı Parametreleri

ParameterSet

```
Parameter set:

Performance:
PerformanceVector [
-----accuracy: 75.12% +/- 5.22% (mikro: 75.13%)
ConfusionMatrix:
True:   2       1
2:      464     179
1:      67      279
]
Decision Tree.confidence        = 1.0E-7
Decision Tree.minimal_gain      = 5.0
Decision Tree.minimal_leaf_size = 3
Decision Tree.minimal_size_for_split    = 4
Decision Tree.maximal_depth     = 9
Decision Tree.number_of_prepruning_alternatives = 10
Decision Tree.apply_pruning     = true
Decision Tree.apply_prepruning  = false
```

Tablo 4.3'de karar ağacı parametrelerinin hesaplanmış optimum değerleri verilmiştir. Tablo 4.3'e göre, Karar Ağacı minimal kazanç parametresi 5, minimal yaprak boyutu parametresi 3, bölünme için minimal boyut parametresi 4, maximal derinlik parametresi 9, önbudama alternatiflerin sayısı parametresi 10 olarak hesaplanmıştır. Optimum sonuçlara göre, budama yapılmasına karar verilirken, ön budama yapılmamasına karar verilmiştir.



Şekil 4.9 Modelde Kullanılan Karar Ağacı Parametreleri

Şekil 4.9 da modelde kullanılan Karar Ağacı Parametreleri gösterilmektedir. Karar Ağacında nitelik seçimi için uygulanan matematiksel metrikler algoritmadan algoritmaya farklılık gösterir. Örneğin, Bilgi Kazanımı, ID3 algoritması için kullanılırken, Gini Endeksi CART Algoritması için, Kazanım Oranı ise C4.5 Algoritması için kullanılır. Bu çalışmada C4.5 Karar Ağacı Algoritması kullanılmıştır ve ağaç Bilgi Kazanım Oranı kriterine göre oluşturulmuştur.

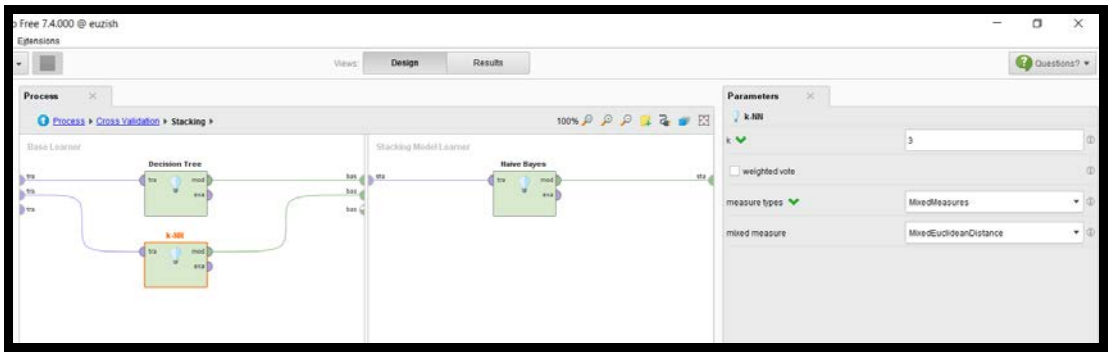
Tablo 4.3'de de bahsedildiği gibi maximal derinlik parametresi 9 olarak hesaplanmış ve budama yapılmıştır. Budama yapılırken kötümser hata hesaplamasında kullanılan güven düzeyini belirten "confidence" parametresi de 1.0E-7 olarak hesaplanmıştır.

Tablo 4.4 K En Yakın Komşuluk Algoritması Parametreleri

ParameterSet

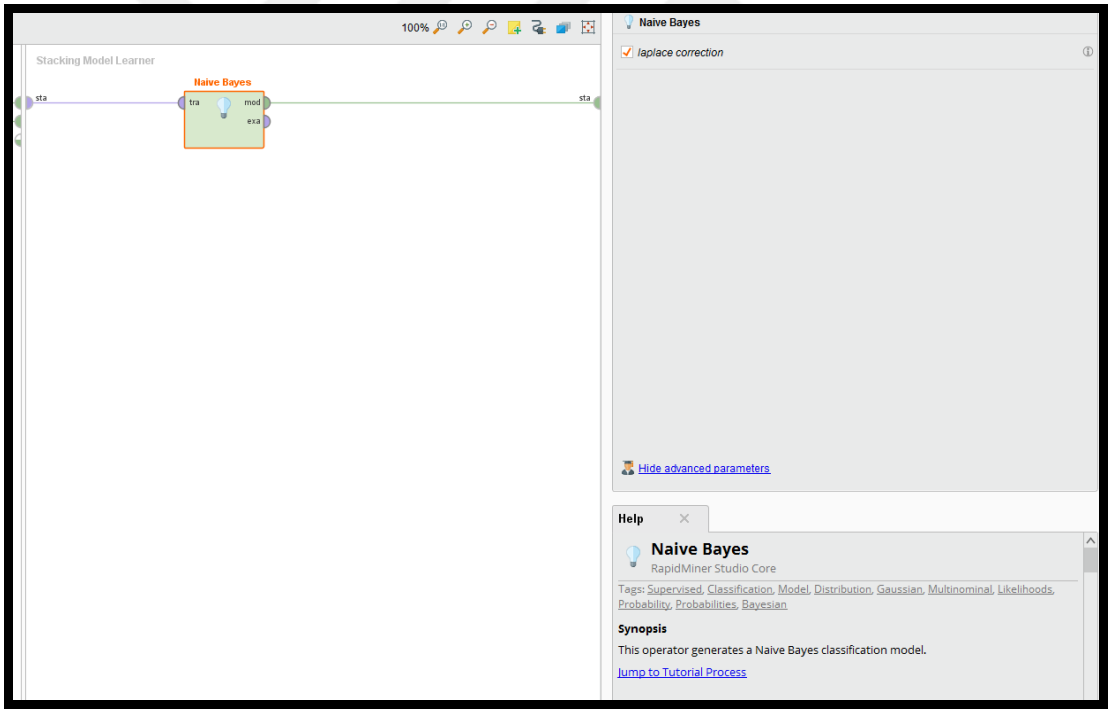
```
Parameter set:

Performance:
PerformanceVector [
*****accuracy: 62.08% +/- 3.55% (mikro: 62.08%)
ConfusionMatrix:
True:   2       1
2:      488     332
1:      43      126
]
k-NN.k  = 3
k-NN.weighted_vote      = false
k-NN.mixed_measure      = MixedEuclideanDistance
k-NN.kernel_type        = radial
```

Tablo 4.4.'de K En Yakın Komşuluk Algoritması Parametreleri gösterilmektedir. Tablo 4.4.2e göre k değeri 4 iken, ağırlıklandırılmış oylama yapılmamıştır.



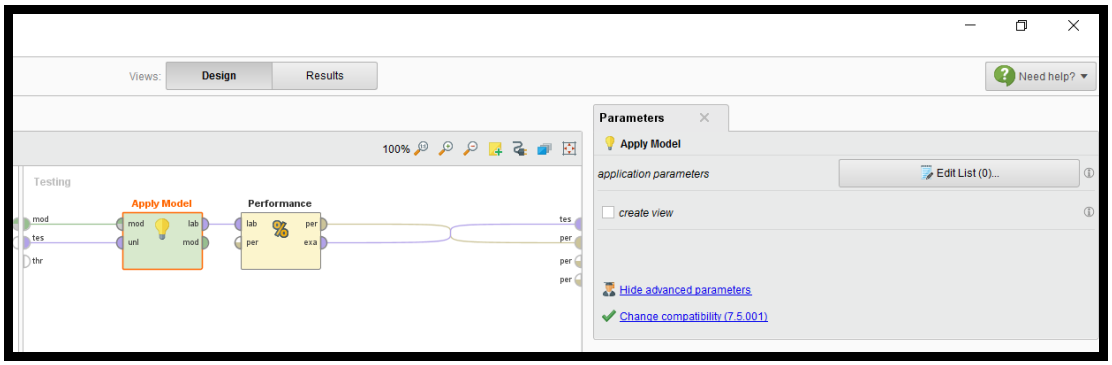
Şekil 4.10 K En Yakın Komşuluk Yöntemi Parametrelerinin Modelde Kullanılması

Şekil 4.10'da hesaplanan optimum K En Yakın Komşuluk Algoritması Parametrelerinin kurulan Stacking Topluluk Modelinde kullanıldığı gösterilmektedir.



Şekil 4.11 Naive Bayes Yöntemi Parametrelerinin Modelde Kullanılması

Şekil 4.11'de hesaplanan optimum Naive Bayes Algoritması Parametrelerinin kurulan Stacking Topluluk Modelinde kullanıldığı gösterilmektedir.



Şekil 4.12 Apply Model Operatörü

Apply model operatörü ile, önceden öğrenilmiş veya eğitilmiş olan model veri seti üzerine uygulanır. Performans operatörü ise, **performans değerlendirmesi için kullanılır. Performans kriterleri değerlerinin bir listesini sunar. Bu performans kriterleri,** öğrenilen modele uygun olarak otomatik bir biçimde belirlenir.

4.5.2 Kurulan İstifleme (Stacking) Modeli

İstifleme (stacked generalization veya stacking) (Wolpert, 1992), bir meta öğrenen kavramını tanıtan çoklu modelleri birleştirmenin bir yoludur. İşlem aşağıdaki gibidir (Kantardzic,2011):

1. Eğitim kümesi ayrık iki kümeye bölünür.
2. İlk bölümde birçok temel öğrenici eğitilir.
3. İkinci bölümdeki temel öğrenciler test edilir.
4. 3'deki tahminler girdi, doğru cevaplar da çıktı olarak kullanılarak daha üst seviyedeki öğrenici eğitilir.

Dikkat edilirse 1'den 3'e kadar olan adımlar çapraz doğrulama ile aynıdır, ancak en iyi öğrenicinin seçilmesi yaklaşımının kullanılması yerine istifleme yönteminde temel öğreniciler çoğu zaman doğrusal olmayan şekilde birleştirilir.

Stacking ise tahmin algoritmalarının optimal birleşimini bulan denetlenen topluluk makine öğrenme algoritmasıdır. Süper Öğrenme veya İstiflenmiş Regresyon olarak da adlandırılan istifleme, temel öğrenicilerin optimum kombinasyonunu bulmak için ikinci düzey bir "meta öğrencisi" eğitmeyi içeren bir algoritma sınıfıdır. Bagging ve boosting yöntemlerinin aksine, stackingin amacı farklı öğrencilerin bir araya getirilmesidir. İstifleme (stacking) kavramı başlangıçta 1992 yılında geliştirilmiş olsa da, istifleme teorisi, 2007'de "Süper Öğrenci" (super learner) başlıklı bir makalenin yayınlanmasına kadar ispatlanmadı. Bu makalede, Süper

Öğrenci topluluğunun öğrenme için asimptotik olarak en uygun (optimal) sistemi temsil ettiği gösterildi. Genel olarak istifleme olarak ifade edilen bazı topluluk yöntemleri vardır, ancak Süper Öğrenci topluluğu, "birinci seviye" veri olarak adlandırılan veriyi oluşturmak için çapraz doğrulama kullanarak ayırt edilir veya meta öğrenme veya "combiner" algoritması üzerinde eğitilmiştir.

Aşağıdaki adımlar, bir istifleme topluluğunun eğitimi ve testinde yer alan özgün görevleri tanımlamaktadır. Bu adımlar istifleme algoritmasının daha iyi anlaşılmasına da yardımcı olur. Her bir adımda parantez içinde belirtilen kısım bu adımların bizim çalışmada karşılık gelen kısmını ifade eder ($H_2O.ai, 2017$).

1)Topluluğu kurun.

Çalışmada k en yakın komşuluk, naif bayes ve karar ağacından oluşan topluluk kurulmuştur.

2)L temel düzeyde algoritmaların model parametre setiyle birlikte bir listesini belirtin.

Temel düzeyde algoritmalar: k en yakın komşuluk ve naif bayesdir. Bu algoritmaların optimize edilmiş parametreleri sırasıyla tablo 5.5 ve şekil 5.9' da gösterilmiştir.

3)Bir meta öğrenme algoritması belirtin.

Meta öğrenme algoritması karar ağacı dır.

4)Topluluğu eğitin.

5)Eğitim kümesindeki L taban algoritmasının her birini eğitin.

6)Bu öğrenicilerin her birinde k katlı çapraz doğrulama yapın ve her L algoritmasından çapraz doğrulama yöntemi ile tahmin edilmiş değerleri toplayın.

Her bir L algoritmasından çapraz doğrulama yöntemi ile tahmin edilmiş N tane değer birleştirilerek yeni bir N x L matris oluşturulabilir. Bu matris, orijinal yanıt vektörü boyunca "birinci seviye" veri olarak adlandırılır. (N = eğitim kümesindeki satır sayısı.)

7)Birinci seviye veri üzerinde meta öğrenme algoritmasını eğitin. "Topluluk modeli", L taban öğrenme modellerinden ve meta öğrenme modelinden oluşur. Meta öğrenme modeli daha sonra test kümesinde tahminler üretmek için kullanılabilir.

8)Yeni veriler üzerinde tahminleme yapın.

9)Topluluk tahminleri oluşturmak için öncelikle temel öğrenicilerden tahminler oluşturun. Sonra bu tahminleri meta öğreniciye gönderin.

Bir topluluk tahmininin değerlendirilmesi, tipik olarak, tek bir modelin tahminini değerlendirmekten çok daha fazla hesaplama gerektirir, bu nedenle topluluklar çok fazla hesaplama yaparak zayıf öğrenme algoritmalarını telafi etmenin bir yolu olarak düşünülebilir. Bir topluluk, denetlenen bir öğrenme algoritmasıdır çünkü eğitilebilir ve ardından tahminler yapmak için kullanılabilir (Rapidminer,2017).

Stacking operatörü iç içe geçmiş bir operatördür. İki alt süreci vardır: Temel öğreniciler (base learners) ve stacking modelini öğrenme alt süreci (Stacking model learners) (Şekil 5.8'de de görüldüğü gibi). Bu çalışmada temel öğreniciler olarak K-NN ve naive bayes algoritmaları eğitilmiştir. Daha sonra her bir algoritmanın 10- katlı çapraz doğrulama yöntemi ile performansları tahmin edilmiştir. Her bir algoritmadan her bir örnek için tahminler yapılmıştır.

K-NN yöntemi tarafından tahmin edilmiş değerler "taban tahmin 0" (base prediction 0) değişkeni tarafından ifade edilir iken, naive bayes yöntemi tarafından tahmin edilmiş değerler "taban tahmin 1" (base prediction 1) değişkeni tarafından ifade edilir.

4.5.2.1.Temel Öğreniciler ve Meta Düzeyde Öğrenici

Temel öğreniciler olarak K-NN ve naive bayes kullanılmıştır. K en yakın komşuluk yönteminde optimum k değerini seçmek çok önemlidir. Çünkü k belirli bir değere kadar arttırıldığında performans yükselirken, belli bir k değerinden sonra performans düşer. Bu çalışmada daha önce de ifade edildiği gibi en iyi performansı veren k değeri optimize edilerek 3 olarak bulunmuştur.

Naive Bayes algoritması, güçlü bağımsızlık varsayımlarıyla Bayes teoreminin uygulanmasına dayanan istatistiksel bir sınıflandırıcıdır. Örneklerin hangi olasılıkla hangi sınıflara ait oldukları bilgisini verir. Naive Bayes sınıflandırıcıya göre; niteliklerin hepsi aynı derecede önemlidir. Naive Bayes algoritmasına göre nitelikler birbirinden bağımsızdır yani bir niteliğin değeri başka bir nitelik değeri hakkında bilgi içermez (Öğüdücü, 2010).

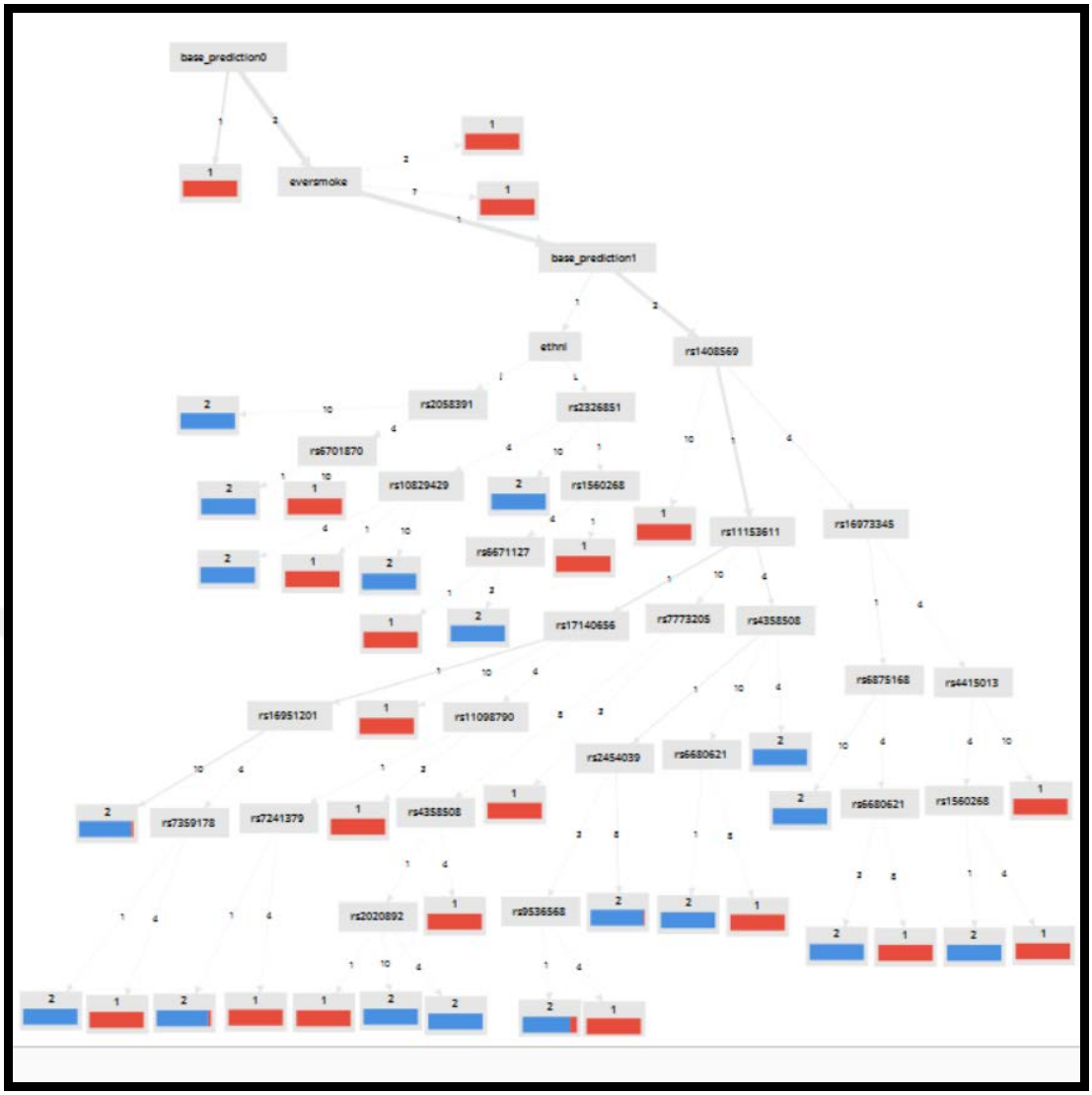
Naive Bayes sınıflandırıcıları, belirli bir sınıfta bir özellik değerinin etkisinin diğer özellik değerlerinden bağımsız olduğunu varsayar. Bu varsayıma, sınıf koşullu bağımsızlık denir. Bu varsayım ilgili hesaplamaları basitleştirmek için yapılır ve bu anlamda "naif" olarak kabul edilir. Kısaca bir özelliğin belirli bir değerinin varlığının diğer herhangi bir özellik değerinin varlığıyla ilgisi olmadığını varsayar.

Örneğin, yeşil, yuvarlak ve çapı 6 cm olan bir top tenis topu olarak sınıflandırılabilir. Bu özellikler birbirine bağlı olsa bile, Naive Bayes algoritması bu özelliklerin tümünü, bu topun bir tenis topu olma olasılığına bağımsız olarak katkıda bulunduğunu varsayar. Bayesian sınıflandırıcıları, büyük veritabanlarına uygulandığında yüksek doğruluk ve hız sergilemiştir (Hoffman ve Klinkenberg,2014). Bu çalışmada da naive bayes algoritması sınıflandırmak için kullanılan bütün niteliklerin birbirinden

bağımsız olduğunu var sayar. Bayes sınıflandırıcılar kolay eğitilir ve çoğu durumda iyi sonuçlar verir.

Fakat sınıf bilgisi verildiğinde niteliklerin bağımsız olduğunu var sayar. Bu da gerçek hayatta her zaman mümkün olmayabilir. Ayrıca değişkenler arası ilişkiyi modelleyemez. RapidMiner'deki Naive Bayes operatörünün yalnızca bir parametresi vardır, o da Laplace düzeltme (Laplace correction) 'dır. Bu uzman parametre, Laplace düzeltmesinin sıfır olasılığının yüksek etkisinin önlenmesi için kullanılıp kullanılmayacağını belirtir. Sıfır olasılıklardan kaçınmak için, eğitim verilerinin çok büyük olduğu ve ihtiyacımız olan her sayıya bir tane eklenmesinin tahmini olasılıklarda önemsiz bir fark yaratacağı varsayımı kullanılabilir. Bu teknik Laplace düzeltme olarak bilinir (Hoffman ve Klinkenberg,2014). Bu çalışmada da laplace düzeltme kullanılmıştır çünkü kullanıldığında performans artmıştır.

Stacking modelini öğrenme alt sürecinde ise meta düzeyde ikinci aşamada karar ağacı yöntemi kullanılmıştır. İkinci aşamada kullanılan karar ağacı algoritması, “base prediction 0” ve “base prediction 1” tahmin değerlerini girdi değişken olarak kullanan, yeni kurulan çapraz doğrulama veri setinde eğitilir. Sonuç olarak da kurulan tüm stacking modelinin tahmin etmesi beklenen “status“ (durum) değişkeninin nihai tahmini gerçekleştirir. Hangi yöntemlerin kullanılacağı daha önce de belirtildiği üzere veri setine uygunluğu ve konuya uygunluğu konusunda literatür titizlikle incelenerek seçilmiştir. Ayrıca denenen başka yöntemler arasında en iyi performans sonucunu veren model olarak Şekil 4.13 ‘deki karar ağacı modeli seçilmiştir.



Şekil 4.13 Analiz Sonucunda Çıkan Karar Ağacı Yapısı

Şekil 4.13'deki modele göre en önemli değişken "base prediction 0" değişkenidir. Bu değişken k en yakın komşuluk tahmin değişkeninin ifade eder.

Bu demektir ki bir kişinin k en yakın komşuluk yöntemi tarafından prostat kanseri olarak sınıflandırılmasını ifade eder. Bu demek oluyor ki stacking modeldeki karar ağacı ilk önce bir kişinin k en yakın komşuluk yöntemi tarafından sınıflandırılıp sınıflandırılmamasına bakar.

O zaman k en yakın komşuluk yöntemi tarafından yani "base prediction 0" değişkeni tarafından prostat kanseri olarak sınıflandırılan bir kişi stacking modelde kullanılan karar ağacı algoritması tarafından da prostat kanseri olarak sınıflandırılmıştır.

Eğer k en yakın komşuluk yöntemi tarafından prostat kanseri olarak tahmin edilmediyse, yani "base prediction 0" değişkeni o kişiyi prostat kanseri değil şeklinde sınıflandırdı ise, eversmoke değişkenine bakılır.

Eversmoke değişkeni bireyin hayatının bir döneminde sigara içip içmeme durumunu ifade eder. Bir kişi hayatının bir döneminde sigara içmiş ise stacking model tarafından o kişi hasta olarak sınıflandırılır. Modele göre hayatının bir döneminde sigara içmiş olmak ile prostat kanseri olmak arasında çok sıkı bir ilişki vardır. Nitekim tahmin değişkeni 0'dan sonra en önemli değişken olarak hayatının bir döneminde sigara içip içmeme fenotip değişkeni görülmektedir. Eğer bir kişi hayatının bir döneminde sigara içmediyse o zaman o kişinin naive bayes yöntemi tarafından yani "base prediction 1" değişkeni tarafından hasta olarak tahmin edilip edilmemesine bakılır. Bu değişken tarafından hasta olarak sınıflandırıldı ise o zaman o kişinin etnik kökenine (ethnicity) bakılır. Kişinin japon veya latin kökenli olup olmamasına göre bakılacak olan genotip değişken farklılık gösterir. Eğer kişi japon kökenli ise rs2058391 SNP'ine bakılırken, kişi latin kökenli ise rs2326851 SNP'ine bakılır. Eğer kişi japon kökenli ise ve rs2058391 SNP'inin allelic profili GG ise hasta değil şeklinde sınıflandırılır. Eğer rs2058391 SNP'inin allelic profili AG/GA ise o zaman rs5701870 SNP'ine bakılır. Eğer rs5701870 SNP'inin allelic profili GG ise bu kişi hasta olarak sınıflandırılırken AA ise hasta değil şeklinde sınıflandırılır.
Eğer bir kişi "base prediction 1" değişkeni tarafından hasta değil şeklinde sınıflandırılırsa o zaman rs1408569 SNP'ine bakılır. Bu SNP'in allelic profili GG ise o zaman o kişi hasta olarak sınıflandırılmıştır. AA ise rs11153611 SNP'ine bakılır ve AG/GA ise rs16973345 SNP'ine bakılır.
Dikkat edilirse karar ağacındaki en önemli değişkenler sırasıyla base prediction0, eversmoke, base prediction1, ethnicity ve rs1408569 SNP'idir. Daha sonraki basamakta rs2058391 ve rs2326851 gelir. Sınıflandırmayı etkileyen en önemli fenotip değişken hayatında bir dönemde sigara içmiş olup olmama durumu (eversmoke) iken onu etnik köken (ethnicity) takip eder. En önemli genotip değişken ise rs1408569 SNP'idir.

4.6 Modelin Performansının Sonuçları

Binominal sınıflandırma çalışmalarında performans aşağıdaki kriterlere göre hesaplanır:
Accuracy (doğruluk)
Precision (kesinlik)
Recall (duyarlılık)
AUC (optimistic) (iyimser AUC eğrisi)
AUC (neutral) (tarafsız AUC eğrisi)

AUC (pessimistic) (kötümser AUC eğrisi)

4.6.1 Karşılaştırma Matrisi

Bir karşılaştırma matrisi, model tahmin sınıfları ile veri setindeki etiketli verilerin gerçek sınıfları arasındaki karşılaştırmaları yapan bir tablodur. (Desphande ve Kotu,2014).

Bir sınıflandırıcının tahmin sonuçlarını sunmanın açık ve net bir yolu, bir karşılaştırma matrisi kullanmaktır. Bu matrise kontenjans tablosu da denir (Brownlee,2018)

Bir ikili sınıflandırma problemi için, tablo 2 satır ve 2 sütundan oluşur. Üst kısmın üstünde gözlenen sınıf etiketleri bulunur ve yan tarafta tahmin edilen sınıf etiketleri bulunur. Her hücre, o hücrenin içine düşen sınıflandırıcı tarafından yapılan tahminlerin sayısını içerir. Aşağıda bizim çalışmamızın karşılaştırma matrisi bulunmaktadır. Çalışmamızda pozitif sınıf hasta olma durumu olarak alınmıştır.

Tablo: 4.5 Karşılaştırma Matrisi

		Gerçek Değerler		
Model		Pozitif (Hasta)	Negatif (Hasta Değil)	
Tahmin	Pozitif	Gerçek Pozitif (a) (Hasta iken hasta olarak sınıflandırılmış)	Yanlış Pozitif (b) (Sağlıklı iken hasta olarak sınıflandırılmış)	PRECISION (KESİNLİK) a/ (a+b)
Sınıfları	Negatif	Yanlış Negatif (c) (Hasta iken sağlıklı olarak sınıflandırılmış)	Gerçek Negatif (d) (Sağlıklı iken sağlıklı olarak sınıflandırılmış)	
		RECALL=SENSITIVITY (DUYARLILIK)	**SPECIFICITY (ÖZGÜLLÜK)**	**ACCURACY (DOĞRULUK)**
		a / (a+c)	**d/ (b+d)**	**(a+d) / (a+b+c+d)**

a – Veride hasta olarak etiketlenmişken model tarafından hasta olarak tahmin edilip sınıflandırılan kişilerin sayısını ifade eder.
b – Veride sağlıklı olarak etiketlenmişken model tarafından hasta olarak sınıflandırılan kişilerin sayısını ifade eder
c – Veride hasta olarak etiketlenmişken model tarafından sağlıklı olarak sınıflandırılmış kişi sayısını ifade eder.
d – Veride sağlıklı olarak etiketlenmişken model tarafından da sağlıklı olarak sınıflandırılmış kişi sayısını ifade eder.

4.6.2. Model Performans Değerlendirme Sonuçları

Kurulan model sonucunda aşağıdaki karşılaştırma matrisi sonuçları çıkmıştır. Bu sonuçlar aşağıda ayrıntılı bir şekilde incelenecektir.

Tablo 4.6 Modelin Karşılaştırma Matrisi

Model		Gerçek Değerler (1:Hasta, 2: sağlıklı)		
		Gerçek 1	Gerçek 2	Sınıf Tahmini
Tahmin Sınıfları	Tahmin 1	340	39	%89,71 → PRECISION (KESİNLİK)
	Tahmin 2	118	492	%81,41
	Sınıf Hatırlama	**%74,23** ↓ **RECALL= SENSİTIVITY (DUYARLILIK)**	**%92,65** ↓ **SPECIFICITY (ÖZGÜLLÜK)**	**%84,13** → **ACCURACY (DOĞRULUK)**

Doğruluk değeri (accuracy), toplam örnek sayısı üzerinden doğru tahminlerin yüzdesi alınarak hesaplanır. Doğru tahmin, tahmini öznitelik değerinin etiket özniteliğinin değerine eşit olduğu örnekler anlamına gelir.

Bu çalışmada kurulan model ile 989 bireyin %84 'ünün hastalık durumu doğru tahmin edilmiştir (Tablo 4.7). Sağlıklı olan 531 bireyden 39'u hasta olarak, hasta olan 458 bireyden 118'i sağlıklı olarak tahmin edilmiştir. Bu değerlere göre sağlıklı olanlar %93 oranında doğru tahmin edilirken (özgüllük), hasta olanların doğru tahmin edilme oranı % 74 'dür (duyarlılık). (Dikkat edilirse değerler virgülden sonra son iki basamak yuvarlatılarak verilmiştir).

Kesinlik (precision), doğru tahmin edilen (gerçek) pozitiflerin sayısının doğru tahmin edilen (gerçek) pozitif ve yanlış pozitiflerin sayısına bölünmesiyle bulunur. Başka bir deyişle, doğru tahmin edilen hastaların sayısı , hasta olarak tahmin edilen tüm sınıf sayısına bölünür. Buna Pozitif Tahmin Edici Değer (PPV) denir. Kesinlik (precision), bir sınıflandırıcıların doğruluğunun bir ölçüsü olarak düşünülebilir. Düşük kesinlik (precision), çok sayıda Yanlış Pozitif varlığının göstergesidir (Brownlee,2018).

Burada 340 hasta hasta olarak tahmin edilirken sağlıklı olan 39 birey de hasta olarak sınıflandırılmıştır. Bu değerlere göre kesinlik (precision) %89,71 dir.

Duyarlılık (recall), doğru tahmin edilen (gerçek) pozitiflerin sayısının doğru tahmin edilen (gerçek) pozitiflerin sayısı ve yanlış negatiflerin sayısının toplamına bölünmesi demektir. Bir başka deyişle, doğru tahmin edilen hasta sayısının , test verilerindeki hasta sayısına bölünmesi olarak da açıklanabilir. Buna Gerçek Olumlu

Oran da denir.Duyarlılık, sınıflandırıcıların bütünlüğünün bir ölçüsü olarak düşünülebilir. Düşük duyarlılık (recall) , birçok Yanlış Negatifleri gösterir.
Burada hasta olan 458 kişiden 340 tanesi doğru olarak tahmin edilirken, 118 tanesi sağlıklı olarak sınıflandırılmıştır (340/458*100). Bu değerlere göre recall (duyarlılık) değeri %74 dür.
Modelin tamamına bakıldığında accuracy (doğruluk), precision (kesinlik) ve recall (duyarlılık) değerleri sırasıyla %84,13, %89,84 ve %74,23'dür. Üç değer de birbirine birbirleriyle çok tutarlıdır.

Tablo 4.7 Birleştirilmiş Modelin Başarım Değerleri

Başarım Kriteri	**Değer**
Doğruluk (Accuracy)	%84,13
Kesinlik (Precision)	%89,84
Duyarlılık (Recall)	%74,23
Özgüllük (Specificity)	%92,65
AUC Optimistik	0,914
AUC	0,828
AUC Pessimistik	0,756

ROC grafikleri, isabet oranları ve sınıflandırıcıların yanlış alarm oranları arasındaki ilişkiyi göstermek için sinyal algılama teorisinde uzun süredir kullanılmaktadır. Bir isabet oranı doğru sınıflandırılmış hedef sayısının sınıflandırılmış hedef sayısına oranıdır. Diğer bir deyişle, bu, gerçek pozitiflerin, veri setinde tanımlanan pozitif sınıfa oranıdır. Bunun duyarlılık (sensitivity) tanımın olduğu karşılaştırma matrisinde görülebilir. Çalışmamıza göre sensitivity, doğru tahmin edilen hasta sayısının , test verilerindeki hasta sayısına bölünmesi ile bulunur .Yanlış alarm oranı yanlış tespit edilen hedeflerin sayısının toplam hedef olmayan (veya negatif) sayısına oranıdır (Karşılaştırma matrisine göre b/b+d şeklinde hesaplanır).

Çalışmamıza göre yanlış alarm oranı, yanlış tahmin edilen (yani sağlıklı olduğu halde hasta olarak sınıflandırılan) hasta sayısının veri setinde sağlıklılara oranı olarak bulunabilir.
Bu, (1-özgüllük) terimi olarak ifade edilebilir çünkü özgüllük (specificity) sağlıklı olarak doğru sınıflandırılmış hedef sayısının test verilerindeki sağlıklı hedef sayısına bölünmesi ile bulunur.

İsabet oranı = Duyarlılık (recall)
= a/ (a+c)

= GP / (GP + YN)

Yanlış alarm oranları = 1- Özgüllük

= 1- (d/ (b+d))

= b/ (b+d)

= 1- (GN/ (YP+GN))

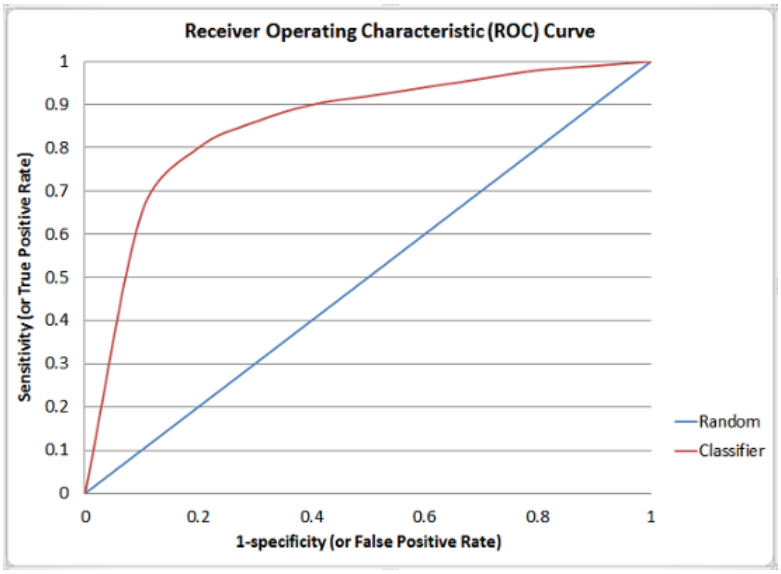
= YP/ (YP+GN)

GP:Gerçek Pozitif GN: Gerçek Negatif

YP: Yanlış Pozitif YN: Yanlış Negatif

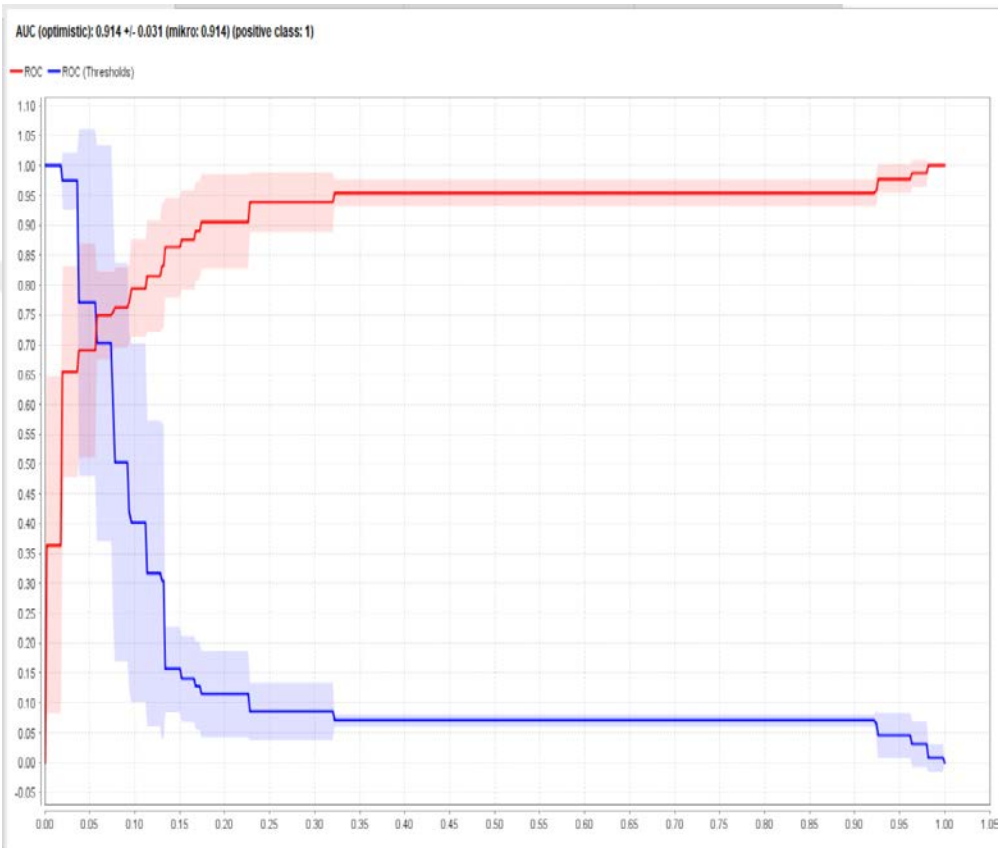
Bir ROC grafiği X ekseninde 1 -Özgüllük (Yanlış Pozitif oran) ve Y ekseni üzerinde duyarlılık (recall=sensitivity) ile oluşturulmuştur. Dolayısıyla bir ROC eğrisi, her yanlış pozitif için algoritma tarafından ne kadar gerçek pozitif algılanıp bulgulanacağına yardımcı olur. Örneğin , 45 derecelik bir açıda düz bir çizgi olan bir ROC eğrisi, tespit edilen her yanlış pozitiflik için karşılık gelen bir gerçek pozitif olduğuna işaret etmektedir. Başka bir deyişle, algoritmanın performansı % 50'dir. Bu "rastgele" performans üzerindeki herhangi bir artış bir gelişme olarak kabul edilir. Bu nedenle, iyi bir sınıflandırıcı bu 45 derece çizgisinin üzerinde bükülür. Sonuç olarak AUC temel olarak sınıflandırıcının performansının statik bir ölçüsüdür.

Örneğin AUC değerinin %70 olma durumunu açıklarsak, 1 grubundan rastgele seçilmiş bir durum O grubundan rastgele seçilmiş bir duruma göre %70 oranında daha yüksek bir değere sahip olacaktır (Desphande ve Kotu,2014).

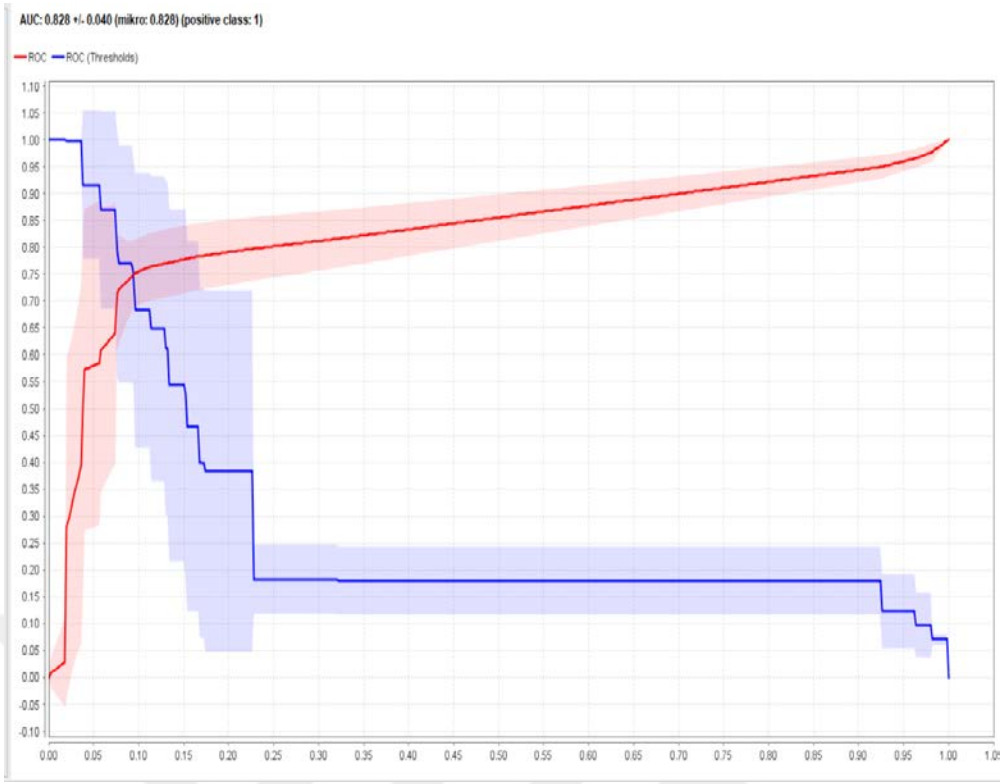


Şekil 4.14 Bir ROC Eğrisi.

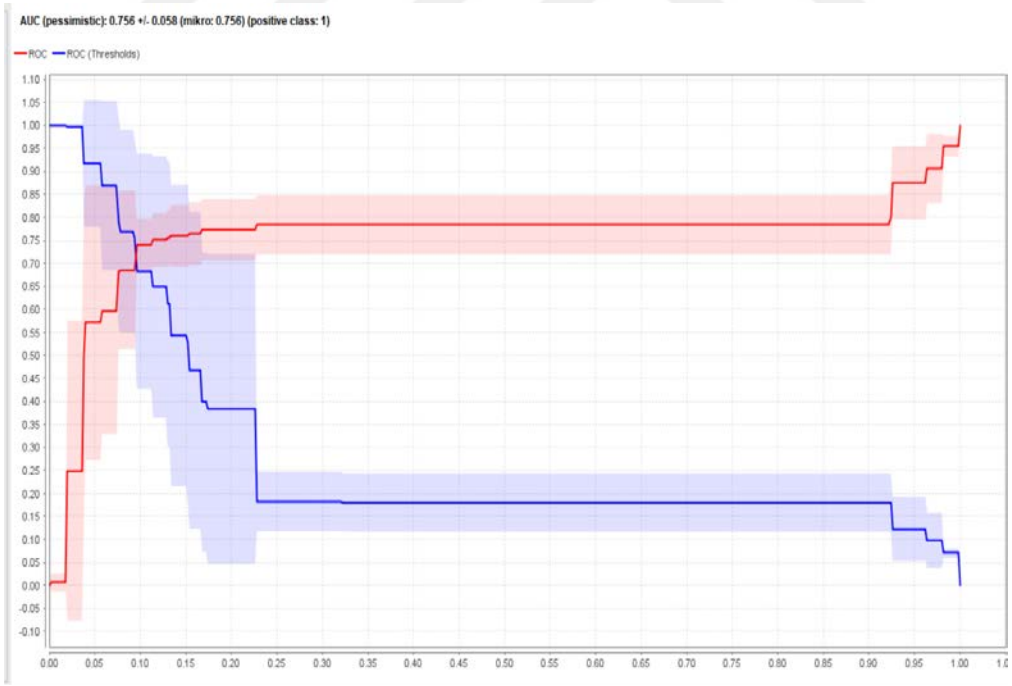
Çalışmada iyimser, kötümser ve tarafsız olmak üzere üç adet ROC eğrisi grafiği çizilmiştir ve altında kalan alanlar hesaplanmıştır. Bu alanlar sırasıyla Şekil 4.15, Sekil 4.16 ve Şekil 4.16'da verilmiştir. Bu şekillere göre sırasıyla en iyimser alan, alan ve en kötümser alanlar 0.914, 0.828 ve 0.756 dır. Bu değerlere göre algoritmanın sınıflandırma performansı en iyimser bakış açısıyla %91 iken en kötümser bakış açısıyla %76 'dır. Bu da demek oluyor ki en kötü ihtimale göre sınıflandırıcının performansı %76'dır.



Şekil 4.15 AUC Optimistik Grafiği



Şekil 4.16 AUC Grafiği



Şekil 4.17 AUC Pessimistik Grafiği

4.7 Modelin İstatistiksel Olarak Anlamlılığı

Bir önceki bölümde model sonuçları ve performans değerleri üzerinde durulmuştur. Modelin performans değerleri ne kadar iyi olursa olsun istatistiksel olarak anlamlı olmadıktan sonra o performans değerlerine güvenemeyiz. Modelin istatistiksel olarak anlamlı olup olmadığını anlamak için yapılması gereken güven aralığının bulunup modelin doğruluk değerinin o güven aralığı içinde kalıp kalmadığına bakılmasıdır. Öncelikle, büyük test kümelerinde doğruluk ölçütü normal olasılık dağılımına sahiptir (ortalama p, varyans p (1-p)/N).

$$Güven\ Aralığı\ p = \frac{2xNxacc+Z_{\alpha/2}\pm\sqrt{Z^2_{\alpha/2}+4xNxacc}}{2(N+Z^2_{\alpha/2})}$$ şeklinde hesaplanır.

Aşağıdaki tablolara göre doğruluk değeri güven aralığı içinde kaldığı için %95 güven aralığında kurulan model anlamlıdır.

Tablo 4.8 Güven Aralığı Tablosu

N	989
p (alt)	0,798
accuracy	0,841
p (üst)	0,856

Tablo 4.9 Güven Aralığı Tablosu

1-α	$Z_{\alpha/2}$
0,99	2,58
0,98	2,33
0,95	1,96
0,90	1,65

4.8 Kurulan Topluluk Model Performanslarının Tekil Algoritmaların Performansı ile Karşılaştırılması

Kurulan topluluk modelin içindeki tekil modellere göre ne kadar fark yarattığını görmek istedik. Bu sebeple içindeki her bir algoritma teker teker veri setine uygulanarak performans değerleri elde edilmiştir. Aşağıdaki tabloda kurulan topluluk

model ile model içindeki her bir algoritmanın karşılaştırmalı performansları bulunmaktadır.

Tablo 4.10: Topluluk Model ile Topluluk Modelde Kullanılan Her Bir Algoritmanın Performanslarının Karşılaştırılması

Performans Kriteri	Karar Ağaçları	K-En Yakın Komşuluk	Naive Bayes	Stacking Topluluk Modeli
Accuracy (Doğruluk)	67,65	59,56	73,62	83,22
Precision (Kesinlik)	64,93	74,06	81,12	85,78
Recall (Duyarlık)	65,94	21,83	56,55	76,42
AUC	0,55	0,65	0,87	0,78

Bu tabloya göre topluluk modelin performans değerlerinin tekil algoritmalı modellerin performans değerinden daha iyi olduğu açıkça görülmektedir. Topluluk modele en yakın performansı naive bayes sınıflandırıcısı vermiştir. Fakat bu tekil modelin de performans sonuçlarına bakıldığında precision (kesinlik) ve AUC değeri iyi olmasına rağmen recall (duyarlılık) değeri topluluk model ile karşılaştırıldığında oldukça düşüktür. Duyarlılığın düşük olması demek daha önce de tartışıldığı üzere yanlış negatif olarak sınıflandırmanın çok olduğu anlamına gelir. Bu da gösterir ki modelin hasta bireyleri sağlıklı olarak sınıflandırma oranı yüksektir. Bütün bu sonuçlara bakıldığında en iyi modelin stacking topluluk modeli olduğu görülmüştür.

4.9 Stacking Model Sonucunda Ortaya Çıkan SNP'lerin Değerlendirilmesi

Tablo 4.11 Modeldeki Önemli SNP'lerin Açıklaması (SNPEDIA, 2018)

SNP Adı	Geni	Gen Tipi	Açıklaması	Hangi Hastalıklarda Bulunduğu
Rs1408569	FAM155A	Protein kodlama	beyin, testisler ve prostat ile ilgili	depresyon
Rs2058391	LIN52	Protein kodlama	böbreklerle, şişmanlıkla, testislerle, beyinle ve prostatla ilgili	(Kalp karıncıkları arasında açıklık)Ventriküler septal bozukluklar, zihinsel yetersizlik, çoklu iskelet anormallikleri, başparmak gelişim sorunu
Rs6701870	CAMTA1	Protein kodlama	beyin, testis prostat ve dalak ve böbreklerde	Beyincik nedenli hareket kusuru, zeka geriliği, Meme kanseri
Rs6671127	GNG12-AS1	Protein kodlama	safra kesesi, prostat, testis, böbrek, kalp, kolon	Gelişim gecikmesi
Rs7773205	NKAIN2	Protein kodlama	En çok beyinde bulunur daha sonra kalp ve testislerde bulunur.	Hipofiz bezi yetmezliği Gelişim gecikmesi
Rs4358508	PDGFRB	Protein kodlama		Safra kesesi, plasenta, yağ dokusu, rahim, yumurtalık, prostat, akciğer, böbrek, prostat, testis Çocukluk dönemi kas hastalığı Beyin çekirdeği beyazlığı Infantile
Rs16951201	LAMA1	Protein kodlama	Bu gen tarafından kodlanan protein, trombosit kökenli büyüme faktörü ailesinin üyeleri için bir hücre yüzeyi tirozin kinaz reseptörüdür. Bu gen, kardiyovasküler sistemin normal gelişimi için gereklidir	Merosin proteininin eksikliğine bağlı doğumsal kas hastalığı
Rs2020892	MNAT1		Yağ, prostat, tiroid, böbrek üstü bezi, böbrek, beyin, rahim	Kulak gelişim bozukluğu Boy kısalığı Kalp damar gelişim bozukluğu 6 parmaklık

Görüldüğü üzere, model sonucunda önemli olarak ortaya çıkan SNP'ler çoğunlukla prostatla ilgili SNP'lerdir. Çok ilginç bir şekilde bazıları da beyinle ilgili Snp'lerdir.

SONUÇ

Uygulama Sonuçları

Bu çalışmada amaç veri madenciliği yöntemleri kullanılarak prostat kanserinin erken ve doğru teşhis edilebilmesi için anlamlı bir model oluşturmaktır. Bu model ile bir bireyin çeşitli genotip ve fenotip özelliklerine dayanarak prostat kanseri olup olmadığını en doğru şekilde sınıflandırmak amaçlanmaktadır. Çalışmada kullanılan veri kümesi NCBI'dan (National Center for Biotechnology Information- Ulusal Biyoteknoloji Bilgi Merkezi) alınmıştır.

Kullanılan veri 989 kişiden oluşmaktadır. Her kişi için genotip değişken olan 200 bin SNP ve 18 adet fenotip değişken vardır. Bu veri üç adımda indirgenmiştir; ilk adımda PLINK analizi yapılmış, ikinci adımda SPOT analizi kullanılmış ve son adımda bilgi kazanım oranı yoluyla ağırlıklandırma kullanılarak veri 418 değişkene indirgenmiştir. Veri, etnik köken, yaş, vücut kitle endeksi, şu anda sigara içme durumu, hayatının bir döneminde sigara içme durumu, likopen alım miktarı, vücuda alınan kalorinin yağ yüzdesi, kalsiyum alım miktarı, ailede prostat kanseri öyküsü, günlük fiziksel aktivite, günlük alkol tüketimi gibi fenotip değişkenleri içermektedir.

Çalışmada Bayes Sınıflandırıcılar, K En Yakın Komşuluk ve Karar Ağaçları yöntemleri Stacking Metodu ile birleştirilerek bir topluluk modeli oluşturulmuştur. Bu model ile prostat kanseri olan ve olmayanların en doğru şekilde sınıflandırılması amaçlanmaktadır. Çalışmada Stacking Metodu kullanılmasının amacı; Stacking Metodunun iki aşamalı bir sınıflandırma yöntemi olmasıdır. İlk adımda kullanılan temel sınıflandırıcılar özellik uzayının belirli bir bölgesini yanlış bir şekilde öğrendiyse, ikinci aşamadaki (meta) sınıflandırıcı bu istenmeyen durumu algılayabilir ve bu yanlış eğitimi düzeltebilir. Oysa ki faklı yöntemlerin kullanılabildiği bir diğer yöntem olan oylamada (voting) sınıflandırma işlemi temel sınıflandırıcılar tarafından en çok oyu alan sınıfa göre yapılır. Oylama tek aşamalıdır. Dzeroski ve Zenko'ya göre (2004) stacking ile daha iyi performans elde edilir. Bagging, boosting ve random forest yöntemlerinde zaten tek bir yöntemin topluluğu kullanılabilir.

Analizin aşamalarına bakıldığında, ilk önce veri süreç sayfasına alınmıştır. Daha sonra verideki nümerik değişkenler polinom değişkenlere dönüştürüldü çünkü kullanılacak sınıflandırma algoritmaları sadece polinom değişkenleri kabul etmektedir. Bir sonraki aşamada veri, kayıp değerlerden arındırılmıştır. Daha sonraki aşamada uygulama kısmında da bahsedildiği gibi doğruluk değerinin %100 çıkmasına sebep olan ve prostat kanseri ile doğrudan ilişkili olan hastalığın ilerleme durumu analizden çıkarılmıştır. Bir sonraki adımda bir bireyin hastalık durumunu

gösteren durum (status) değişkeni etiket değişken (label) olarak atanmıştır. Bilindiği üzere etiket değişkeni kurulan modelde sınıflandırma kriteri olarak alınır. Çalışmada da bir bireyin hastalık durumunu gösteren status değişkenine göre sınıflandırma yapılmaktadır.

En son adımda Çapraz Doğrulama (Cross Validation) operatörü ile modelin performansı tahmin edilmiştir. Böylece analiz için ayrı bir test veri setine gerek kalmadan tek bir veri seti eşit büyüklükte on alt kümeye bölünerek dokuz parçası eğitim veri seti bir parçası test veri seti olarak kullanılarak on kez performans değeri üretilmiş ve sonuç performans değeri üretilen on performans değerinin ortalaması olarak hesaplanarak bulunmuştur. Çapraz doğrulama operatörü iç içe geçmiş yani alt süreçleri olan bir operatördür; eğitim ve test olmak üzere iki alt süreçten oluşur. Eğitim sürecinde,veri örneklerinin %90 ı kullanılarak alt süreçlerinde topluluk modelinin kurulduğu stacking operatörü kullanılmıştır. Test sürecinde ise eğitim sürecinde oluşturulan model kalan %10 veri örneği üzerinde test edilmiştir. Test aşamasından sonra modelin performansı değerlendirilmiştir. Modeli kuran stacking operaörü de çapraz doğrulama operatörü gibi iç içe geçmiş bir operatördür ve iki alt süreci vardır; temel öğreniciler ve meta öğrenme. Temel öğreniciler aşamasında Naive Bayes ve K En Yakın Komşuluk kullanılırken, meta öğrenme düzeyinde Karar Ağaçları kullanılmıştır. Stacking modelinin süreçlerinde kullanılan tüm algoritmalar daha önce optimize edilmiştir ve modelde optimum sonuçlar kullanılmıştır.

Stacking modelinin tahmin etmesi beklenen “status” değişkeninin tahmini aşamasında, temel öğreniciler olan K En Yakın Komşuluk ve Naive Bayes Algoritmalarının tahmin değerleri meta öğrenme aşamasında girdi değişken olarak kullanılmıştır.

Kurulan modele göre en önemli fiziksel değişken K En Yakın Komşuluk yönteminin tahmin değerini veren değişkendir. Bunu sırasıyla **eversmoke** (hayatının bir döneminde sigara içme durumu) ve naive bayes yönteminin tahmin değişkeni ve **etnik köken** takip eder. En önemli genetik değişkenlerin ise sırasıyla rs1408569,rs2058391 ve rs2326851, rs5701870 ve rs10829429 olduğu tespit edilmiştir.

Modelin performans sonuçlarına bakıldığında, Accuracy (Doğruluk), Precision (Kesinlik) ve Recall (Duyarlılık) değerleri sırasıyla %84,13, %89,84 ve %74,23’dür. Bu değerlere göre, prostat kanserinin tahmininde sadece fiziksel ve genetik değişkenler kullanan model uygun sonuçlar vermektedir.

Tartışma

Bu çalışmada, herhangi bir tıbbı tahlil kullanılmadan, sadece bireyin fiziksel ve genetik değişkenlerine bakılarak prostat kanseri olup olmadığının en uygun veri madenciliği yöntem topluluğu kullanılarak tahmin edilmesi amaçlanmaktadır. Literatürde hem fiziksel (fenotip) hem de genetik (genotip) değişkenler içeren veri seti ile yapılan çalışmalar parmakla gösterilecek kadar azdır. Bu tarz çalışmalara Yücebaş (2016) ve Yücebaş ve Son (2014) değinmiştir.

Yücebaş (2016), çalışmasında SNP'lerin karmaşık hastalıklarla ilişkilendirilmesi ve teşhisinde veri madenciliği yöntemlerinden k-nn,nb, ka ve destek vektör makinasının prostat kanseri veri kümesi üzerindeki performansdları karşılaştırılmıştır. Bu çalışmada 1261 kişi için 2710 SNP'li ve 12 fenotipli bir veri kullanılmıştır. Bu veri Afro Amerikan, Japon ve Latin etnik kökenlerini içermektedir. Yöntemlerden Destek Vektör Makinası kesinlik ve ROC eğrisi altında kalan alan bakımından en yüksek başarımı verirken, NB ise duyarlılık ölçütü en yüksek yöntem olmuştur. Çalışmada sadece başarım performansı değerleri karşılaştırıldığı için prostat kanseri teşhisinde en önemli genotip veya fenotip değişkenlerin hangileri olduğuna değinilmemiştir.

Yücebaş ve Son (2014)'un birlikte yaptıkları çalışmalarında ise Destek Vektör Makinesi ve Karar Ağacı yöntemlerini birleştirerek prostat kanseri hastalarını sınıflandırmayı amaçlamışlardır. Çalışmalarında 4650 hasta ve 4795 kontrol grubu olmak üzere, Afro Amerikan, Latin ve Japonları içeren çok etnikli bir veri kümesi kullanmışlardır. Bu veri kümesi de bizim çalışmamızdaki gibi hem genotip hem de fenotip değişkenler içermektedir. Aynı zamanda 2710 SNP ve 20 fenotip içermektedir. Yücebaş ve Son'un çalışmalarında her bir yöntemin ve birleştirdikleri modelin sonuçları paylaşılmıştır. Birleştirdikleri model tekil yöntemlerden çok daha iyi sonuç vermiştir. Modelin doğruluk değeri %93,81, kesinlik (precision) %96,55 ve duyarlılık (recall) %90,92'dir. Yani model %93,81 oranla hasta olanları doğru tahmin edebilmektedir. Yüksek kesinlik ve hassasiyet değerleri de sağlıklı iken hasta olarak sınıflandırılan ve hasta iken sağlıklı olarak sınıflandırılan kişi sayısının azlığını gösterir.

Yücebaş ve Son'un yaptığı çalışma (2014) bizim çalışmamızla karşılaştırıldığında, Yücebaş ve Son'un çalışmasının daha iyi performans sonuçları vermesinin sebebi kullanılan veri kümesinin daha çok kişi içermesinden kaynaklandığı düşünülmektedir. Yücebaş ve Son'un çalışmasında kullanılan karar ağacına bakıldığında, en önemli değişken in etnik köken olduğu görülmektedir..İlgili çalışmadan, etnik kökeni vücut kitle edndeksi, onu da sigara içme durumu takip etmektedir. Bizim çalışmamızda da sigara içme durumu ve etnik köken en önemli

değişkenler olarak belirlenmiş olup sonuçlar Yücebaş ve Son'un çalışmasında elde edilen sonuçlarla tutarlıdır.

Konu ile ilgili literatürde farklı çalışmaların yer aldığı görülmektedir. Kaya ,Çolak ve Özdemir (2013), Gülkesen v.d (2010) ve Elshazly ve arkadaşları (2013) bu konuda çalışmalar yapmışlardır. Yapılan üç farklı çalışmada prostat kanseri şüphesi duyulan vakalara kan testi yapılarak bakılan Prostat Spesifik Antijeni (PSA) değerleri yardımıyla, çeşitli veri madenciliği yöntemleri ile prostat kanseri olan veya olmayan vakaların tahminlerinin yapılmış olduğu görülmektedir.

Kaya, Çolak ve Özdemir (2013), 203 erkek birey için PSA değerleri yardımıyla prostat kanserini Yapay Sinir Ağları (YSA) modelleri yardımıyla tahmin etmeyi amaçlamışlardır. Çalışma sonucunda, değişik YSA modellerinin, PSA Yardımıyla prostat kanserini tahmin etmede iyi sonuçlar verebildiği belirtilmiştir.

Gülkesen ve arkadaşları (2010) prostat kanserinin tahmini hakkında yaptıkları çalışmada, PSA değerinin tanı koymada doğruluğunun arttırılması için prostat kanserinin tahmininde ikili lojistik regresyon (LR), Karar Ağaçları (KA) ve Yapay Sinir Ağları (YSA) modellerinin başarılarını karşılaştırmışlardır. İlgili çalışmada,997 hastadan biyopsi alınmış, PSA, PSAD ve f/t değişkenlerini kullanmışlardır. PSAD VE f/t de PSA gibi prostat kanserinin tanısında kullanılan tıbbi değerlerdir. Çalışma sonucuna göre, işlem karakteristik eğrisi (ROC) altında kalan alan (AUC) Lojistik Regresyon (LR) için 0,717, Yapay Sinir Ağları (YSA) için 0,515 ve Karar Ağaçları (KAÇ için 0,629 'dur. Bu da demektir ki Lojistik Regresyon yönteminin sınıflandırma performansı 0,717 iken, Yapay Sinir Ağları yönteminin sınıflandırma performansı 0,515 ve Karar Ağaçları Yönteminin sınıflandırma performansı ise 0,629'dur.Gülkesen ve arkadaşlarının çalışmasında en iyi sınıflandırma performansına sahip yöntem Lojistik Regresyon olmuştur.

Elshazly ve arkadaşları (2013), Prostat kanserinin biyopsiye ihtiyaç duymadan erken ve doğru teşhisi için Çoklu Sınıflayıcı Forrest yöntemi kullanmışlardır. Elshazly ve arkadaşları ilgili çalışmalarında, kullandıkları topluluk metodun, tek bir random forest testinden ve tek bir karar ağacı yönteminden çok daha iyi sonuç verdiğini göstermişlerdir.

Genelde literatürde PSA değerleri yardımıyla, çeşitli Veri madenciliği Yöntemleri ile prostat kanseri olan veya olmayan vakaların tahminlerinin yapılmış olduğu görülmektedir. Bunun sebebi hem genetik hem de fiziksel veri içeren veri setlerinin azlığıdır. PSA içeren veriler hastanelerden elde edilebilir fakat diğer veriler özel veri tabanlarından izinle elde edilebilmektedir. Bu yüzden çalışmamızın literatüre önemli bir katkı sağladığı düşüncesindeyim. Çalışmamızın ayrıca herhangi bir teste gerek duyulmaksızın direk genetik ve fiziksel değişkenleri belirlenmiş olan bir bireyin

hastalık durumunu tahmin etmesi de öenmli bir artıdır. Günümüzde genetik o kadar gelişmiştir ki çeşitli şirketler web sitelerinde SNP'leri kullanarak hastalık taraması yapmaktadırlar.(23andMe:https://www.23andme.com/en-int/;MyHeritage: https://www.myheritage.com/;Ancestry.com:https://www.ancestry.com/, FamilyTreeDNA: https://www.familytreedna.com/)

Bu şirketler hangi hastalıklara yatkın olduğumuzu söyleyebilmektedirler. Bu şirketler sayesinde bireyin yatkınlığa sahip olduğu hastalıklar konusunda önceden önlem alıp o konuda kontrollerini eksiksiz yaptırması ve hastalığının erken teşhis edilmesi açısında büyük önem arz etmektedir.

Sınırlılıklar

Tez konusuna ilk karar verme aşamasında verinin İzmir'in bir üniversitesinin üroloji bölümünden alınması kararlaştırılmıştı, fakat veri toplama aşamasında alınacak verinin miktarının ve değişken sayısının analiz için yetersiz olduğu anlaşılmıştır. Bu sebeplerden dolayı veri çeşitli prosedürlerle NCBI'dan istenmiştir. Bu çalışmadaki en büyük kısıt verinin NCBI'dan istenmiş olması, dolayısıyla farklı etkin kökenleri içermesidir. Gelecekteki çalışmalarda geniş bir veri tabanına erişilmesi halinde aynı uygulama ülkemiz için de rahatlıkla yapılabilecektir.

Öneriler

Bu çalışma Türkiye'de yaşayan erkekler üzerindek yapılabilse literatürümüz için daha gerçekçi sonuçlar vereceği kesindir. Ayrıca veri 989 vakadan oluşmaktadır, esasen bu sayı analiz için yeterlidir. Ancak söz konusu çalışmanın daha büyük bir veri seti ile yapılması halinde daha büyük performans değerlerinin yakalanabileceği ve dolayısıyla teşhise daha büyük bir katkı sağlanacağı düşünülmektedir.

Stresin günümüzde sağlığı olumsuz etkilediği kanıtlanmıştır. Çalışmada kullanılan değişkenlerin tümü daha önce ayrıntılı açıklandığı gibi prostat kanseri ile ilişkilidir. Fakat fiziksel değişkenler arasında stresi ölçen veya kişinin duygu durumunu ifade eden psikolojik değişkenlerin de yer alması halinde çalışmanın isabeti artabilir. Sonuç olarak bu gibi çalışmaların fiziksel ve genetik değişkenler yanında psikolojik değişkenleri de içeren veri kümeleri ile yapılması önerilebilir.

KAYNAKÇA

Akamatsu S, Takahashi A, Takata R, v.d.: 2012	"Reproducibility, performance, and clinical utility of a genetic risk prediction model for prostate cancer in Japanese", **PloS one**,7 (10), e46454.
Akman, M.: 2010	"Veri Madenciliğine Genel Bakış ve Random Forests Yönteminin İncelenmesi: Sağlık Alanında Bir Uygulama", **Yayımlanmamış Doktora Tezi,** Ankara Üniversitesi, Ankara.
Akpınar, H**.:** 2000	"Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği". **İ.Ü. İşletme Fakültesi Dergisi,** 29, 1-22.
Alapont,J.,Bella Sanjuán,A., Ferri,C., Hernández-Orallo, J., Llopis-Llopis, ve M. J.D., Ramírez-Quintana,J. :2005	"Specialised Tools for Automating Data Mining for Hospital Management", (Çevrimiçi) http://www.dsic.upv.es/~abella/papers/HIS_DM.pdf
Alpaydın E. :2000	"Zeki veri madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri", **Bilişim 2000 Eğitim Semineri Bildiriler** s.5
Aly M, Wiklund F, Xu J, v.d.: 2011	"Polygenic risk score improves prostate cancer risk prediction: results from the Stockholm-1 cohort study", **Eur Urol**,60 (1),21–28.
American Cancer Society.:2012	**Cancer Facts and Figures .**
Anonymous. 1999	"Texas Medicaid Fraud and Abuse Detection System recovers $2.2 million, wins national award", **Health Management Technology**, vol. 20, no. 10
Anunciacao O V.D.: 2010	"A Data Mining Approach for the Detection of High-Risk Breast Cancer Groups". In: Rocha, M.P, v.d. Editors", **Advances in Bioinformatics**, Berlin Heidelberg: Springer, pp. 43–51.
Atalay,M., Çelik,E.: 2017	"Büyük Veri Analizinde Yapay Zekâ ve MakineÖğrenmesi Uygulamalari - Artificial Intelligence and Machine Learning Applications in Big Data Analysis," **Mehmet Akif Ersoy Üniversitesi Sos. Bilim. Enstitüsü Derg**., pp. 155–172.
Andriole GL, Bostwick DG, Brawley OW, v.d.: 2010	"Effect of dutasteride on the risk of prostate cancer", **N Engl J Med**, 362 (13),1192–1202.

Bae, S.J.M., Li, Z.M., Shin, M. H., Kim, D. H., Lee M.S. v.d.: 2015	"Cigarette Smoking and Prostate Cancer Risk: Negative Results of the Seoul Male Cancer Cohort Study",**Asian Pac J Cancer Prev**, 14 (8), 4667-4669.
Baglietto, L., Severi, G., English, D.R. v.d.: 2006	"Alcohol consumption and prostate cancer risk: results from the Melbourne Collaborative Cohort Study". **Int J Cancer**, 119 (6),1501–1504.
Bansal A, Murray DK, Wu JT, v.d.: 2000	"Heritability of prostate-specific antigen and relationship with zonal prostate volumes in aging twins", **J Clin Endocrinol Metab**,85 (3),1272–1276.
Bensen JT, Xu Z, Smith GJ, v.d.: 2013	"Genetic polymorphism and prostate cancer aggressiveness: a case-only study of 1,536 GWAS and candidate SNPs in African-Americans and European-Americans", **The Prostate**,73 (1),11–22.
Berson,A, Smith,S., Thearling,K.: 1999	**Buildind Data Mining Applications for CRM**, Mcgraw Hill, 510, USA, 1999
Bhatia, N. ve Vandana.: 2010	"Survey of nearest neighbor techniques" **International Journal of Computer Science and Information Security**, 8 (2):302-305.
Bilgin,D., Çamurcu,Y.,Şentürk, A.: (2006)	**Veri Madenciliği Kavram ve Teknikler**, Ekin Kitabevi, Bursa.
Bilgin, T., Çamurcu,Y.: 2008	"Çok Boyutlu Veri Görselleştirme Teknikleri," **Akad. Bilişim***, Çanakkale Onsekiz Mart Üniversitesi*, pp. 107–11.
Bostwick, D.G., Burke, H.B., Djakiew, D., Euling, S., Ho, S.M., v.d.: 2004	"Human prostate cancer risk factors ",**Cancer** (ÇEVRİMİÇİ) , https://www.ncbi.nlm.nih.gov/pubmed/15495199101 (10), 2371-490
Boyle ,P., Ferlay, J.: 2004	"Cancer incidence and mortality in Europe", **Annals of Oncology**, 16,481-488.
Breiman.L.:1996	"Bagging predictors", **Machine Learning,** 24 (2),123-140.
Brownlee, J.: 2018	"What is a Confusion Matrix in Machine Learning" (Çevrimiçi) https://machinelearningmastery.com/confusion-matrix-machine-learning/ , 01.09.2018
Calle, E.E., Rodriguez, C., Walker-Thurmond K.v.d.: 2003	"Overweight, obesity and mortality from cancer in a prospectively studied cohory of U.S adults" **N Engl J Med** ,348,1625-1638.
Cansino, A.J.R., Martinez-Pineiro, L.: 2006	"Molecular biology in prostate cancer",**Clinical and Translational Oncology**, 8, 148–52.

Carlis, J.V. ; Konstan, J.A.;1998	"Interactive Visualization of Serial Periodic Data", **In UIST'98 Conference Proceedings,** New York, NY: ACM Press, USA 29-38.
Carter BS, Beaty TH, Steinberg GD, v.d.: 1992	"Mendelian inheritance of familial prostate cancer", **Proc Natl Acad Sci USA** ,89 (8),3367–3371.
Carter, B.S, Bova, G.S., Beaty, T.H., Steinberg, G.D., Childs, B. v.d.: 1993	"Hereditary prostate cancer: epidemiologic and clinical features", **Journal of Urology**, 150:797.
Chan, J.M., Stampfer, M.J., Giovannucci, E., v.d al.: 1998a	"Plasma insulin-like growth factor-I and prostate cancer risk: a prospective study", **Science,** 279, 563–566.
Chan, J.M., Giovannucci, E., Andersson, S.O., Yuen, J., Adami, H.O.: 1998b	"Dairy products, calcium, phosphorous, vitamin D, and risk of prostate cancer (Sweden)",**Cancer Causes Control**,9 (6),559-566.
Chen, P., Zhang, W., Wang, X.,Zhao,K., Negi, D.S. v.d.: 2015	"Lycopene and Risk of Prostate Cancer A Systematic Review and Meta-Analysis",**Medicine**, 94 (33), e1260.
Christy,T.: 1997	"Analytical tools help health firms fight fraud", **Insurance & Technology,** vol. 22, no. 3, pp. 22-26.
Cingiz, M.Ö., Albayrak, A. ve Amasyalı, F.:2013	"Sınıflandırıcı Topluluklarının Gürültülü Verilere Karşı Gürbüzlüğünün Değerlendirilmesi, Evaluatıon Of Robustness Of Ensemble Learners To Noisy Data" **IEEE**.
Clinton,S.K.,Giovannucci , E.: 1998	"Diet, nutrition, and prostate cancer" **,Annual review of nutrition.**
Dakins,D.R.,: 2001	"Center takes data tracking to heart", **Health Data Management**, vol. 9, no. 1, pp. 32-36.
Dasu,T., Johnson,T.: 2003	**Exploratory DataMining and Data Cleaning,** JohnWiley & Sons, 2003.
De Kok JB, Verhaegh GW, Roelofs RW, v.d.: 2002	"DD3 (PCA3), a very sensitive and specific marker to detect prostate tumors",**Cancer Res**,62 (9),2695–2698.
Değirmenci ,M.: 2010	"Prostat Kanseri Hücre Hattında Oktreotid ve Gossypol (At-101) Uygulamasının Sitotoksik ve Apoptotik Etkileri",**Yayımlanmamış Uzmanlık Tezi**, Ege Üniversitesi, İzmir.

Dennis, L.K., Snetselaar, L.G., Smith, B.J., Stewart, R.E, Robbins, M.E.: 2004	"Problems with the assessment of dietary fat in prostate cancer studies",**Am J Epidemiol**,160,436-444.
Dennis, L.K.: 2000	"Meta-analysis for combining relative risks of alcohol consumption and prostate cancer", **Prostate** ,42 (1),56-66.
Deshpande, B., Kotu, V.:2014	**Predictive Analytics and Data Mining.**
Dietterich, T.G.:2000	"An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization", **Machine Learning**, 40, 139-158.
Dietterich,T.: 1997	"Machine-learning research: four current directions",**AI Magazine** 18 (4):97–136.
Doad, P.K., Bartere, M.M.: 2013	"A Review : Study of Various Clustering Techniques", **International Journal of Engineering Research & Technology,** 2 (11):3141-3145.
Doumpos,M., Zopounidis,C.: 2002	"Multi-Criteria Decision Aid Classification Methods, Holland",**Academic Publishers.**
Duda, R.O., Hart, P.E., Stork, D.G.:2000	**Pattern Classification**, New Jersey,John Wiley & Sons.
Dunham, M. H.: 2003	**Data Mining Introductory and Advanced Topics,** New Jersey: Prentice Hall.
Dzeroski, S. ve Zenko, B.: 2004	"Is Combining Classifiers with Stacking Better than Selecting the Best One?", **Kluver Academic Publishers**, 54,255-273.
Easton D.F., Eeles, R.A.: 2008	"Genome-wide association studies in cancer" **Hum. Mol. Genet ,**17 (R2): R109–R115. doi: 10.1093/hmg/ddn287
Eeles RA, Olama AA, Benlloch S, v.d..: 2013	"Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array", **Nat Genet**,45 (4),385–391.
Eeles R, Goh C, Castro E, v.d.: 2014	"The genetic epidemiology of prostate cancer and its clinical implications", **Nature reviews Urology**,11 (1),18–31.
Eeles, R., Durocher, F., Edwards, Teare,D., E aston, D. v.d.: 1997	"Does the hereditary prostate cancer gene, HPC1 contribute to a large proportion of familial prostate cancer results from the CRC/BPG UK", **Texan & Canadian Consortium**? **Am J Hum Genet** ,61, a64

Ekin, R.G., Zorlu, F.: 2013 "Türkiye verilerine göre prostat kanseri taranmalı mı ?", **Üroonkoloji e-Bülten**, 2,71-75.

Elshazly H.I., Elkorany A. M, Hassanien A.E.: 2013 Ensemble-based classifiers for prostate cancer diagnoss. **IEEE,** 978-1-4799-3370-9, (çevrimiçi) https://ieeexplore.ieee.org/document/6736475

EUPEDİA. : 2018 "Genes and Mutations Associated with Cancer", (Çevrimiçi), https://www.eupedia.com/genetics/cancer_related_snp.shtml , 08.05.2018

Faloutsos, C.: 1996 "Searching Multimedia Databases by Content", **Kluwer Academic Publishers**, Boston, MA.

Fenokulu.: 2015 (Çevrimiçi) http://www.fenokulu.net/portal/Sayfa.php?Git=KonuKategorileri&Sayfa=KonuBaslikListesi&baslikid=143&KonuID=623, 08.05.2018

Gams, M., Bohanec, M.,Cestnik, B.: 1994 "Aschema for using multiple knowledge",**Computational Learning Theory and Natural Learning Systems**, 2.,157–170 Cambridge, Massachusetts: MIT Press.

Ganesh, S.: 2002 "Data Mining: Should it be included in the 'Statistics' cirriculum?", **The Sixt International Conference on Teaching Statistics**, Cape Town, South Africa, 7–12 July.

Ge, G., Wong,G.W.: 2008 "Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles," **BMC Bioinformatics,** 9 (1), 275.

GenomTürkiye.: 2018 (Çevrimiçi) http://www.genomturkiye.com/genetik-hakkinda-bilgi.html, 08.05.2018

Giovannucci, E., Ascherio, A., Rimm, E.B, Colditz,G.A, Stampfer, M.J.: 1995 "Physical activity, obesity, and risk for colon cancer and adenoma in men", **Ann Intern Med**,122 (5),327-34.

Gong ,Z., Neuhouser, M.L, Goodman, P.J v.d., 2006 "Obesity, diabetes, and risk of prostate cancer: results from the prostate cancer prevention trial",**Cancer Epidemiol Biomarkers Prev**, 15,1977–1983.

Guy M, Kote-Jarai Z, Giles GG, v.d..: 2009 "Identification of new genetic risk factors for prostate cancer", **Asian journal of andrology**,11 (1),49–55.

Gülkesen, K.H., Köksal, I.T., Bilge, U., Saka, O.: 2010 "Comparison of methods for prediction of prostatecancer in Turkish men with PSA levels of 0-10 ng/mL", **J Buon,**15 (3): 537-42.

Hallick J.N.: 2001 "Analytics and the data warehouse", **Health Management Technology**, vol. 22, no. 6, pp. 24-25.

Haenszel, W. ve Kurihara, M.:1968 "Studies of Japanese migrants. Mortality from cancer and other diseases among Japanese in the United Steates", **J Natl Cancer Inst**, 40,43-68.

Hand, D.J**.:** 1998 "Data Mining: Statistics and More?", **The American Statistician,** 52,112-118.

Helfand, B. T., Catalona, W. J., Xu, J.: 2015 "A Genetic- Based Approach to Personalized Prostate Cancer Screening and Treatment", **Current Opinion in Urology**, 25 (1), 53-58.

Hickey, K., Do, K.A., Green, A.: 2001 "Smoking and prostate cancer",**Epidemiol Rev,** 23, 115–125

Hsu FC, Sun J, Wiklund F, v.d..: 2009 "A novel prostate cancer susceptibility locus at 19q13", **Cancer Res**,69 (7), 2720–2723.

Hall, M.A.,Holmes, G.:2003 "Benchmarking attribute selection techniques for discrete class data mining", **IEEE Transactions on Knowledge and Data Engineering,** 15 (6) 1437-1447.

Hofmann, M., Klinkenberg, R.: 2014 **Rapidminer Data Mining Use Cases and Business Analytics Applications**, Data Mining and Knowledge Discovery Series

Huncharek,M. Haddock,K. S., Reid,R., Kupelnick,B.: 2010 "Smoking as a Risk Factor for Prostate Cancer: A Meta-Analysis of 24 Prospective Cohort Studies", **American Journal of Public Health,** 100 (4).

Ienco, D., Pensa, G., Meo, R.: 2012 "From context to distance: learning dissimilarity for categorical data clustering", **ACM Transactions on Knowledge Discovery,**6 (1): 1-27.

Ishak MB, Giri VN. A:. 2011 "Systematic review of replication studies of prostate cancer susceptibility genetic variants in high-risk men originally identified from genome-wide association studies. Cancer epidemiology, biomarkers & prevention", **A publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology**,20 (8),1599–1610.

Islami, F., Moreira, D.M., Boffetta, P., Freedland, S.J.: 2014 "A systematic review and meta-analysis of tobacco use and prostate cancer mortality and incidence in prospective cohort studies", **Eur Urol**, 66, 1054–64.

Jemal, A., Murray, T., Ward, E. v.d.:2005 "Cancer statistics".,**CA Cancer J Clin,** 55 (1),10-30.

Jiang,X.,Barmada,M.M., Visweswaran,S.: 2010 "Identifying Genetic Interactions in Genome-Wide Data Using Bayesian Networks", **Genet Epidemiol**,34 (6): 575–581.

Hardy, J. ve Singleton, A.: 2009 "Genomewide Association Studies and Human Disease". **N Engl J Med,** 360:1759- 1768.

Kader, A.K, Sun, J., Reck, B.H., v.d.: 2012 "Potential impact of adding genetic markers to clinical parameters in predicting prostate biopsy outcomes in men following an initial negative biopsy: findings from the REDUCE trial", **Eur Urol**,62 (6),953–961.

Kantardzic, M.: 2011 **Data Mining: Concepts, Models, Methods, and Algorithms:** Second Edition.

Kavzoğlu, T.,Çölkesen,İ.: 2010 "Karar Ağaçları İle Uydu Görüntülerinin Sınıflandırılması: Kocaeli Örneği Classification of Satellite Images Using Decision Trees: Kocaeli Case**", Electronic Journal of MaPTechnologies Harita Teknolojileri Elektronik Dergisi**, 2 (21), 36–4536

Kaya, M. O., Çolak, C., Özdemir. E.: 2013 "Araştırma Makalesi Prostat Spesifik Antijeni Yardımı ile Prostat Kanserinin Değişik Yapay Sinir Ağı Modelleri ile Tahmini", **E.Üniversitesi, İ., Fakültesi, T., & Bilişimi,** T. 19–22.

Koh, K.A., Sesso,H.D., Paffenbarger Jr,R.S., Lee, I.M.: 2006 "Dairy products, calcium and prostate cancer risk",**British Journal of Cancer**, 95, 1582 – 1585.

Koh, H.C,Tan,G.: 2005 "Data Mining Application in Healthcare", **Journal of Healthcare Information Management**, vol. 19, no. 2

Kolar,H.R.: 2001 "Caring for healthcare", **Health Management Technology**, vol. 22, no. 4, pp. 46-47.

Kote-Jarai, Z., Easton, DF., Stanford, JL.,v.d.: 2008. "Multiple novel prostate cancer predisposition loci confirmed by an international study. the PRACTICAL Consortium. Cancer epidemiology, biomarkers & prevention", **American Association for Cancer Research American Society of Preventive Oncology**,17 (8):2052–2061.

Koyuncugil, A. S., Özgülbaş, N.: 2009 "Veri Madencili ğ i : T ı p ve Sa ğ l ı k Hizmetlerinde Kullanımı ve Uygulamaları Data Mining : Using and Applications in Medicine and Healthcare", **Bilişim Teknolojileri Dergisi**, *2* (2), 21–32. http://doi.org/10.17671/btd.01471

Kresse, W., Danko, D.M.:2012 **Springer Handbook of Geographic Information,** Berlin: Springer-Verlag.

Kuonen, D.: 2004 "Data Mining and Statistics: What is the Connection?", **The Data Administration Newsletter**, http://www.tdan.com/view-articles/5226/ (Erişim tarihi, 1.10.2018).

Kushi, L.H, Byers T, "American Cancer Society guidelines on nutrition and

Doyle, C. v.d.: 2006 — physical activity for cancer prevention: reducing the risk of cancer with healthy food choices and physical activity", **CA Cancer J Clin**, 56, 254–281.

Landis, S.H., Murray, T., Bolden, S., Wingo, PA.: 1999 — "Cancer Statistics", **CA Cancer J Clin**, 49,8-31.

Lange EM, Salinas CA, Zuhlke KA, v.d..: 2012 — "Early onset prostate cancer has a significant genetic component", **The Prostate**,72 (2),147–156.

Laura J., Marc J.v.d. :2002 — "Gene expression profiling predicts clinical outcome of breast cancer. **Nature**, 415 (6871),530-536, ISSN 0028-0836. URL http: //dx.doi.org/10.1038/415530a

Lee, I.M., Sesso, H.D., Paffenbarger, R.S. Jr.:2001 — "A prospective cohort study of physical activity and body size in relation to prostate cancer risk"**,Cancer Causes Control**, 12,187–193.

Lehrer, S., Diamond, E.J., Stagger, S., Stone, N.N., Stock ,R.G.:2002 — "Increased serum insulin associated with increased risk of prostate cancer recurrence", **Prostate**, 50, 1–3.

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

Liu, Y., Hu, F., Li, D., Wang, F., Zhu, L. v.d.: 2011 — "Does physical activity reduce the risk of prostate cancer? A systematic review and meta-analysis", **Eur Urol**, 60,1029-1044.

Loeb S, Catalona WJ.:2014 — "The Prostate Health Index: a new test for the detection of prostate cancer", **Therapeutic advances in urology**,6 (2),74–77.

Lose F, Batra J, O'Mara T, v.d.: 2011 — "Common variation in Kallikrein genes KLK5, KLK6, KLK12, and KLK13 and risk of prostate cancer and tumor aggressiveness," **Urol Oncol**.

Lunenfeld, B.: 2002 — "The ageing male: demographics and challenges", **World J Urol** ,20, 11-16.

Ly, D., Reddy, C.A., Klein, E.A., Ciezki, J.P.: 2010 — "Association of body mass index with prostat cancer biochemical failure", **J Urol**,183 (6),2193-2196.

Mary,O., Mat,K.: 2004 — "Application of Data Mining Techniques to Healthcare Data", **Infection Control and Hospital Epidemiology.**

MacLean, C.H., Newberry, S.J., Mojica, W.A et.al.: 2006 — Effects of omega-3 fatty acids on cancer risk: a systematic review,**JAMA**, 295,403-415.

Marc J. Laura J. v.d.: 2002 — "A gene-expression signature as a predictor of survival in breast cancer",**New England Journal of Medicine**,347 (25),1999-200.

McCaughan, Y.S.: 2012 "Potential for prostate cancer prevention through physical activity", **World J Urol**, 30,167-179.

Michael,H. v.d.: 2010 "Smoking as a Risk Factor for Prostate Cancer: A Meta-Analysis of 24 Prospective Cohort Studies." **American Journal of Public Health ,**100 (4).

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T.: 2006 "YALE: Rapid prototyping for complex Data Mining tasks", **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,** Vol. 2006 935–940.

Milley,A.: 2000 "Healthcare and data mining", **Health Management Technology**, vol. 21, no. 8, pp. 44-47.

Moss, L.T., Atre,S.: 2003 **Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications**, Addison-Wesley Publishing, 576, USA.

Nahleh, Z.A.: 2006 "Hormonal therapy for male breast cancer: A different approach for a different disease", **CancerTreatment Reviews,** 32,101-105.

Nam RK, Zhang WW, Klotz LH, v.d.: 2006 Variants of the hK2 protein gene (KLK2) are associated with serum hK2 levels and predict the presence of

Oğuzlar, A.: 2003 "Veri Önişleme", **Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi**, Sayı 21 s. 67–76.

Oza, N. C,. Tumer, K.: 2008 "Classifier Ensembles: Select Real-World Applications", **Information Fusion** ,9 (1), 420.

Öğüdücü, Ş.G.: 2018 **Temel Sınıflandırma Yöntemleri** (Çevrimiçi) http://web.itu.edu.tr/~sgunduz/courses/verimaden/slides/d3.pdf, 01.09.2018

Pal M., Mather P.M.: 2003 "An assessment of the effectiveness of decision tree methods for land cover classification", **Remote Sensing of Environment,** 86, 554-565 .

Park, S.Y., Murphy, S.P., Wilkens, L.R v.d.: 2007 "Fat and meat intake and prostate cancer risk: the multiethnic cohort study", **Int J Cancer**, 121,1339-1345.

Paul, G., Watson,H.J.: 1998 **Decision Support in the Data Warehouse**, New Jersey, Prentice Hall PTR, s.4.

Piatetsky-Shapiro,G., Fawley,W.J.: 1991 **Knowledge Discovery in Databases,** AAAI/MIT Pres.

Pilia G, Chen WM, Scuteri A, v.d.: 2006 "Heritability of cardiovascular and personality traits in 6,148 Sardinians", **PLoS Genet**, 2 (8),e132.

Ping Chen, MD, v.d.: 2015 — "Lycopene and Risk of Prostate Cancer A Systematic Review and Meta-Analysis", **Medicine,** 94 (33).

Platz, E.A., Rimm, E.B., Willett, W.C, Kantoff, P.W., Giovannucci, E.: 2000 — "Racial variation in prostate cancer incidence and in hormonal system markers among male health professionals", **J Natl Cancer Inst**, 92,2009-2017.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Sham, P. C.: 2007 — "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses", **The American Journal of Human Genetics,** *81* (3), 559–575.

Quinlan, J.R.: 1993 — **C4.5: Programs for Machine Learning**,San Mateo, CA:Morgan Kaufmann.

Quinones, L.A., Irrazabal, C.E., Rojas, C.R. v.d.: 2006 — "Joint effect among p53, CYP1A1, GSTM1 polymorphism combinations and smoking on prostat cancer risk: an exploratory genotype-enviroment interaction study",**Asian J Androl,**8 (3), 349-355.

Reddy M.V., Wang,H., Liu, S., Bode, B., Reed, J.C.: 2011 — "Association between type 1 diabetes and GWAS SNPs in the southeast US Caucasian population".

Reiter, R.E., ve Dekernion, J.B.: 2005 — **Epidemiology of prostate cancer, etiology and prevention,** Campbell's Urology, Ankara.

Ren S, Xu J, Zhou T, v.d.: 2013 — "Plateau effect of prostate cancer risk-associated SNPs in discriminating prostate biopsy outcomes", **The Prostate**,73 (16),1824–1835.

Ridinger,M.: 2002 — "American Healthways uses SAS to improve patient care", **DM Review**, vol. 12, no.139.

Roberts, C.K, Vasiri, N.D, Barnard, R.J, 2002 — "Effect of diet and exercise intervention on blood pressure, insulin, oxidative stress, and nitric oxide availability",**Circulation**,106,2530–2532.

Ryten,M., Trabzuni, D., Hardy, J.:2009 — "Genotypic analysis of gene expression in the dissection of the aetiology of complex neurological and psychiatric diseases". **Oxford Journals Life Sciences Briefings in Functional Genomics,** 8 (3): 194-198.

Savaş, S.,Topaloğlu,N., Yılmaz,M.: 2012 — "Veri madenciliği ve Türkiye'deki uygulama örnekleri," **İstanbul Ticaret Üniversitesi Fen Bilim. Dergisi.**, 21, 1–23.

Schwartz, GG. v.d,: 1995 — "1,25- Dihydroxy-16-ene-23-yne-vitamin D3 and prostate cancer cell proliferation in vivo", **Urology** 46, 365 – 369

Shmueli, G., Patel, N.R., Bruce, P.C.:2010 — **Data mining: for Business Intelligence,** New Jersey,John Wiley & Sons**.**

Schoonen, W.M., Salinas, C.A., Kiemeney, L.A. v.d.: 2005	"Alcohol consumption and risk of prostate cancer in middle-aged men",**Int J Cancer**,113 (1),133–140.
Schuerenberg, B.K.: 2003	"An information excavation", **Health Data Management**, vol. 11, no. 6, pp. 80-82.
Seni, G., Elder, J.F.: 2010	**Ensemble Methods in DataMining: Improving Accuracy Through Combining Predictions,** Chicago: University of Illinois.
Sesso, H.D., Paffenbarger, R.S. Jr., Lee, I.M.: 2001	"Alcohol consumption and risk of prostate cancer: the Harvard Alumni Health Study",**Int J Epidemiol,** 30 (4),749–755.
Siegel, R., Naishadham, D., Jemal, A.: 2013	Cancer statistic, **CA Cancer Journal**, 2013, 63,11-30.
Nedir: 2016	(Çevrimiçi) https://www.nedir.com/snp, 08.05.2018
SNPEDIA. :2018	"About SNPedia", (Çevrimiçi) https://www.snpedia.com/index.php/Testing**,** 08.05.2018
SNPedia.: 2018	(Çevrimiçi) https://www.snpedia.com/index.php/Prostate_cancer, 01.09.2018.
Sutcliffe, S., Giovannucci, E., Leitzmann, M.F. v.d.:2007	"A prospective cohort study of red wine consumption and risk of prostate cancer", **Int J Cancer**, 120 (7),1529–1535.
Şimşek,U.T.: 2006	"Veri Madenciliği ve Müşteri İlişkileri Yönetiminde (CRM) Bir Uygulama" **Yayınlanmamış Doktora Tezi**, İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü.
Tan, P.N., Steinbach, M., Kumar, V.: 2014	**Introduction to data mining**, London:Pearson Education Limited.
Taşcı, E., Onan, A.:2016	"K- En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi", **XVIII.Akademi Bilişim Konferansi**.
Tomar,D., Agarwal,S.:2013	"A survey on data mining approaches for healthcare", **Int. J. Bio-Science Bio-Technology**, vol. 5, no. 5, pp. 241–266, 2013.
Tumer, K., ve Ghosh, J.:1996	"Error correlation and error reduction in ensemble classifiers. Connection Science, Special Issue on Combining Artificial Neural Networks", **Ensemble Approaches**, 8 (3) 4,385-404.

Tüzüntürk,S.: 2010	"Veri Madenciliği ve İstatistik ," **Uludağ Üniversitesi İktisadi ve İdari Bilim. Fakültesi Dergisi**, vol. 1, no. 2001, pp. 65–90, 2010.
Velicer, C.M., Kristal, A., White, E.: 2006	"Alcohol use and the risk of prostate cancer: results from the VITAL Cohort Study", **Nutr Cancer**,56 (1),50–56.
Vignal,A. v.d.: 2002	"A review on SNP and other types of molecular markers and their use in animal genetics", **Genet. Sel. Evol**, 34: 275-305.
Wright,M.E., Chang,S.C., Schatzkin, A. v.d.: 2007 684.	"Prospective study of adiposity and weight change in relation to prostate cancer incidence and mortality", **Cancer** ,109,675–
Woolf, C.: 1960	"An investigation of the familial aspects of carcinoma of the prostate Cancer", 13, 739-744.
Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D.: 2008	**Top 10 algorithms in data mining**. Knowledge and Information Systems ,vol. 14.
Xu, G., Zong, Y., Yang, Z.: 2013	**Applied Data Mining,** New York: CRC Press
Xu Z, Bensen JT, Smith GJ, v.d.: 2011	"GWAS SNP Replication among African American and European American men in the North Carolina-Louisiana prostate cancer project (PCaP)" ,**The Prostate**, 71 (8),881–891.
Xu J, Sun J, Kader AK, v.d.: 2009	"Estimation of absolute risk for prostate cancer using genetic markers and family history", **Prostate**,69 (14),1565–1572.
Yücebaş, S. C., Son, Y.A.: 2014	" A Prostate Cancer Model Build by a Novel SVM-ID3 Hybrid Feature Selection Method Using Both Genotyping and Phenotype Data from dbGaP", *9* (3), 1–8.
Yücebaş, S.C.: 2016	"Prostat Kanseri Teşhisinde Veri Madenciliği Yöntemlerinin Başırım Karşılaştırması", **Çanakkale 18 mart Üniversitesi**
Zeegers, Jellema, A., Ostrer, H.:2003	"Empiric risk of prostate carcinoma for relatives of patients with prostate carcinoma: a metaanalysis", **Cancer**,97,1894–1903.
Zhao C.M., ve Luan, J. :2006	"Data Mining: Going Beyond Traditional Statistics", ***New Directions for Institutional Research,*** No. 131, 7–16.

Zheng S.L, Sun J, Wiklund F. v.d.: 2008 "Cumulative association of five genetic variants with prostate cancer", **N Engl J Med**,358 (9),910–919.

Zhou, N. ve Wang, L.: 2007 "Effective selection of informative SNPs and classification on the HapMap genotype data", **BMC Bioinformatics**, 8,484.

Zhi-Hua, Z.: 2003 "Three Perspectives of Data Mining", **Artificial Intelligence**, 143:139-146 .

EKLER

Ek.1

İndirgeme Yöntemi	DEĞİŞKEN SAYISI	Doğruluk (ACCURACY)	Kesinlik (PRECISION)	Duyarlılık (RECALL)	AUC (Optimistik)	AUC	AUC (Pessimistik)
BİLGİ KAZANIMINA GÖRE AĞIRLIKLANDIRMA (WEIGHT BY INFORMATION GAIN)	100	%76,75	%75,63	%73,99	0,863	0,689	0,657
	200	%80,19	%79,44	%77,72	0,910	0,736	0,699
	250	%79,58	%77,86	%78,83	0,907	0,726	0,696
	300	%79,48	%77,68	%78,83	0,897	0,712	0,682
	400	%81,60	%82,66	%73,63	0,879	0,762	0,718
	500	%81,09	%80,19	%78,81	0,887	0,736	0,692
BİLGİ KAZANIMI ORANINA GÖRE AĞIRLIKL-ANDIRMA (WEIGHT BY INFORMATION GAIN RATIO)	100	%74,63	%81,28	%59,14	0,941	0,583	0,523
	200	%75,74	%79,83	%63,54	0,933	0,633	0,567
	250	%78,37	%81,36	%69,41	0,920	0,741	0,644
	300	%76,95	%80,78	%66,16	0,943	0,704	0,642
	400	%84,13	%89,84	%74,23	0,914	0,828	0,756
	500	%78,67	%79,14	%72,92	0,916	0,751	0,679
Principal Component Analysis (Temel Bileşenler Analizi)		%48	%44	%51			
*Varyans eşiği değeri: kümülatif varyansı eşik değerden büyük olan bütün bileşenler veri setinden atılır.	805 * (varyans eşiği= 0,95)	%52,07	%39,71	%8,51	0,905	0,500	0,076
	294 (varyans eşiği =0,60	%55,41	%54,94	%21,82	0,879	0,500	0,183
	146 (varyans eşiği= 0,4)	%57,43	%56,60	%32,75	0,857	0,500	0,258

EK.2

Aşağıdaki tablolarda değişken çiftleri arasındaki ilişkilerden örnekler gösterilmiştir. İndirgenen Verinin Covariance Matrisine baktığımızda; Değişkenlerin aralarında bir korelasyon gözükmemektedir. Bu yüzden indirgenen veri güvenilirdir.

First Att...	Second ...	Covaria...
ethni = L	rs67018...	-0.015
ethni = L	rs17376...	-0.007
ethni = L	rs877518	-0.009
ethni = L	rs64241...	0.009
ethni = L	rs66806...	0.011
ethni = L	rs66770...	0.019
ethni = L	rs27455...	0.008
ethni = L	rs16826...	-0.005
ethni = L	rs12127...	-0.016
ethni = L	rs46464...	0.008
ethni = L	rs46464...	0.008
ethni = L	rs13383...	-0.007
ethni = L	rs855821	0.004
ethni = L	rs12085...	-0.002
ethni = L	rs21498...	-0.002
ethni = L	rs28154...	0.002
ethni = L	rs10518...	-0.008
ethni = L	rs12037...	-0.003

First Att...	Second ...	Covaria...
ethni = L	rs12031...	-0.007
ethni = L	rs66720...	0.027
ethni = L	rs11807...	0.001
ethni = L	rs17130...	0.002
ethni = L	rs10494...	0.029
ethni = L	rs13428...	-0.016
ethni = L	rs13905...	-0.001
ethni = L	rs98879...	-0.017
ethni = L	rs12081...	-0.019
ethni = L	rs46576...	0.018
ethni = L	rs5367	0.003
ethni = L	rs5361	0.002
ethni = L	rs39174...	0.003
ethni = L	rs12145...	-0.008
ethni = L	rs12141...	0.003
ethni = L	rs10911...	-0.001
ethni = L	rs16824...	-0.001
ethni = L	rs10159...	-0.011

First Att...	Second ...	Covaria...
ethni = L	rs37302...	0.002
ethni = L	rs10494...	0.002
ethni = L	rs41492...	0.016
ethni = L	rs12080...	0.000
ethni = L	rs66906...	0.001
ethni = L	rs12027...	-0.028
ethni = L	rs44839...	-0.009
ethni = L	rs15602...	0.005
ethni = L	rs22871...	-0.029
ethni = L	rs46706...	0.009
ethni = L	rs67209...	-0.005
ethni = L	rs873171	0.018
ethni = L	rs27067...	-0.009
ethni = L	rs67524...	0.001
ethni = L	rs17006...	0.012
ethni = L	rs17436...	-0.003
ethni = L	rs10117...	0.000
ethni = L	rs12623...	-0.014

First Att...	Second ...	Covaria...
packyrs_...	d_calc_cat	0.016
packyrs_...	currsmoke	0.013
packyrs_...	eversmo...	-0.002
ethanol_...	d_lyco_cat	0.007
ethanol_...	p_fat_cat	0.001
ethanol_...	d_calc_cat	-0.004
ethanol_...	currsmoke	-0.005
ethanol_...	eversmo...	-0.002
d_lyco_cat	p_fat_cat	-0.004
d_lyco_cat	d_calc_cat	0.000
d_lyco_cat	currsmoke	-0.005
d_lyco_cat	eversmo...	-0.003
p_fat_cat	d_calc_cat	-0.005
p_fat_cat	currsmoke	-0.008
p_fat_cat	eversmo...	0.002
d_calc_cat	currsmoke	0.039
d_calc_cat	eversmo...	-0.001
currsmoke	eversmo...	-0.011

ÖZGEÇMİŞ

Özin KALEMCİ 1983 yılında İstanbul'da doğmuştur. İlkokulu Kütahya'da bitirmiştir.

2001 yılında Bornova Anadolu Lisesi'nden mezun olmuştur. 2006 yılında Hacettepe Üniversitesi Fen Fakültesi Matematik Bölümün'ü bitirmiştir. 2008 yılında Kanada'da MBA Masterını tamamlamıştır. İki yıl, Yıldız Teknik Üniversitesi İktisadi ve İdari Bilimler Fakültesi İşletme Bölümün'de araştırma görevlisi olarak çalışmıştır. Dört yıl öğretim görevlisi olarak çalışmıştır.

KALEMCİ, şu anda matematik öğretmenliği yapmaktadır.