TÜRKİYE CUMHURİYETİ
KARADENİZ TEKNİK ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK VE TIP BİLİŞİMİ ANABİLİM DALI

KARSİNOGENEZDE MUTASYONLAR ARASI İLİŞKİLERİN VERİ MADENCİLİĞİ METOTLARI İLE TESPİTİ

Uğur TOPRAK

YÜKSEK LİSANS TEZİ

Prof. Dr. Kemal TURHAN

TRABZON – 2015

TÜRKİYE CUMHURİYETİ
KARADENİZ TEKNİK ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK VE TIP BİLİŞİMİ ANABİLİM DALI

KARSİNOGENEZDE MUTASYONLAR ARASI İLİŞKİLERİN VERİ MADENCİLİĞİ METOTLARI İLE TESPİTİ

Uğur TOPRAK

YÜKSEK LİSANS TEZİ

Prof. Dr. Kemal TURHAN

TRABZON – 2015

ONAY

Bu tez Yüksek Lisans Tezi Standartlarına Uygun Bulunmuştur

Prof. Dr. Kemal TURHAN

Biyoistatistik ve Tıp Bilişimi Anabilim Dalı Başkanı

Karadeniz Teknik Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ve Tıp Bilişimi Anabilim Dalı Yüksek Lisans öğrencisi Uğur TOPRAK'ın hazırladığı "**Karsinogenezde Mutasyonlar Arası İlişkilerin Veri Madenciliği Metotları ile Tespiti"** başlıklı tez KTÜ Lisansüstü Eğitim - Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca, kapsam ve bilimsel kalite yönünden değerlendirilerek oy birliği ile Yüksek Lisans Tezi olarak kabul edilmiştir.

Danışman: Prof. Dr. Kemal TURHAN _____

Yüksek Lisans Sınavı Jüri Üyeleri

Prof. Dr. Kemal TURHAN _____

Prof. Dr. Ersan KALAY _____

Yrd. Doç. Dr. Aybar Can ACAR _____

Tarih: 24/04/2015

Bu tez KTÜ Sağlık Bilimleri Enstitüsü Yönetim Kurulu'nun …./…/…. tarih ve …sayılı kararıyla onaylanmıştır.

Prof. Dr. Ali Osman KILIÇ

Sağlık Bilimleri Enstitüsü Müdürü

BEYAN

Bu tez çalışmasının KTÜ Sağlık Bilimleri Enstitüsü tez yazım kılavuzu standartlarına uygun olarak yazıldığını, tezin akademik ve etik kurallara bağlı kalınarak gerçekleştirilmiş özgün bir bilimsel araştırma eseri olduğunu, tezde yer alan, bu tez çalışmasıyla elde edilmeyen tüm bilgi ve yorumlara kaynak gösterdiğimi, kaynakların kaynaklar listesinde yer aldığını, tezin çalışılması ve yazımı aşamalarında patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

24. 04. 2015

Uğur TOPRAK

TEŞEKKÜR

Yüksek Lisans öğrenimim ve tez çalışmalarım boyunca hiçbir zaman desteğini esirgemeyen, yanında çalışmaktan gurur duyduğum tez danışmanım ve çok değerli hocam Prof. Dr. Kemal TURHAN başta olmak üzere bilgilerinden faydalandığım Asya Yazılım Proje Yöneticisi Serbülent ÜNSAL, Ögr. Gör. Dr. Burçin KURT'a, KTÜ Tıp Fakültesi Biyoistatislik ve Tıp Bilişimi Anabilim Dalı Teknisyeni Serkan KALAYCI'ya, tüm KTÜ Sağlık Bilimleri Enstitüsü personeline, çok sevdiğim değerli arkadaşım Melih KAYA'ya, çalışmalarımda bana yardımcı olan arkadaşım Uzman Biyokimyager Feyza Demir'e ve hayatım boyunca sevgilerini ve desteklerini her an yanında hissettiğim aileme sevgi, saygı ve teşekkürlerimi sunarım.

Uğur TOPRAK

İÇİNDEKİLER

TABLOLAR DİZİNİ

ŞEKİLLER DİZİNİ

RESİMLER DİZİNİ

KISALTMA, SİMGE ve FORMÜLLER

Kısaltmalar

AIC	Akaike Information Criterion
ARA	Arıtılmış Regresyon Ağaçları
BDe	Bayesian-Dirichlet equivalent
BIC	Bayesian Information Criterion
DNA	Deoksiribonükleik asit
RNA	Ribonükleik asit
TCGA	The Cancer Genom Atlas
VBK	Veritabanlarında Bilgi Keşfi
YDG	Yönlendirilmiş Düz Grafik
YSA	Yapay Sinir Ağları

1. ÖZET

Karsinogenezde Mutasyonlar Arası İlişkilerin Veri Madenciliği Metotları ile Tespiti

Veri madenciliği, büyük miktardaki veriler arasından kolaylıkla öngörülemeyen anlamlı örüntülerin çıkarılmasını sağlayan çalışma alanına denilmektedir. Günümüzde yaygın bir şekilde kullanılan veri madenciliği metotlarından biri de birliktelik analizidir. Bu analizin amacı mevcut veri kümesinde birlikte görülen parametreleri tespit ederek, karar verme sürecine faydalı olacak örüntüler ortaya çıkarmaktır. Ayrıca büyük miktarlardaki veri kümelerini budayarak, üzerinde analiz yapması kolay ve anlamlı veri kümeleri elde etmek için kullanılmaktadır. Bayesci ağlar, değişkenleri temsil eden düğümler ve nedensel ilişkileri gösteren oklardan oluşan grafiksel modellerdir. Bayesci ağlar, yönlü döngüsüz grafik oluşturmak ve değişkenlerin koşullu olasılık değerlerini bulmayı esas alır. Bu iki işlem öğrenme olarak tanımlanır. Bayesci ağlarda yapı ve parametre öğrenme olmak üzere iki tür öğrenme yöntemi vardır.

Bu tez çalışmasında The Cancer Genom Atlas (TCGA) veritabanından elde edilen küçük hücreli dışı akciğer kanserine ait genetik mutasyonların evrimsel geçmişi ilk paragrafta belirtilen veri madenciliği teknikleri kullanılarak araştırılmıştır. Öncelikle TCGA veritabanından elde edilen yüksek miktardaki genetik veriden birliktelik kuralı analizi ile en sık görülen mutasyonların birlikte görülme kuralları çıkarılmıştır. Bu kurallar yardımıyla belirlenen en çok birlikte görülen on yedi mutasyonun atasal sıralanması Bayesci ağlarla yapılmıştır. Bayesci ağlar oluşturulurken iki farklı yapı öğrenme algoritması ve bu algoritmaların kullandığı dört farklı skorlama metodu kullanılmıştır. Mevcut veri seti ile oluşturulan Bayesci ağların tepe tırmanma ve tabu arama gibi yapısal öğrenme algoritmalarına göre değil, bu algoritmaların içinde mutasyonların ağdaki en iyi olasılıksal yerini belirleyen skorlama metoduna göre farklılık gösterdiği ortaya konulmuştur. Ayrıca mutasyonların veri setindeki görülme sıklıklarının, hastalığın birikimli yapısı hakkında önemli bilgiler içerdiği görülmüştür.

Anahtar Sözcükler: Bayesci ağlar, veri madenciliği, yapısal öğrenme

2. SUMMARY

Determining Relations Between Mutations in Carcinogenesis With Data Mining Methods

Data mining can be described as an extraction of meaningful patterns from large amounts of data. Recently, one of the widely used data mining metods is association analysis. This analysis, aims to identify parameter pairs which has seen together in data set and uncover helpful patterns that will be useful to desicion-making process. Association analysis also used to prune data sets in large quantities to perfrom this analysis easily and obtain meaningful smaller data sets. Bayesian networks are the graphical models that nodes represents variables and causal relationships shown with arcs. Bayesian networks, bases on creating direcred acyclic graphs and finding conditional probablity values of variables. These two processes are defined as a learning process. There are two learning methods in Bayesian networks which are structure learning and parameter learning.

In this thesis study, data mining techniques used on non-small cell lung adenocarcinoma's genetic mutation data to investigate evolutionary history of mutations which obtained from The Cancer Genom Atlas (TCGA). First, association rules analysis applied on dataset obtained from TCGA which contains large amount genetic of data to extract most common association rules. Seventeen genetic mutations determined with association rules to build ancestral sequence with Bayesian networks. Bayesian networks created with two different structure learning algorithms and four different scoring methods which are used by these algorithms. Bayesian networks created with obtained data showed us these networks differs not according to structure learning algorithms but according to scoring function which shows us best probable location of mutations in network. In addition, prevalence of the mutations in data set gave us important information about the cumulative nature of the disease.

Key Words: Bayesian networks, data mining, structure learning

3. GİRİŞ ve AMAÇ

Günümüzde gelişen ve ilerleyen teknolojinin etkisiyle hemen her alanda yaygın olarak kullanılmaya başlanan bilgisayar sistemleri, dijital veri erişimi ve depolama tekniklerinin de hızlı bir şekilde gelişmesini etkilemiş kayıt altına alınabilir ve saklanabilir veri miktarı oldukça artmıştır. Bu çok sayıda ve türdeki verileri tutmak için veritabanları kullanılmaktadır. Toplanan bu veriler, ileride büyük veri kümelerinden bilgiye ulaşma ihtiyacını gidermek için analiz ve tahminler yapılmak üzere zemin teşkil etmektedir. Bu bağlamda veri madenciliği, veri kümelerinin analiz edilerek önceden bilinmeyen fakat potansiyel olarak anlamlı bilginin ortaya çıkarılması işi olarak tanımlanır (1).

Veri madenciliğini klasik istatistikten ayıran özellik, büyük veri kümeleri içinde gizli kalmış örüntülerin keşfedilmesine imkan tanımasıdır. Klasik istatistikte ise bilinen parametreler arasındaki ilişkiler incelenmektedir. Örneğin, bebek bezi alan müşterilerin %35'nin cips de aldığı ilişkisi kolayca öngörülüp, tespit edilebilecek bir ilişki değildir (2). Bu noktada veri madenciliği uygulamaları ile veriler arasındaki bu tarz ilişkiler ortaya konulabilmektedir (2).

Son yıllarda özellikle veri madenciliği ve bilgi keşfine dair çalışma ve uygulamalar oldukça hızla artmaktadır. Bunların nedenleri arasında veri toplama tekniklerinin gelişmesi, veri ambarlarının yapısı ve erişilebilirliği, veriye erişim yollarının kolaylaşması, şirketler bazında rekabetin artması ve ticari veri madenciliği yazılımlarının geliştirilmesi ve bilgisayarların hesaplama hızındaki artış gösterilebilir (3).

Veri madenciliğinde, veri kümelerinin analizi sürecinde istatistik, makine öğrenmesi, veri tabanı yönetim sistemleri, örüntü tanımlama ve yapay zeka gibi birçok yöntemden faydalanılabilir. Ayrıca bu alanlarda kullanılan algoritmalardan yararlanılabileceği gibi tamamen özgün algoritma ve yöntemlerde geliştirilebilir. Bu tez çalışmasında bahsedilen yöntemlerden iki tanesi üzerinde durulacaktır. Bunlar; Birliktelik kuralı analizi (association rule analysis) ve Bayesci ağlar (Bayesian networks)'dır.

Veri madenciliğinde en sık kullanılan yöntemlerden biri olan birliktelik analizi ilk kez 1993 yılında Agrawal ve arkadaşları tarafından tanıtılmıştır. Bu yöntem "sık gözlenen nesne kümeler (frequent itemsets)" ve "birliktelik kuralları (association rules)" olarak adlandırılan büyük veri kümelerinde ilginç ilişkilerin bulunmasında kullanılmaktadır. Birliktelik analizinin uygulama sürecinde dikkat edilmesi gereken iki nokta bulunmaktadır. Bunlardan ilki, üzerinde çalışılan veri kümelerinin çok büyük olması durumunda veri içerisindeki ilginçlik ya da ilişkileri belirlemenin hesaplama açısından karmaşık ve zaman alıcı bir hal alması; diğeri, veriler arasındaki önemli gibi görülen bazı ilişkilerin rassal olarak ortaya çıkmış olması nedeniyle yanıltıcı sonuçların elde edilmesidir. Bu sebeplerden dolayı algoritma ve veri setleri bu ön koşullar göz önünde bulundurularak seçilir (3,4).

Yakın zamanda, veri kümelerindeki sık gözlenen nesne kümelerin ve birliktelik kurallarının keşfedilmesine dair çalışmalara literatürde oldukça sık karşılaşılmaktadır. Bu çalışmalar, birliktelik analizi ile ortaya çıkarılan örüntülerin sayısı çok fazla olduğunda uygulamaların sonucunda çok karmaşık çıktılar elde edildiğini göstermiştir. Özellikle destek değeri düşük belirlendiğinde büyük veritabanlarından binlerce örüntü çıkarılabilir. Bu durumda mevcut örüntülerin ilginçlik düzeyleri incelenerek önemsiz olanlar elenerek, sadece belirli düzeyin üzerindeki ilginçlikler değerlendirilmektedir (5).

Bu tez çalışmasında mevcut genetik veritabanı içerisinden ilginçlik seviyesi yüksek olan ilişkileri ortaya çıkarmak için birliktelik kuralı analizi kullanılmıştır.

Çalışmada kullanılan bir diğer yöntem ise Bayesci ağlardır. Bayesci ağlar, değişkenlerin üzerindeki olasılıksal dağılımın yönlendirilmiş düz grafikler (directed acyclic graph) vasıtasıyla gösterildiği modellerdir (6). Esasları Bayes teoremine dayanan Bayesci ağlar ilk olarak 1985 yılında Judea Pearl tarafından yapılan çalışmada kullanılmıştır (7). Çok geniş bir uygulama alanı olan Bayesci ağlar biyolojide gen ağı tahmini (8), tıpta akciğer kanseri yayılım süreçleri (9), gen kümelerinin tespiti (10) ve tüpteki yanlış kanın bulunması (11) vb. gibi farklı alanlarda yaygın bir biçimde kullanılmaktadır.

Bayesci ağların temelini bilgisayar bilimleri, grafiksel modeller, olasılık teorisi ve istatistik teorisi alanları oluşturmaktadır. Bayesci ağlar, ortak olasılık dağılımını oluşturma ve koşullu olasılıkların elde edilmesi işlemleri için olasılık teorisinden

faydalanır. Değişkenlerin yerel olasılık dağılımlarından yola çıkarak ortak bir olasılık dağılımı oluşturulur. Değişkenler arasındaki bağımlılık yapılarını görsel olarak ortaya koymak için grafiksel modeller kullanılır. Grafiksel modeller, değişkenler arası ilişkiler ve koşullu bağımsızlık durumlar hakkında çıkarımsamlar yapma imkanı sağlar. İstatistik teorisi ise Bayesci ağların oluşturulmasında son derece önem arz etmektedir. Bayesci ağın içerisinde bulunan değişkenler arası ilişkilerin tespiti aşamasında çeşitli istatistiksel yöntemler kullanılmaktadır (12).

Bayesci ağlar, değişkenler arasındaki ilişkisel belirsizlikleri yok etmek için oldukça kullanışlı bir yöntem olmakla beraber, grafiksel gösterimi sayesinde hem dolaylı hem de nedensel ilişkilerin görsel olarak kolaylıkla keşfedilmesini sağlar. Böylece bir durumun gerçekleşmesi ya da gerçekleşmemesine etki eden nedensel faktörler tespit edilebilir. Bayesci ağlar, değişkenler arası karmaşık ilişkileri görsel olarak ortaya koyarak belirsiz yapıların yorumlanmasını da kolaylaştırmaktadır (12).

Bayesci ağların oluşturulması ve mevcut veriler hakkında olasılıksal çıkarımların yapılabilmesi için öğrenme algoritmaları kullanılmaktadır. Öğrenme algoritmaları yapı öğrenme ve parametre öğrenme olmak üzere iki ayrı başlık altında tanımlanmaktadır. Değişkenler arasındaki bağımlılık ve bağımsızlık ilişkilerinin ortaya çıkarılabilmesi için yapı öğrenme algoritmaları kullanılmaktadır. Parametre öğrenme algoritmaları ise değişkenlerin koşullu olasılık dağılımındaki değerlerinin elde edilmesinde kullanılmaktadır (12).

Yapı öğrenme algoritmaları ise kısıtlama tabanlı, skor tabanlı ve karma algoritmalar olmak üzere üç kısımda incelenmektedir. Kısıtlama tabanlı algoritmalar mevcut gözlenen veri setindeki değişkenler arası koşullu bağımsızlık ilişkilerini (Markov koşulları) inceler. Bu tez çalışmasında ise skor tabanlı algoritmalar kullanılmıştır.

Skor tabanlı algoritmalar, Bayesci ağları öğrenmek için veriyi ne kadar iyi tahmin ettiğini ölçmek üzere Bayesci ağa bir skor ataması yapar. Bu metot skor değeri en yüksek ağın elde edilmesi esasına dayanır. Bu algoritmalar içerisinde en yaygın kullanılan yöntemler hırslı arama (greedy search), genetik algoritmalar ve benzetimli tavlama (simulated annealing)'dır. Tepe tırmanma ve tabu araştırma algoritmaları hırslı arama algoritmalarının en önemlileridir. Bu algoritmaların kullandığı istatistiksel

skorlama metotlarından bazılarına ise Akaike Bilgi Kriteri (Akaike Information Criterion (AIC)), Bayeci Bilgi Kriteri (Bayesian Information Criterion (BIC)), Bayesci-Dirichlet eşitlik (Bayesian-Dirichlet equivalent (BDe)) ve K2 örnek verilebilir (13).

Veri madenciliğinde Bayesci ağlardan yararlanmanın sağladığı birçok üstünlük bulunmaktadır. Öncelikle değişkenler arasındaki nedensel ilişkilerin ortaya çıkarılmasını sağlar. Üzerinde çalışılan problemin olası sonuçlarına ilişkin kestirim dağılımlarının hesaplanması için kullanılır. Olasılık kuramını temel aldığı için genelde tutarlı sonuçlar üretir. Araştırılan problemin önsel dağılımları modellenebilir. Uzman görüşü kullanılarak doğrudan oluşturulabileceği gibi uzman görüşü olmadan da veriler kullanılarak oluşturulabilir. Bu sebeple uygulamasında esneklikler sağlar. Anlaşılması ve yorumlanması kolaydır. Verilerde kayıp gözlemlense dahi çıkarımsamalar yapılabilir (13).

Bu tez çalışmasında TCGA kanser veritabanından elde edilen küçük hücreli dışı akciğer kanserinde görülen genetik mutasyonların geriye dönük atasal ilişkilerinin veri madenciliği metotları ile çıkarılması hedeflenmiştir.

Çalışmada sırasıyla konu ile ilgili bilgi ve kavramlardan bahsedilmiştir. Öncelikle kısaca kanserin mekanizmaları anlatılmıştır. Ardından biyoinformatik alanı ve TCGA veri portalı kabaca incelendikten sonra bilgi ve veri kavramları hakkında genel bilgiler verilmiş, veri tabanlarında bilgi keşfi süreci ve sürecin adımları açıklanmıştır. Bilgi keşfi sürecinin en önemli adımı olan veri madenciliği konusu anlatılmış ve kullanılan yöntemler sunulmuştur. Bununla beraber bilgi keşfi sürecinde uygulanan veri önişleme basamaklarından ve veri madenciliğinin problem alanlarından söz edilmiştir. Çalışmanın temelinde olasılıksal bir tahmin süreci olduğu için, kısaca olasılık ve Bayes teoremi ve Bayesci ağlardan bahsedilmiştir.

Bir sonraki bölümde bu çalışmada kullanılan veri madenciliği yöntemlerinden birliktelik kuralları, Bayesci ağ yapısı ve yapı öğrenme algoritmaları konuları detaylandırılmıştır. Ayrıca yapılan uygulamaların amaç, kapsam, kullanılan algoritma ve programlar gibi ayrıntılar verilmiştir.

Altıncı bölümde bu tez çalışması kapsamında yapılan uygulamaların sonucunda elde edilen bulgular ve değerlendirmeleri sunulmuştur.

Son bölümde ise çalışmadan elde edilen sonuçlar incelenerek katkıları tartışılmıştır. Ayrıca ileride yapılacak benzer çalışmalar için tavsiyeler sunulmuştur.

4. GENEL BİLGİLER

Bu kısımda tezin kapsadığı alanlar hakkında bir literatür özeti sunulacaktır. İçerik olarak öncelikle kanseri daha iyi anlayabilmek için karsinogenez hakkında bilgiler sunulacak. Hücre döngüsü ve kanserde etkili mutasyonlardan bahsedilecek. Veri madenciliği metotları incelenecek. Bayes Teoremi ve olasılık kavramları hakkında bilgiler verilecek. Ardından çalışmada kullanılan verileri skorlamak için kullanılan yöntemler açıklanacak. Son olarak küçük hücreli dışı akciğer kanserinde mutasyonlar arası atasal ilişkiyi çıkarımsamak için Bayes ağlarını oluşturmada kullanılan yapısal öğrenme algoritmaları incelenecektir.

4.1. Karsinogenez

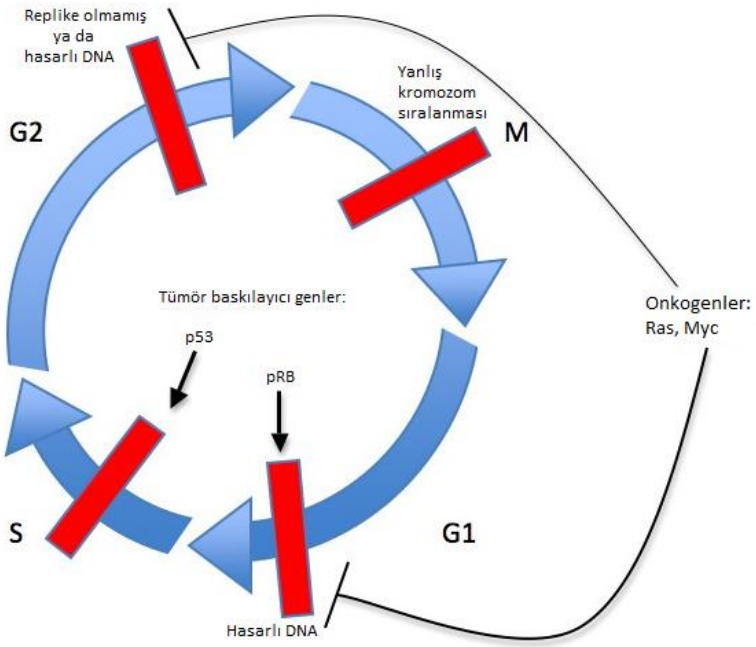
Kanser, birçok etkenin bir araya gelmesiyle hücrelerin kontrolsüz bir biçimde aşırı çoğaldıkları, immün sistemin kontrolünden çıkıp nihai olarak da dokuları istila ederek (metastaz) yapmalarına yol açan, metabolik ve davranışsal değişiklikler geçirdikleri, çok aşamalı bir süreçtir (14,15). Kanseri tanımlamak için tümör hücrelerinde bazı karakteristik özellikler aranır. Bunlar, otonomi (bağımsız çoğalabilme), kontrolsüz çoğalma (kontak inhibisyon kaybı), apoptozisin (kontrollü ölüm mekanizması) baskılanması, anjiyogenez, ölümsüzlük, invazyon ve metastaz kabiliyetidir (16).

Hücrelerin izlediği bu farklılaşmaya genetik materyalinde meydana gelen değişiklikler neden olur. Bu değişiklikler; genomik dayanıksızlık, kromozom kaybı, kromozomların yeniden düzenlenmesi veya insan kromozomlarının lokuslarına yabancı Deoksiribonükleik asit (DNA) dizilerinin girmesi şeklinde özetlenebilir (17,18). Çevresel ve kalıtsal faktörlerin ektisiyle bu değişiklikler ortaya çıkmaktadır.

4.1.1. Hücre Döngüsü

Hücrelerin yaşam süreleri boyunca büyüdüğü ve bölündüğü süreç hücre döngüsü olarak ifade edilir. Normal hücrelerde hücre döngüsü, bazı karmaşık sinyal mekanizmaları ile kontrol edilir. Ayrıca bu süreçte oluşabilecek hataları düzeltmeye yönelik tamir mekanizmaları bulunmaktadır. Eğer bu hatalar düzeltilemez ise hücre kontrollü bir intihara yönlendirilir. Bu işlem ortaya çıkacak etkileri gelecek nesillerden arındırmak için bir korunma mekanizmasıdır. Kanser hücrelerinde ise genetik mutasyonlar sonucu hücre döngüsü kontrolü için gerekli düzenlemeyi yapan sinyal

mekanizmaları görevlerini yerine getiremez hale gelirler ve bunun sonucunda hücre bölünmeleri kontrolsüz bir şekilde gerçekleşmeye başlar (19).



Şekil 1. Hücre döngüsü kontrol mekanizması (G1: Gap 1, S: DNA sentezi, G2: Gap 2, M: Mitoz) *Nature Education'dan adapte edilmiştir (20).

Hücre döngüsü G1 fazının sonunda, S fazında, G2 fazının sonunda ve M fazında olmak üzere farklı noktalarda denetlenmektedir. Bu denetim noktalarının her birinde, ilgili proteinlerce döngünün ilerlemesine veya durmasına karar verilmektedir. Bu kararların verilmesinin kontrolü, hedef proteinleri seçip fosforile eden bir enzim ailesinden olan protein kinazlar ve hücre döngüsünün işlerliğini kontrol eden siklin adlı proteinler tarafından yapılır (17).

Hücre döngüsünün ilk kontrol noktası geç G1 fazında yer almaktadır. Hücrenin G1 fazını terk edip S fazına geçebilmesi için DNA'nın hasarsız olması gerekmektedir. Aksi takdirde hasar onarılana kadar hücre bu kontrol noktasında durur. Hasar onarılamayacak boyutta ise birkaç saat alan G1 noktasındaki bu durma p53'ün girmesiyle uzatılır (21, 22).

Kanser hücrelerindeki genetik bozukluğun en önemli nedenlerinden biri hücre döngüsünün kontrolünün herhangi bir basamakta bozulmasına neden olan mutasyonlardır. Örneğin G1 kontrol noktasında görev alan kinazları ya da siklinleri kodlayan genlerde meydana gelen mutasyonların hücrelerin malign dönüşümünde önemli rol alabileceği düşünülmektedir (17). Ayrıca tamir mekanizmalarından kaçmış hasarlı DNA veya replike olmamış DNA varsa G2 fazda kontrol edilir.

Kromozom takımlarının yeni hücrelere geçmesini sağlamak üzere kromozomların mitoz sırasındaki dizilimlerini kontrol eden M kontrol noktası bir diğer önemli kontrol noktasıdır. Eğer, hatalı dizilim sonucu kromozomların yeni hücrelere dağılımında bir problem olursa hata giderilinceye kadar mitoz fazının metafazında döngüsü durur (23).

Normal hücreler hataları saptayan ve yok etmeye çalışan bazı mekanizmalara sahiptirler. Bu mekanizmalara:

- DNA sentezinde oluşan hasarlı nükleotidleri hasarsız moleküllerle değiştirme,
- Tamiri olanaksız DNA hasarında ya da
  - yanlış,
  - eksik,
  - gereksiz

  olarak fazla transkribe olmuş DNA söz konusu olduğunda ise apoptoz adını verdiğimiz programlı hücre ölüm mekanizmasının devreye girmesi örnek olarak verilebilir.

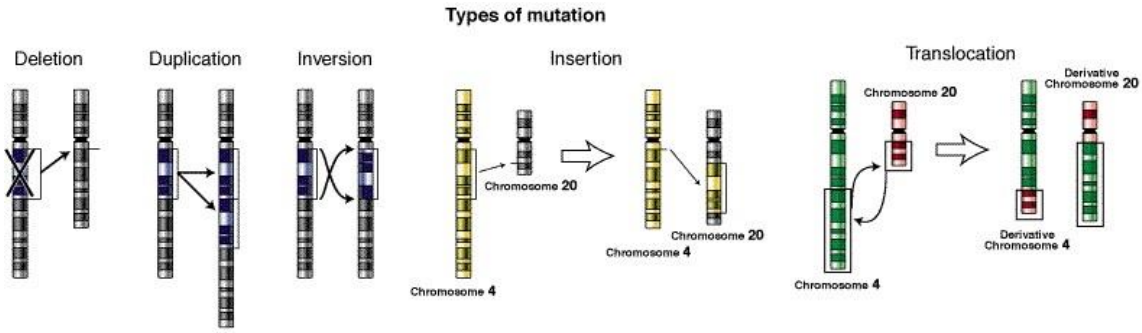
Sonuç olarak, kontrol noktaları genetik defektlerin düzeltilmesi için hataların saptanmaya çalışıldığı önemli yerlerdir. Böylece, DNA'sını sadece doğru ve tam olarak replike etmiş hücrelerin döngüde ilerlemesine izin verilir (24).

Tüm bunlara rağmen kanser dokusunda hızlı bir hücre çoğalması ve dokunun büyümesi söz konusudur. Kanser dokusunun büyümesinde en önemli temel faktör genetik defektlerdir (25).

4.1.2. Genetik Defektler

Mutasyon bir canlının genetik materyalinde çevresel ya da kalıtımsal nedenlerle meydana gelen kalıcı değişmelerdir (26). Kanserle ilgili genleri üç sınıfta incelemek mümkündür: onkogenler, tümör baskılayıcı genler ve DNA tamir genleri.

Normal hücrelerde hücre bölünmesinin kontrollü bir şekilde gerçekleştirilmesinden protoonkogenler sorumludur. Büyüme faktörleri hücre membranındaki resptörlere bağlanarak, normal hücre döngüsü olaylarını kontrol ederler. Bu süreçte etkinlik gösteren genlerden herhangi birisinde aktivasyonu arttırıcı yönde bir değişiklik ya da mutasyon meydana gelirse bu durum genellikle hücre çoğalmasının kontrolünün kaybolmasına neden olur. Hücre bölünmesinin düzenlenmesi için, bu genlerin ve/veya bu genlerin ürünlerinin pasif hale geçirilmiş durumda olmaları gereklidir. Kontrolsüz hücre çoğalmasının uyarılmasına neden olan genler, protoonkogen mutasyonlarının sonucunda onkogen adını verdiğimiz genler tarafından oluşturulur (17). Onkogenlere örnek olarak büyüme faktörleri, büyüme faktörü reseptörleri, sinyal iletim proteinleri ve transkripsiyon faktörleri verilebilir. En çok bilinen onkogenlerden bazıları: *RAS, c-MYC, RET, MET, CDK4, BCR/ALB, BCL2*'dir (27).



Şekil 2. Mutasyon oluşma şekilleri (28)

Anti onkogenler diğer adlarıyla tümör baskılayıcı genler hücre bölünmesini durdurma görevini yaparlar. Tümör baskılayıcı genlerin inaktivasyonları sonucu tümör gelişimi söz konusu olur (29). Pro veya antiapoptotik proteinler, hücre siklusu kontrol proteinleri, DNA tamir proteinleri, hücrelerarası iletişimde görevli proteinler tümör baskılayıcı genlere örnek olarak verilebilir. En iyi bilinen tümör baskılayıcı gen p53

geni olup diğerleri: *RB1, WT1, NF1, NF2, DCC, BRCA1, BRCA2* dir (27). 1971 yılında Alfred Knudson'ın "ikili vuruş" hipotezinde de önerdiği gibi tümör baskılayıcı genin her iki allelinde oluşan mutasyonlar kanser gelişimini başlatmakta iken onkogenin tek kopyasında meydana gelecek bir mutasyon kanser gelişimini başlatabilmektedir (17, 29).

DNA tamir genleri hücre döngüsü süresince oluşan mutasyonların onarılması sürecinde rol almaktadır. Baz kesip çıkarma (Base Excision Repair), hasarın direkt geriye döndürülmesi, yanlış eşleşme tamiri (Mismatch Repair), nükleotid kesip çıkarma (Nucleotide Excision Repair), homolog birleşmesi gibi mekanizmalar en iyi bilinen DNA tamir mekanizmalarıdır (30).

Kanserle ilişkili mutasyonlarda ana neden olan onkogenler, anti onkogenler ve DNA onarım genleridir. Bununla beraber mutasyonlar aynı zamanda karsinojenleri aktive ya da deaktive eden genlerde, hücre sinyal genlerinde, hücre döngüsü kontrol noktaları genlerinde, hücresel farklılaşma genlerinde, metastaz genlerinde, replikasyona bağlı hücresel senesans genlerinde de olabilmektedir (31).

4.2. Biyoinformatik

Geçtiğimiz 20-30 yıl içerisinde birçok bilim dalında çeşitli seviyelerde gerçekleşen gelişmeler bu alanlardaki ihtiyaçları da farklılaştırmıştır. Özellikle teknoloji ve biyoloji alanında geliştirilen yeni teknikler sayesinde daha kapsamlı çalışmalar yapılmaktadır. Hızlı DNA dizi analiz yöntemleri ile çok çeşitli türlerin genomlarının DNA dizilerinin belirlenmesi bu çalışmalara bir örnek olarak gösterilebilir (32). Ancak çok büyük bir hızla ve miktarda artarak biriken bu biyolojik verileri saklamak ve analiz etmek giderek zorlaşmaktadır. Bu bağlamda bilgisayar bilimleri hızlı bir ivmeyle artan bu verilerin saklanması, düzenlenmesi, birleştirilmesi, analizi ve kolayca erişilmesinde büyük katkı sağlamaktadır. Biyolojik bilimler başta kimya olmak üzere matemetik, fizik gibi temel bilimlerden biyolojik verilerin analizinde yararlanmaktadır. Bu yönüyle verilerin analizinde çok yüksek derecede hesaplama gücüne ve hızına ihtiyaç duyulmaktadır. Bilgisayar bilimleri temelde biyolojik verilerin hesaplanabilir daha farklı bir ifade ile sayısallaştırılmış forma dönüşmesinde en etkili araç durumundadır. Son dönemde insan genom projesi gibi çalışmalarda katkısı çok yüksektir. Ancak gelecek potansiyeli çok yüksek olacaktır. Bu nedenle biyolojik verilerin hesaplanması

için ihtiyaç duyulan hesaplama hızı ve işlem yapabilme becerisine ulaşabilmek için iki bilim dalının ortaklaşa gelişmesi gerekmektedir. Bu ihtiyaca cevap vermek amacı biyoinformatik bilim dalının doğal olarak ortaya çıkmasına neden olmuştur (33).

Biyoinformatik, biyoloji, bilgisayar, matematik, istatistik ve genetik alanlarını kapsayan, biyolojik dizi verilerini, gen içeriklerini ve sıralamalarını analiz eden disiplinler arası bir bilim dalıdır. Bu sayede moleküler düzeydeki yapıları ve işlevlerini tahmin etmeyi amaçlamaktadır. Bu bağlamda biyoinformatik, teknolojinin gelişimi ile paralel ilerlemektedir. Biyoinformatik, mikrodizin teknolojileri, hesaplamalı biyoloji ve sistem biyoloji gibi birçok alt araştırma başlıklarını da kapsamaktadır (34).

Mikrodizin analizi; genlerin diğer genlerle beraber nasıl bir ilişki içerisinde olduklarını inceleyerek, hastalıkların etiyolojilerinde rol alan genlerin tespit edilmesinde etkin bir şekilde kullanılmaktadır. Mikrodizin teknolojileri ve biyoinformatik araçları sayesinde yüzlerce genin analizini aynı anda yapmak mümkün hale gelmiştir. Son zamanlarda gelişen yeni teknolojilerle yerini RNA sekanslama tekniklerine bırakmaya başlamıştır (35).

Hesaplamalı biyoloji; gerçek hayatın bilgisayar ortamında benzetimi üzerinde durmaktadır. Bunlara modelleme çalışmaları örnek olarak verilebilir. En yaygın kullanılan alanlardan biri de ilaç geliştirme çalışmalarında yararlanılan proteinlerin üç boyutlu yapısını modelleme çalışmalarıdır. Ayrıca çoklu dizi hizalaması ve veri tabanlarında benzer dizileri bulmaya yönelik yeni ve etkili algoritmaların keşfedilmesi için etkin yöntemleri araştırır (36).

Sistem biyoloji; canlıların genler, proteinler ve biyokimyasal tepkimelerin etkileşen ve bütünleşmiş bir ağ yapısı olarak algılanarak incelenmesini amaçlayan çok yeni ve çok disiplinli bir bilim dalıdır. Bu etkileşimler sayısal ve biyoinformatik araçlarıyla tanımlanmaya çalışılmaktadır (37).

Biyoinformatiğin araştırma alanlarından biri de dizi ve gen analizidir. Genlerin bulundukları sistem içerisindeki yapılarının analizi, ağ yapıları, evrimleri ve matematiksel olarak modellenmeleri konularını inceler. Tüm bu işlemlerin yapılabilmesi için gerekli bilgiyi depolayan online veritabanları bulunmaktadır. The Cancer Genome Atlas (TCGA) veritabanı buna bir örnek teşkil etmektedir. TCGA, kanserden sorumlu mutasyonları kataloglamak için genom sekanslama ve

biyoinoformatik tekniklerini kullanan bir veri portalıdır. Yüksek çıktılı genom analiz tekniklerini kullanarak bu hastalığın temel dinamiklerini daha iyi anlamayı, teşhis, tedavi ve kanserden korunma olanaklarını geliştirmeyi hedeflemektedir (38, 39). Veritabanlarının yanı sıra farklı işlemler için kullanılan bazı programlar vardır. Örneğin, dizi eşleme (sequence alignment) yöntemiyle belli bir oranın üzerinde benzerlik gösteren homolog genlerin tespiti için geliştirilen BLAST ve FASTA gibi yazılımlar bulunmaktadır (40).

4.3. The Cancer Genome Atlas (TCGA) Veri Portalı

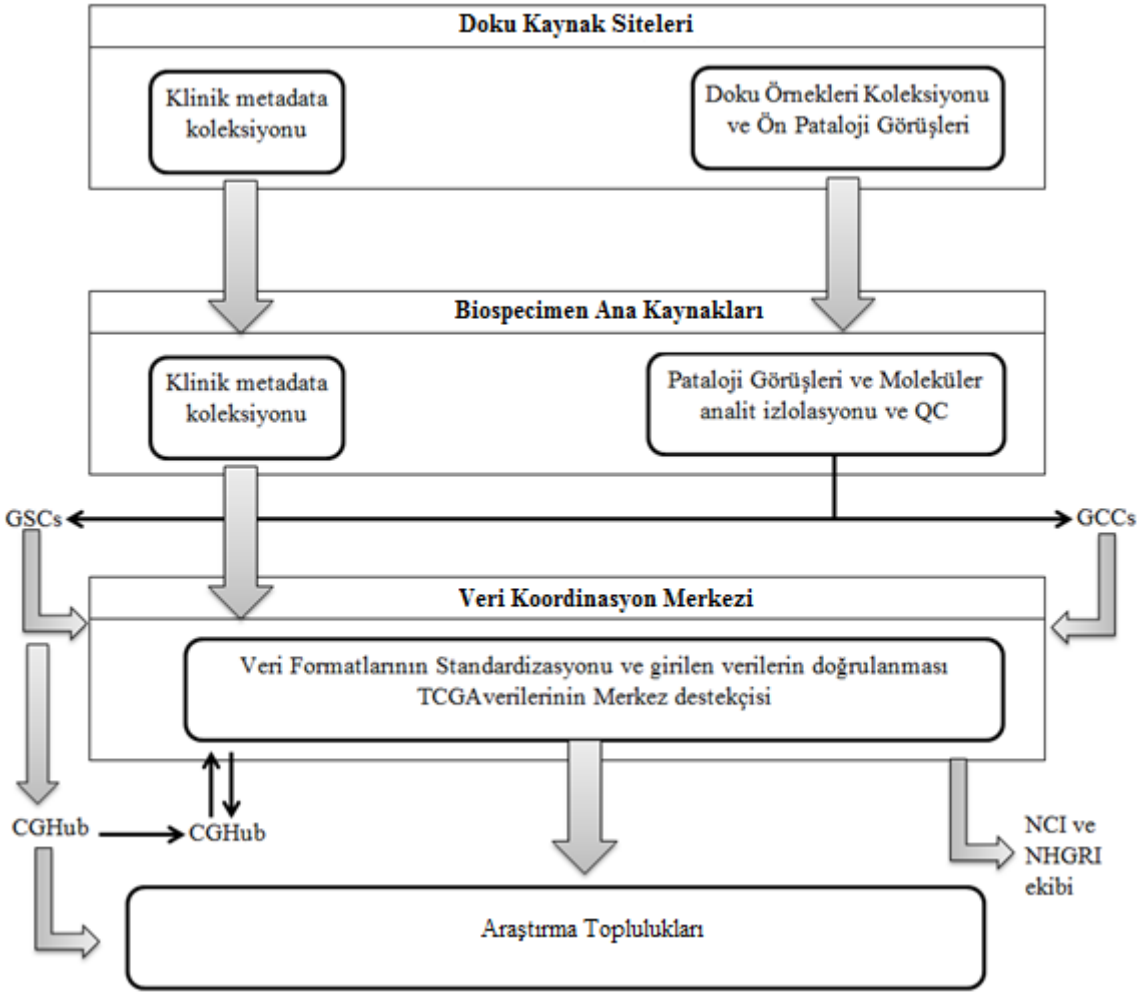
Kanser insanlığın karşı karşıya kaldığı en karmaşık hastalık olarak kabul edilmektedir. Tanımlanmış 200'den fazla çeşidi ve her çeşidi bir birinden ayrı karakterize özgün moleküler yapısı ile kanser, neredeyse kanser olan insanlar adedince tedavi metotları gerektirmektedir. Kanser, hücrelerin yapısındaki genomun dinamik değişiklikleri ile ilerler. Her bir kanser türünde görülen somatik mutasyonlar, kopya sayısı varyantları, değişken gen ekspresyonu profilleri ve farklı epigenetik değişiklikler vb. genetik bozukluklar aslında her bir kanser tipine özgün bir şekilde gelişim göstermektedir. Bu sebepler daha iyi tanı, tedavi ve kanserin önlenmesi konularının üzerine gidilmesi gerektiğini göstermektedir. Bu konular ise kanserin iç dinamiklerinde yaşanan değişikliklerin daha iyi anlaşılması ihtiyacını beraberinde getirmiştir. Genom dizileme ve biyoinformatik alanlarında gerçekleşen gelişmeler kanserli hücrelerin genom yapısının anlaşılmasına yeni bir ışık tutmuştur (38).

Amerika'da Ulusal Sağlık Enstitüleri (National Institutes of Health-NIH) tarafından 2005 yılında The Cancer Genome Atlas – TCGA projesi başlatılmıştır. Daha sonra 2006 yılında Ulusal Kanser Enstitüsü (National Cancer Institute) ve Ulusal İnsan Genom Araştırma Enstitüsü (National Human Genome Research Institute) tarafından yürütülen bu projenin pilot aşaması üç yıl sürmüştür. Bu süre içerisinde 50 milyon dolarlık bir yatırım alan TCGA öncelikle beyin, akciğer ve rahim olmak üzere sadece üç tip kanserin karakterizasyonuna yoğunlaşmıştır. 2009'dan itibaren yelpazesini genişleten bu proje şu an yaklaşık 20-25 farklı kanser türünün genomik karakterizasyon ve sekans analiz verilerini tutmaktadır (38).

TCGA, kanserden sorumlu mutasyonların karakterizasyonu için genom sekanslama ve gelişmiş biyoinoformatik teknikleri kullanarak, yüksek çıktılı genom

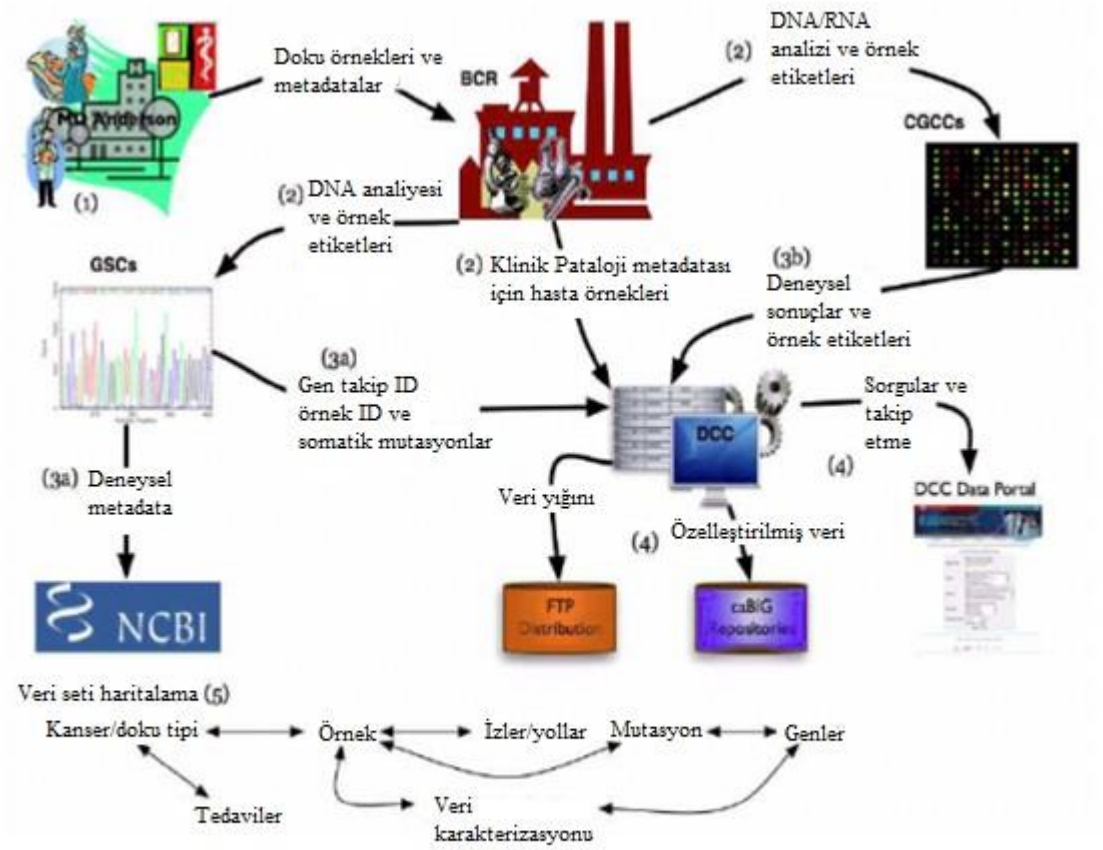
analiz teknikleri yardımıyla bu hastalığın temel dinamiklerini daha iyi anlamayı, teşhis, tedavi ve kanserden korunma olanaklarını geliştirmeyi hedeflemektedir (38).

TCGA yapı olarak çok iyi organize olmuş bazı araştırma merkezlerinin ortak çalışması olarak görülebilir. Aşağıdaki şekilde TCGA birimlerinin organizasyonu verişmiştir.



Şekil 3. TCGA'nın organizasyon yapısı (38).

TCGA'nın kanser genom profillerini daha kapsamlı olarak analiz edebilmek için mikrodizilerde ve yeni nesil sekanslama yöntemleri üzerinde kullandığı bazı yüksek çıktılı teknoloji platformlar bulunmaktadır. Bunlardan bazıları: RNA sekanslama, MikroRNA sekanslama, DNA sekanslama, SNP tabanlı platformalar, dizi tabanlı DNA metilasyon sekanslama bunlara örnek olarak verilebilir (38).



Şekil 4. TCGA'de verilerin görselleştirilmesi süreci (38)

TCGA, kanserin genomik yapısını anlamada kapsamlı bir resim ortaya koymuş ve sofistike yüksek-çıktılı teknolojiler yardımıyla geliştirilen biyoinformatik araçları kanserlerin genomik mimarisindeki benzerlik ve farklılıkların tespitinde büyük katkılarda bulunmuştur. Ayrıca dünyadaki diğer araştırmacılara kanser, genetik ve epigenetik hakkında ve kanser biyobelirteç (biomarker) adayları ve hedeflenmiş tedavilerin geliştirilmesinde çok büyük bilgi yardımı sağlamıştır. Bununla beraber, elde

edilen bu bilgi potansiyeli ile yakın zamanda kişilselleştirilmiş kanser ilaçları üretmek için gerekli desteği verecek durumdadır. TCGA bilim adamlarının bir sonraki hedefi ise daha güçlü biyoinformatik araçları geliştirerek verilerdeki gürültüyü ortadan kaldırmak analiz çözünürlüğünü arttırmak ve daha ileride tüm bu bulgularla kanserin tanı, teşhis ve tedavisini çok kolay bir hale getirmektir (38).

4.4. Veri, Veritabanı ve Veri Madenciliği Kavramları

4.4.1. Veri

Veri, bir bilgi sistemine girilen veya diğer cihazlardan alınan, ölçüm, sayım, deney, gözlem ya da araştırma yolu ile elde edilmiş, yapısal olmayan, işlenmemiş girdilerin tümüdür. Veri metin, sayı, görüntü gibi çeşitli formatlarda olabilir. Veriler filtreleme, formatlama gibi işlemler uygulanarak anlaşılır ve yoruma hazır hale getirilebilirler (41, 42).

4.4.2. Enformasyon

Enformasyon, dünyaya ilişkin durumların ham veri şartlarını aşacak bir biçimde işlenmesi durumudur. Örneğin; kilomuz bir veridir fakat biz düzenli aralıklarla kilomuzu ölçüp sonuçları tarih sırasına konulmuş bir tablo haline getirirsek enformasyona dönüştürmüş oluruz. Özetle seçilmiş verilerin anlamlı bir biçimde yapılandırılmasına enformasyon denir (43).

4.4.3. Bilgi

Bilgi, verilerin enformasyon sürecinden geçtikten sonra yorumlanması aşaması olarak tanımlanabilir. Kurum, kuruluş ve organizasyonlar için bilgi; müşteriler, ürünler, süreçler, hedefler, hatalar ve başarılar hakkında sahip olunan enformasyonların değerlendirilmesidir (44). Bilgiyi depolamak için şu an hala sadece beyin vardır. Bilgisayarlar büyük oranda veri ve enformasyonları saklamak için kullanılmaktadır. Enformasyonları ve daha sonra bilgileri birbirleri ile kendi kendine ilişkilendirebilen, insan beyni özellikleri gösteren bilgisayar teknolojilerine erilişmeye çalışılmaktadır.

4.4.4. Veritabanı ve Veritabanı Yönetim Sistemleri

Veritabanları, büyük miktardaki veriyi kolay ve verimli bir şekilde depolamayı, düzenlemeyi ve gerektiği zaman bilgiye ulaşabilmeyi amaçlayan, aynı zamanda birden çok kullanıcıya erişim olanağı tanıyan sistemlerdir (45, 46). Veritabanları verilerin

düzenli olarak bir arada tutulduğu, güncellenebilen, taşınabilen ve farklı platromlardan erişim imkanı sunabilen, veriler arasında kural ve ilişkilerin tanımlanabildiği kayıt ve dosyalar bütünüdür. Veritabanı yazılımları ile büyük miktarlardaki verileri, depolamak ve işleyebilmek mümkün hale gelmiştir.

Tüm alanlarda gittikçe veri miktarı ve veri tipi arttıkça verilerin depolanması ihtiyacı da artmıştır. Bununla beraber bu büyük miktardaki verileri el ile depolamak ve kontrol etmekte imkansız hale gelmiştir. Bu yüzden karmaşıklaşan bu verileri daha düzenli ve verimli bir şekilde depolamak için veri tabanı yönetim sistemleri geliştirilmiştir. Bir veritabanını tanımlamak, kullanmak, değiştirmek ve veritabanı sistemleri ile ilgili her türlü işletimsel gereksinimleri karşılamak için kullanılan yazılımlara Veritabanı Yönetim Sistemleri (Database Management Systems) adı verilir (46). Veritabanı modellerine göre ağ, hiyerarşik veya nesne yönelimli olmak üzere farklı Veritabanı Yönetim Sistemleri mevcuttur. Bunlar içinde en yaygın kullanılanı ilişkisel veritabanı modelidir (47).

İlişkisel veritabanı, tablolar şeklinde saklanır. Bu veritabanı yönetim sisteminde; veri alışverişi için özel işlemler kullanılır. Tablolar arasındaki ilişkiler matematiksel bağlantılarla temsil edilir. Her tablo bir ilişkiye veya bir varlığa karşılık gelmektedir. Tablonun sütunları nitelikleri; satırları ise bu niteliklerin aldığı değerleri ifade eder. Her bir satır bir varlığın ya da ilişkinin somut örnekleri olarak da düşünülebilir. İlişkisel veri tabanı kavramlarından birincil anahtar, aday anahtar ya da yabancı anahtar tablolarda oluşan varlık örneklerinin ilişkilerini tanımlamayı sağlar. Bu ilişkiler sayesinde veriler verimli ve tutarlı şekilde yönetilebilir (47, 48).

4.4.5. Veri Ambarı

Veri ambarı; farklı kaynaklarda tutulan verilerin ortak bir çatı altında birleştirilerek, verilerin zaman boyutunda birbiri ile etkileşimini sağlayan, tutarlı ve doğru verilerin yer aldığı sistemler olarak tanımlanır. Veri ambarı veri kaynaklarında tutulan verilerin, hatalardan arındırılarak üzerinde işlem yapılmaya uygun hale dönüştürülmesini sağlar. Bu şekilde ilişkili veriler sorgulanabilir ve analiz edilebilir hale getirilir. Veri ambarlarında veriler üzerinde uygulanan işlemlerden başlıcaları veri madenciliği, çok boyutlu analizler, kampanya yönetimi, istatistiksel analizler, sorgulama ve raporlamadır (49).

Veri ambarı oluşturma aşamasında verilere temizleme, birleştirme, dönüştürme ve indirgeme gibi bir dizi işlem uygulanır. Oluşturulduktan sonra veriler düzenli olarak güncelleştirilir (49).

4.4.6. Veritabanlarında Bilginin Keşfi Süreci

Geçtiğimiz çeyrek asırda yaşanan bilimsel gelişmeler, bilgisayar bilimlerinin uygulama alanlarını oldukça genişletmiş ve birçok disiplinle iç içe girmesini sağlamıştır. İhtiyaç duyulan yapılarda veritabanları oluşturularak, farklı türlerdeki veriler sistematik bir şekilde tutulmaktadırlar. Bilişim sistemlerinin gelişmesiyle beraber, veritabanlarında tutulan verilerde ve veri türlerinde büyük artış olmuştur. Organizasyonların kullandığı ürün-stok bilgileri, insanların demografik bilgileri, gen verileri, sağlık durumları gibi çok çeşitli alanlardan farklı türlerdeki verilere örnek olarak verilebilir.

Büyük veri yığınlarının veritabanları yardımıyla kolayca saklanabilmesinin yanında bu verileri analiz etmek ve yorumlamak için bir takım işlemler gerekmektedir. Bu işlemler için bilinen yöntemlerin, istatistiksel metotların ve raporlama araçlarının yetersiz kaldığı durumlar vardır. Bu durumlarda Veritabanlarında Bilgi Keşfi – VBK (Knowledge Discovery in Databases) olarak isimlendirilen süreçlerle gerekli bilgi elde edilebilmektedir (50). VBK yeni, nitelikli, doğru, yorumlanmış ve anlaşılır bilgilere ulaşmaya olanak sunar.

Veritabanı içerisinde açık bir şekilde görülemeyen, önceden ne olabileceği konusunda bir bilgi olmaksızın, veritabanı içerisinde gizli örüntüleri arama işlemine keşif denir. Büyük veritabanlarında kullanıcının hızlıca bulamayacağı ve bulmak için gerekli doğru soruları bile düşünemeyeceği saklı örüntüler olabilir. Keşfin asıl amacı bilgi ve kalite seviyesi zengin örüntüleri bulmaktır (51).

VBK süreci, bilginin kullanıcıya başarılı sonuçlar sunabilmesi için bazı işlem basamakları içerir. Bu basamaklar aşağıdaki gibidir (50):

- Problemin tanımlanması
- Verilerin hazırlanması
- Modelin kurulması ve değerlendirilmesi
- Modelin kullanılması
- Modelin izlenmesi

4.4.6.1. Problemin Tanımlanması

Başarılı sonuçlar elde etmek için öncelikle problem açık ve detaylı bir şekilde tanımlanmalıdır. Düzgün tanımlanmayan problemlere çözüm getirmekte mümkün olmayacaktır. Bu nedenle VBK sürecinin başarıya ulaşmasında problemi tanımlama önemli yer tutmaktadır.

4.4.6.2. Verilerin Hazırlanması

Bilgi keşfi sürecinin önemli ve zaman alan basamaklarından biri de verilerin hazırlanması sürecidir. Bu süreçte üzerinde veri madenciliği uygulamaları kullanılabilecek veri kümeleri bulmak ya da oluşturmak gereklidir. Veriler ön bir işleme tabi tutularak program tarafından kullanımı daha kolay hale getirilmelidir. Bu aşama oldukça iyi tasarlanmalıdır. Aksi halde ilerleyen bölümlerde sık sık bu aşamaya geri dönmek zorunda kalınacaktır (52). Bu nedenden dolayı verilerin hazırlanmasında dikkat edilmesi gereken bazı alt basamaklar vardır. Bunlar verilerin toplanması, değer biçme, birleştirme ve temizleme, seçme ve dönüştürme işlemleridir (Şekil 5).

4.4.6.2.1. Verilerin Toplanması

Problemin tanımlanmasının ardından problemin ortaya koyduğu sorulara yanıt verecek verilerin gerekli olan kaynak ya da kaynaklardan toplanması işlemidir. Bunun için öncelikle varsa elektronik ortamda bulunmayan veriler elektronik ortama aktarılmalıdır. Farklı veritabanları veya dosyalardaki veriler üzerinde işlem yapılabilecek tek bir veritabanında birleştirilmelidir. Veri madenciliği metotlarının etkin

uygulanabilmesi için verilerin çok boyutlu veri küplerinde toplanması daha çok tercih edilen bir yoldur.

4.4.6.2.2. Değer Biçme

Verilerin toplanması aşamasında farklı kaynaklardan gelen veriler arasında uyumsuzluklarla karşılaşılabilir. Toplanan veriler arasındaki uyumsuzluklar ölçü birimi, veri formatı ve kodlamadan kaynaklanan farklılıklar vb. olabilir. Bu aşamada, uyumsuzlukların tespiti, ne oranda olduğu ve giderilmesi üzerinde durulur.

4.4.6.2.3. Temizleme ve Birleştirme

Değer biçme aşamasında belirlenen uyumsuzlukların giderilmesini ve değişik kaynaklardan elde edilen verilerin birleştirilerek tek bir veritabanında toplama işlemleri bu aşamada gerçekleştirilir. Araştırmanın sonuçlarının olumsuz yönde etkilenmemesi için hatalı girilmiş, tutarsız veya gürültülü veriler düzeltilmelidir.

Veri temizlemek için kullanılabilecek farklı yöntemler bulunmaktadır. Örneğin, eksik verilerin bulunduğu bir kayıtta birden fazla boş alan var ise kayıt silinebilir ya da boş alan sayısal bir değer ise diğer kayıtların ortalaması alınarak boş alana atanır ve böylece kayıp veri telafi edilebilir. Gruplama yapmanın daha etkin sonuçlar verdiği durumlar olabilir. Örneğin, bir sınıftaki öğrencilerin başarı durumlarını gösteren bir alanda zayıf, orta, iyi gibi sınırlar belirleyerek veri kategorize edilerek sadeleştirilebilir. (53).

4.4.6.2.4. Seçme

Geliştirilmesi hedeflenen veri madenciliği modeline göre verilerin seçiminin yapıldığı aşamadır. Tahmin edici bir model geliştirmek için bu aşamada bağımlı ve bağımsız olmak üzere değişkenlerin ve modelin eğitiminde kullanılacak veri setinin seçimidir (54).

Modele, modeldeki değişkenlerin ağırlıklarının azalmasına sebep olabilecek sıra numarası ve kimlik numarası gibi anlamlı olmayan değişkenlerin karışmamasına özen gösterilmelidir. Veri madenciliği algoritmalarından bazıları bu tip değişkenleri otomatik olarak filtrelese de, uygulamada bu işlemin yazılıma bırakılmaması daha efektif sonuçlar alınmasına yardımcı olacaktır (54).

4.4.6.2.5. Dönüştürme

Seçilen verilerin veri madenciliği modelinde kullanılacak algoritmaya uygun formata dönüştürülme aşamasıdır. Bunun yanında modelde kullanılacak parametrelerin uygun bir biçime dönüştürülmesini de içine alır (41).

4.4.6.3. Modelin Kurulması ve Değerlendirilmesi

Önceki adımlar başarıyla tamamlandıktan sonra bilgi keşfi sürecinin temelini oluşturan model kurma aşamasını geçilir. Bu adımda veri madenciliği uygulamaları başlar.

Veri madenciliği (Data Mining), farklı şekillerde tanımlanabilir. Veri madenciliği, büyük miktarda veri kümeleri arasından, saklı kalmış, değerli, kullanılabilir bilgileri ortaya çıkarmak ve stratejik karar destek sağlamak amacıyla kullanılan yöntemlerdir (55). Başka bir ifadeyle, veri tabanı teknolojileri, istatistik, yapay zeka, makine öğrenmesi, örüntü tanıma, veri görselleştirilmesi gibi pek çok teknik alan arasında birleştirici bir köprü vazifesi yapan disiplinlerarası bir alandır (56). VBK sürecinin önemli bir basamağı olan bu aşama son zamanlarda VBK yerine kullanılmaktadır; ancak VBK sürecinin bir alt basamağını oluşturmaktadır.

Modelin kurulum aşamasında, problemin tanımlanması basamağındaki noktalar dikkate alınarak süreç başlatılmalıdır. Verilerin hazırlanması aşaması tekrar gözden geçirilip eğer gerek varsa veriler yönteme uygun hale getirilebilir. Kullanılacak veri madenciliği modeli belirlendikten sonra modelde koşulacak uygun algoritmalara karar verilmelidir. Tüm veri madenciliği modelleri için geliştirilen değişik algoritmalar mevcuttur. Bu algoritmalar veri tabanı üzerinde denenmeden hangisinin en uygun olduğunu kestirmek oldukça zordur. En uygun model ve algoritma bulunana kadar bu süreç tekrarlanır.

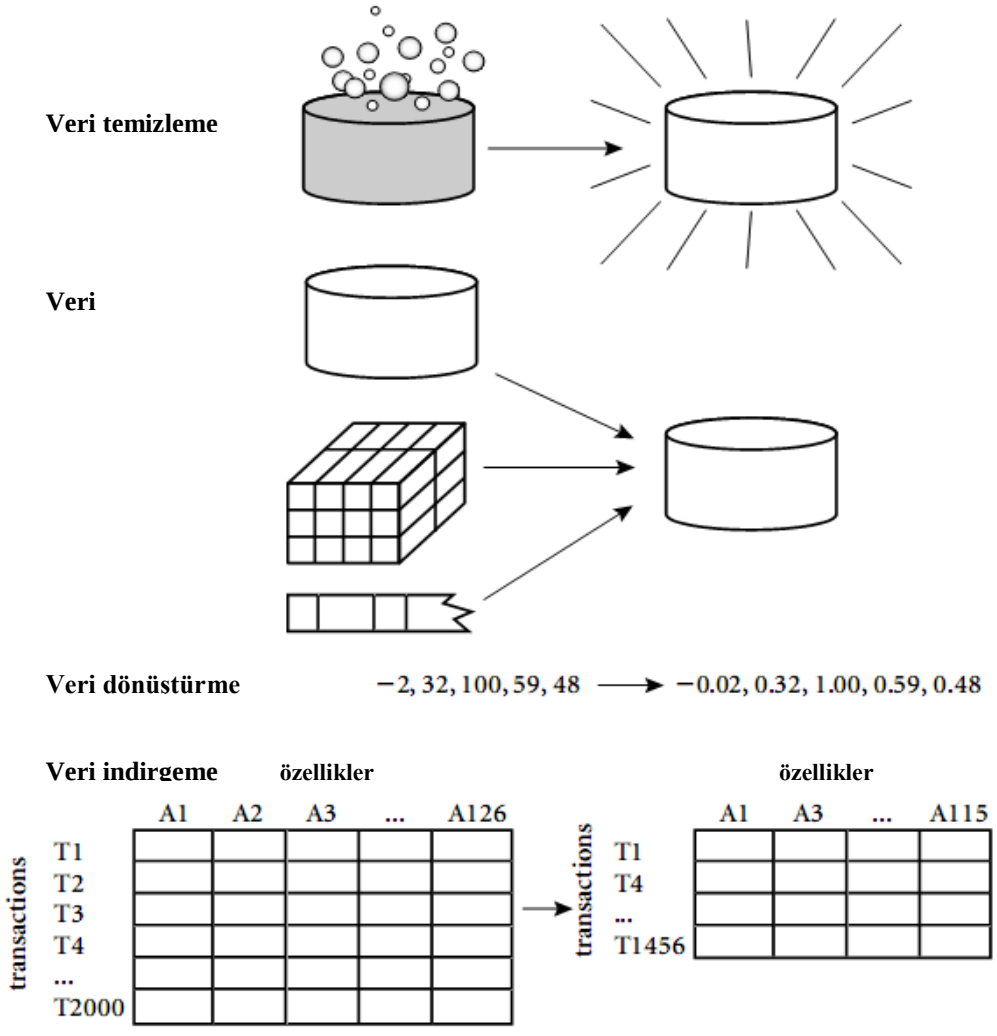
Modeli değerlendirmek için elde edilen sonuçların güvenilir, nitelikli, faydalı ve yeni olup olmadığına bakılır. Bu işlem modelin etkinliğini ortaya koyar. Modeli değerlendirmek için çeşitli yöntemler bulunmaktadır.

4.4.6.4. Modelin Kullanılması

Geliştirilen ve performansı onaylanan model kullanım amacına göre tek başına bir uygulama olabileceği gibi başka bir uygulamanın bir aracı olarak da kullanımı mümkündür.

4.4.6.5. Modelin İzlenmesi

Kullanıma geçen modeller zaman içerisinde gerek girdi ve çıktılarda gerçekleşen gerek kullanılan sistemin özelliklerinde yaşanan değişiklerden dolayı yeniden düzenlenmesi gerekebilir. Bu nedenlerden dolayı kullanım aşamasında modellerin izlenmesi ve takibi unutulmamalıdır.



Şekil 5. Veri önişleme adımları (2)

4.4.7. Veri Madenciliği

Veri madenciliği, büyük miktarlarda toplanan verilerin karmaşık analizler neticesinde anlamlı ve nitelikli bilgiye dönüştürülmesi sürecinin çekirdeğini oluşturmaktadır. Disiplinlerarası bir alan olan veri madenciliğinin tam olarak bir tanımını yapmak mümkün olmamakla beraber kullanım ve uygulama alanlarına göre farklı tanımları ortaya koyulmuştur.

Önceki bölümde belirtilen tanımdan farklı bir tanıma göre, sahip olunan verilerden elde edilen gizli, çok net anlaşılamayan ve önceden bilgi sahibi olunmayan fakat kullanışlı bilgi potansiyeline sahip çıkarımların elde edilmesidir (58). Diğer bir tanımlamaya göre veri madenciliği; büyük miktarlardaki verilerin arasından ilgili konuya dair geleceğin tahmin edilmesine yardımcı olacak nitelikli, anlamlı ve faydalı bağlantı ve kuralların araştırılması ve analizidir. Bununla beraber veri madenciliği, çok büyük veri kümelerinde veriler arasındaki ilişkileri inceleyerek birbirleri arasında fark edilemeyen, gizli kalmış bağlantıları bulmaya yardımcı olan veri analiz tekniğidir (59).

Veri madenciliği tanımlarına yakından bakıldığında, bu tanımların tümünde ortak noktaların olduğu göze çarpacaktır. Bunlar "çok fazla" miktarda veri ve verilerden "anlamlı" bilgilerin çıkarılmasıdır (41).

Veri madenciliği tek başına problemler için bir çözüm olarak görülmemelidir. Çözüme ulaşmak için gereken süreçte bir karar-destek aracı görevi yapmaktadır. Veri madenciliği; kullanıcılarına, veri tabanlarında veriler arasındaki örüntü ve şablonların bulmasına ve yorumlamasına yardımcı olmaktadır (2).

4.4.7.1. Veri Madenciliğinin Kullanım Alanları

Veri madenciliği sahip olduğu nitelikler ve sunduğu faydalar göz önüne alındığında geniş bir uygulama alanı ortaya çıkacaktır. Veri madenciliğinin bu çıkarımsal gücünden sağlık, ilaç, pazarlama, bankacılık, sigorta ve reklamcılık gibi birçok sektörde faydalanılmaktadır. Güvenlik sektöründe kimlik sahtekarlıklarının belirlenmesi, pazarlama sektöründe müşterilerin sosyo-ekonomik durumları ve satın alma analizi, bankacılıkta kredi kartı taleplerinin değerlendirilmesi gibi vb. birçok örnek vermek mümkündür.

4.4.7.2. Veri Madenciliğinde Kullanılan Modeller

Verinin bilgiye dönüşmesinde kullanılan modeller iki başlık altında incelenir. Bunlar tahmin edici (predictive) ve tanımlayıcı (descriptive) modellerdir.

Önceden sahip olunan verilerden yola çıkarak bir model geliştirilip sonra bu model yardımıyla karşılaşılan yeni durumlardan ilgili probleme yönelik sonuçlar elde etmek tahmin edici modellerin hedefidir. Örneğin, bir banka vermiş olduğu kredilere ilişkin tüm verileri bir veri tabanında tutuyor olabilir. Bu verilerde kredi alan müşterilerin sahip olduğu özellikler bağımsız değişkenler olarak tanımlanırken alınan kredinin geri ödenip ödenmediği bilgisi ise bağımlı değişken olarak değerlendirilir. Bu verilerden yola çıkılarak kurulan model yardımıyla daha sonraki kredi taleplerinde müşteri özellikleri kullanılarak kredinin geri ödenip ödenmemesi ile ilgili tahminler yapılabilmektedir (56).

Tanımlayıcı modeller ise verilerin sahip olduğu ve karar verme sürecinde etkili olabilecek örüntülerin ortaya çıkarılmasını sağlamaktadır. Böylece veri hakkında karara varılabilir. Tanımlayıcı olarak kullanılan veri madenciliği tekniklerine kümeleme ve birliktelik kuralları örnek olarak verilebilir.

Veri madenciliği teknikleri, veri yapıları ve örüntü biçimlerine göre bazı alt kategorilere ayırılmışlardır (2).

- Sınıflandırma ve Regresyon (Classification and Regression)
- Kümeleme Analizi (Cluster Analysis)
- Birliktelik Analizi (Association Analysis)
- İstisna Analizi (Outlier Analysis)
- Evrimsel Analiz (Evolution Analysis)
- Tanımlama ve Ayrımlama (Characterization and Discrimination)

4.4.7.2.1. Sınıflandırma ve Regresyon (Classification and Regression)

Veri madenciliğinde yaygın bir şekilde kullanılan modellerden birisi sınıflandırmadır. İnsanın doğasında olan sınıflandırma, günlük hayatta sık bir şekilde bilinçli ya da bilinçsiz bir şekilde kullanılmaktadır. Fakir-zengin, çalışkan-tembel, kısa-uzun gibi insanları ya da nesneleri belirli özelliklerine göre kategorilere ayırarak sınıflandırma yapmaktayız.

Bu yöntem de veriler için önceden tanımlanmış gruplar vardır. Veriler kendileri için ayrılmış gruplara atanır ve yeni bir değişkenle karşılaşıldığında hangi sınıfa yerleştirileceğine karar verilir. Önceden belirlenmiş sınıfların bulunduğu bu tekniklere denetimli öğrenme teknikleri denir.

Sınıflandırma ve regresyon metodunda tahmin edilen bağımlı değişken, diğer tahmin edici değişkenler bağımsız değişkenlerdir. Burada bağımlı değişkenin veri tipi yöntemi belirler. Eğer bağımlı değişken kategorik bir değerse sınıflama, sürekli bir değişkense regresyon problemi olarak değerlendirilir (2). Bu iki metot önemli veri sınıflarını ortaya çıkaran veya verilerin gelecek eğilimlerini tahmin eden modelleri kurabilen veri analiz yöntemleridir. Sınıflama ve regresyon modellerinde yaygın olarak kullanılan yöntemlerin bazıları; karar ağaçları, saf-Bayes, lojistik regresyon, genetik algoritmalar, k-en yakın komşu, yapay sinir ağlarıdır.

• **Karar Ağaçları (Desicion Trees):** Bir obje ile hedef değeri arasındaki ilişkilerin tahmin edilmesi ve açıklanması konusunda kolay anlaşılan kurallar oluşturabilen ağaç görünümünde basit fakat başarılı bir tekniktir. Ağaç yapısına benzer bir akış diyagramına sahip olan karar ağaçlarında her bir düğüm bir özelliğe denk gelir, her bir dalı bir ölçütün sonuçlarını gösterir ve yaprakları da sınıf dağılımlarını temsil eder (2). Regresyon, kümeleme ve özellik seçimi gibi veri madenciliği işlemlerinde kullanılan karar ağaçları, kolaylıkla sınıflandırma kurallarına dönüştürülebilir. Karar ağaçlarının kullanımının yaygın olmasının nedenlerinden bazıları şöyledir; kolay okunabilirliği, nominal, nümerik veya metin girdileri kullanabilmedeki esnekliği; hatalı veya kayıp değer içeren veri kümeleri üzerindeki tolerans yeteneği; düşük işlem gücü ile yüksek tahmin performansları gösterebilmesi ve büyük veri kümeleri için kullanışlılığıdır (60).

• **Saf-Bayes:** Bayes karar teorisini esas alan olasılıksal bir sınıflandırma metodudur. İzlenen değişkenlerin birbirinden bağımsız olduğu varsayımı ile çalışmaktadır (61).

• **Lojistik Regresyon:** Bağımlı ve bağımsız değişkenler arasındaki en iyi uyuma sahip ilişkiyi en az değişken kullanarak ortaya çıkarmaya çalışan modellerdir. Örneğin; bir hastada kolon kanseri teşhisi konulabilir veya konulmayabilir, ilaç ya da radyasyon tedavisine cevap vermiş veya vermemiş olabilir. Böyle iki muhtemel farklı neticeyi içeren veri kümlerine iki değerli (binary) kümeler denilmektedir. İki değerli değişkenler 0 ve 1 ile kodlanmışlardır. Bağımlı değişkenin iki kategori içeren değişken olduğu durumlarda doğrusal regresyon uygulanamaz. Bunun için çok değişkenli istatistiksel metotlardan biri olan lojistik regresyon analizi kullanılmalıdır. Çeşitli nedenlerden dolayı varsayım bozulmalarının olduğu durumlarda diskriminant ve çapraz tablo analizlerine bir alternatif olarak uygulanmaktadır (62).

• **Genetik Algoritmalar:** Darvin'in evrim teorisinden esinlenerek ortaya çıkan genetik algoritmalar çoğunlukla optimizasyon problemlerinde kullanılmaktadır. Doğal seleksiyon ve en iyinin kurtulması ya da korunumu ilkelerinin programcılığa uyarlanmasıyla elde edilen arama tekniğidir. Günümüzde kullanılan biçimi J.H. Holland tarafından 1975 yılında ortaya konulmuştur (63). Genetik algoritmalarda adaylar başlangıçta eşit uzunlukta vektörler olarak tanımlanır ve bu vektörlerden rastgele seçilerek belirlenen bir popülasyon oluşturulur. Bu oluşturulan vektörlere kromozom adı verilir ve yeni nesiller oluşturarak değişikliğe uğrarlar. Kromozomun üzerindeki genler, n boyutlu vektörlerin bir boyutuna karşılık gelir. Yeni nesiller oluştukça kromozomlar hedef fonksiyonuna yerleştirilerek sonuçlar hesaplanır ve etkinliği ya da iyilik derecesi ölçülmüş olur. Yeni bir nesil oluşturulurken tekrarlama, çaprazlama, mutasyon gibi fonksiyonlardan faydalanılarak bazı kromozomlar tekrar üretilir. Böylece en iyi sonuçlar ya da nesil bulunana kadar süreç devam eder (63).

• **K-En Yakın Komşu:** Günümüzde bilgisayar teknolojilerinin gelişmesiyle hafıza tabanlı metotlar da önem kazanmış bir veri madenciliği yöntemidir. Bu metotlar arasında oldukça yaygın bir şekilde kullanılan k-en yakın komşu algoritmasıdır. Algoritma ilgili veri için en yakın komşuları inceleyerek, veri tabanındaki kayıtların birbirlerine olan uzaklıklarını değerlendirir ve bu bilgiye göre bir sınıflama yapar. Yeni

bir kayıt sınıflandırılacağı zaman, k tane yakın komşunun tutumları incelenir ve ortalamaları alınarak nesnenin davranışı olarak kabul edilir (64).

• **Yapay Sinir Ağları (YSA):** YSA kavram olarak 1900'lü yılların ilk yarısında ortaya çıkmış fakat 1980'lere kadar bilgisayarla kullanılmamıştır. YSA, insan beyni fonksiyonları incelenerek modellenmiştir. İnsanlar tüm yaşam süreleri boyunca edindikleri tecrübelerle sürekli öğrenme süreci içerisindedirler. Bu tecrübelerle beyindeki sinaptik bağlar güçlenir, çeşitlenir ve yeni bağlar oluşur. YSA da aynı şekilde çalışır. Dışarıdan gelen uyaranlara yani girdilere karşı dinamik olarak yanıt oluşturma yoluyla veriyi işleyen ve birbiriyle bağlantılı elemanlardan oluşan bir yapı sergiler. YSA öğrenme şekillerine göre denetimli (supervised) ve denetimsiz (unsupervised) olmak üzere iki başlıkta incelenir. Denetimli öğrenme biçiminde ağa girdi-çıktı bilgileri verilerek ayrıntılı bir eğitim sağlanmaktadır. Denetimsiz öğrenmede ise ağa yalnızca girdiler verilerek, çıktıları ağın oluşturması istenmektedir (63).

• **Destek Vektör Makineleri-DVM (Support Vektor Machines-SVM):** Destek vektör makineleri (DVM) parametrik olmayan ve istatistiksel öğrenme yapısına sahip bir sınıflandırma tekniğidir (65). İkili sınıflandırmalar için geliştirilen DVM az sayıdaki örneklem verisi ile en doğru sınıflandırmayı elde etmeyi amaçlar (66). İlk olarak iki sınıflı doğrusal verileri sınıflandırmak için tasarlanmış olan bu yöntem daha sonra çok sınıflı ve doğrusal olmayan verileri sınıflandırmak için geliştirilmiştir. İki sınıfı birbirinden ayırabilen hiper düzlemin belirlenmesini temel alır (65).

• **Arıtılmış Regresyon Ağaçları-ARA (Boosted Regression Trees-BRT):** Makine öğrenmesi ve istatistiksel yöntemleri birleştiren Arıtılmış Regresyon Ağaçları (ARA) son zamanlarda kullanılmaya başlayan yeni bir veri madenciliği tekniğidir. Tahmine dayalı veri madenciliği problemlerinde yüksek başarı gösteren ARA, yöntem olarak karar ağaçları grubundan regresyon ağaçları ve denetimli öğrenme gerçekleştiren bir makine öğrenmesi algoritması olan artırma (boosting) algoritmalarındaki avantajlarının bir araya getirilmesini benimser (63).

4.4.7.2.2. Kümeleme Analizi (Cluster Analysis)

Belirli sınıflara ayrılan veri seti üzerinde, her bir sınıfın kendi içerisindeki benzerliklerini maksimize edip, diğer sınıflar ile olan benzerliklerini minimize edilmesini esas alan verilerin denetimsiz olarak gruplandırıldığı yöntemdir (49). Diğer bir ifade ile heterojen yapıya sahip veri uzayını daha homojen yapıya sahip alt gruplara ayırma işlemi olarak tanımlanabilir. Kümeleme analizinde sınıflandırmadan farklı olarak verilerin ayrılacağı gruplar önceden belirli değildir, verinin dağılımına göre sonradan oluşturulur. Çalışmanın amacına ve veri tipine bağlı olarak literatürdeki birçok kümeleme algoritmalarından biri seçilebilir.

4.4.7.2.3. Birliktelik Analizi (Association Analysis)

Aynı veri tabloları içerisindeki kayıtların birlikte bulunma ya da olayların beraber gerçekleşme olasılıklarını veren yönteme birliktelik kuralları denir. Birliktelik kuralları market-sepet araştırması olarak da bilinmektedir. Müşterilerin sepetlerindeki her bir ürün veritabanında bir nesne her bir alışveriş de bir işlemdir. Sepetler analiz edilerek hangi ürünlerin birlikte alındıkları keşfedilip, müşterilerin satın alma eğilimleri belirlenir. Elde edilen bu bilgilere göre marketteki raf düzeni değiştirilir, güncellenir ve müşterilere daha fazla seçenek sunulmuş olur.

Birliktelik kuralları oluşturulurken kullanılan bazı kıstaslar mevcuttur. Bunlardan biri destek değeri; diğer ise güven değeridir. Destek ve güven değerlerini sağlayan birliktelikler değerlendirilmeye alınır. Bir birliktelik kuralı,

$A \Rightarrow B$ (Destek = %5, Güven = %60)

şeklinde ifade edilir. Bu ifade de A ürünü satın alanların %60'ı aynı zamanda B ürününü de almıştır. A ve B ürünlerinin tüm satışlar içinde beraber satın alınma oranı ise %5'tir. Burada güven değerinin hesaplanması için (A ve B'nin bulunduğu satır sayısı) / (A'nın bulunduğu satır sayısı) formülü kullanılır. Destek değerini hesaplamak için ise (A ve B'nin bulunduğu satır sayısı) / (Toplam satır sayısı) formülü kullanılır (2, 67).

Birliktelik kurallarının genetik veri üzerinde kullanımına dair literatürde birçok çalışma mevcuttur. Mikro dizin verilerinden sık görülen sinyal yolaklarının tespiti vb. gibi çalışmalar buna bir örnek teşkil edebilir (68).

4.4.7.2.4. İstisna Analizi (Outlier Analysis)

Bir veritabanında verilerin genel dağılımlarına uygunluk göstermeyen farklı veriler bulunabilir, bunlara sıra dışı (outlier) veriler denir. Veri madenciliği tekniklerinin bir çoğu istisnaları gürültü ve ya aşırı durumlar olarak değerlendirir ve göz ardı eder. Ancak bazı durumlarda istisna veriler diğer verilere oranla daha çok bilgi içerir. İstisnaları analiz etmek için iki yöntem kullanılır (2).

İstatistik Tabanlı Yöntem: Verilerim dağılım analizi ya da standart sapma hesabı gibi istatistiksel metotlarla muhtemel istisna noktalar belirlenir. Fakat büyük miktarlarda veriler üzerinde hesaplama gücü gerektirdiği için performansları limitlidir.

Yoğunluk Tabanlı Yöntem: Bu yöntemde her noktanın çevresindeki komşuları ile olan uzaklığı Öklit uzaklığı kullanılarak hesaplanır. Böylece yeterince yakın olmayan noktalar yani istisnalar belirlenmektedir.

4.4.7.2.5. Evrimsel Analiz (Evolution Analysis)

Nesnelerin davranışlarında zamanla değişen düzenlilik ve eğilimlerini modelleyen evrimsel analiz, zaman serileri analizi, ardışık veya periyodik örüntü eşleştirme, benzerlik analizi gibi teknikleri kullanır. Farklı yöntemlerle birlikte kullanılsa da asıl amacı verilerin zamanla ilişkisini analiz emektir (2).

4.4.7.2.6. Tanımlama ve Ayrımlama (Characterization and Discrimination)

Gösterdikleri özelliklere göre genelleştirilmiş sınıflara ayrılan veriler, küme elemanlarının ortak özellikleri ya da veri kümesinin diğer kümelerden farklılıklarını ortaya koyacak şekilde yapılmaktadır.

Tanımlama: Bir veri kümesinin içerdiği elemanların sahip olduğu genel özelliklerin özeti şeklindedir. Bu işlemin sonuçları veri kümelerini görselleştiren pasta grafikler, çubuk grafikler, eğriler gibi çok çeşitli şekillerde olabilir.

Ayrımlama: Bir veri kümesinin sahip olduğu elemanların genel özelliklerini, diğer veri kümelerinin elemanları ile karşılaştırarak sonuçlar ortaya çıkarılması işlemidir.

4.8. Olasılık Teorisi

Bir olayın gerçekleşip gerçekleşmeyeceğinin matematiksel değeri veya olabilirlik yüzdesine olasılık denir (69). İki farklı olaydan birinin diğerinin olmasını ya da olmamasını etkilemediği durumlarda bu olaylara bağımsız olaylar denir (70).

$$P(A \cap B) = P(AB)$$

olarak ifade edilir.

$$P(A \cap B) = P(A)P(B)$$

şeklinde gösterilebilir.

Koşullu olasılık, bir olayın gerçekleştiğinin bilinmesi durumunda diğer olayın olma olasılığıdır (71). P(A|B) biçiminde gösterilir ve B olayı gerçekleştiğinde A olayının olması olasılığı biçiminde okunur.

Bir örnek ile açıklayacak olursak: Bir firmada çalışan 100 kişiye, üst düzey yöneticilere çok yüksek ücretler ödenmesini onaylayıp onaylamadıkları sorulmuş ve aşağıdaki tabloda verilen sonuçlar elde edilmiştir.

Tablo 1. Çalışanların verdiği cevaplara göre iki yönlü sınıflanması (71)

Cinsiyet	Onaylıyor	Onaylamıyor	Toplam
Erkek	15	45	60
Kadın	4	36	40
Toplam	19	81	100

Burada tablo incelenecek olursa seçilen herhangi bir çalışanın erkek ya da kadın, onaylıyor ya da onaylamıyor olması mümkündür. İşte bu dört olayın olasılıklarına bileşen ya da marjinal olasılık denir. Bu basit olasılık satır ya da sütün toplamlarının genel toplama bölünmesi ile elde edilir.

Yukarıdaki probleme ilişkin marjinal olasılıkları hesaplayacak olursak:

$$P(Erkek) = \frac{\text{Erkeklerin sayısı}}{\text{Tüm çalışanların sayısı}} = \frac{60}{100} = 0{,}60$$

$$P(Kadın) = \frac{40}{100} = 0{,}40$$

$$P(Onaylıyor) = \frac{19}{100} = 0{,}19$$

$$P(Onaylamıyor) = \frac{81}{100} = 0{,}81$$

Sonuçları elde edilir.

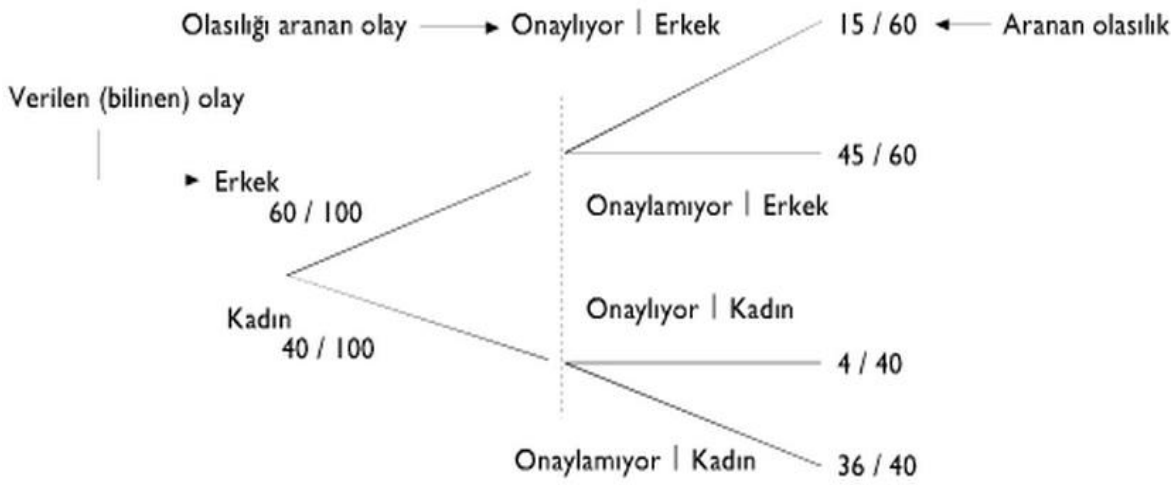
Buna göre 100 çalışana ait bu sonuçlardan rastgele seçilen bir kişinin erkek olduğu biliniyorsa onaylama olasılığını bulalım. P(Onaylıyor | Erkek) = ?

Yukarıda elde ettiğimiz bilgilerden faydalanarak aranan P(Onaylıyor | Erkek) koşullu olasılığı;

$$P(Onaylıyor \mid Erkek) = \frac{\text{Onaylayan Erkeklerin sayısı}}{Toplam\ Erkek\ sayısı} = \frac{15}{60} = 0{,}25$$

olarak bulunur.

Aynı çözümü bir ağaç diyagramı üzerinde göstererek anlaşılmasını kolaylaştırabiliriz.



Şekil 6. Koşullu olasılıklar için ağaç diyagramı (71)

4.8.1. En Çok Olabilirlik (Maximum Likelihood)

İstatistik ve ekonometride en sık kullanılan nokta tahmin yöntemlerinden biri olan en çok olabilirlik fonksiyonu; belli bir örneklem değerlerinin gerçekleşme olabilirliğini en yüksek yapan anakütle parametrelerini bulmaya çalışır. En çok olabilirlik tahmin edicileri olabilirlik fonksiyonunu en yükseğe çıkaran tahmin ediciler olarak tanımlanır. Başka bir ifade ile rastsal bir olayın gerçekleşmesi, o olayın gerçekleşme olasılığı en yüksek olay olmasındandır. Ki-kare testi, Bayesci yöntemler ve çeşitli ölçüt modelleri gibi birçok istatistiksel çıkarım yöntemi, temelde en çok olabilirlik yaklaşımına dayanmaktadır (72).

4.8.2. Bayesci Yaklaşım ve Bayes Teoremi

Günümüzde istatistik sosyal bilimlerde, tıp ve eğitim bilimleri gibi temel bilimler ve daha birçok alanda yaygın olarak kullanılmaktadır. İstatistikte yer alan önemli konulardan birisi de karar teorisidir. İstatistiksel karar teorisi, karar alıcının karşısındaki alternatif durumlar arasından tercih yapabilmek için uygun bilgiyi kullanmasına dayanan istatistiktir (73).

Bayesci yaklaşım, belirsizlik halinde karar alma durumunda alternatif hareket seçeneklerinin sonuçlarının tahmini olarak bilinmektir. Durumlar hakkında tam bir belirsizlik ve bilgisizlik söz konusudur. Kısmi ihtimal durumunda karar almada, karar alıcının durumların gerçekleşme olasılıkları hakkında bilgi sahibi olduğu veya kısmi bir

ihtimalin olduğu varsayılmaktadır. Bayesci yaklaşımda, problemin formüle edilmesi ve durumların gerçekleşme olasılıklarının saptanmasında karar alıcının konu ile ilgili bilgisi, geçmiş tecrübeleri ve sezgisinin sayılarla ifadesi olan olasılıklar önemli kısmı teşkil eder ve bu olasılıklara sübjektif olasılıklar denir. Bu tip olasılıklarla çalışan olasılık kuramına sübjektif olasılık kuramı ya da Bayesci istatistik denilmektedir (74).

Bayesci yaklaşım, mevcut bütün verinin ve olasılıkların kullanılmasına imkan vererek belirsizliklerin miktarını azaltmaya çalışır. Olasılık belirsizliğin bir diğer ifadesi şeklinde düşünülebilir. Sadece belirsizliğin oranı değişerek olasılığı artırır ya da azaltır. Birçok durumda karar aşamasından önce yeni bilginin daha önce bilinen bilgi ile birleştirilebildiği bir yöntem gerekmektedir. Böylelikle verilecek karar sahip olunan bilgilerin tümüne dayandırılabilecektir. Bayes Teoremi, elde edilebilecek ek bilgilerin daha önceden bilinen bilgiler ile birleştirilerek karar almada kullanılmasına imkan sağlar (75). Bayesci yaklaşım için kısaca, bilinene şartlı olarak bilinmeyen hakkında bir olasılık ifadesi türetmektir denilebilir.

Matematiksel istatistiğin önemli bir kavramı olan Bayes teoremi, koşullu olasılıklar ile marjinal arasındaki ilişkiyi göstermek için kesinlik içermeyen bir bilginin tahmininde gözlemleri ve sübjektif verileri kullanan ve herhangi bir durumu modellemek için evrensel doğruları ve gözlemleri kullanarak sonuçlar üretmeyi hedefleyen istatistiksel bir yöntemdir (76, 77).

4.9. Bayesci Ağların Temel Kavramları

Koşullu bağımsızlık ve Grafiksel modeller kavramları Bayesci ağları daha iyi anlamayı sağlayan önemli başlıklardır.

4.9.1. Koşullu Bağımsızlık

Koşullu bağımsızlık, grafik teorisi için oldukça önemli bir kavramdır. Grafikte bulunan düğümle arasında kenar yapılarının oluşturulması ve değişkenler arası ilişkilerin ortaya çıkarılması gibi konularda koşullu bağımsızlık ilişkilerinden faydalanılır. Koşullu bağımsızlık, değişkenlerin çok değişkenli olasılık dağılımının hesaplanmasında kullanılan parametre sayısını indirgemeyi sağlar. Örneğin, elimizde A, B ve C gibi üç farklı olay olduğunu varsayalım. $P(C)\neq 0$ olmak üzere, C olayı biliniyorken A ve B olayları birbirinden bağımsız ise, bu duruma C olayı verildiğinde A ve B olayları koşullu bağımsızdır denir. İstatistiksel perspektiften yorumlandığında,

koşullu bağımsızlık bulunuyorken B olayında meydana gelecek bir değişiklik A olayını etkilemeyecektir (78).

4.9.2. Grafiksel Modeller

Biyoistatistik, biyoinformatik, ekonometri, sosyometri vb. gibi istatistiksel metotların uygulandığı birçok alanda değişkenler arasındaki ilişkisel yapıların belirlenmesinde kullanılan grafiksel bir yöntemdir. Grafiksel modeller, değişkenlerin aralarındaki ilişkilerin tespit edilmesi ve bu ilişkilerin yorumlanması için gereken birçok istatistiksel ölçüt ve karmaşık yapıyı kolaylaştırarak değişkenler arasındaki belirsizlikleri azaltıp daha anlaşılır ve yalın bir hale getirmektedir. Grafiksel modeller iki türden oluşmaktadır. Bunlar; yönsüz grafikler ve yönlü grafikler (78).

Yönsüz Grafikler

Yönsüz grafikler düğümler arasındaki kenarların herhangi bir yön belirtmediği grafiklerdir. İki düğüm arasında bulunan yönsüz bir kenar yalnızca iki değişken arasındaki ilişkinin varlığını gösterir. X ve Y şeklinde iki düğüm bulunduğunu varsayalım. X ve Y değişkenleri arasındaki ilişki X–Y şeklinde gösterilir. Fakat hangi değişkenin hangisini etkilediği hakkında bir bilgi vermez (78).

Yönlü grafikler

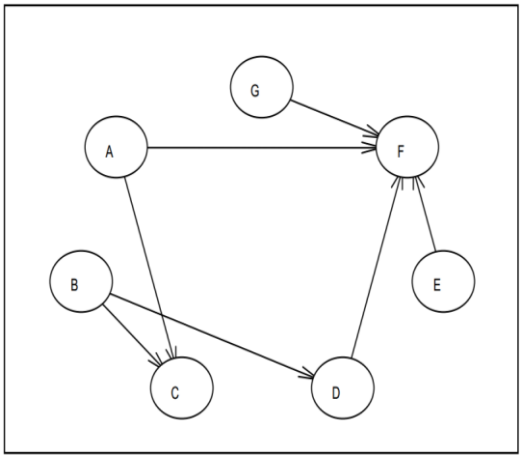
Düğümler arasındaki kenarlar yön belirtiyorsa bu tip grafiklere yönlü grafikler denir. Yönlü grafiklerdeki düğümler arası kenarların yönleri nedensellik ilişkilerini ifade eder. X→ Y şeklindeki bir yönlü kenar X'ten Y'ye doğru bir nedensellik ilişkisinin bulunduğunu göstermektedir. Yani bu yönlü kenar ilişkisi incelendiğinde X değişkeninin Y değişkenini etkilediği ya da X, Y'nin nedenidir şeklinde yorumlanabilir. Burada nedensellik, bir değişkenin farklı durumlarında değişme gözlendiğinde bir diğer değişken için de değişimin söz konusu olmasıdır (78).

4.9.3. Bayesci Ağlar

Bayeci ağlar, birden fazla değişken içeren veri kümeleri için değişkenler arası nedensellik ve koşullu bağımsızlık ilişkilerini görsel olarak tasvir eden grafiksel modellerdir. Bir Bayesci ağı oluşturan üç ana bileşenden söz edilebilir. Bunlar;

rastlantısal değişkenler kümesi, yönlü döngüsel olmayan grafik ve yerel olasılık dağılımlarının çarpımsal şekliyle ifade edilen bir ortak olasılık dağılımıdır (78).

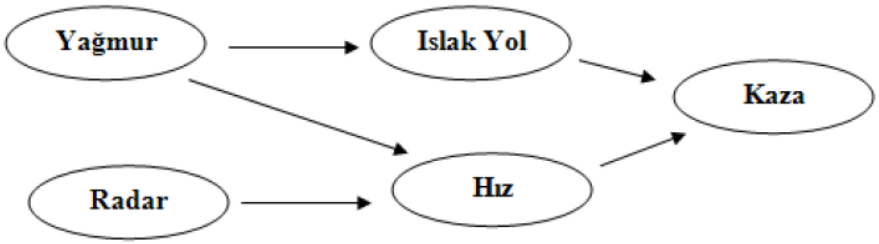
Bayesci ağlarda bulunan her değişken bir düğüme karşılık gelir. İki düğüm arasında bulunan yönsüz bir kenar sadece değişkenler arasındaki ilişkinin varlığını gösterir. Fakat yönlü bir kenar bulunuyorsa, değişkenler arasında nedensellik ilişkileri olduğu anlamına gelmektedir. Bayesci ağlarda düğümler arasındaki ilişkiler Şekil 7'de görüldüğü gibi ebeveyn, çocuk ve soy dışı olarak adlandırılırlar (78).



Düğüm	Ebeveyn	Çocuk	Soy Dışı
A	Ø	C,F	B,D,E,G
B	Ø	C,D	A,E,F,G
C	A,B	Ø	D,E,F,G
D	B	F	A,C,E,G
E	Ø	F	A,B,C,D,G
F	A,D,E,G	Ø	A,B,C
G	Ø	F	A,B,C,D,E

Şekil 7. Bayesci ağ yapısında düğümler arası ebeveyn, çocuk, soy dışı ilişkileri

Bir Bayesci ağın yapısı mevcut veriler üzerinden öğrenildikten sonra ağın içerisindeki koşullu olasılık değerlerini tahmin etme sürecine, parametre öğrenme denir. Parametre öğrenme sürecinde en yüksek olabilirlik tahmini literatürde oldukça yaygın bir biçimde kullanılmaktadır. Bu yöntem Bayesci ağlardaki her koşullu olasılık değeri için en yüksek olabilirlik tahminini elde etmeyi esas alır. Bayesci tahmin metodunda her olasılık dağılımı için önsel dağılımı kullanarak sonsal bir dağılım oluşturulur. Elde edilen sonsal dağılımdan faydalanarak parametre tahmini gerçekleştirilir. Bayesci ağlarda parametre tahmini için kullanılan çeşitli öğrenme algoritmaları bulunmaktadır (79). Örnek bir ağ yapısı Şekil 8'de verilmiştir.



Şekil 8. Bayesci ağ yapısı örneği

Tablo 2. Bayesci ağ koşullu olasılıklar tablosu

Kaza	Islak Yol	Hız	p(Kaza/Islak Yol, Hız)
Evet	Evet	Yüksek	0.18
Evet	Evet	Düşük	0.04
Evet	Hayır	Yüksek	0.06
Evet	Hayır	Düşük	0.01
Hayır	Evet	Yüksek	0.82
Hayır	Evet	Düşük	0.96
Hayır	Hayır	Yüksek	0.94
Hayır	Hayır	Düşük	0.99

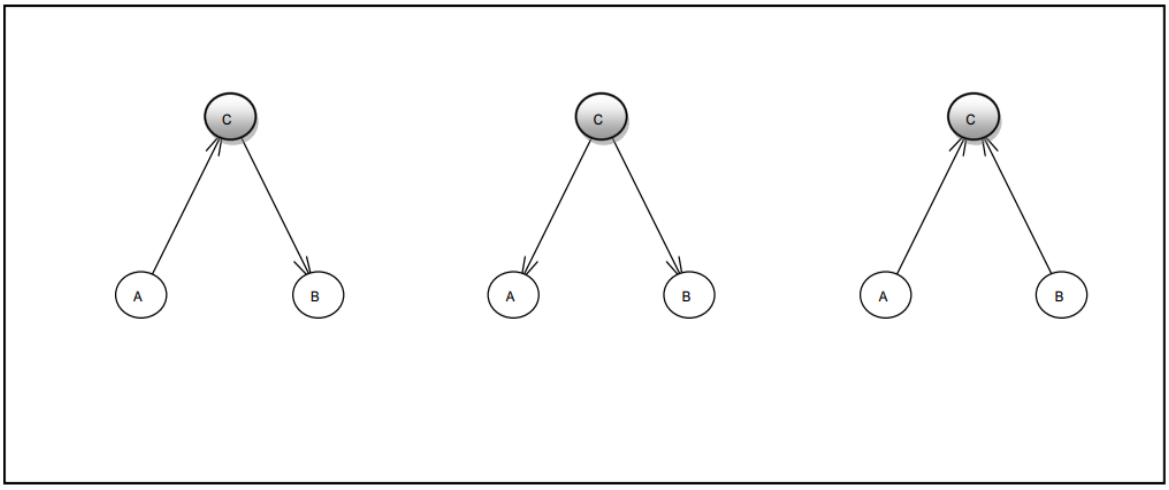
4.9.4. D-ayrılık

Bayesci ağlarda düğümler arasındaki koşullu bağımsızlıkları ortaya çıkarmak için kullanılır. Örneğin, X, Y, Z şeklinde çeşitli değişkenlerden oluşan düğümler olduğunu varsayalım. X ve Y düğümleri arasında bulunan herhangi bir yolun içerisinde bulunan bir v düğümü için iki durum geçerlidir. Bu durumlardan biri sağlanıyorsa Z, X'i Y'den d-ayırmaktadır (80). Bu durumlar:

1. v düğümünün yakınsayan kenarları varsa ve v düğümü ya da torunları Z kümesi içerisinde değilse
2. v düğümünün yakınsayan kenarları varsa ve v düğümü ya da torunları Z kümesi içerisindeyse

Burada bahsedilen yakınsama kavramı, bir düğüme karşılık gelen iki yönlü kenarın bulunmasıdır. Başka bir deyişle bir düğümün en az iki ebeveyni buluyorsa, o

düğüm yakınsayan kenarlar içermektedir. Olasılıksal olarak Z'nin X'i Y'den d-ayırması, Z bilindiğinde X ve Y düğümlerinin koşullu bağımsız olduklarını ifade eder.



Şekil 9. Bayesci ağlarda koşullu bağımsızlık ilişkilerini gösteren d-ayırılık kuralları

Şekildeki Bayesci ağ örneklerinde koşullu bağımsızlık ilişkileri gösterilmektedir. En soldaki örnek C düğümü bilindiğinde A ve B düğümlerinin koşullu bağımsız olduğunu ifade eder. Ortadaki örnekte ise C düğümü bilindiğinde yine A ve B düğümlerinin koşullu bağımsızlığını gösterir. Sağdaki örnekte ise C düğümü bilindiğinde A ve B düğümlerinin koşullu bağılı olduğunu gösterir. Bu yapı yukarıda bahsedilen kurallar ile açıklanabilir. Soldaki ve ortadaki Bayesci ağ örnekleri için A ve B düğümleri arasındaki yolda C kenarı yakınsamadığı için A ve B, C bilindiğinde koşullu bağımsızdır. Fakat sağdaki örnekte C kenarına doğru yakınsayan iki kenar olduğu için A ve B düğümleri arasında koşullu bağımlılık ilişkisi bulunmaktadır (78).

4.10. Model Seçim ve Değerlendirmede Bilgi Kriterleri

İstatistiksel analizlerde en çok karşılaşılan zorluklardan biri de, uygun modeli seçmek, kestirmek ve boyutunu belirlemektir. Bu zorluk modeldeki parametre sayısı ile doğru orantılı olarak artmaktadır. Model değerlendirme ise gözlemlenen verileri daha iyi anlamak amacı ile uygulanmaktadır.

Araştırmacı, modelin istatistiksel tanımlanabilirliği veya değerlendirilmesi olarak isimlendirilen teknik vasıtasıyla modelin kalitesini inceler ve doğru modele erişmek için araştırmalarına daha doğru yön verebilir. Özellikle son yıllarda model

seçimi ya da model değerlendirme alanlarının önemi artmıştır. Bu yöndeki çalışmaların literatürdeki artışı göze çarpmaktadır. Mevcut veri kümesi ile ilgili model kurarken karşılaşılan en önemli problem, farklı modeller arasından en uygun olanının nasıl seçileceğini ortaya çıkarmaktır. Eldeki veri kümesini tanımlamak için karşılaştırılan modellerden birinin seçiminde parametre yalınlığını gösteren kriterler vardır. Yalınlıkta genel kabul, daha basit ve yalın bir modelin, daima karmaşık bir modele tercih edilmesidir. En iyi modelin, en az karmaşık veya en yüksek bilgiye sahip model olduğu kabul edilir (81).

Model seçiminde kullanılan bazı yöntemler şöyledir: Akaike Bilgi Kriteri (Akaike Information Criterion)-AIC, Bayesci Bilgi Kriteri (Bayesian Information Criterion)-BIC, Bayesci-Dirichlet Eşitlik (Bayesian Dirichlet equivalent)-BDE, K2.

Akaike Bilgi Kriteri: Farklı boyutlardaki modellerin karşılaştırılmasında kullanılan güçlü bir model seçme kriteridir. AIC, bir örneklem verisinden elde edilen parametre tahminlerinin farklı örneklemler için de kullanılmasından kaynaklanan doğruluk kaybının bir ölçümü olarak tanımlanabilir. Buna göre AIC, orijinal örneklemden elde edilen parametre kestirimlerinin farklı örneklemler için çapraz-geçerliliğin bir ölçümü olarak da kullanılabilmektedir (82).

Bayesci Bilgi Kriteri: Değişik boyutlara sahip modellerden en iyisini seçme probleminin Bayes yaklaşımıyla çözümünün denenmesidir. AIC ile çok benzerlik gösterir. Farklı olarak modelin iyilik derecesini belirlemekte kullanılan parametreleri kullanılan ceza terimi, parametre sayısındaki değişim ve karmaşıklık arttıkça modele karşı daha sert bir yaklaşım sergiler yani daha yüksek ceza uygular (83).

Bayeci Dirichlet Eşitlik: Bayesci sonsal olasılık kavramı üzerine kurulmuş bu model skorlama metodu, Dirichlet dağılımının önsel olasılıkları maksimize etmeye çalışmasıdır. Birçok noktadan Bayeci bilgi kriteri ile benzerlik gösterir. Modeli cezalandırma prensibini BIC'ten miras almıştır (84).

K2: Bir arama ve skorlama yöntemi olan K2, mevcut başlangıç verileri üzerinden modelin sonsal olasılıklarını maksimize etmeye ve model karmaşıklığını azaltmaya çalışır. Bayesci ağlarda (Bayesian networks) oldukça yaygın kullanılan bir yapısal öğrenme tekniğidir (85).

5. GEREÇ ve YÖNTEM

5.1. Çalışmada Kullanılan Yöntemlerin Detayları

Bu tez çalışması kapsamında gerçekleştirilen çalışmalarda veri madenciliği ve bayesci ağ oluşturma teknikleri kullanılmıştır. Bu teknikler sırasıyla incelenecektir.

5.1.1. Birliktelik Analizi Kuralı

Birliktelik kuralları, sahip olunan veriler incelenerek aynı işlem içerisinde birlikte karşılaşılan nesnelerin tespit edildiği kurallardır. Geçmiş veriler içerisindeki birliktelik kuralları belirlendikten sonra gelecekte karşılaşılması muhtemel olaylar hakkında tahmin imkanı verir. Oldukça yaygın bir kullanım alanına sahip olan bu yöntem birçok alan ve disiplin tarafından verdiği sonuçlar açısından dikkat çekmektedir. Bilinen en yaygın kullanım alanı market sepeti uygulamalarıdır.

Market sepeti analizi, müşterilerin satın aldıkları ürünler gözden geçirilerek satın alma alışkanları ortaya çıkarılır. Bu şekilde marketteki ürünlerin raflara dizilimi, indirimli ürünleri belirlenmesi vs. gibi konularda müşterilerin alışkanlıkları incelenerek, market yöneticileri tarafından etkili satış ve pazarlama politikaları geliştirilebilir. Mağazadaki tüm ürünlerin kümesi evren olarak varsayılırsa, bir işlemde her bir ürünün bulunma ya da bulunmama durumu ikili (boolean) bir değişkenle ifade edilir. Bütün alışveriş sepetleri de bu değerlerden oluşan birer vektör olarak ifade edilir. Hangi ürünlerin sıklıkla birlikte alındığını veya birbiriyle ilişkili olduğunu gösteren satın alma örüntüleri vektörler analiz edilerek elde edilebilir (2).

İlk olarak 1993 yılında Agrawal ve diğerleri market sepeti analizi ile birliktelik kuralları çıkarımını ele almışlardır (86). Çalışmada, X ve Y'nin nesne kümesi $X \Rightarrow Y$ (X birliktelik Y) şeklinde gösterilmiştir. Bu gösterim birliktelik kurallarının matematiksel ifadesidir. Kuralın sol tarafı öncül veya gövde (antecedent) ve sağ tarafı artçıl veya baş (consequent) şeklinde tanımlanmaktadır. Oluşturulan kuralların gücünü göstermek için destek (support) ve güven (confidence) olmak üzere iki ölçüt tanımlanmıştır. Kurallar, kullanıcının amacına göre belirlediği minimum destek (min_destek) ve minimum güven (min_güven) eşik değerleriyle oluşturulur. Market sepeti analizinde, müşteriler tarafından satın alınan ürünler nesnelerdir. Birçok nesneyi içinde bulunduran tek bir satın alma ise işlemi (transaction) ifade eder.

5.1.1.1. Birliktelik Kuralları Temel Kavramlar

Birliktelik kuralları, bir işlemde birlikte gerçekleşen olayların birbiriyle olan ilişkili olan değişkenlerin arasındaki anlamlı bağlantıları analiz eder. Agrawal, Imielinski ve Swami birliktelik kurallarının matematiksel notasyonunu 1993 yılındaki çalışmalarında açıklamışlardır (86).

Bu modele, göre nesneleri $I=\{i_1, i_2, \ldots i_m\}$ şeklinde ikili (binary) özelliklerin kümesi varsayalım. Her bir nesneyi (ürün) bir i temsil etsin. D veritabanındaki her bir hareket (transaction) veya işlem T ile gösterilsin ve $I \supseteq T$ koşulunu yerine getirecek şekilde nesne kümesi (itemset) ifade etsin. Eğer i_k nesnesi alınmış ise t_k = 1, alınmamışsa t_k = 0'dır. Yani, $X \subseteq I$ için, X' teki her bir I_k'ya karşılık gelen t_k değeri t_k=1'dir.

$X \Rightarrow I_j$ şeklindeki bir birliktelik kuralından bahsederken; X, I içindeki bazı nesnelerden oluşan bir kümedir, yani I'nın altkümesi ve I_j ise, I içinde bulunan ve X içinde bulunmayan bir elemandır. Güven faktörü sağlandıktan sonra $X \Rightarrow I_j$ kuralı T işlemler kümesi için geçerli olabilmektedir. Güven değeri (confidence-c) $0 \le c \le 1$ aralığındadır ve T içindeki tüm X'lerin ne kadarının I_j'yi içerdiğinin göstergesidir. Kuralın gücünü güven seviyesi göstermektedir. $X \Rightarrow I_j \mid c$ biçiminde ifade edilebilir.

Güven değeri gibi sayısal bir ifade olan destek değeri, kuralın geçerli olabilmesi için son derece önemli olan diğer bir faktördür. Destek değeri, X ve I_j'nin T içerisinde kaç işlemde birlikte bulunduğunu gösterir.

Bir birliktelik kuralının geçerliliğinden söz edebilmek için öncelikle destek ve güven değerlerinin hesaplanmış olmaları gerekmektedir. Destek, kuralın öncül ve artçılının birlikte olduğu işlemlerin sayısının, veritabanındaki bütün işlemlerin sayısına oranı olarak tanımlanır. Güven ise, kuralın öncül tarafındaki ürün ya da ürünlerin, artçıl tarafındaki ürün ya da ürünlerle birlikte görülme olasılığıdır. Matematiksel olarak $A \Rightarrow B$ kuralı için güven ve destek değerlerini hesaplanması aşağıdaki gibidir (87):

$$\text{Destek } (A \Rightarrow B) = \text{destek } (A \cup B) \tag{5.1}$$

$$\text{Güven } (A \Rightarrow B) = \text{destek } (A \cup B) \,/\, \text{destek } (A) \tag{5.2}$$

Destek ve güven değerleri aldıkları değerlere göre kuralların kullanışlılığını ve olabilirlik oranını ifade eder. Güçlü kurallar, destek ve güven değerleri yüksek olan yani birlikteliğin fazla olduğu kurallardır. Kuralların minimum destek eşiği (min-destek) ve minimum güven eşiği (min-güven) kullanıcı tarafından çalışmanın başında belirlenir. Kurallar oluşturulduktan sonra eşik değerlerinden yüksek destek ve güven değerlerine sahip olanlar dikkate alınır.

Bir marketteki müşterilerden A ürününü satın alanların aynı zamanda B ürününü de beraber aldıklarında çıkarılan bir birliktelik kuralı aşağıdaki gibi ifade edilir.

$A \Rightarrow B$ (destek = %10, güven = %50)

Eşik değerleri kullanıcı tarafından belirlenen bu birliktelik kuralı, destek eşiğinin %10'dan, güven eşiğinin de %50'den büyük olması durumunda geçerlidir ve bu değerler kuralın gücünü göstermektedir. Eşitlik (5.1) ve eşitlik (5.2) dikkate alınarak kural incelenirse, analiz edilen tüm alışverişler içerisinde, A ve B ürünlerinin %10 oranında birlikte satın alındığı ve B ürünü A ürününü satın alanların %50'sin de görüldüğü söylenebilir. Nesnelerin bulunduğu kümeye nesneküme denir ve bir nesneküme k tane elemana sahipse, bu küme 'k-nesneküme' biçiminde gösterilir. Eğer min_destek değerini sağlayan bir nesneküme varsa bu küme yaygın nesneküme (frequent itemset) olarak adlandırılır ve L_k biçiminde ifade edilir (2). Birliktelik kuralı analizi iki aşamadan meydana gelen bir süreç olarak tanımlanabilir (88).

İlk olarak destek değeri minimum eşiğin üzerinde olan tüm nesnekümelerin tespitidir. Bu koşulu sağlayan nesnekümeler yaygın nesnekümeler olarak nitelendirilir. Birliktelik kuralı çıkarımı için kullanılan algoritmaların performansı bu aşamada değerlendirilir.

İkincisi ise belirlenen yaygın kümelerden güçlü birliktelik kuralları çıkarmaktır. Bu aşama yaygın nesnekümeleri yardımıyla min_güven eşiğinin üstünde kalan kuralların tespit edilmesidir.

Birliktelik kurallarının çıkarımında destek ve güven değerlerinin yanı sıra basit bir korelasyon ölçütü olan lift (ilgi, kaldırma) değeri de kullanılır. Destek ve güven değerlerinin kuralın çıkarımında yetersiz kaldığı durumlarda lift değer alınabilir (2).

Market sepeti örneğinden gidecek olursak pekmez alan birisi yüksek bir olasılıkla tahin de alacaktır. Ancak bu ilginç bir kural olarak görülmez çünkü bilinen bir durumdur (89).

$A \Rightarrow B$ (destek, güven, korelasyon)

Destek ve güven faktörlerinin yanı sıra, kuralın korelasyonunun da hesaplandığı durumlarda çıkarılan lift değeri; korelasyon değeri ile kuralın içerdiği neneler arasındaki ilişkileri incelemek için kullanılabilir.

$$\text{Lift } (A \Rightarrow B) = \frac{güven(A \Rightarrow B)}{destek(B)} = \frac{destek(A \Rightarrow B)}{destek(A).destek(B)} = \frac{destek(A \cup B)}{destek(A).destek(B)} \quad (5.3)$$

Lift değerinin 1'den küçük olması, A ile B'nin beraber görülmeleri arasında negatif yönde bir ilişki olduğunu; 1'den büyük olması durumunda, A ile B'nin beraber görülmeleri arasında pozitif yönde bir ilişki olduğu ifade eder. Lift değerinin 1 çıkması durumunda ise iki nesnenin birbirinden bağımsız olduğu anlamına gelir.

Bu kavramları daha anlaşılır olabilmesi için bir örnek üzerinde inceleyelim. Bir bilgisayar mağazasından alışveriş yapan müşteri kayıtlarının 10 bin tanesi analiz edilmiş olsun. Bu kayıtların 7500 tanesinde A ürünü, 6 bin tanesinde de B ürünü satın alınmış olsun. Bunun yanında 4 bin kayıtta ise her iki ürün birlikte alınmış olsun. Bu durum için yapılan analiz de minimum destek değeri %25, güven değeri ise %50 olarak belirlenmiş olsun. Bu veriler ışığında aşağıdaki birliktelik kuralı elde edilmiş olur.

$B \Rightarrow A$ (destek = %40, güven = %66, 0.83)

Destek, güven ve lift değerlerinin hesaplanması ise aşağıda matematiksel notasyonu ile gösterilmiştir.

$$\text{Destek } (B \Rightarrow A) = Destek(B \cup A) = \frac{4000}{10000} = 0.40$$

$$\text{Güven } (B \Rightarrow A) = \frac{Destek(B \cup A)}{Destek(B)} = \frac{0.40}{0.60} = 0.66$$

Lift (B $\Rightarrow$ A) = $\frac{Güven(B \Rightarrow A)}{Destek(A)} = \frac{0.66}{0.75} = 0.88$ olarak bulunur.

Oluşturulan birliktelik kuralının destek ve güven değerleri önceden belirlenen min_destek ve min_güven eşik değerlerinin üzerinde olduğu için güçlü bir kuraldır. Ancak A ürününün destek değeri olan %75, kuralın eşik değeri olan %66'dan büyüktür. Bu durum lift değerinin 1'den küçük çıkmasına neden olur ve kuraldan görüleceği üzere A ve B arasında negatif yönde bir korelasyon vardır (2). Buradan ürünlerden birinin satışı diğerinin satışını olumsuz yönde etkilemektedir. Bu yüzden birbirinden uzak raflarda konumlandırılmalı ve beraber satış indirimi yapılmamalıdır şeklinde yorumlanabilir.

Lift değeri hem öncülün, hem de artçılın destek değerleri dikkate alınarak hesaplanmalıdır. Lift değerinin simetrik bir yapıya sahip olması onun bir dezavantajıdır. Buna göre lift (A $\Rightarrow$ B) ile lift (B $\Rightarrow$ A) arasında bir fark yoktur (89).

Birliktelik kuralları analizinin başarılı ve başarısız olduğu bazı durumlar bulunmaktadır. Başarılı olduğu durumlar; kolay anlaşılır ve yalın sonuçlar vermesi, farklı boyutlardaki verilere uygulanabilmesi ve performansının kayıt sayısı ve kombinasyon seçiminden diğer yöntemler (genetik algoritmalar, yapay sinir ağları vb.) kadar etkilenmemesi başlıca avantajları arasında yer almaktadır. Başarısız olduğu noktalar ise; problemin boyutu arttıkça gerekli hesaplama miktarının üstel olarak artması, veritabanında çok az rastlanan örüntüleri göz ardı etmesi ve eşik değerlerini aşan kural sayısına kısıtlama getirildiği takdirde kullanıcı için önemli bazı kuralların gözden kaçırılması olarak belirtilebilir (2).

5.1.1.2. Birliktelik Kuralı Çıkarımında Kullanılan Algoritmalar

1993 yılında Agrawal ve diğerleri tarafından ortaya atılan birliktelik analizinin daha kolay, hızlı ve verimli bir şekilde yapılabilmesi için geliştirilen değişik algoritmalar vardır. Bu algoritmalar; AIS algoritması, SETM algoritması, Apriori algoritması, AprioriTid algoritması, FP-Growth algoritması ve diğer birçok yöntem bulunmaktadır. Bu tez çalışmasında belirtilen yöntemler arasından Apriopri algoritması kullanıldığı için bu yöntemden detaylı bahsedilecektir.

Apriori Algoritması

1994 yılında Agrawal ve Srikant tarafından geliştirilen apriori algoritması; en iyi bilinen ve en çok kullanılan birliktelik kuralları çıkarımı algoritmasıdır. Kullandığı bilgileri bir önceki adımdan aldığı için "prior" anlamına gelen "Apriori" kelimesi algoritmaya ismini vermiştir. Yaygın nesnekümenin bir özelliği olarak; bir yaygın nesnekümenin tüm alt kümleri de yaygın nesneküme olmalıdır (89).

Apriori algoritması, her bir işlem sonrası bir önceki geçişte oluşturulan yaygın nesnekümeleri bir araya getirerek aday nesnekümeler oluşturur ve aralarından yaygın olmayanları silerek sürekli olarak yaygın nesnekümeler elde eder. Bu işlemleri yapmak için yaygın nesnekümeleri bulan algoritmalar veri üzerinde birçok defa tarama yapmalıdırlar.

Apriori algoritmasında, ilk taramada öncelikle her bir nesnenin destek değeri hesaplanır ve bu değerler daha önceden belirlenmiş olan min_destek eşik değeriyle kıyaslanır. Böylece destek eşik değerinin altında kalanlar elenir ve üzerinde kalanlar yaygın nesnekümeler olarak belirlenir. Apriori algoritması, "Bir yaygın nesnekümenin tüm alt kümeleri de yaygın nesneküme olmalıdır." özelliğini uygularken öncelikle oluşturulan *k* boyutlu bir nesneküme oluşturulur. Bu özelliği göz önünde bulundurarak *k*-1 boyutlu nesnekümeler birleştirilir ve elde edilen alt kümlerden yaygınlık kriterlerini sağlamayanlar silinir. Böylece geriye yalnızca yaygın nesnekümeler kalır.

Apriori nesneküme ya da işlemdeki nesnelerin alfabetik sırada olduğunu kabul ederek işlem yapar. F_k, *k* boyutlu bir yaygın nesneküme ve C_k da nesnekümelerin adayları varsayalım. Apriori algoritması öncelikle mevcut veritabanını bir boyutlu nesnekümeler için tarar ve önceden belirlenen minimum güven eşik değerinin üzerinde olan kayıtları tespit eder. Daha sonra 3 adımdan oluşan bir döngüyü veritabanı üzerinde tekrar eder (90). Bu aşamalar sırasıyla:

1. *k* boyutlu yaygın nesnekümelerden, *k*+1 boyutlu yaygın nesneküme adaylarının oluşturulması, C_{k+1}'in üretilmesi,
2. Veritabanının taranarak her bir aday yaygın nesneküme için destek değerinin hesaplanması,
3. Min_destek değerini sağlayan nesnekümelerin F_{k+1}'e eklenemesidir.

Apriori algoritması F_k yaygın nesnekümesinden C_{k+1} üretmek için Şekil 3'te verilen algoritma yapısındaki üçüncü satırda bulunan apriori-gen fonksiyonunu kullanır. İki aşamalı olan bu fonksiyonun işleyiş yapısı aşağıdaki gibidir.

Birleşme işlemi: *k* boyutlu ve *k*-1 yaygın elemana sahip olan P_k ve Q_k kümeleri birleştirilerek, *k*+1 boyutlu yeni başlangıç için aday yaygın nesnekümeler, R_{k+1} oluşturulur. Bu kümede nesneler $nesne_1 < nesne_2 < \ldots < nesne_k < nesne_{k'}$ şeklinde sıralanır.

$$R_{K+1} = P_k \cup Q_k = \{nesne_1, ..., nesne_{k-1}, nesne_k, nesne_{k'}\}$$
$$P_k = \{nesne_1, nesne_2, ..., nesne_{k-1}, nesne_k\}$$
$$Q_k = \{nesne_1, nesne_2, ..., nesne_{k-1}, nesne_{k'}\}$$

Budama işlemi: *k* boyutlu R_{k+1} nesne kümelerinin yaygın olup olmadığı kontrol edilir ve bu şartı sağlamayan nesnekümeler ayıklanarak yeni C_{k+1} oluşturulur. Bunun nedeni daha önce bahsettiğimiz gibi C_{k+1}'in *k* boyutlu nesnekümelerinden daha yaygın olmayanlar, *k*+1 elemanlı yaygın nesnekümelerin alt kümesi olamazlar.

Algorithm *Apriori*

F_l= (1 elemanlı sık geçen nesnekümeler)
for($k = 1; F_k \neq \phi; k++$) do begin
    C_{k+1} = apriori-gen(F_k); // Yeni adaylar
    for all transactions $t \in$ Database do begin
        C'_t = subset(C_{k+1}, t); // T içindeki adaylar
        for all candidate $c \in C'_t$ do
            $c.count++$;
        end
        $F_{k+1} = \{C \in C_{k+1} \quad | c.count \geq$ minimum support $\}$
    end
end
Answer $\cup_k F_k$;

Resim 1. Apriori algoritması (2)

Apriori algoritmasında *t* işlemi içerisindeki tüm yaygın neseküme adaylarını bulmak için beşinci satırdaki alt küme fonksiyonu kullanılır (Resim 1). Bu işlemlerden sonra Apriori veritabanını tarayarak yalnızca bahsedilen şekilde üretilmiş adayların frekanslarını hesaplar (90). Ayrıca, yaygın nesekümelerin boyutu en fazla k_{max} ise, Apriori veritabanında maksimum k_{max+1} defa tarama yapar.

Apriori algoritması aday kümelerin sayısını azaltmada yüksek başarı gösterir. Fakat; çok fazla yaygın nesneküme olması veya destek değerinin çok düşük tanımlanması gibi durumlar karmaşıklığa yol açabilir. Bu durumlar yüksek sayıda aday küme oluşturulması ve çok sayıdaki aday nesnekümelerin kontrol edilmesi için veritabanının defalarca taranması gibi zorluklar ortaya çıkarabilir. Hesaplamanın zorluğunu kavrayabilmek için gerçekte 100 boyutlu yaygın bir nesnekümenin aday nesnekümelerinin sayısının 2^{100} olduğu unutulmamalıdır.

5.1.2. Bayesci Yaklaşım

Bayeci yaklaşım, model parametreleri için verinin içermiş olduğu objektif bilgi ile kişisel, bilimsel ya da geçmişten gelen ön bilginin birleştirilmesine ve parametrelere ilişkin çıkarımsamaların yapılmasına imkan sağlayan istatistiksel çıkarımsama ve karar verme yoludur (91). Başka bir ifade ile tam bir bilgisizlik ve belirsizlik halinde karar alma durumunda muhtemel hareket seçeneklerinin neticeleri tahmini olarak bilinmesi ve buna göre karar verilmesidir (73). Bilinen tüm verilerin kullanılarak belirsizlik miktarının azaltılmaya çalışılmasıdır (74).

Klasik yaklaşım ve Bayesci yaklaşım, istatistik biliminin gelişim sürecinde son derece öneme sahip iki farklı temel yaklaşımdır. Bu yaklaşımlar karşılaştırıldığında birçok farklılıklarının olmasının yanında, Bayesci yaklaşımın sağladığı avantajların üstünlüğü oldukça fazladır (75, 91, 92).

Bayesci yaklaşımın önemli bir özelliği olan bilinmeyene ait önsel bilginin probleme eklenebilmesidir. Bu şekilde bir yandan somut gözlem ve verilerle ilişkilendirilirken diğer yandan somut olmayan bilgi ve inançlarla ilişkilendirilir. Gözlemsel olmayan ve bilinmeyene ait olan bu bilgi, insanların o konudaki tecrübe, beklenti ve öngörülerine bağlı olarak oluşturulabilir. Böylece mevcut somut ve soyut tüm ipuçları kullanılarak bilinmeyene ulaşılmaya çalışılmaktadır (93). Ayrıca Bayesci

yaklaşım yeni ek bilgilerle karşılaşıldıkça, olasılıkların yeniden hesaplanarak kararların güncellenmesi imkanı sunar (94).

İstatistiksel çıkarımsamaların amacı değişkenler hakkında tahminler yürütmektir. Klasik istatistikte değişkenlerin sabit ve bilinmeyen değerler olduğu ve gözlenen verinin ise rastlantı vektörü olduğu kabul edilir (95).

Klasik istatistikte bilinmeyen değişkenler tahmin etmek için genellikle en çok olabilirlik ya da en küçük kareler metodu kullanılmaktadır. Bayesci istatistik ise hipotez testlerine, nokta ve güven aralığı tahminlerine bir alternatif imkanı sunar. Bayesci istatistik, temel olarak problemle ilgili tüm bilgilerin çözümlemede kullanılması fikrini taşır. Böylece Bayeci uygulamalar, ilgili değişken için çıkarımsama yapmayı hedefler (94).

Klasik yaklaşımda aralık tahminleri, aralığın değişkeni içermesi ihtimali üzerine yapılırken, Bayesci yaklaşım ise değişkene uygun olan aralığa düşmesi ihtimali ile ilgilidir. Bayesci yaklaşımda aralığın içerisindeki her bir noktanın sahip olduğu olasılık yoğunluğu, aralık dışında bulunanlardan daha büyük olduğunda, bu aralığa en yüksek sonsal yoğunluk aralığı denir (94).

5.1.3. Bayes Teoremi

1763 yılında İngiliz rahip ve matematikçi olan Thomas Bayes tarafından yazılan, ancak ölümünden belli bir süre sonra arkadaşı Richard Price tarafından yayınlanan "An Essay Towards Solving a Problem in the Doctrine of Changes" isimli makale günümüzde kullanılan Bayes teoreminin temelini oluşturmaktadır (76, 92, 96).

Matematiksel istatistiğin önemli bir kavramı olan Bayes teoremi, koşullu olasılıklar ile marjinal olasılık arasındaki ilişkiyi göstermek için kesinlik içermeyen bir bilginin tahmininde gözlemleri ve sübjektif verileri kullanan ve herhangi bir durumu modellemek için evrensel doğruları ve gözlemleri kullanarak sonuçlar üretmeyi hedefleyen istatistiksel bir yöntemdir (75, 76).

Koşullu olasılık tanımından çıkarılan Bayes teoreminin A ve B olayları için ifadesi aşağıdaki eşitlikte gösterildiği gibidir.

$$P(A, B) = P(A \cap B) = P(A)P(B|A) \qquad 5.4.$$

Eşitlik 5.4. incelendiğinde P(A, B) = P(A∩B), A ve B olaylarının birlikte gerçekleşme olasılığını, P(A), A olayının gerçekleşme olasılığını ve P(B|A) ise A olayının gerçekleşme olasılığı bilindiğinde B olayının gerçekleşme olasılığını göstermektedir. Bu eşitlikten yola çıkılarak öncelikle P(A) ≠ 0 olmak üzere A (veya B) olayının gerçekleşme olasılığı bilinirken B (veya A) olayının koşullu gerçekleşme olasılığı bulunabilir. Bu olasılık eşitlik 5.5.'te gösterilmiştir.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad \text{veya} \qquad P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad (5.5.)$$

Buradaki bileşik ya da marjinal olasılıklar, çarpım kuralı ile tek bir olay için Bayes formülünde ifade edildiğinden aşağıda gösterilen Bayes teoremini elde etmiş oluruz.

$$\mathbf{P(A|B)} = \frac{P(B \mid A)\,P(A)}{P(B \mid A)\,P(A) + P(B \mid \overline{A})\,P(\overline{A})} \qquad \mathbf{(5.6.)}$$

$$= \frac{P(B \mid A)\,P(A)}{P(B)} \qquad \mathbf{(5.7.)}$$

Eşitlik 5.6.'da kullanılan $\overline{A}$ ifadesi, A'nın tümleyeni olarak kabul edilmiştir. Yani P(A), A olayının gerçekleşme ihtimalini ifade ederken P($\overline{A}$) ise A olayının gerçekleşmeme ihtimalini gösterir. P($\overline{A}$) = 1-P(A) şeklinde de gösterilir.

Bayes teoremi N tane olay için genişletilebilir. Örneklem uzayı U' yu oluşturan N tane olay $A_1, A_2, \ldots, A_N$ ise $A_1 \cup A_2 \cup \ldots \cup A_N = U$ olarak ifade edilir. Bu olaylar ayrık olaylar olduklarından dolayı iki olay aynı anda gerçekleşemez ve olaylardan biri kesinlikle gerçekleşmek zorundadır. Tüm ayrık olay çiftleri $A_i \cap A_k = \emptyset$ biçiminde gösterilir. Buradaki ifadeleri incelersek i = 1, . . . , N, k = 1, . . . , N ve i ≠ k'dır. Bu

durumu Bayes teoreminde yerine yazdığımızda aşağıdaki eşitlikler elde edilmiş olur (97).

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \cdots + P(A_N)P(B|A_N)}$$

$$= \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^{N} P(A_i)P(B|A_i)}$$

Bayes formülünde analize başlamadan önce bilindiği varsayılan, i = 1, . . . , N olmak üzere, $P(A_i)$ marjinal olasılıkları önsel (prior) olasılıklardır ve örneklem uzayını oluşturan olaylar A_1, A_2, . . . , A_N, olmak üzere N tane gözlenemeyen olaylardır. A_i (i = 1, . . . , N) bilindiğinde B'nin koşullu olasılığı $P(B| A_i)$, gözlemlenen A_1, A_2, . . . , A_N olaylarının olabilirlik (likelihood) olarak tanımlanan koşullu olasılığıdır. B bilindiğinde A'nın sonsal (posterior) olasılığı $P(A_i | B)$ biçiminde ifade edilir (80).

Bayes teoremi ya da Bayesci istatistik, verilerin içerdiği objektif bilgi ile önceden bilinen önsel bilgi etkili bir şekilde birleştirilerek istatistiksel çıkarımsamalara daha geniş bir perspektif getirilebilir. Bayesci istatistik, önsel bilgileri formüle ederek değişkenlerin sonsal dağılımını tahmin etmeye çalışır (98).

5.1.3.1. Bayesci Bilgi Güncellemesi

Klasik istatistikte çözüm bulunamayan problemlere Bayesci istatistiğin ardışık yapısının verdiği imkanlar ile çözümler bulunabilir. Ayrıca, bulunan çözümlerin örneklem büyüklüğüne bağlı kalmadan küçük örneklemlerde de yorumlanabilmesi avantajını sunar. Bayesci yaklaşımı diğer metotlardan ayıran en önemli özellik, benzer çalışmalardan elde edilmiş bilgilerin ya da uzman görüşlerinin analiz sürecine katılabileceği fikridir (96).

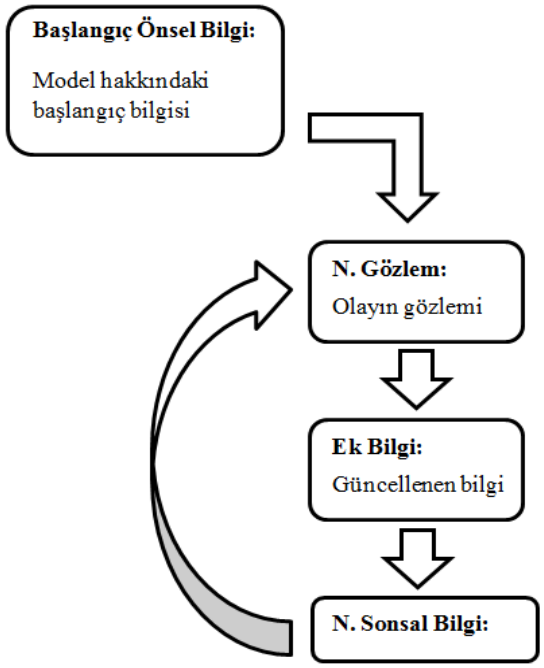
Bayesci yaklaşım, probleme ilişkin önsel bilgiyi ve verilerden elde edilen ek bilgiyi Bayes teoremi ile bir araya getirmektedir. Sürecin başında önsel olasılık dağılımı sübjektiftir ve kişi ve durumlara göre değişkenlik gösterebilir. Bu nedenle probleme ilişkin doğru önsel olasılıkların belirlenmesi önemli bir işlemdir (98).

Bayesci bilgi güncelleme sürecinde, tahmin etmede kullanılan bilgiler matematiksel olasılık dağılımları biçiminde ifade edilir. Bu tahmin etme süreci; önsel dağılım, ek bilgi (olabilirlik) ve sonsal dağılım olmak üzere üç olasılık dağılımından meydana gelir (98).

Önsel (Prior) dağılım; herhangi bir ek bilgi gözlemlenmeden önce eldeki başlangıç bilgisinin dağılımı olarak ifade edilir. Bir a olatının önsel olasılığı P(A) biçiminde gösterilir.

Ek Bilgi (Olabilirlik) (Likelihood) dağılımı; mevcut önsel bilgiyi güncellemek için kullanılan bilginin dağılımıdır ve B ek bilgisi ise bir A olayının olabilirliği P(B|A) koşullu olasılığı ile gösterilir.

Sonsal (Posterior) dağılım; ek bilgi gözlemlendikten sonra eldeki bilginin dağılımıdır. Gözlemlenen B ek bilgisiyle A olayının sonsal olasılığı P(A|B) koşullu olasılığı biçiminde ifade edilir.



Şekil 10. Bayesci Bilgi Güncellemesi Akış Şeması

5.1.4. Bayesci Ağlar

Bayesci ağlar, 1990'lı yılların başarında kullanılmaya başlanmıştır. İnanç ağları olarak da bilinen ve grafiksel modeller ailesine ait olan Bayesci ağlar, veritabanlarındaki rastlantı parametreleri arasındaki olasılıksal ilişkileri kodlamaya yardımcı olur. Bayesci ağlar, araştırılan problemin kesin olmayan tanım kümesi hakkındaki bilgiyi ifade etmek için kullanılır. Sahip oldukları nedensel ve olasılıksal özelliklerden dolayı, veri bilgisi ve uzman görüşleri bu ağlar vasıtasıyla kolaylıkla birleştirilebilir (99).

Olasılıksal grafik modellerinde düğümler (nodes) rastlantı değişkenlerini ve düğümler arasındaki bağlar (edges) ise rastlantı değişkenleri arasındaki olasılıksal bağımlılık durumlarını ifade eder. Bu bağımlılıkların analizinde çoğunlukla bilinen istatistiksel ve sayısal yöntemler kullanılır. Değişkenler arasındaki bağların yönlendirilmemiş olduğu grafiksel modeller "Markov ağları" olarak isimlendirilir. Eğer değişkenler arasındaki bağlar yönlendirilmiş ise bu tür grafiksel modellere Bayesci ağlar denir (99).

Bayesci ağlar, yönlendirilmiş düz grafik (directed acyclic grap – DAG) olarak bilinen ve istatistik, makine öğrenmesi, yapay zeka vb. gibi birçok alanda oldukça yaygın bir biçimde kullanılan grafiksel bir model yapısına sahiptir (99). Bayesci ağlarda öğrenme için geliştirilen yöntemler henüz oldukça yeni olmasına rağmen birçok veri modelleme probleminde verdiği yüksek başarılarla dikkat çekmektedir (100).

5.1.4.1. Bayesci Ağlarda Temel Kavramlar

Bu bölümde Bayesci ağlarda kullanılan temel kavramların kısaca ele alınacaktır.

5.1.4.1.1. Yönlendirilmiş Döngüsüz Grafik

YDG yapısını oluşturan düğümler kümesi ve yönlendirilmiş bağlar kümesi şeklinde iki küme bulunmaktadır. Rastlantı değişkenlerini temsil eden düğümler, daire biçiminde gösterilir. Bağlar ise değişkenler arasındaki doğrudan bağımlılıkları ifade eder ve düğümler arasında çizilen oklarla gösterilir.

Bir YDG üzerinde X_i ve X_j iki düğüm olmak üzere, X_i düğümünden X_j düğümüne doğru çizilen bir bağ, bu değişkenler arasındaki istatistiksel bağımlılığı ifade eder. Böylelikle, iki düğüm arasındaki bağ, X_j'nin aldığı bir değerin X_i'nin aldığı bir

değere bağlı olduğunu gösterir. Başka bir deyişle, X_i'nin X_j'yi etkilediği sonucu çıkarılabilir. YDG üzerindeki bu yapıda X_i düğümü, X_j düğümünün ebeveyni (parent); X_j düğümünün ise X_i düğümünün çocuğu (child) biçiminde tanımlanır. Bunun yanında ağdaki bir düğümden çıkan yol üzerinde yer alan düğümler kümesine soy ya da torun düğümler (descendants), bir düğüme gelen yol üzerinde yer alan düğümler ise ata (ancestors) düğümler olarak adlandırılır. YDG yapısı, herhangi bir düğümün kendisinin soy ya da aya düğümü olmasına izin vermez. Bunun nedeni ise düğümlerin çok değişkenli olasılık dağılımlarının çarpanlara ayrılması içindir (99, 101).

5.1.4.1.2. Koşullu Bağımsızlık

Bayesci ağlarda değişkenler arasında geçerli olan koşullu bağımsızlık özelliği, bir değişkenin ebeveynleri hakkında bilgi sahibi olunduğunda ya da durumları bilindiğinde bu değişkenin, soy değişkenleri dışındaki değişkenlerden bağımsız olması şeklinde tanımlanmaktadır. Koşullu olasılık, çok değişkenli olasılık dağılımının hesaplanmasında kullanılan değişken sayısının azaltılmasına yardımcı olur. Değişken sayısındaki bu azalma, Bayesci ağa yeni bir bilgi girdiğinde sonsal olasılıkların hesaplanmasını oldukça kolaylaştırır (99).

5.1.4.1.3. Koşullu Olasılık Dağılımı / Tablosu

Bayesici ağların nicel kısmı olarak da tanımlanan, ağdaki her bir düğüme ait koşullu olasılık dağılımları olarak ifade edilir. Bu koşullu olasılıklar her bir değişken için yalnızca ebeveynlerine bağlı olarak tanımlanır ve değişkenler kesikli olduğunda, koşullu olasılık dağılımları tablo şeklinde gösterilir. Bu tablolar, her bir değişkenin ebeveynlerinin aldığı değerlere göre değerleri ve bu değerleri alma olasılıkları bulunur. Bu şekildeki tablolara koşullu olasılık tablosu (conditional probablity tables) ismi verilir (99, 102).

5.1.4.1.4. Bayesci Ağların Tanımı ve Özellikleri

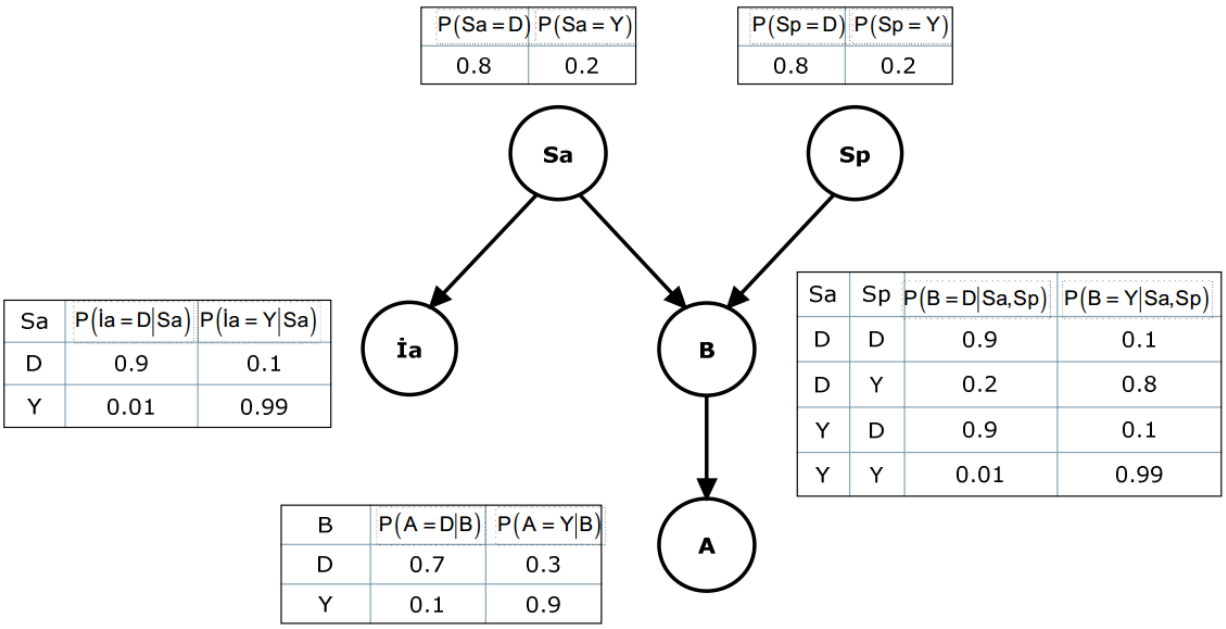
Bu kısımda bir Bayesci ağın yapısı oluşturan temel özellikleri ve matematiksel notasyonu üzerinde durulacaktır. Bir Bayesci ağ, $V = \{X_1,\ldots,X_n\}$ biçiminde rastlantı değişkenleri kümesine dair çok değişkenli olasılık dağılımını gösteren bir DAG'dır. Bayesci bir ağ, G ve Θ olmak üzere iki kısımdan oluşur ve $BN = (G, \Theta)$ şeklinde ifade edilir. Bayesci ağların ilk bileşeni G, düğümlerin $X_1,\ldots,X_n$ rastlantı değişkenlerini ve düğümler arasındaki bağlar ise bu değişkenler arasındaki doğrudan bağımlılıkları

gösteren bir grafik yapısını belirtmektedir. İkinci bileşen olan Θ ise ağdaki her bir X_i rastlantı değişkeniyle ilgili koşullu olasılıkların dağılımını ifade etmektedir. Bir X_i rastlantı değişkeni için koşullu olasılık dağılımı, X_i'nin G'deki ebeveynlerinin kümesi π_i olmak üzere $\theta_{X_i|\pi_i} = P_{BN}(X_i|^{\pi_i})$ biçiminde ifade edilir. Bayesci bir ağda rastlantı değişkeninin ebeveyni yoksa bu rastlantı değişkeni için koşullu olasılık dağılımı, marjinal olasılık dağılımına eşittir. Değişkenlerin her birine ait marjinal ya da koşullu olasılık dağılımlarına genel bir ifade ile yerel olasılık dağılımları (local probability distributions) denir. Bayesci ağın yapısı ve bu değişkenlerden faydalanarak, $V = \{ X_1,\ldots,X_n \}$ için çok değişkenli bir olasılık dağılımı tanımlanır. Bu çok değişkenli olasılık dağılımına zincir kuralı (chain rule) denir ve aşağıdaki eşitlikten elde edilir (102, 103).

$$P(\cap_{k=1}^{n} X_k) = \prod_{k=1}^{n} P(X_k \mid \cap_{j=1}^{k-1} X_j) \qquad (5.10.)$$

Bayesci ağlarda, bir düğümün ifade ettiği değişken gözlenmiş ise bu düğüm bilgi düğümü (evidence node); gözlenmemiş ise gizli düğüm (hidden / latent node) olarak isimlendirilir (99).

Bayesci ağlar ve özelliklerinin daha iyi anlaşılabilmesi için somut bir örnek takibeden ölümde verilecektir. Problem olarak bir kişinin belinin incinmesine sebep olan olayları inceleyelim. Bazı parametreleri kısaltarak kullanalım. Bel incinmesi olayı "Bel (B)" değişkeni ve bel incinmesinin neden olduğu bel ağrısı "Ağrı (A)" değişkeni ile gösterilsin. Bu bel incinmesinin nedeni bir spor egzersizinin yanlış uygulanması olabilir. Bu olayı "Spor (Sp)" temsil etsin. Başka bir neden ise kişinin işyerinde kullandığı sandalyenin ortopedik olmaması ya da konforsuz olması olabilir. Bu durum "Sandalye (Sa)" ile gösterilsin. Aynı problem söz konusu olduğunda, bu kişinin iş arkadaşlarında da benzer bel sorunlarına sahip olup olmadıkları araştırılabilir. Bu ise "İş Arkadaşları (İa)" olarak temsil edilsin. Bu problemdeki tüm değişkenler iki (binary) yapıya sahip ve "Doğru (D)" ve "Yanlış (Y)" değerlerini almaktadır (104). Bu problem için oluşturulan Bayesci ağ Şekil 12.'de verilmiştir.



Şekil 11. Sırt incinmesi örneği için oluşturulan Bayesci ağ

Tüm düğümler için koşullu olasılık tabloları ilgili düğümün yanında gösterilmiştir. Örneği incelediğimizde Sa ve Sp, B düğümünün ebeveynleri olarak belirtilmiştir. A düğümü B düğümünün çocuğu ve Sa, İa'nın ebeveynidir. Burada koşullu bağımsızlık varsayımından faydalanarak Bayesci ağ yorumlanabilir. Örneğin, Sa ve Sp değişkenleri başlangıçta marjinal olarak bağımsızdır fakat B değişkeni verildiğinde bu değişkenler koşullu bağımlı hale gelir. Sa değişkeni verildiğinde ise İa ve B değişkenleri koşullu bağımsızlık gösterirler. A değişkeni B bilindiğinde Sa ve Sp değişkenlerinden koşullu bağımsız olur. Bayesci ağlardaki koşullu bağımsızlık özelliği, zincir kuralından faydalanılarak değişkenlerin çok değişkenli olasılık dağılımının daha yalın bir biçimde ifade edilmesini sağlar. Yukarıdaki Şekil 12'den örnek verecek olursak:

$$P(Sa,Sp,İa,B,A) = P(Sa)P(Sp)P(İa|Sp,Sa)P(B|İa,Sp,Sa)P(A|B,İa,Sp,Sa)$$

şeklinde gösterilebilen çok değişkenli olasılık dağılımı, koşullu bağımsızlık özelliği yardımıyla

$$P(Sa,Sp,İa,B,A) = P(Sa)P(Sp)P(İa|Sp)\ P(B|İa,Sp,Sa)P(A|B)$$ biçiminde yazılabilir.

Böylece modeldeki parametre sayısı 2^5-1 = 31'den 10'a indirgenmiş olur. Parametre sayısındaki bu azalma modeldeki çıkarımsamaların, hesaplamaların ve öğrenmenin gerçekleştirilmesinde oldukça büyük kolaylık sağlar. Ayrıca daha az parametreye sahip olan model yanlılık ve varyans etkilerine karşı daha dayanıklı olacaktır (99).

5.1.4.1.5. D-Ayrılık

D-ayrılık, Bayesci ağlar ile modellenen çok değişkenli olasılık dağılımlarının yapısını incelemek için kullanılan grafiksel modellerin bir özelliğidir. Bu özellik Markov koşulundan faydalanılarak elde edilen koşullu bağımlılık ve bağımsızlık ilişkilerini ortaya çıkarmak için kullanılır (99). Bayesci ağlarda, ağa yeni bir bilgi dahil olduğunda ağdaki değişkenlerin bağımsızlık durumlarına karar vermek için d-ayrılık özelliğinden faydalanılır (102).

A ve B nedensel bir ağda yer alan iki değişken olmak üzere, bu değişkenler arasındaki tüm yolları kapatan, kesen ya da bloke eden bir C değişkeni bulunuyorsa ve

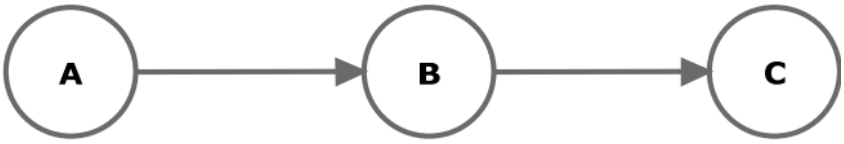
- A ile B arasında yer alan dizisel (serial) ya da yayılan (diverging) bir bağ varsa ve C'nin durumu biliniyorsa ya da
- A ile B arasında birleşen (converging) bağ varsa ve C ya da C'nin soy düğümleri hakkında herhangi bir bilgi bulunmuyorsa,

A ile B değişkenleri d-ayrı (d-seperated) olarak tanımlanır. A ile B d-ayrı biçiminde bir ilişkiş yoksa d-bağlı (d-connected) olarak ifade edilir (102, 105).

Grafiksel modellerde bulunan düğümler dizisel, yayılan ve birleşen bağlar olmak üzere üç şekilde d-ayrılır.

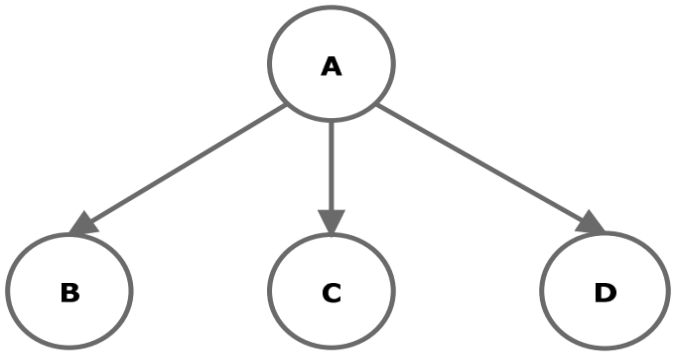
Dizisel bağlar: Şekilde gösterilen biçimde ifade edilen bağ türüdür. Burada A'nın B üzerinde ve dolaylı olarak C üzerinde bir etkisi olduğu sonucu çıkarılabilir. Yani A değişkeni hakkındaki bir bilgi güncellemesi ya da değişikliği dolayısıyla B ve C'nin gerçekleşme olasılıklarını da etkileyecektir. Aynı şekilde C hakkındaki bir bilgi de B ve A'nın gerçekleşme olasılıkların da değişikliğe neden olacaktır. Bununla beraber, B biliniyorsa bu yol kesintiye uğrayıp bloke olur ve A ile C bağımsız hale gelirler. Bu durum "B bilindiğinde A ve C d-ayrıdır" biçiminde ifade edilir. Buradan iki

değişken arasındaki bağlantıyı sağlayan değişkenin ya da düğümün durumu bilinmediği sürece, bilginin bu yol boyunca iletilebileceği sonucu çıkarılabilir (106, 107).



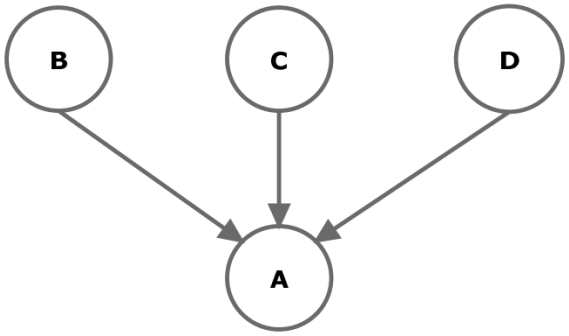
Şekil 12. Dizisel bağ

Ayrılan Bağlar: Şekilde gösterilen bağ türüne göre, A'nın durumu bilinmiyorsa ağa giren yeni bir bilgi A'nın tüm çocukları arasında yayılabilir. A'nın durumu biliniyorsa B, C ve D d-ayrı olarak tanımlanır (106, 107).



Şekil 13. Ayrılan bağ

Birleşen Bağlar: Şekilde gösterilen bağ türünde A hakkında bilinen hiçbir bilgi yoksa ebeveynleri (B, C ve D) d-ayrıdır. Yani ebeveynlerin herhangi birine ait bir bilgi diğer ebeveynlerle ilgili olasılıkları değişikliğe uğratmaz. Ayrıca A değişkeni hakkında bilgi bulunuyorsa ebeveynlerden herhangi biri hakkında sahip olunan bir bilgi diğer ebeveynler hakkında bilgi verebilir. Özetle, çocuk değişken ya da o değişkenin soy değişkenleri hakkında bilinen bir bilgi bulunuyorsa bu bilgi birleşen ağ boyunca iletilebilme özelliğine sahiptir (106, 107).



Şekil 14. Birleşen bağ

5.1.4.2. Bayesci Ağlarda Çıkarımsama

Bayesci ağlar örneklem uzayındaki değişkenlerin çok değişkenli olasılık dağılımını çarpanlara ayırarak tanımlar. Bayesci ağlardan çıkarımsama yapılırken ilgilenilen değişkenlerin marjinal dağılımları hesaplanır. İlgisiz değişkenlerden ise değişkenler üzerinden çok değişkenli olasılık dağılımlarının toplamı alınarak çıkarımsama yapılır. Bayesci ağlarda çıkarımsama yaparken iki yol kullanılır. Bunlardan ilki, tanım kümesinde ilgilenilen bir düğümün ebeveyn düğümleri yardımıyla bağlı olan bilgi düğümlerinden hesaplanan öngörüsel destek (predictive support) değeridir. Bu yöntem yukarıdan aşağıya çıkarımsama (top down reasoning) olarak isimlendirilir. Diğer bir metot ise ilgili düğüm için çocuk düğümleri vasıtasıyla bağlı olan bilgi düğümlerinden hesaplanan tanısal destek (diagnostic support) değeri bulunur. Bu yöntem ise aşağıdan yukarıya çıkarımsama (bottom up reasoning) olarak tanımlanır (108).

5.1.4.3. Bayesci Ağlarda Öğrenme

Bayesci ağ yapısı oluşturulurken ağdaki parametrelerin belirlenmesi için genellikle uzman görüşlerinden faydalanılır. Fakat her zaman araştırmanın problemine ya da veri yapısına uygun uzman görüşüne erişmek mümkün olmamaktadır. Yapılan çalışmaların birçoğunda problem alanı ile ilgili yeterli uzman görüşü elde edilemediği için Bayesci ağların oluşturulmasında veritabanlarından yararlanılmaktadır. Bu sorun Bayesci ağlarda öğrenme problemi olarak adlandırılır ve önsel bilgi verildiğinde ağ yapısının ve parametrelerin tahmin edilmesi olarak ifade edilir. Öğrenme kavram olarak başlı başına geniş ve zengin bir araştırma alanıdır. Bu alanda geliştirilen birçok yöntem

bulunmaktadır fakat yöntemlerin birçoğu her tür veri için uygun değildir. Bu nedenle uygulama aşamasında problemlerle karşılaşmak mümkündür. Ortaya çıkan bu problemlerin çözümleri çoğu zaman bilinmesine rağmen çözüm için gerekli zaman ve hesaplama hızını sağlamak mümkün olmamaktadır. Bayesci ağlarda, ağ yapısının öğrenilmesi, parametrelerin öğrenilmesinden daha zor bir problemdir. Bunun yanında, gizli düğümler ya da kayıp veri gibi kısmi bilgi durumu ile karşılaşıldığında başka zorluklar ortaya çıkmaktadır (99, 102). Bu tür zorluklarla başa çıkabilmek için kullanılan farklı öğrenme algoritmaları bulunmaktadır.

5.1.4.3.1. Bayesci Ağlarda Öğrenme Algoritmaları

Bayesci ağlarda, ağın yapısı öğrenildikten sonra ağın içerisindeki koşullu olasılık değerlerini tahmin etme sürecine parametre öğrenme denir. Parametre öğrenme algoritmalarını genel olarak üç başlık altında toplamak mümkündür. Bunlar en çok olabilirlik tahmini, Bayesci tahmin ve düzgünleştirilmiş tahmindir (78).

Literatürde parametre öğrenme sürecine ilişkin kullanılan en yaygın yöntem en çok olabilirlik tahmini yöntemidir. Bu yöntem Bayesci ağlarda modeldeki her koşullu olasılık değeri için en yüksek olabilirlik tahminine ulaşmaya çalışır. Bayesci tahmin metodu ise tüm yerel olasılık dağılımları için ayrı ayır birer eşlenik önsel dağılım kullanarak sonsal bir dağılım oluşturur. Oluşturulan bu sonsal dağılımdan faydalanarak parametrelerin tahmin edilmesi sağlanır. Düzgünleştirilmiş tahmin yöntemi ise dağınık biçimde sonuçlanan parametrelerin yerel dağılım değerlerini cezalandırılmış bir şekilde tahmin etme yolunu izler (78).

5.1.4.3.1.1. Yapı Öğrenme Algoritmaları

Bayesci ağlarda yapı kavramı düğümler arasındaki nedensellik ve koşullu bağımsızlık ilişkilerini gösteren yönlendirilmiş döngüsüz grafik ile (YDG) temsil edilir. Yapı öğrenme algoritmaları veri seti hakkında herhangi bir uzman görüşü veya önsel bilgiye sahip olunmadığı durumlarda YDG'yi otomatik bir biçimde oluşturmak için kullanılan metotlardır. Özellikle çok miktarda düğüm sayısının bulunduğu Bayesci ağ yapısını oluşturmak için uzman görüşlerini kullanmak uygulamada oldukça fazla zorluklar taşımaktadır. Böyle durumlarda ağdaki yapı öğrenme işlemini otomatik bir biçimde gerçekleştirebilecek algoritmalara başvurulması oldukça kolaylık sağlar. Bir

Bayesci ağın yapısını oluştururken temel hedef mevcut veri yapısını en iyi biçimde temsil eden bir YDG meydana getirmektir (109).

Bu algoritmalar, YDG oluştururken tüm muhtemel Bayesci ağlar için parametreleri hesaplayarak gerçek veriye en uygun olan Bayesci ağı şeçmeye çalışır (91). Fakat veri setindeki değişken sayısı arttıkça YDG sayısı da artar. Robinson tarafından 1977 yılında gerçekleştirilen çalışmada düğüm sayısı ile orantılı olarak oluşacak Bayesci ağ sayısı hesaplanmıştır (110).

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{(n-i)!n!} 2^{i(n-i)} f(n-1) \qquad (5.11.)$$

Düğüm sayısına göre elde edilebilecek tüm Bayesci ağların sayısı ise aşağıdaki Tablo 3'te gösterilmiştir.

Tablo 3. Farklı düğüm sayılarına göre Bayesci ağ sayıları

Düğüm Sayısı	Bayesci Ağ Sayısı	Düğüm Sayısı	Bayesci Ağ Sayısı
1	1	13	$1{,}9.\ 10^{31}$
2	3	14	$1{,}4.\ 10^{36}$
3	25	15	$2{,}4.\ 10^{41}$
4	543	16	$8{,}4.\ 10^{46}$
5	29281	17	$6{,}3.\ 10^{52}$
6	$3{,}8.\ 10^{6}$	18	$9{,}9.\ 10^{58}$
7	$1{,}1.\ 10^{9}$	19	$3{,}3.\ 10^{65}$
8	$7{,}8.\ 10^{11}$	20	$2{,}35.\ 10^{72}$
9	$1{,}2.\ 10^{15}$	21	$3{,}5.\ 10^{79}$
10	$4{,}2.\ 10^{18}$	22	$1{,}1.\ 10^{87}$
11	$3{,}2.\ 10^{22}$	23	$7{,}0.\ 10^{94}$
12	$5{,}2.\ 10^{26}$	24	$9{,}4.\ 10^{102}$

Tabloda görüldüğü üzere bir Bayesci ağdaki düğüm sayısı arttıkça oluşturulabilecek YDG sayısı da yüksek miktarda artış göstermektedir. Bu kadar fazla sayıdaki YDG arasından en iyi seçimi yapabilmek için çok yüksek miktarlarda

hesaplama gücü ve hızı gerekmektedir. Bayesci ağ yapısı bu işlemlerin haricinde uzman bilgisinden faydalanılarak da oluşturulabilir. Genellikle bu tip bilgilerin elde edilmesi çok güç ya da mümkün olmadığı için ağ yapısını oluşturma sürecini otomatize ederek kendiliğinden oluşturabilecek algoritmalara ihtiyaç duyulmaktadır. Bu problemlerin önüne geçebilmek için yapı öğrenme algoritmaları son derece önem arz etmektedir. Bu algoritmaların yardımıyla herhangi bir uzman bilgisi olmadan ve daha az sayıda işlem gücü harcayarak Bayesci ağ yapısı oluşturmak mümkündür (109).

Yapı öğrenme algoritmalara genel olarak üç ana başlık altında incelenmektedir. Bunlar kısıtlama tabanlı, skor tabanlı ve karma algoritmalardır.

5.1.4.3.1.1.1. Kısıtlama Tabanlı Algoritmalar

Kısıtlama tabanlı algoritmalar koşullu bağımsızlık testlerinden faydalanarak ağdaki düğümler arasındaki ilişkileri en iyi şekilde yansıtabilecek yapıyı kurmayı hedefler. Buradaki "kısıt" kavramı, düğümler için belirlenen koşullu bağımsızlıkları ifade etmektedir. Bu algoritma Bayesci ağ içerisindeki her düğüm için Markov örtülerinin çıkarılması esasına dayanır. Markov örtüsü ise bir düğümün ebeveynleri, çocukları ve ebeveynlerinin diğer çocuklarını ifade eder. Yaygın bir bicinde kullanılan çeşitli kısıtlama algoritmaları bulunmaktadır. Büyüme-daralma , artımsal ilişki, aralıklı artımsal ilişki, hızlı artımsal ilişki ve en az en çok ebeveynler ve çocuklar kullanılan algoritmalara örnek olarak verilebilir (111).

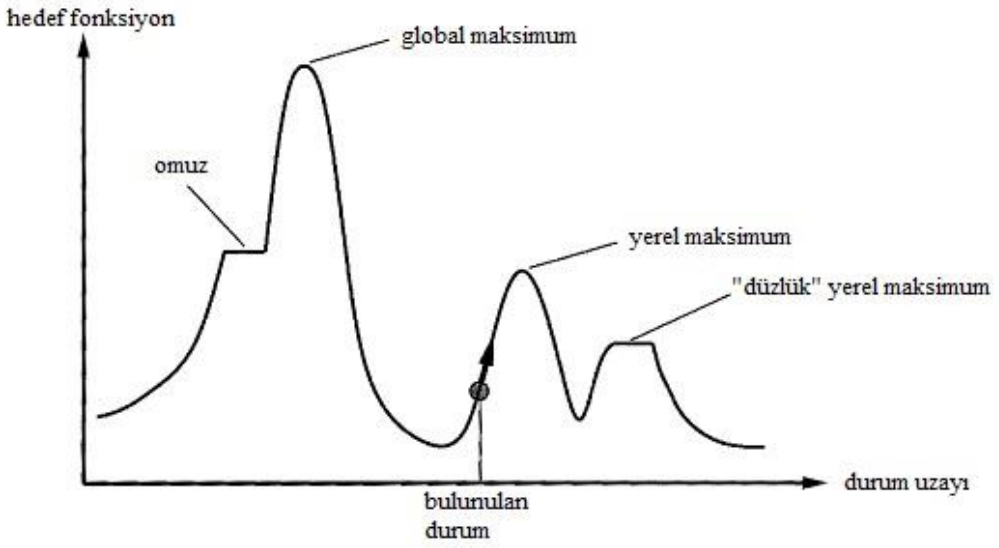
5.1.4.3.1.1.2. Skor Tabanlı Algoritmalar

Çok değişkenli bir veri seti için nedensellik yapısının doğruluğunu ifade eden skor fonksiyonunu maksimize edecek şekilde bir Bayesci ağ oluşturan yöntemler, skor tabanlı algoritmalar olarak adlandırılır. Skor tabanlı algoritmalar bir skor fonksiyonu ve bir arama yöntemi kullanarak veri setine uygun en yüksek skor değerine sahip Bayesci ağ yapısını elde etmeyi amaçlar. Skor tabanlı algoritmalar bu işlemlerden arama için sezgisel bir optimizasyon algoritması kullanır ve skorlama için özel skorlama fonksiyonlarından faydalanır. Bu algoritmalardaki işlevsel alt yapının temeli, buluşsal bir optimizasyon tekniği kullanarak oluşturulan Bayesci ağlar içerisinden en yüksek skor değerine sahip olan yapıyı elde etmeye dayanmaktadır. Yani modeli veri setine yakınsamak olarak da ifade edilebilir (112).

Skor tabanlı algoritmaların bu süreçleri yürütürken takip ettikleri üç farklı işlem bulunur. Bunlar kenar ekleme, kenar silme ve kenar yönünü tersine çevirmedir. Her aşamada skor fonksiyonunun değerini maksimize eden işlem yapılır. Bu işlemler Bayesci ağın yapısına en yüksek skor değerini veren YDG'ye ulaşılıncaya kadar tekrar edilir (112).

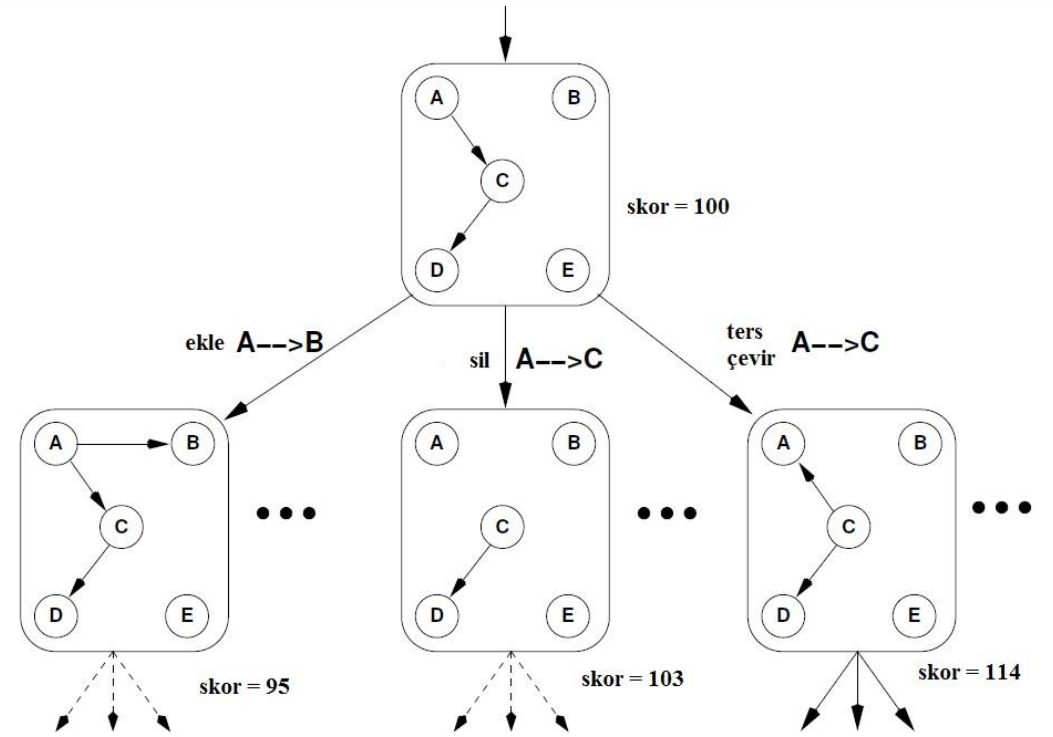
Bir Bayesci ağ oluşturmak için kullanılan skor tabanlı yöntemler arasında literatürde en yaygın olanları tepe tırmanma algoritması, tabu araştırma algoritması, genetik algoritma, benzetimli tavlama, karınca kolonisi optimizasyon algoritması ve parçacık sürü optimizasyon algoritması olarak gözlemlenir. Bu algoritmalar arasından ise tepe tırmanma ve tabu araştırma algoritmaları incelenecektir.

Tepe Tırmanma (Hill Climbing): Bayesci ağlarda yapı öğrenme algoritmaları arasında en sık kullanılan ve bir açgözlü arama (greedy search) yöntemi olan tepe tırmanma algoritması matematiksel bir optimizasyon tekniği olarak da bilinir. Bu metot açgözlü tepe tırmanma buluşsal optimizasyon tekniğini kullanarak en yüksek skor değerine sahip skor fonksiyonunu bulmayı hedefler. Başka bir ifade ile algoritma rastgele bir çözümle başlar ve iteratif bir şekilde tüm örneklem uzayını ele alarak sürekli bir şekilde modeli daha doğru ifade eden çözümleri araştırır (113).



Şekil 15. Tepe tırmanma algoritması ile global maksimum aranması (114)

Bayesci ağların yapısının oluşturulmasında kullanılan tepe tırmanma algoritması mevcut düğümlerin muhtemel yerel kombinasyonlarını kullanır. Her bir tepe tırmanma adımında ele alınana düğümün ön belleği ya da önsel olasılığı göz önünde bulundurularak düğümler arası bire bir değişimler yapılır. Tekli değişimler yapmak modelin uygunluğunu oldukça arttıran önemli bir yöntemdir. Tekli değişim, bir düğümün konumu ve ok yönünde yapılan değişikliği ifade ettiği gibi düğümlerin değerlerinde gerçekleşen çoklu değişimleri de dikkate alarak gerçekleştirilir. Bu işlemler en yüksek skoru oluşturan düğüm kombinasyonları keşfedilene kadar tekrarlanır (Resim 2). Aşağıdaki Şekil 17'de tepe tırmanma algoritmasının Bayesci bir ağın yapısının oluşturulmasında izlediği süreçler gösterilmiştir (115).

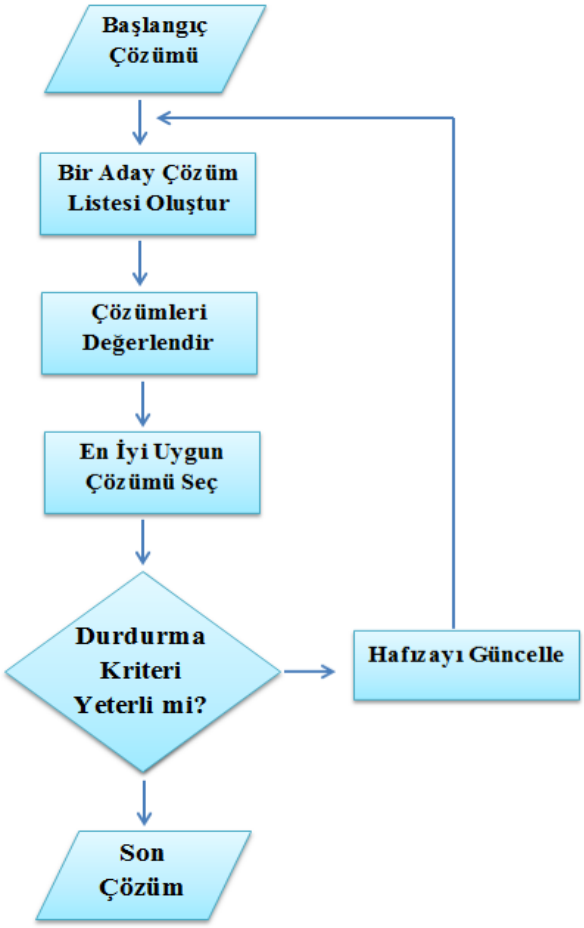


Şekil 16. Bayesci ağların yapısının oluşturulmasında tepe tırmanma araması süreci (115)

```
function HILL-CLIMBING(problem) returns a solution state
         inputs: problem, a problem
         static: current, a node
                 next, a node
         current <– MAKE-NODE(INITIAL-STATE[problem])
         loop do
                 next– a highest-valued successor of current
                 if VALUE[next] < VALUE[current] then return current
                 current *–next
         end
```

Resim 2. Tepe tırmanma Algoritması (117)

Tabu Araştırma (Tabu Search): Tabu araştırma algoritması, 1986 yılında F.Glover tarafından optimizasyon problemlerinin çözümü için geliştirilmiş iteratif bir araştırma algoritmasıdır. Klasik optimizasyon yöntemlerinin ötesine giden Tabu araştırma algoritmaları, kaynak planlaması, telekomünikasyon, finansal analiz, programlama, moleküler mühendislik, genetik, biyoinformatik, lojistik gibi birçok alanda yaygın olarak kullanılmaktadır. Bu yöntem kısaca, son çözüme götüren adımın, dairesel hareketler oluşturmaması için bir sonraki döngüde tekrarının yasaklanması veya cezalandırılması esasına dayanır (117). Tabu araştırma algoritması, rastgele bir çözüm kümesi seçerek arama işlemine başlayarak skorlama fonksiyonunu maksimum yapacak bir önceki çözümün komşuluğundaki çözüm kümesinden faydalanarak yapı öğrenme işlemini yerine getirir. Bir diğer ifade ile algoritmanın bölgesel optimalliği aşmak amacıyla kullandığı temel prensip, değerlendirme fonksiyonu tarafından her iterasyonda en yüksek değerlendirme değerine sahip hareketin bir sonraki çözümü oluşturmak amacıyla seçilmesi esasına dayanmaktadır (118, 119).



Şekil 17. Tabu araştırma algoritması akış diyagram

Tabu araştırma önceden incelenmiş olan bir yol olmadığı sürece her çözümü araştırabilen bir yöntemdir. Bu şekilde yerel minimumdan uzaklaşılarak istenilen çözüme erişebilmektir. Algoritma öncelikle yerel minimuma doğru hareket ederek başlar ve daha önce yapılmış hareketlere tekrar dönüş yapmayı engellemek için bir veya daha fazla tabu listesi oluşturur. Bu listenin amacı, önceden yapılmış bir hareketin tersine dönmesini engellemektir. Tabu listeleri zamansal bir sıralama yapısına sahiptir ve Tabu araştırma hafızasını şekillendirir. Başlangıçta hedefi çözüm uzayında kaba, yüzeysel araştırmalar yapmak olan hafızanın rolü algoritma ilerledikçe farklılaşabilir. Düğümler için aday konumlar tespit edildikçe algoritma, yerel optimum sonuca ulaşmaya daha fazla odaklanır (120). Bir örnek Tabu araştırma algoritması aşağıda verilmiştir (120).

Çözüm=Başlangıç çözümü,

En_İyi_Çözüm=Çözüm,

Tabu_Listesi (Boş),

Durdurma_Kriteri,

Kontrol=FALSE,

Repeat

– **Eğer** *Çözüm>En_İyi_Çözüm* **ise** *En_iyi_Çözüm=Çözüm*

– **Eğer** *Durdurma_Kriteri'* ne ulaşılmış **ise** *Çözümü Tabu_Listesine* ekle

• **Eğer** *Tabu_Listesi* dolu **ise** ilk gireni listeden çıkar,

çözümlerin içinden başka birini *Yeni_Çözüm* olarak seç

• **Eğer** *Yeni_Çözüm* bulunamadıysa ***veya***

(***Eğer*** geliştirilen *Yeni_Çözüm, Uzun_Dönem_Hafıza* da bulunuyor **ise** *Yeni_Çözümü* rasgele üret)

• **Eğer** *Yeni_Çözüm, Tabu_Listesi'* nde yok **ise** *Çözüm=Yeni_Çözüm*

Değil ise *Kontrol=TRUE*

Until *Kontrol=TRUE*

5.1.4.3.1.1.3. Karma Algoritmalar

Karma algoritmalar kısıtlama tabanlı ve skor tabanlı algoritmaların birlikte kullanıldığı yöntemlerdir. Bu algoritmalarda ilk olarak kısıtlama tabanlı yöntemler uygulanır ve öğrenilen bu yapı üzerinden skor tabanlı yöntemlerin uygulanması ile Bayesci ağlar elde edilir.

Karma algoritmalarda işleyen bu iki aşamalı yapı genel olarak kısıtlama ve maksimize etme şeklindedir. Kısıtlama aşamasında mevcut düğümlerle ilişkili olan başka bir düğüm kümesi oluşturulur ve daha sonra maksimize etme aşamasına geçilir. Burada sadece her bir düğümle ilişkisi olan çözüm kümesi üzerinden skor tabanlı metotlar uygulanır ve kenarların yönlendirilmesi ile Bayesci ağı temsil eden YDG'ler oluşturulur (121).

5.1.4.3.2. Bayesci Ağlarda Model Seçim ve Değerlendirmede Bilgi Kriterleri

Bayesci ağlarda en sık karşılaşılan sorunlardan biri veri setine uygun modeli seçmek, kestirmek ve boyutunu belirlemektir. Bu işlemin zorluğu modelde yer alan parametre miktarı ile orantılı olarak artmaktadır. Model değerlendirme ise ağdaki modellerin incelenerek en iyisini seçme işidir (81).

Kullanıcı Bayesci ağdaki modellerin istatistiksel tanımlanabilirliği veya değerlendirilmesi sonucu ortaya çıkan modelin kalitesini inceler ve en doğru modele ulaşmak için araştırmalarına yön verebilir. Model seçim ve değerlendirme alanlarında artan çalışmalar oldukça dikkat çekmektedir. Sahip olunan veri kümesi ile ilgili kurulan modellerde en yaygın karşılaşılan sorun ortaya çıkan farklı modeller arasından en uygun olanının nasıl seçileceğidir. Bir Bayesci ağda oluşabilecek muhtemel düğüm sayılarından bahsetmiştik. Bunlar göz önüne alındığında bu durumun oldukça büyük zorluklar içerdiği görülmektedir. Bu problemi aşmak için geliştirilen bazı kriterler bulunmaktadır. Bu kriterler Bayesci ağlarda öğrenme algoritmaları ile birlikte çalışarak uygun modelin ortaya çıkarılmasına yardımcı olurlar (81).

Model seçiminde en sık kullanılan metotlar: Akaike Bilgi Kriteri (Akaike Information Criterion)-AIC, Bayeci Bilgi Kriteri (Bayesian Information Criterion)-BIC, Bayesci-Dirichlet Eşitlik (Bayesian Dirichlet equivalent)-BDE, K2.

Akaike Bilgi Kriteri: Akaike bilgi kriterinin kurucusu olan Hirotugu Akaike; 1973, 1974, 1977 ve 1981'de yayınladığı çalışmalarla istatistiksel veri modelleme, istatistiksel model tanımlanabilirliği ve değerlendirmesi ile ilgili alanların öncüsü niteliğinde gösterilen araştırmacılardan biri olarak kabul edilmektedir. Ayrıca "Bilgi kriterleri" kavramı AIC'in çıkarımında kullanılan KullBack-Leiber Bilgisinden gelmektedir. KullBack-Leiber Bilgisi; model ve gerçek dağılım arasındaki uzaklığın objektif bir ölçümü varsa, iyi bir çıkarımsama yöntemi bu uzaklığı olabildiğince küçük yapmalıdır, biçiminde tanımlanmaktadır (122, 123).

Farklı boyutlu modellerin karşılaştırılmasında güçlü bir model seme kriteri olan AIC; bir örneklem verisinden elde edilen parametre tahminlerinin farklı örneklemler için de kullanılması nedeniyle oluşan doğruluk kaybının bir ölçümü olarak yorumlanmaktadır. Buna göre AIC, gerçek örneklemden elde edilen parametre

tahminlerinin farklı örneklemler için çapraz-geçerliliğin bir ölçümü olarak kullanılmaktadır (123).

$$AIC = -2\ \log L(\boldsymbol{\theta}) + 2k \qquad \text{veya} \qquad AIC = -2\ \ln(L(\hat{\theta})) + 2k \qquad (5.12.)$$

AIC; θ , k boyutlu bilinmeyen parametreler vektörü, $\hat{\theta}$, θ 'nın en çok olabilirlik kestiricisi (Likelihood) ve $L(\hat{\theta})$, k bilinmeyen parametreli olabilirlik fonksiyonu olmak üzere tanımlanmıştır. Burada en küçük AIC değerine sahip olan model en iyi olarak görülür. Formülde ilk terim parametre tahmininde en çok olabilirlik (Maximum likelihood) kullanıldığında uyum farkları yahut model yanlılığının bir ölçümüdür. İkinci terim ise karmaşıklık güvenilirliği azalttığı için cezanın (penalty) bir ölçümü veya birinci terimdeki yanlılığı telafi etmenin bir yöntemidir (81).

Bayesci Bilgi Kriteri: Schwarz, 1978 yılında yayınladığı çalışmasıyla değişik boyutlara sahip modellerden en iyisini seçme sorununa Bayesci bir yaklaşımla getirdiği çözümle tanınmaktadır. Bu sebeple Schwarz-Bayes bilgi kriteri olarak da bilinmektedir. Akaike bilgi kriteri ile benzer noktaları vardır. BIC, AIC'ten farklı olarak modeldeki parametreleri, parametre sayısı ve modeldeki karmaşıklık arttıkça daha fazla cezaya çarptırır. Bu durum modeldeki iç tutarlığı korumak için kullanılan sezgisel bir yaklaşımdır (124).

$$BIC = -2\ \log L(\hat{\theta}) + k\ \log(n) \quad \text{veya} \quad BIC = -2\ \ln(L(\hat{\theta})) + k\ \ln(n) \qquad (5.13.)$$

BIC; θ , k boyutlu bilinmeyen parametreler vektörü, $\hat{\theta}$, θ 'nın en çok olabilirlik kestiricisi (Likelihood) ve $L(\hat{\theta})$, k bilinmeyen parametreli olabilirlik fonksiyonu ve n toplam gözlem sayıdır. Modeldeki aday dağılımların her biri için BIC değeri hesaplanır ve en küçük BIC değerine sahip olan dağılım verileri en iyi uyum sağlayan model olarak seçilir(124, 125).

Bayeci Dirichlet Eşitlik: 1991 yılında Buntine tarafından temelleri atılan bu yöntem, ismini 1995 yılında Heckerman ve diğerleri tarafından çalışma sonucu almıştır (126, 127). Dirichlet dağılımı, önsel multinominal dağılımları düzeltilmiş hiper-parametreler ile birleştirecek şekilde seçilir. Hiper-parametreler ise eşdeğer örneklem

büyüklüğü (Dirichlet parametreleri toplamı eşit olan düğümler) kullanarak bulunabilir. Bu metot, verilen bir YDG'nin sonsal olasılığını Dirichlet önsel olasılıkları ile maksimize etmeyi amaçlar (128, 129). BDe skorlama fonksiyonu aşağıdaki eşitlikte verilmiştir.

$$\mathrm{BD}(B,T) = \log(P(B)) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{\Gamma(N'_{ij})}{\Gamma(N_{ij}+N'_{ij})}\right) + \sum_{k=1}^{r_i}\log\left(\frac{\Gamma(N_{ijk}+N'_{ijk})}{\Gamma(N'_{ijk})}\right)\right) \quad (5.14.)$$

K2: Cooper ve Herkovits 1992 yılında bir Bayesci ağ içerisindeki en iyi ağı bulmak için bir algoritma sunar ve bu algoritma K2 algoritması olarak isimlendirilir. BDe skorunun özel bir durumu olarak ifade edilmektedir. K2 algoritması Bayesci ağ yapısını veri seti üzerinden öğrenen açgözlü bir arama algoritmasıdır. K2 algoritması, mevcut verilerden oluşturulacak bir Bayesci ağın sonsal olasılığını maksimize eden modeli seçmeyi hedefler. K2, Ağ yapısını oluşturacak düğümleri sıralayarak modelin hesaplanmasındaki karmaşıklığı azaltır. Bu sıralama, her bir değişken X_i için X_i'den küçük numaralı düğüm, ebeveyn kümesine eklenir. Sıralama süreci yeni düğümlerin eklenmesiyle tam bir ağ modeli elde edilene kadar tekrarlanır (130, 131). K2 skorlama fonksiyonu aşağıdaki eşitlikte verilmiştir.

$$\mathrm{K2}(B,T) = \log(P(B)) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left(\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\right) + \sum_{k=1}^{r_i}\log(N_{ijk}!)\right) \quad (5.15)$$

5.2. Genetik Verilerde Veri Madenciliği Uygulamaları

Bu bölümde veri madenciliği yöntemlerinin genetik verilere uygulanması ile belirlenen birtakım araştırma alanları irdelenmiş ve çeşitli sonuçlar elde edilmiştir. Problemin belirlenmesinde en önemli etken genetik veri ve veritabanlarının karmaşıklığı olarak belirlenmiştir. Önceki bölümlerde bahsedilen teorik bilgiler ışığında TCGA veri portalından elde edilen genetik bilgiler üzerinde çeşitli veri madenciliği

metotları uygulanmıştır. Bu uygulamalardan elde edilen ilginç sonuçlara dikkat çekmek ve bunların detaylı analizini yapma amaçlanmıştır.

5.2.1. Uygulamanın Amaç ve Kapsamları

Genetik, veri türünün oldukça çeşitli olduğu alanlardan birisidir. Genetik verinin artış hızı karşısında kullanılan geleneksel veri analiz metotları yetersiz kalmaktadır. Bu nedenle genetik veritabanlarından yeni bilgilerin keşfedilmesi, veritabanlarında depolanan verinin etkin kullanımı son derece önem arz etmektedir. Veri madenciliği teknikleri, özellikle bilinen yöntemlerle elde edilemeyen, önceden herhangi bir tahmin yapılamayan örüntü ve bilgilerin açığa çıkarılmasında etkili bir rol oynamaktadır.

Bu tez çalışmasında TCGA veri potalından elde edilen küçük hücreli dışı akciğer adenokarsinomuna ait genetik veriler kullanılarak, bu kanser türünde ortaya çıkan mutasyonların atasal sıralaması ve birbirleri ile olan ilişkileri veri madenciliği metotları ile ortaya çıkarılmak istenmektedir. Böylece kişiselleştirilmiş kanser tedavilerinde hastaya özgü yaklaşımları kolaylaştıracak sonuçlar hedeflenmiştir.

5.2.2. Uygulamalarda Kullanılan Algoritma ve Programlar

Bu tez çalışmasının temel kaynağı olan genetik verilerin elde edilmesinde TCGA veri portalı, elde edilen verileri kullanıma hazır hale getirmek, analizleri yapıp sonuçları almak için Python (132) ve R (133) programları kullanılmıştır.

Çalışmada veri madenciliği metotlarından birliktelik kuralları, Bayesci ağları oluşturmak için tepe tırmanma ve tabu arama algoritmaları ve model seçimi için AIC, BIC, BDe ve K2 kriterleri kullanılmıştır. Birliktelik kuralları analizinde en yaygın kullanılan Apriori algoritması kullanılmıştır. Sıralı algoritmalar arasında uygulandığı örneklem büyüklüğüne göre en iyi performans gösteren algoritmalardan birisi Apriori algoritmasıdır.

Verilerin uygun hale getirilmes için Python programlama dili kullanılmıştır. Pyhton, son derece kolay okunabilen, nesne yönelimli, modüler ve etkileşimli bir programlama dilidir. Şuan dünyadaki Google, NASA, CERN büyük kurumlar bu dili kullanmaktadır (132). Verilerin analizi için R programından faydalanılmıştır. R bir programlama dili ve aynı zamanda istatistiksel hesaplama ve grafik programıdır (133). R istatistiki yazılım geliştirme ve veri analizi alanlarında oldukça yaygın

kullanılmaktadır (134). R çok geniş istatistiki (doğrusal ve doğrusal olmayan modelleme, klasik istatistik testleri, zaman serileri analizi, sınıflandırma, kümeleme ve diğer) ve grafik çizim teknikleri sunmaktadır (133). Bu tez çalışmasında R programının "arules","bnlearn" ve "Rgraphviz" paketleri yardımıyla grafiksel modeller oluşturulmuştur.

5.2.3. Verilerin Hazırlanması

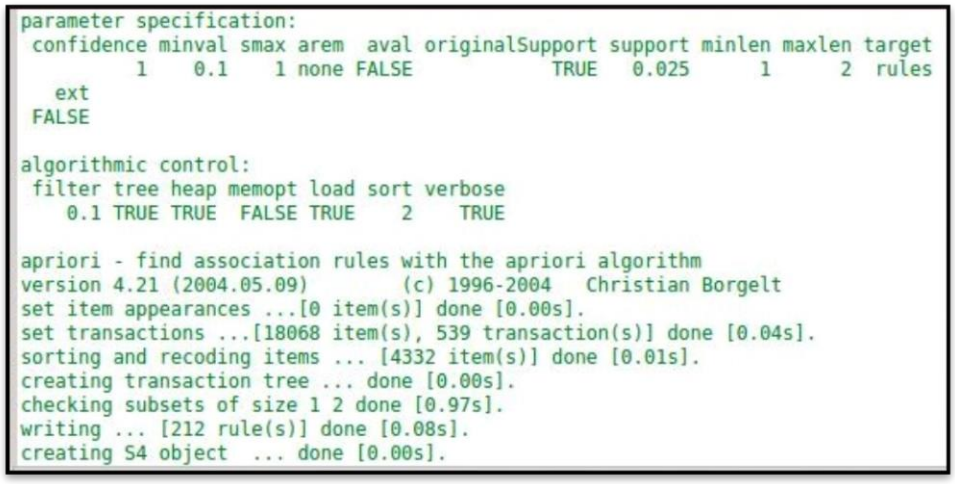
Bu tez çalışmasında TCGA veri portalından alınan küçük hücreli dışı akciğer adenokarsinom genetik verileri kullanılmıştır. TCGA birçok kanser türü için açık kaynaklı bilgi sağlamaktadır. Bu bilgiler vaka tabanlı olup, yüzlerce farklı doku ve hastadan alınan bilgiler doğrultusunda oluşturulmuştur (38). Elde edilen bu bilgiler TCGA'nın kendine özgü kategorileri ve bilgileri içerecek şekilde depolanmıştır. Bu bilgileri çalışmada kullanılacak formata getirmek için Python programı kullanılmıştır. Genetik mutasyon verileri bulunup bulunmama durumları göz önünde alınarak bir matrise aktarılmıştır. Bu durumlar ise 0 ve 1 olmak üzere ikili bir biçimde gösterilmiştir. Tablo 4'de veri matrisinin yapısı gösterilmektedir.

Tablo 4: Hastalara göre mutasyonların bulunma durumunu gösteren veri matrisi

	TCGA-05-4244	**TCGA-05-4249**	**TCGA-05-4250**
5S_rRNA	0	0	0
A1BG	0	1	0
A1CF	0	0	0
A2M	0	0	0
A2ML1	0	0	0
A4GALT	0	0	0
A4GNT	0	0	0
AAAS	0	0	0
AACS	0	0	0
AACSP1	0	0	0
AADAC	0	0	0
AADACL2	0	0	0
AADACL3	0	0	0
AADACL4	0	0	0
AADAT	0	0	0
AAED1	0	0	0
AAGAB	0	0	0
AAK1	0	0	0
AAMP	0	0	0
AANAT	0	0	0
AARD	0	0	0
AARS	0	0	0

5.2.4. Birliktelik Kuralları Analizi ile Sık Görülen Kuralların Tespit Edilmesi

Düzenlenmiş veri seti birliktelik analizi için hazır hale getirilmiştir. 539 adet hasta ve 18068 adet mutasyon verisinin bulunduğu veri matrisi üzerinde Apriori algoritması ile birliktelik kuralı madenciliği (association rule mining) yöntemi uygulanmıştır. Burada verileri budamak için Apriori algoritması destek ve güven değerlerini kriter olarak alır. Destek değeri 0.025 ve güven değeri 1 olarak belirlemiştir. Apriori algoritması doğası gereği veri setindeki mutasyonları çiftler halinde seçerek beraber görülme sıklıklarına göre analiz edecektir. Burada güven değerinin 1 olarak belirlenmesi tüm veri seti boyunca sürekli olarak birlikte görülen mutasyonları tespit etmek içindir. Böylece tüm mutasyonların %100 beraber görüldüğü kayıtlar filtrelenmiş olur. Apriori algoritması R programı üzerinde çalışan "arules" paketi ile verilere uygulanmıştır. Resim 3'te programın çıktı ekranı görüntülenmektedir. Sonuçta 212 kural elde edilmiştir.



Resim 3. R programı ile birliktelik analizi

5.2.5. Bayesci Ağların Oluşturulması

Birliktelik kuralları analizi yapıldıktan sonra geriye kalan 212 mutasyonu kullanarak Bayesci ağ oluşturulmuştur. R programında bulunan "bnlearn" paketi yardımıyla skorlama tabanlı Bayesci ağ oluşturma algoritmalarından tepe tırmanma ve tabu araştırma algoritmaları kullanılmıştır. Bu algoritmaların çalışma prensipleri gereği

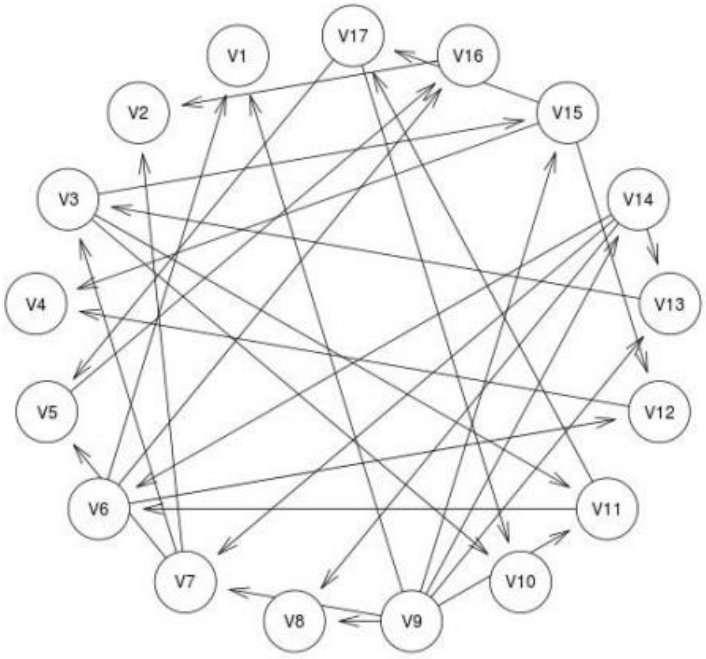
verilerin skorlanması gerekmektedir. Bayesci ağ öğrenme algoritmaları verilerin skorlanmasında AIC, BIC, BDe ve K2 kriterlerini kullanarak ağ yapısını kurmaktadır. Ayrıca verilerin görülme oranları dikkate alınarak oluşturulacak Bayeci ağlarda mutasyonların frekansı fazla olandan az olana doğru ilerlemesi koşulunu sağlayacak bir "blacklist" oluşturulmuştur. Böylece frekansı az olan mutasyon atasal sıralamada üst kısımlara çıkamayacaktır. Bunun sebebi ise daha fazla görülen mutasyon sayısınn daha önce oluştuğu hipotezinden yola çıkarak belirlenmiştir. Bayesci ağı oluşturmak için çıkarılan 212 kural setinden 20 adet mutasyon seçilmiştir. Bu seçim sırasında Online Mendelian Inheritance in Man (OMIM) veri tabanında küçük hücreli dışı akciğer kanserinde görülen mutasyonlar esas alınmıştır (135). Resim 4'te R programı çıktı ekranında gösterilmiştir.

```
bn.hc <- tabu(df2, score = "k2", blacklist = blacklist)
plot(bn.hc)
z <- c("NKX2-1","CASP8","PPP2R1B","CYP2A6","DOK2","DLEC1","EML4","FASLG",
"ERBB2","PARK2","MPO","ERCC6","MET","PIK3CA", "BRAF", "ALK", "EGFR", "STK11", "KRAS", "TP53")
g <- graphviz.plot(bn.hc)
names(z) <- nodes(g)
nAttrs <- list()
nAttrs$label <- z
attrs <- list(node = list(shape = "ellipse"))
attrs$node$width<-1.5
attrs$node$fontsize<-12
par(mfrow = c(1, 1))
plot(g, nodeAttrs=nAttrs, attrs = attrs)
```

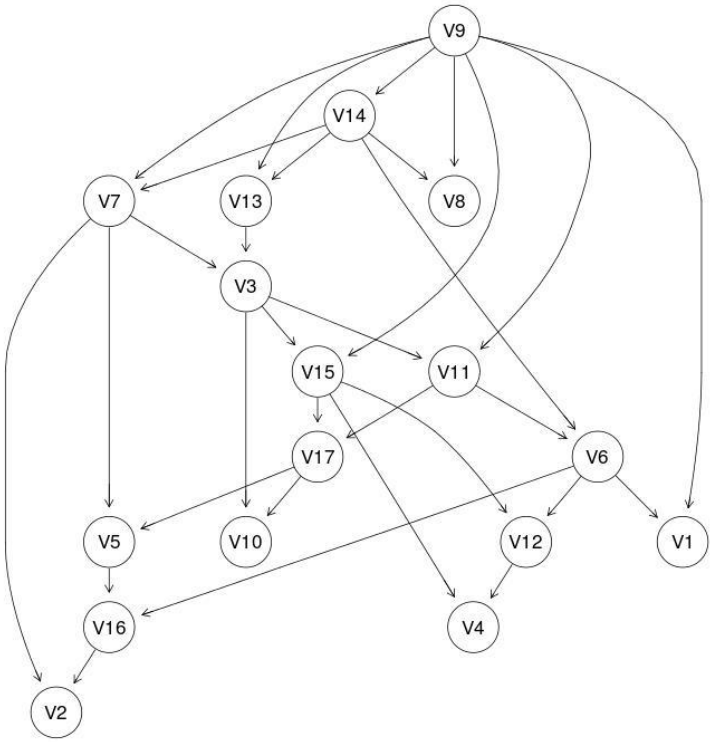
Resim 4. Bayesci ağ kurmak için kullanılan R programı çıktı ekranı

5.2.6. Bayesci Ağların Grafiksel İfadesi

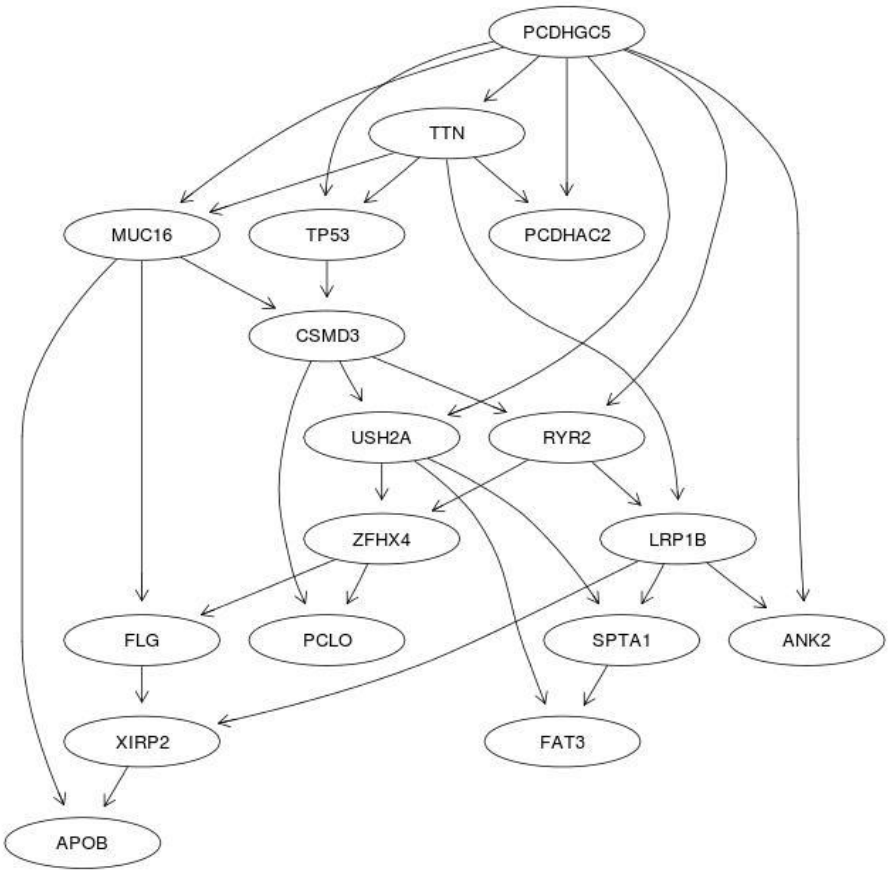
Bayesci ağların grafiksel ifadesinde yönlendirilmiş döngüsüz grafik (YDG) kullanılmaktadır. Mevcut verilerden oluşturulan Bayesci ağı, bir YDG biçiminde ifade etmek için R programının "Rgrapviz" paketinden yararlanılmıştır. Bu paket program elde edilen karmaşık ağ görünümünü daha hiyerarşik bir biçimde ifade etmeyi sağlar. Ayrıca oluşturulan ağı düzenleme imkanı sunmaktadır. Aşağıda verilen şekillerde örnek bir Bayesci ağın düzenlenme süreci gösterilmiştir.



Şekil 18. Bayesci ağ yapısının oluşturulması



Şekil 19. Bayesci ağın hiyerarşik yapısının kurulması



Şekil 20. Düğümlerin tayini ve etiketlenmesi

6. BULGULAR

6.1. Birliktelik Kuralları Analizi ile Kural Çıkarımına Dair Bulgular

TCGA veri portalından temin edilen büyük çaptaki genetik veri setindeki anlamlı ilişkileri ortaya çıkarabilmek için birliktelik kuralları analizi kullanılmıştır. Bu yöntemin uygulamasında Apriori algoritmasından faydalanılmıştır. R programı içerisinde bulunan "arules" paketi kullanılarak Apriori algoritması veriler üzerinde uygulanılmıştır. 539 adet hasta ve 18068 adet mutasyon verisinin bulunduğu veri setinden destek değeri 0.025 ve güven değeri 1 esas alınarak 212 kural üretilmiştir. Destek değeri, veriler arasındaki anlamlı ilişkileri içeren ve daha kolay işlem yapılabilecek küçük bir veri seti oluşturmak için gerekli optimum değeri denenerek 0.025 belirlenmiştir. Güven değeri ise bir mutasyonun görüldüğü tüm diğer işlemlerde birlikte görülen mutasyonları tespit etmek için 1 olarak belirlenmiştir. Böylece, veri setinde görülme oranları yüksek birliktelikler ön plana çıkarılmıştır. Tablo 5'te elde edilen sonuçlardan bir kısmı gösterilmiştir.

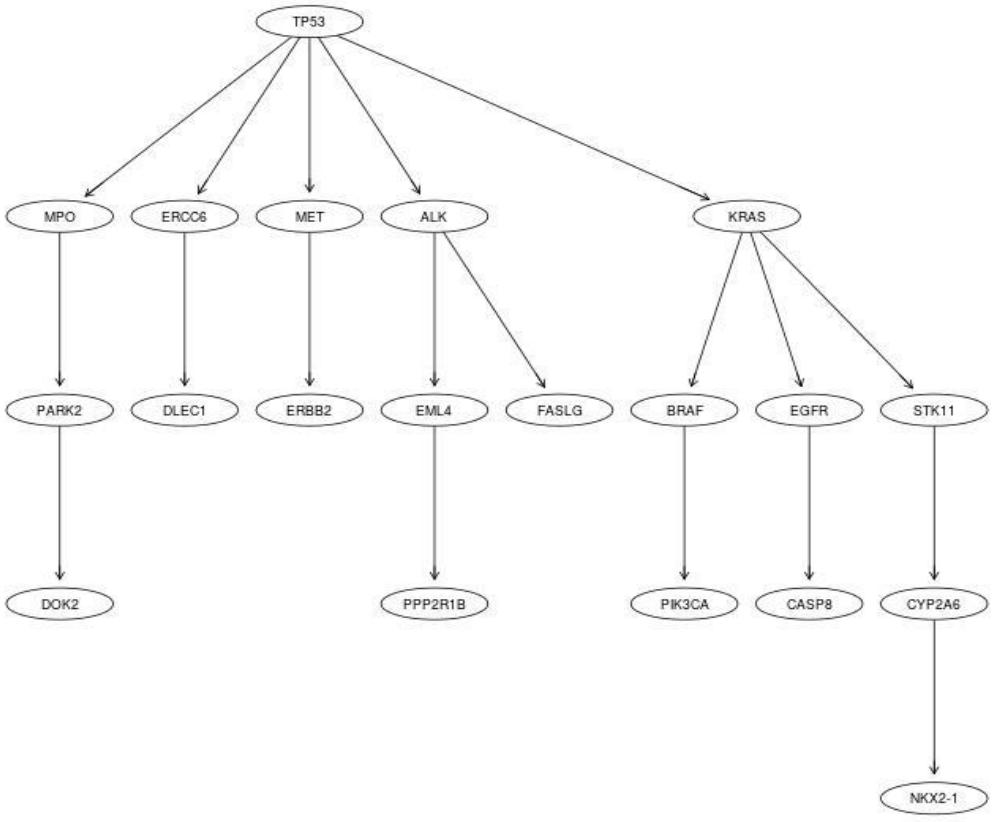
Tablo 5: Birliktelik kuralları analizi ile çıkarılan kurallar

Mutation	rules	support	confidence	lift
CGNL1	abParts	0.025974026	1	1.6283987915
ZWINT	CSMD3	0.025974026	1	2.2742616034
MAGED1	abParts	0.025974026	1	1.6283987915
TRIM56	abParts	0.025974026	1	1.6283987915
NBEAL2	abParts	0.025974026	1	1.6283987915
PKP1	abParts	0.025974026	1	1.6283987915
OR6X1	PCDHAC2	0.025974026	1	2.0111940299
ZNF16	abParts	0.0278293135	1	1.6283987915
DNAAF3	MUC16	0.025974026	1	2.0187265918
CYP7A1	abParts	0.025974026	1	1.6283987915
TPSD1	abParts	0.025974026	1	1.6283987915
FAM160A2	abParts	0.0296846011	1	1.6283987915
ZNF644	abParts	0.0278293135	1	1.6283987915
APOA4	abParts	0.025974026	1	1.6283987915
COL13A1	TP53	0.025974026	1	1.8780487805
COL13A1	abParts	0.025974026	1	1.6283987915
FASTKD5	abParts	0.025974026	1	1.6283987915
RASIP1	abParts	0.025974026	1	1.6283987915
HBD	abParts	0.025974026	1	1.6283987915
MBD1	CSMD3	0.025974026	1	2.2742616034

6.2. Oluşturulan Bayeci Ağlara Dair Bulgular

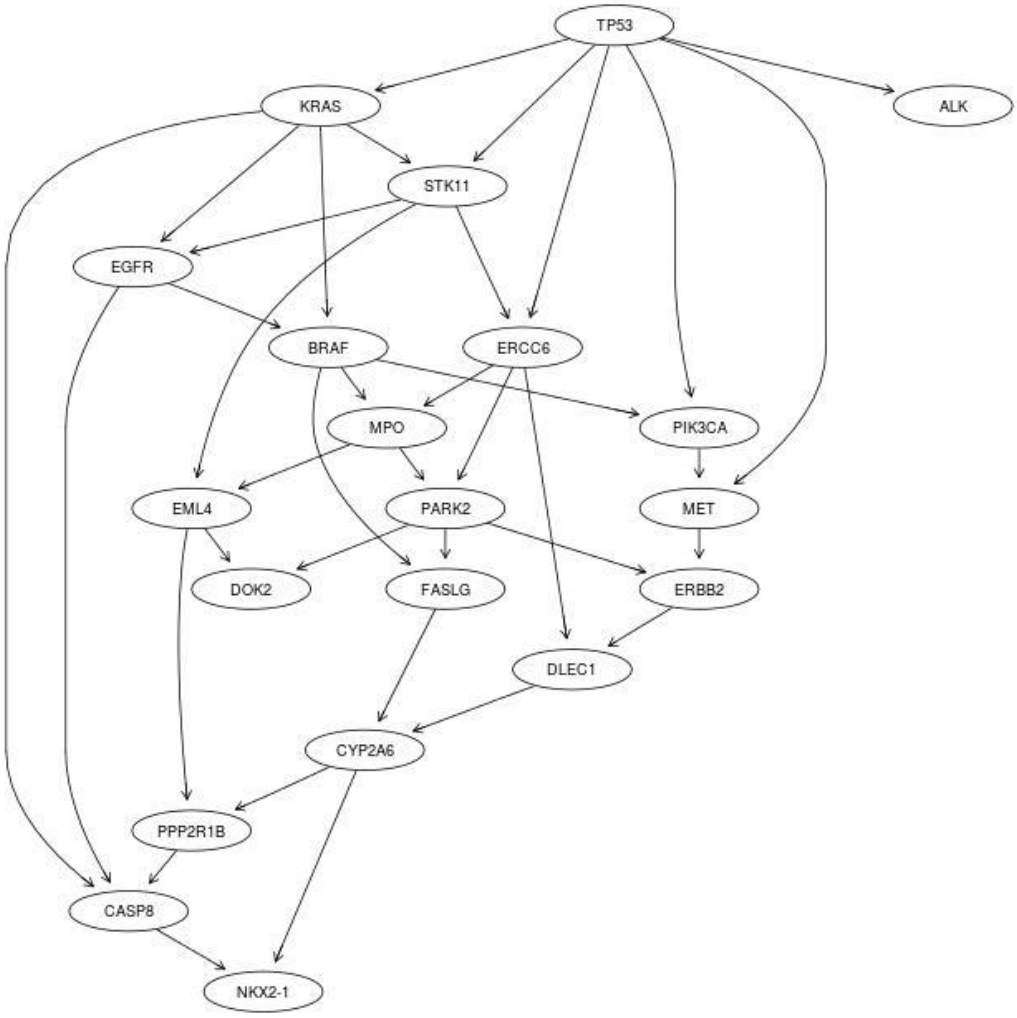
Bayesci ağı oluşturmak için gerekli veri seti birliktelik kuralları yardımıyla budanıp, veri büyüklüğü oldukça indirgenmiştir. Burada Bayesci ağı oluşturmak için R programının "bnlearn" ve oluşturulan ağların YDG biçimde ifade edilmesi için "Rgraphviz" paketleri kullanılmıştır.

Bayesci ağı oluşturmak için iki adet öğrenme algoritması seçilmiştir. Bunlar tepe tırmanma ve tabu arama algoritmalarıdır. Bu algoritmalar Bayesci ağı oluştururken muhtemel birçok yaklaşımı deneyerek giderler. Bu nedenle çok sayıda farlı yapıya sahip ağ oluşur. Bu ağlardan en uygununu seçmek için bir skorlama metodu ile kombine edilmeleri gerekmektedir. Bu tez çalışmasında AIC, BIC, BDe ve K2 olmak üzere dört adet skorlama tabanlı model seçim metodu kullanılmıştır. Bu metotlarla elde edilen sonuçlar aşağıda gösterilmiştir.



Şekil 21. Tepe tırmanma algoritması ve AIC model seçim yöntemi ile oluşturulan Bayesci ağ

Tepe tırmanma algoritması ve AIC model seçim yöntemi ile oluşturulan yapı incelendiğinde Bayesci ağın dizisel ve ayrılan bağlardan oluştuğu gözlemlenmektedir. Ayrıca burada her bir mutasyonun sadece bir atası olacak şekilde yerleşimi söz konusudur. Veri setinde en sık görülen TP53 tümör baskılayıcı geni en üst kısımda yer almakta ve KRAS, EGFR gibi bilinen genlerle olan doğrusal ilişkisi olduğu gözlemlenmektedir. Bu YDG'ye göre her bir mutasyonun oluşumu için sadece ebeveyninin ortaya çıkması yeterli görülmektedir (Şekil 21).



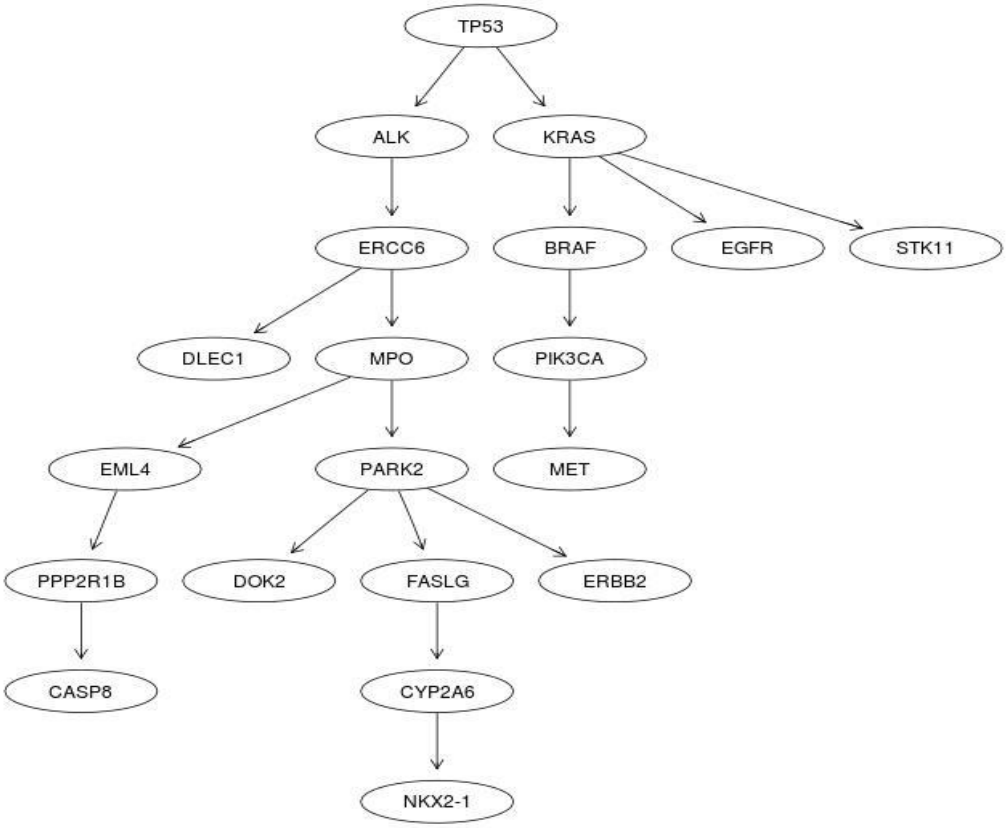
Şekil 22. Tepe tırmanma algoritması ve BDe model seçim yöntemi ile oluşturulan Bayesci ağ

Tepe tırmanma algoritması ve BDe model seçim yöntemi ile oluşturulan yapı incelendiğinde Bayesci ağın dizisel, ayrılan ve birleşen bağlardan oluştuğu gözlemlenmektedir. Mutasyonları ifade eden düğümler birden fazla ata düğümü olabilecek şekilde bir yapı göstermektedir. Yapı olarak AIC'ten oldukça farklı olan BDe de özellikle ALK mutasyonunun yalnız kalması dikkat çekmektedir. Bu Bayesci ağ da AIC'e göre KRAS ve EGFR ilişkileri benzer alırken MPO, ERCC6, MET gibi ilişkiler yer değiştirmiştir. Nedensellik yapısının oldukça karmaşık olduğu bu ağ, aynı zamanda kanserin genetik alt yapısının kompleksitesini ifade etmektedir (Şekil 22).

Şekil 23. Tepe tırmanma algoritması ve BIC model seçim yöntemi ile oluşturulan Bayesci ağ

Tepe tırmanma algoritması ve BIC model seçim yöntemi ile oluşturulan yapı incelendiğinde Bayesci ağın dizisel bir bağlardan oluştuğu gözlemlenmektedir. Bu yöntem mutasyonları bir birinden bağımsız olarak ele almıştır. Yanlızca bir ilişkisel yapı barındıran bu Bayesci ağ, nedensel yapıdan oldukça uzak bir yapı sergilemektedir. Bu YDG'ye göre bir mutasyon dışında her bir mutasyonun oluşumu bağımsız bir şekilde gerçekleşmektedir (Şekil 23).

Tepe tırmanma algoritması ve K2 model seçim yöntemi ile oluşturulan yapı incelendiğinde Bayesci ağın dizisel ve ayrılan bağlardan oluştuğu gözlemlenmektedir. Bu ağ yapısında AIC ve BDe de olduğu gibi TP53, KRAS ve EGFR yolu korunmuş fakat diğer mutasyonlar farklı biçimlerde konumlanmıştır. Nedensellik yapısının oldukça karmaşık olduğu bu ağ, aynı zamanda kanserin genetik alt yapısının kompleksitesini ifade etmektedir (Şekil 24).

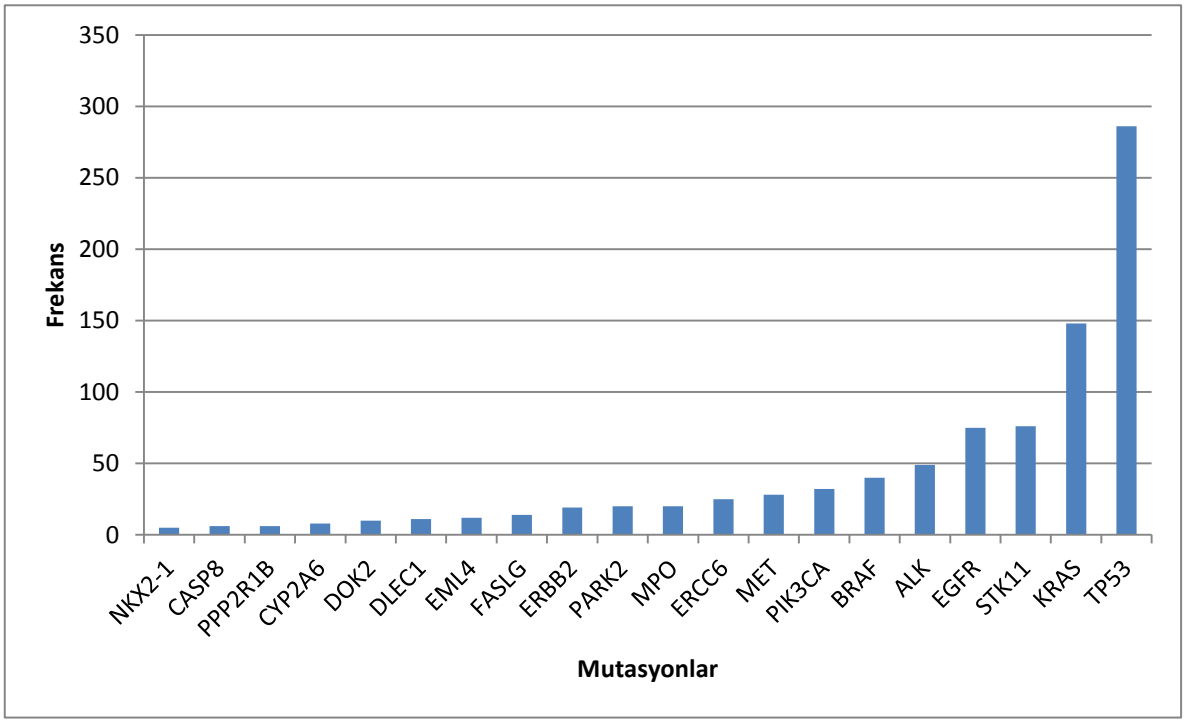


Şekil 24. Tepe tırmanma algoritması ve K2 model seçim yöntemi ile oluşturulan Bayesci ağ

Tabu araştırma algoritması ile oluşturulan modeller, BDe algoriması dışında tepe tırmanma algoritması ile birebir benzerlik göstermektedir. Buradan mevcut veri seti ile oluşturulan Bayesci ağların yapısını belirlemede skorlama algoritmalarının daha etkin bir rol oynadığı söylenebilir. Tabu araştırma algoritması ile elde edilen sonuçlar Ek 1'de verilmiştir.

Tüm mutasyonların gelişimleri sürecinde etkilendikleri ve/veya etkiledikleri mekanizmalar bulunmaktadır. Yukarıdaki şekillerde gösterildiği üzere, farklı yapısal öğrenme algoritmaları ve istatistiksel skorlama metotları ile oluşturulan ve mutasyonlar arası atasal ilişkileri gösteren YDG'ler ortaya konulmuştur. Bu YDG'ler bize mutasyon frekanslarından yola çıkarak küçük hücreli dışı akciğer adenokarsinomunun gelişimi sürecinde mutasyonların atasal ilişkileri ile ilgili farklı öngörüler sunmuştur. Örneğin, hangi mutasyon ya da mutasyonların bir birlerini etkilediği, belirli bir mutasyonun oluşumu için gerekli ön koşulların neler olabileceği ya da hangi bir mutasyondan sonra ortaya çıkacak mutasyonu tahmin etme vb. gibi ön görüler ortaya koymuştur.

Tablo 6: Mutasyonların toplam veri setinde görülme yüzdeleri



Veri setinde en çok görülen mutasyonlar sırasıyla TP53, KRAS, STK11 ve EGFR şeklinde gözlemlenmiştir (Tablo 6). Mutasyonların görülme düzeylerinde oldukça büyük farklar olduğu görülmektedir. Özellikle TP53 ve KRAS'ın diğer mutasyonlara kıyasla oldukça fazla mutasyon oranına sahip görülmektedir. Bu durum kanserin gelişim sürecinde baskın mutasyonları belirleme ve özellikle bu kanser türünü karakterize eden bozulmaları tespit etmede önemli bilgiler sağlamaktadır.

7. TARTIŞMA ve SONUÇ

Günümüzde tıp dünyasında sağlık çalışanları ve hastalara veri madenciliği teknikleri ile hizmet sunumu gittikçe artan bir talep haline gelmiştir. Özellikle kanser gibi ölüm oranı oldukça yüksek bir hastalığın genetik alt yapısı birçok yönden halen keşfedilmemiştir. Bu sebeple bu tez çalışmasında açık kaynaklı birçok kanser türüne ilişkin genetik verilerin bulunduğu TCGA veritabanı kullanılarak, küçük hücreli dışı akciğer adenokarsinomuna dair çıkarımsamalarda bulunulmuştur. Bu çıkarımsamaları elde etmek için günümüzde yaygın olarak kullanılan veri madenciliği tekniklerinden birliktelik analizi ve Bayesci ağlar kullanılmıştır. Bu yöntemleri uygulayarak, ham halde bulunan genetik veri yığınının biyoinformatik araçları yardımıyla irdelenerek bu verilere karşı farklı bir bakış açısı getirmek hedeflenmiştir. Çalışmada uygulanan tekniklerle;

- Genetik verilerin birliktelik analizi ile incelenerek, yoğun bir şekilde beraber görülen mutasyonların tespiti,
- Bayesci ağlarla, küçük hücreli dışı akciğer adenokarsinomunda görülen mutasyonların muhtemel atasal ilişkilerinin çıkarımı,

gerçekleştirilmiştir.

Genetik verilerin incelenmesinde verinin çok büyük boyutlarda olması çoğu zaman veri hakkında tahmin yürütmede büyük bir engel teşkil etmektedir. Birliktelik analizi ile genetik veriler incelenirken özellikle hizalama yöntemi kullanıldığı zamanlarda benzer bölgelerin tespiti ve genlerin başka genlerle beraber görülme sıklıkları hakkında bilgiler kolaylıkla elde edilebilir. Yapılan bu çalışmada da TCGA veritabanından alınan verilerde küçük hücreli dışı akciğer adenokarsinomunda görülen tüm mutasyonlar arasından en sık görülenler saptanmıştır. Bunun için birliktelik analizinden oldukça yaygın kullanılan Apriori algoritması kullanılmıştır. Apriori algoritması aday kümelerin sayısını azaltmada yüksek performans gösterdiği için tercih edilmiştir. Yapılan başka bir çalışmada ise Creighton ve Hanash gen ekspresyonu veritabanları üzerinde bilinmeyen birçok birlikteliğin keşfedilmesinde veri madenciliği tekniklerinden birliktelik analizini kullanmışlardır (136).

Bayesci ağlar, yakın zamanda oldukça popüler hale gelen geriye yönelik genetik programlama konularında da yaygın bir biçimde kullanılmaktadır. Perrin ve diğerleri (137), Murphy ve Mian (138), farklı zamanlarda yaptıkları çalışmalarda geriye dönük genetik yapının çıkarımını Bayesci ağları kullanarak yapmışlardır. Yapılan başka çalışmalar incelendiğinde Bayesci ağlarla genetik materyalde gerçekleşen evrimsel mutasyonların topolojik yapısı incelenmiştir (139). Yine başka bir çalışmada Bayesci ağlarda BDe skorlama yöntemiyle moleküler gen düzenleyici ağların yeniden oluşturulmasıyla önemli yolakların tespiti yapılmıştır (140).

Bayesci ağlar farklı alanlarda değişkenler arasında belirsizlikleri gidermek ve çıkarımsamalarda bulunmak üzere oldukça yaygın bir şekilde kullanılmaktadır. Bayesici ağları oluşturmak için kullanılan yapı öğrenme algoritmaları çok değişkenli bir veri seti içersindeki belirsiz koşullu bağımsızlık ilişkilerini bir uzman yardımına ihtiyaç duymadan çıkarabilmektedir. Bu tez çalışmasında ise Bayesci ağlar, birliktelik analizi ile çıkarılan kurallar doğrultusunda oluşturulmuştur. Bayesci ağların oluşturulmasında skorlama tabanlı yapısal öğrenme algoritmaları kullanılmıştır. Tepe tırmanma ve tabu araştırma algoritmalarının mevcut veriye uygulanması sonucu aynı yapıya sahip Bayesci ağlar elde edilmiştir. Burada oluşturulan ağların yapısını öğrenme algoritmalarından ziyade istatistiksel skorlama metotlarının etkilediği gözlemlenmiştir. Elde edilen Bayesci ağlar, küçük hücreli dışı akciğer adenokarsinomunun genetik yapısında gerçekleşen mutasyonlar arası ilişkilere muhtemel farklı bakış açıları sunmaktadır. Böylece bu hastalığın altında yatan önemli dinamiklerin incelenmesinde uzmanlara ve ilaç geliştiricilerine daha geniş bir perspektif sunmak hedeflenmiştir.

Misra ve diğerleri tarafından yürütülen bir çalışmada akciğer ve beyin kanserinde sık görülen mutasyonların aralarındaki ilişkiyi incelemiştir (141). Bu çalışma sonuçlarından biri olarak bazı mutasyonların görülmelerini başka bir ya da birden fazla mutasyonun daha önceden görülmesi ile ilişkilendirilmiştir. Bayesci ağlar ve BIC skorlama metodu kullanılarak oluşturulan YDG'ler bir bakıma ilgili kanser çeşitlerinde görülen mutasyonların ortaya çıkma önkoşullarını göstermiştir.

Bu tez çalışmasında kullanılan model seçme kriterleri göz önünde bulundurulduğunda AIC, BDe ve K2'nin BIC'e göre daha anlamlı ağlar oluşturduğu söylenebilir. Oluşturulan ağ yapılarının validasyonu Kyoto Encyclopedia of Genes and

Genomes (KEGG) veri tabanında bulunan küçük hücreli dışı akciğer kanseri yolak haritasıyla örtüşen bölümler üzerimden yapılmaya çalışılmıştır (142). Benzerliklerin bulunduğu alanlar daha detaylı ve gelişmiş bir şekilde işlendiği takdir de gelecek vadetmektedir. Bu çıktıların kesin doğruluğu laboratuvar çalışmaları ve online kaynaklarla doğrulanmak üzere başka bir çalışma alanı olarak belirlenmiştir.

Ölümcül bir hastalık olarak kanser, genetik yapısındaki dinamik değişiklikler nedeniyle son derece önem arz eden güncel bir konudur. Kanserin bu değişken yapısı genetik materyalde gerçekleşen mutasyonlar ve bu mutasyonların birikimi ile ilgilidir. Kanserin yapısında bulunan çok sayıda mutasyonun karakterizasyonu ve gelişimi sürecinde erken tespiti ile ilgili yapılan çalışmalar literatürde mevcuttur (143). Bu tez çalışmasında mevcut veriler üzerinde sık görülen mutasyonların frekansları incelenmiştir. Sonuç olarak mutasyonların birbirlerine yakın frekans değerlerine sahip olmaları hastalığın birikimli bir yapı sergilediğini kanıtlar niteliktedir. Bu yapı bize kanserin büyüme tahminleri, ilaç direnci vb. gibi konularda fikir sağlayacaktır. Ayrıca bu konuda gelecek çalışmalar için bireylere göre mutasyon kümeleri oluşturarak bu kümeler içerisindeki baskın mutasyon örüntülerini belirme işlemleri yapılabilir.

Yapılan bu tez çalışmasıyla genetik verilerden önceden görülemeyen ve klasik yöntemlerle elde edilmesi oldukça zor olan sonuçlar elde edilmiştir. Fakat bu sonuçların anlamlandırılması ve gerek araştırmacı gerekse sağlık çalışanları için yararlı olabilmesi için alan uzman desteği kesinlikle gerekmektedir. Bu nedenle genetik veriler üzerinde uygulanacak olan veri madenciliği ve biyoinformatik çalışmaları disiplinlerarası bir şekilde yürütülmelidir. Bu tip çalışmaların planlanması aşamasında gerekli alan uzmanları çalışmanın alan yeterlilikleri kapsamında güvenirliği sağlamak için organize bir biçimde hareket etmelidir.

Çalışmada kullanılan genetik veriler ön işlemeye tabi tutulmuştur. Böylece boş işlemlerden ve kayıp verilerden arındırılmıştır. Veri madenciliği analizlerinin başarısını etkileyen birinci derecede önemli unsur veri kalitesidir. Bu sebeple yapılacak çalışmalarda verilerin eksik-yanlış, tekrarlı ya da kayıp verilerden mümkün olduğunca arındırılması gerekmektedir. Veritabanlarında tutulan verinin başarılı bir şekilde analizi için verilerin elde edilmesi ve veritabanına aktarılması sürecine özen gösterilmelidir.

İlerleyen ve gelişen bilim ve teknolojinin hemen her alanda kullanılması kaçınılmaz hale gelmiştir. Genetikte ise teknoloji büyük miktarlarda bilgiyi öncelikle elde etme, depolama, depolanan bilgiye hızlı erişme, analizler yapma ve karar destek sistemleri gibi insan gücü ile yapılması çok zor işlemlere çözümler sunarak araştırmacılara kolaylıklar sağlamaktır. Bilginin etkin kullanımı, en iyi teknolojilerin seçilerek elde edilen sonuçların arttırılmasında, bu bilgilerin analizi ve yorumlanmasında, ayrıca hastalıkların genetik yapısının anlaşılmasıyla beraber tanı, teşhis ve tedavi metotlarının da geliştirilmesinde yararlı olacaktır.

Bu tarz çalışmaların oldukça sınırlı olduğu ülkemizde, çalışmaların sayısını ve etkinliğini arttırmak için araştırmacılar daha disiplinlerarası alanlara teşvik edilerek yönlendirilmelidir. Ayrıca yenilikçi çalışmaların artması için araştırmacıların, hesaplamalı biyoloji (computational biology), biyoinformatik (bioinformatics), tıp bilişimi (medical informatics) vb. gibi disiplinler arası alanlara özendirilmesi ve desteklenmesi gerekmektedir.

8. KAYNAKLAR

1. Hand D, Mannila H, Smyth P (2001). Principles of Data Mining, The MIT Press, Cambridge, 546p.

2. Han J, Kamber M (2000). Data Mining Concepts and Techniques. Morgan Kaufmann Publishers, 1st Ed., San Francisco, USA.

3. Larose DT (2004). Discovering Knowledge in Data: An Introduction to Data Mining, Wiley-Interscience, New York, 222p.

4. Tan P, Steinbach M, Kumar V (2006). Introduction to Data Mining, Addison-Wesley, Boston, 769p.

5. Jaroszewicz S, Simovici DA (2004) Interestingness of frequent itemsets using Bayesian networks as background knowledge, Proceedings of the 10th ACM SIGKDD Conference on Knowledge Dicovery and Data Mining, August 20-25, 2004, New York, USA, pp.178-186.

6. Margaritis D (2003). Learning Bayesian Network Model Structure from Data. PhD thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA. CMU-CS-03-153.

7. Pearl J (1985). "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning" (UCLA Technical Report CSD-850017), Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, CA. pp. 329–334. Retrieved 2009-05-01.

8. Dawy Z, Yaacoub E, Nassar M, Abdallah R, Zeineddine HA (2011). A multiorganism based method for Bayesian gene network estimation. BioSystems, 103, 425–434.

9. Ahmad FK, Safaai DM, Abdullah S (2011). Synergy network based inference for breast cancer metastasis, Procedia Computer Science 3 (2011) 1094–1100.

10. Alterovitz G, Tuthill C, Rios I, Modelska K, Sonis S (2011). Personalized medicine for mucositis: Bayesian networks identify unique gene clusters which predict the response to gamma-D-glutamyl-L-tryptophan (SCV-07) for the

attenuation of chemoradiation-induced oral mucositis, Oral Oncology 47 951–955.

11. Doctor NJ, Strylewicz G (2010). Detecting ‘wrong blood in tube’ errors: Evaluation of a Bayesian network approach, Artificial Intelligence in Medicine, 50 75–82.

12. Chow CK, Liu CN (1968). Approximating discrete probability distributions with dependence trees, IEEE Transactions on Information Theory, IT-14 (3): 462–467.

13. Heckerman D (1995). Bayesian networks for data mining, Data Mining and Knowledge Discovery .1, pp.79-119.

14. Merlo LM, Pepper JW, Reid BJ, ve çal. ark. (2006). Cancer as an evolutionary and ecological process (Evrmsel ve ekolojik bir süreç olarak kanser). Nat Rev Cancer 6: 924-935.

15. Ruddon RW (2007). Cancer biology. Oxford University Press

16. Park Moon-Taek, and Su-Jae Lee. (2003) "Cell cycle and cancer." Journal of biochemistry and molecular biology 36.1: 60-65.

17. Fışkın K (2002). Genetik Kavramlar. Altıncı baskıdan çeviri. Öner, C. Sayfalar 635-651.

18. Aktaş S.H.: Kemoterapinin Kolon Kanseri, Meme Kanseri Ve Mide Kanserinde Vegf Düzeylerine Etkisinin İn Vivo Ve İn Vitro İncelemesi. Ankara Üniversitesi, Biyoteknoloji Enstitüsü, Yüksek Lisans Tezi, Ankara, 2010

19. Cyclacel. (2014) [online]. http://www.cyclacel.com/research_science_cell-cycle.shtml

20. Cell Cycle. (2015) [online]. http://www.nature.com/scitable/topic/genetics-5

21. Cabadak H (2008). Hücre Siklusu ve Kanser. ADÜ Tıp Fakültesi Dergisi. 9 (3): 51-61

22. Kastan M, Bartek J (2004). Cell cycle check points and cancer. NATURE. Vol 432, 316-323.

23. Howard Hughes Medical Institute. (2015) [online]. http://www.hhmi.org/biointeractive/eukaryotic-cell-cycle-and-cancer

24. Hartwell LH, Michael BK (1994). "Cell cycle control and cancer." Science 266.5192: 1821-1828.

25. Kopnin BP (2000). Targets of oncogenes and tumor supressors: Key for understanding basic mechanisms of carcinogenesis. Biochemistry. Vol 65. No 1, pp. 2-27.

26. Mutation, (2004). Herder Lexikon der Biologie

27. Akbulut H ve Akbulut KG (2005). Tıbbi Onkoloji Kitabı. Ankara Üniversitesi Tıp Fakültesi Antıp Yayınları, Editör İçli F. Sayfa 23

28. Types of Mutations. (2015) [online]. http://mumtazticloft.com/images/TypesOfMutations.png

29. Knudson AG (1971). Mutation and Cancer: Statistical Study of Retinoblastoma. Proc Natl Acad Sci U S A. 68(4): 820–823.

30. Wood RD, Mitchell M, Sgouros J, Lindahl T (2001). Human DNA Repair Genes. SCIENCE. Vol 291. p.1284-89.

31. Pierotti MA, Sozzi G, Croce CM. (2003) Oncogenes. In: Kufe DW, Pollock RE, Weichselbaum RR, et al, eds. Cancer Medicine. 6th ed. Hamilton, Ontario: BC Decker Inc;: 73–85.

32. Lander ES, LM Linton, B Birren, et al. (2001). Initial Sequencing and Analysis of the Human Genome. Nature 409(6822): 860-921

33. Altschul SF, Bogusk, MS, Gish W & Wootton JC (1994). "Issues in Searching Molecular Sequence Databases." Nature Genet. 6:119-129

34. Vilda Purutçuoğlu, Ezgi Ayyıldız (2014). Biyoinformatik Alanında İstatistik. Nobel Akademik Yayıncılık, 1. Baskı. Ankara, Türkiye

35. Stears RL, Martinsky T, and Schena M (2003). "Trends in microarray analysis." Nature medicine 9.1: 140-145.

36. Waterman, MS (1995). Introduction to computational biology: maps, sequences and genomes. CRC Press.

37. Sistem Biyoloji. (2015) [online]. http://www.ibgizmir.deu.edu.tr/tr/index.php/arastirma/temel-arastirmalar/sistem-biyolojisi

38. The Cancer Genome Atlas (2015) [online]. http://cancergenome.nih.gov/

39. NIH Launches Cancer Genome Project Washington Post Dec 14, (2005)

40. ATALAY RÇ (2002). "NEDEN BİYOİNFORMATİK?." Avrasya Dosyası, Moleküler Biyoloji ve Gen Teknolojileri Özel, Sonbahar 8.3,129-141

41. Engin YZ (2013). Karadeniz Teknik Üniversitesi Farabi Hastanesi Biyokimya Laboratuvarı Test Sonuçlarından Veri Madenciliği Yolu İle Örüntü Çıkarma. Yüksek Lisans Tezi, Karadeniz Teknik Üniversitesi, Sağlık Bilimleri Enstitüsü, Trabzon.

42. Wikipedia. (2011) [online]. http://tr.wikipedia.org/wiki/Veri

43. Bollenger AS, Smith RD (2001). Managing Organizational Knowledge as a Strategic Asset. Journal of Knowledge Management, Vol:5, No:1, s.9.

44. Bailey C, Martin C (2000). How Do Managers Use Knowledge About Knowledge Management?. Journal of Knowledge Management, Vol:4, No:3 2000, s.237.

45. Veritabanı. (2014) [online]. http://megep.meb.gov.tr/mte_program_modul/moduller_pdf/Veritaban%C4%B1%20Haz%C4%B1rlama.pdf

46. Wikipedia. (2011) [online]. http://en.wikipedia.org/wiki/Database

47. Veritabanı Yönetim Sistemi. (2015) [online]. http://web.firat.edu.tr/mbaykara/vtys.pdf

48. Akpınar H (2000). Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. İstanbul Üniversitesi İşletme Fakültesi Dergisi, C.29, S.1, s: 1-22

49. Düzgünoğlu S (2006). Veri Ambarı ve OLAP Teknolojilerinden Yararlanılarak Karar Destek Amaçlı Raporlama Aracı Gerçekleştirimi. Yüksek Lisans Tezi. Hacettepe Üniversitesi Bilgisayar Mühendisliği AD., Ankara.

50. Swift RS (2001). Accelerating Customer Relationships. Prentice Hall, New Jersey.

51. Famili F, Shen W, Weber R, Simoudis E (1997). Data Preprocessing and Intelligent Data Analysis. Intelligent Data Analysis, 1(1-4):3-23.

52. Roiger RJ, Geatz MW (2003). Data Mining A Tutorial-Based Primer. Addison Wesley, USA.

53. Doğan Ş (2007). Veri Madenciliği Kullanarak Biyokimya Verilerinden Hastalık Teşhisi. Yüksek Lisans Tezi, Fırat Üniversitesi, Fen Bilimleri Enstitüsü, Elazığ.

54. Koyuncugil, AS, and Özgülbaş N (2009). "Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları." INTERNATIONAL JOURNAL OF INFORMATICS TECHNOLOGIES 2.2

55. Özekes S (2003). "Veri madenciliği modelleri ve uygulama alanları."

56. Hand DJ (1998). Data mining: statistics and more?. The American Statistician, Vol:52 no:2, 112–118.

57. Ceran G (2006). Esnek Akıs Tipi Çizelgeleme Problemlerinin Veri Madenciligi Ve Genetik Algoritma Kullanılarak Çözülmesi. Yüksek Lisans Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.

58. Savaş S, Topaloğlu N, and Yılmaz M(2012). "Veri madenciliği ve Türkiye'deki uygulama örnekleri."

59. Ergün E (2008). Ürün Kategorileri Arasındaki Satış İlişkisinin Birliktelik Kuralları Ve Kümeleme Analizi İle Belirlenmesi Ve Perakende Sektöründe Bir Uygulama. Doktora Tezi, Afyon Kocatepe Üniversitesi, Sosyal Bilimler Enstitüsü, Afyon.

60. Rokach L, Maimon O (2008). Data Mining With Decision Trees Theory And Applications. Singapore: World Scientific.

61. Sezgin E, and Çelik Y(2013). "Veri madenciliğinde kayıp veriler için kullanılan yöntemlerin karşılaştırılması." Akademik Bilişim Konferansı, Antalya.

62. Bircan H (2004). "Lojistik regresyon analizi: Tıp verileri üzerine bir uygulama." Kocaeli Üniversitesi Sosyal Bilimler Enstitüsü Dergisi 2: 185-208.

63. Nabiyev VV (2003). Yapay Zeka. Seçkin Yayınevi, Ankara.

64. Çalışkan SK, and Soğukpınar (2008). İ. "K× KNN: K-MEANS VE K EN YAKIN KOMŞU YÖNTEMLERİ İLE AĞLARDA NÜFUZ TESPİTİ." EMO Yayınları: 120-124.

65. Cortes C, and Vapnik V (1995). "Support-vector networks." Machine learning 20.3: 273-297.

66. Foody GM and Mathur A (2004). "Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification." Remote Sensing of Environment 93.1: 107-117.

67. Agrawal R, Shafer JC (1996). Parallel Mining of Association Rules. IEEE Transactions on Knowledge and Data Engineering, Vol:8 pp:962–969.

68. Yıldız B, and Şelale H (2011). "Mining Frequent Patterns from Microarray Data." The 6th International Symposium on Health Informatics and Bioinformatics,(HIBIT 2011), İzmir, Turkey.

69. Wikipedia. (2015) [online]. http://tr.wikipedia.org/wiki/Olas%C4%B1l%C4%B1k

70. Olasılık. (2015) [online]. http://www.acikders.org.tr/file.php/103/LectureNotes/MIT14_30s09_lec03.pdf

71. Koşullu Olasılık. (2015) [online]. http://notoku.com/bilesen-marjinal-ve-kosullu-olasiliklar/

72. Maksimum olabilirlik. (2015) [online]. http://kisi.deu.edu.tr//hamdi.emec/Yaz%20Okulu/Ekonometri1/EYOY-S%C4%B1n%C4%B1rlamalar.pdf

73. Thompson GE (1972). Statistics for Decisions, 1st Edition, USA: Little, Brown and Company.

74. Kurtuluş K (1998). Pazarlama Araştırmaları, 6. Basım, İstanbul: Avcıol Basım.

75. Winkler RL (2003). Bayesian Inference and Decision, 2nd Edition, USA: Probabilistic Publishing.

76. Ekici O (2009) İstatistikte Bayesyen ve Klasik Yaklaşımın Farklılıkları, Balıkesir Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 12(21): 89-101.

77. Link WA and Barker RJ (2010). Bayesian İnference For Ecological Application, Elsevier Academical Publication, California.

78. Koller D, and Friedman N (2009). Probabilistic graphical models: principles and techniques. MIT press.

79. Scutari M (2011). Measures of Variability for Graphical Models, PhD thesis, Universita degli Studi di Padova, Dipartimento di Scienze Statistiche.

80. Geiger D, Verma TS, and Pearl J (2013). "d-separation: From theorems to algorithms." arXiv preprint arXiv:1304.1505

81. Bozdoğan H (1987). Model selection and Akaike's information criterion (AIC): the general theoryandits analytic alextensions. Psychometrika, 3, 345-370

82. Bandalos DL (1993) Factors influencing cross-validation of confirmatory factor analysis models, Multivariate Behavioral Research, 28(3), 351-374

83. Burnham KP and Anderson DR (2004). "Multimodel Inference Understanding AIC and BIC in Model Selection" Sociological Methods & Research, 33: 261-304

84. Yu J, et al. (2004). "Advances to Bayesian network inference for generating causal networks from observational biological data." Bioinformatics 20.18: 3594-3603.

85. Chen X, Anantha G, and Lin X (2008). "Improving Bayesian network structure learning with mutual information-based node ordering in the K2 algorithm." Knowledge and Data Engineering, IEEE Transactions on 20.5: 628-640.

86. Agrawal R, Imielinski T, Swami A (1993). Mining Association Rules Between Sets of Items in Large Databases. ACM SIGMOD Conference on Management of Data, Washington.

87. Webb GI (2003). Association Rules. In Nong Ye (Edt.), The Handbook Of Data Mining (pp. 26–39). New Jersey: Lawrence Erlbaum Associates,Inc.

88. Agrawal R, Srikant R (1994). Fast Algorithms for Mining Association Rules. VLDB Conference, Santiago, Chile.

89. Dunham MH (2002). Data Mining Introductory and Advanced Topics. Pearson Education Inc., Upper Saddle River, New Jersey.

90. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z, Steinbach M, Hand DJ, Steinberg D (2008). Top 10 Algorithms in Data Mining. Knowledge And Information System, Vol:14, s:1–37.

91. O'Hagan A (1986). Probability methods and measurement. Chapman and Hall, London.

92. Wade PR (2000). Bayesian methods in conservation biology, Conservation Biology, 14(5):1308-1316.

93. Mc Charty MA (2007). Bayesian Methods For Ecology, Cambridge UniversityPress, New York.

94. Kayabol K, (2008). İmge Kaynaklarının Ayrılmasında Bayesci Yaklaşımlar, Doktora Tezi, İstanbul Üniversitesi Fen Bilimleri Enstitüsü Elektrik-Elektronik Mühendisliği.

95. Güner A (2014). Bayesci Yaklaşımda Eşlenik Aileleri Önseli ile Jeffreys Önselinin Karşılaştırılması. Yüksek Lisans Tezi, Eskişehir Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, Eskişehir.

96. Saçıldı Saçaklı İ (2011). Gelişmiş ve Gelişmekte Olan Piyasalarda Hisse Senedi Getiri Volalitelerinin Bayes Yaklaşımı ile Analizi, Doktora Tezi, İstanbul

97. Savchuk VP and Tsokos CP (2011). Bayesian Theory and Methods and Applications, Atlantis Press, Paris.

98. Freund, RJ, Wilson, WJ and Sa P (2006). Regression Analysis, 2rd ed., Academic Press, Elsevier, USA.

99. Bolstad, WM (2007). Introduction To Bayesian Statistics, Hoboken, Nj: Wiley-Interscience.

100. Ben-Gal I (2007) Bayesian Networks. Encyclopedia of Statistics in Quality & Reliability. Ruggeri, F., Faltin, F., Kenett, R. (eds), Wiley & Sons, 1800p.

101. Cheng J, Bel D, Liu W (1998). Learning Bayesian Networks from Data: An Efficient Approach Based on Information Theory, Technical Report, Department of Computer Science, University of Alberta, 41p.

102. Jensen FV (2001) Bayesian Networks and Decision Graphs, Springer-Verlag, New York, 268p.

103. Boettcher SG and Dethlefsen C (2003). deal: A package for learning Bayesian networks. Journal of Statistical Software. 8(20), pp.1-40

104. Ersel D (2002). Birliktelik Analizinde Özgün Bir Birleşik İlginçlik Ölçümü. Doktora Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

105. Borgelt C and Kruse R (2002). , Graphical Models: Methods for Data Analysis and Mining, Wiley, New York, 368p.

106. Koski T and Noble JM (2009). Bayesian Networks - An Introduction, John Wiley & Sons, West Sussex, 356p.

107. Neapolitan RE (2004). Learning Bayesian Networks, Prentice Hall, Englewood Cliffs, 693p.

108. Myllymaki P, Silander T, Tirri H, Uronen P (2002). B-course: A web-based tool for Bayesian and causal data analysis. International Journal on Artificial Intelligence Tools, 11(3), pp.369-387.

109. Jensen FV, Nielsen TD (2007). Bayesian Networks and Decision Graphs, Information Science and Statistics.

110. Robinson RW (1977). "Counting unlabeled acyclic digraphs." Combinatorial mathematics V. Springer Berlin Heidelberg, 28-43.

111. Yaramakala S, Margaritis D (2005). Speculative Markov Blanket Discovery for Optimal Feature Selection, In Proceedings of the 5th IEEE International Conference on Data Mining, pp 809-812.

112. Gamez JA, Mateo JL, Puerta JM (2011). Learning Bayesian Networks by hill-Climbing: efficient methods based on progressive restriction of the neighborhood, Data Mining Knowledge, 22, 106-148.

113. Bart S, and Gomes CP (2006). "Hill-climbing Search." Encyclopedia of Cognitive Science.

114. Hill Climbing. (2015) [online].www.studyblue.com

115. Hill Climbing. (2015) [online]. http://www.inside-r.org/packages/cran/abn/docs/search.hillclimber

116. Hill Climbing. (2015) [online]. http://en.wikibooks.org/wiki/Algorithms/Hill_Climbing

117. Colin, R. Reeves, 1995, Modern Heuristic Techniques for Combinatorial Problems, London, McGraw - Hill

118. Tabu Arama Algoritması. (2015) [online]. http://www.ibrahimcayiroglu.com/Dokumanlar/IleriAlgoritmaAnalizi/IleriAlgoritmaAnalizi-9.Hafta-TabuAramaAlgoritmasi.pdf

119. Cura T (2006). "Tabu Arama Algoritması Ile Tek Makine Programlama Probleminin Çözümü", Yönetim Dergisi:İstanbul Üniversitesi İşletme Fakültesi İşletme İktisadı Enstitüsü, ss.25-34.

120. Tabu Search. (2015) [online]. http://www.cs.sandia.gov/opt/survey/ts.html

121. Tsamardinos I, Brown LE, Aliferis CF (2006). The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm, Machine Learning, 65 31-78.

122. Haughton DMA, Oud HLJ, Jansen RARG (1997). Information and other criteria in structural equation model selection, Communuciation in Statistics, 26(4), 1477-1516.

123. Bandalos DL (1993). Factors influencing cross-validation of confirmatory factor analysis models, Multivariate Behavioral Research, 28(3), 351-374

124. Burnham KP and Anderson DR (2004). "Multimodel Inference Understanding AIC and BIC in Model Selection" Sociological Methods & Research, 33: 261-304.

125. Koehler AB and Murphree ES (1988). “A Comparision of the Akaike and Schwarz Criteria for Selecting Model Order” , Applied Statisitcs, 37: 187-195

126. Buntine WL (1991). Theory refinement on Bayesian networks. In B. D’Ambrosio, P. Smets and P. Bonissone, (eds.), Proc. of the Seventh Int. Conf. Uncertainty in Artificial Intelligence, pages 52–60.

127. Heckerman DG and Chickering D (1995). Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning, 20, pages 197–243.

128. Scanagatta M, Campos CP, and Zaffalon M (2014). "Min-BDeu and Max-BDeu Scores for Learning Bayesian Networks." Probabilistic Graphical Models. Springer International Publishing, 426-441.

129. Maomi U(2012). "Learning networks determined by the ratio of prior and data." arXiv preprint arXiv:1203.3521

130. Cooper F, Herskovits E (1992). A Bayesian method for the induction of probabilistic networks from data. Machine Learning;9(4):309-47.

131. OLMUŞ H and ERBAŞ S (2012). "Bayes ağlarda kümeleme metotunu kullanarak meme kanseri tanısının modellenmesi." Turkiye Klinikleri Journal of Biostatistics 4.1: 10-19.

132. Python. (2015) [online]. http://tr.wikipedia.org/wiki/Python_(programlama_dili)

133. R Project. (2015) [online]. http://www.r-project.org/about.html

134. Fox J, and Andersen R (2005). "Using the R statistical computing environment to teach social statistics courses." Department of Sociology, McMaster University.

135. OMIM (2015) [online]. http://www.omim.org/

136. Creighton C, and Hanash S (2003). "Mining gene expression databases for association rules." Bioinformatics 19.1: 79-86.

137. Perrin, Bruno-Edouard, et al. "Gene networks inference using dynamic Bayesian networks." Bioinformatics 19.suppl 2 (2003): ii138-ii148.

138. Murphy K, and Mian S (1999). Modelling gene expression data using dynamic Bayesian networks. Vol. 104. Technical report, Computer Science Division, University of California, Berkeley, CA.

139. Yan LJ, and Cercone N (2010). "Bayesian network modeling for evolutionary genetic structures." Computers & Mathematics with Applications 59.8: 2541-2551.

140. Yu J et al. (2002). "Using Bayesian network inference algorithms to recover molecular genetic regulatory networks." International Conference on Systems Biology. Vol. 2002.

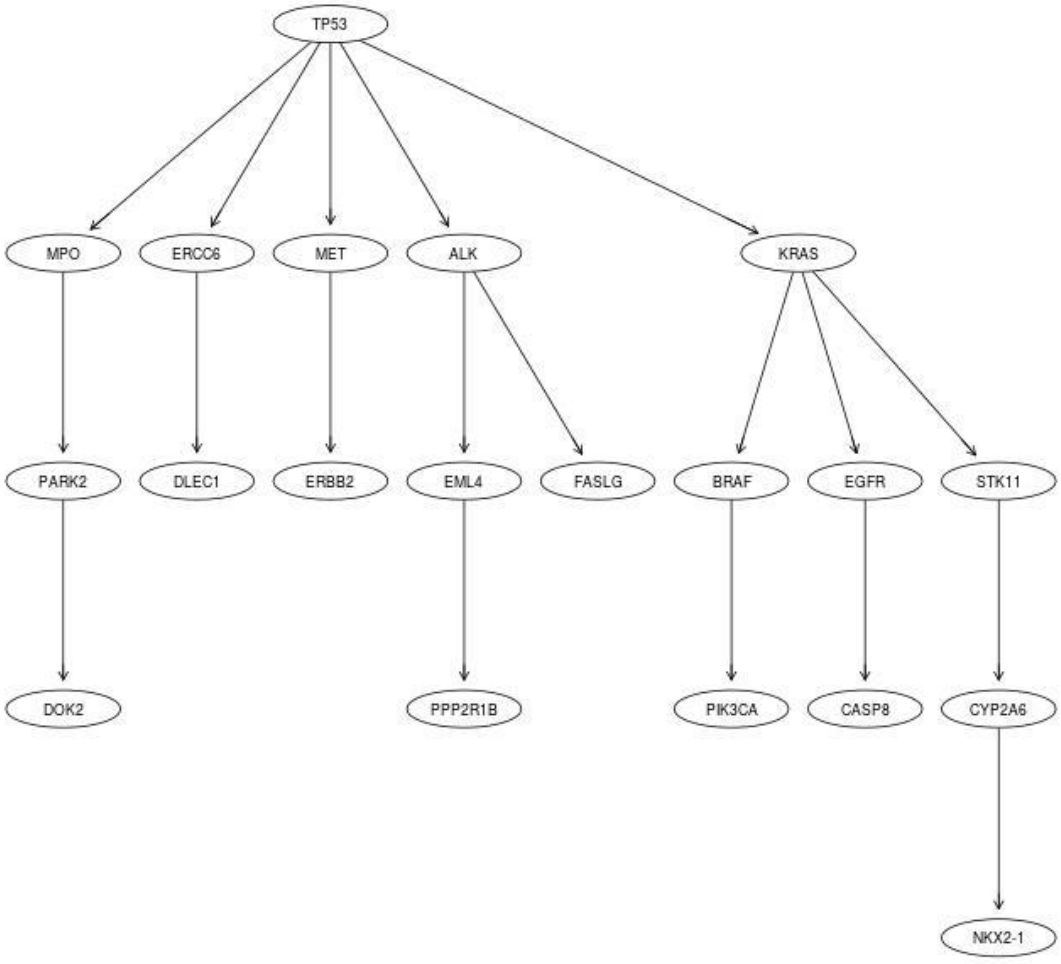
141. Misra N, Szczurek E (2012). and Vingron M. "Learning Gene Networks Underlying Somatic Mutations in Cancer." 22nd Annual Workshop on Mathematical and Statistical Aspects of Molecular Biology.

142. KEGG. (2015) [online]. http://www.genome.jp/kegg/pathway.html

143. Loeb LA, Loeb KR, and Anderson JP (2003). "Multiple mutations and cancer." Proceedings of the National Academy of Sciences 100.3: 776-781.

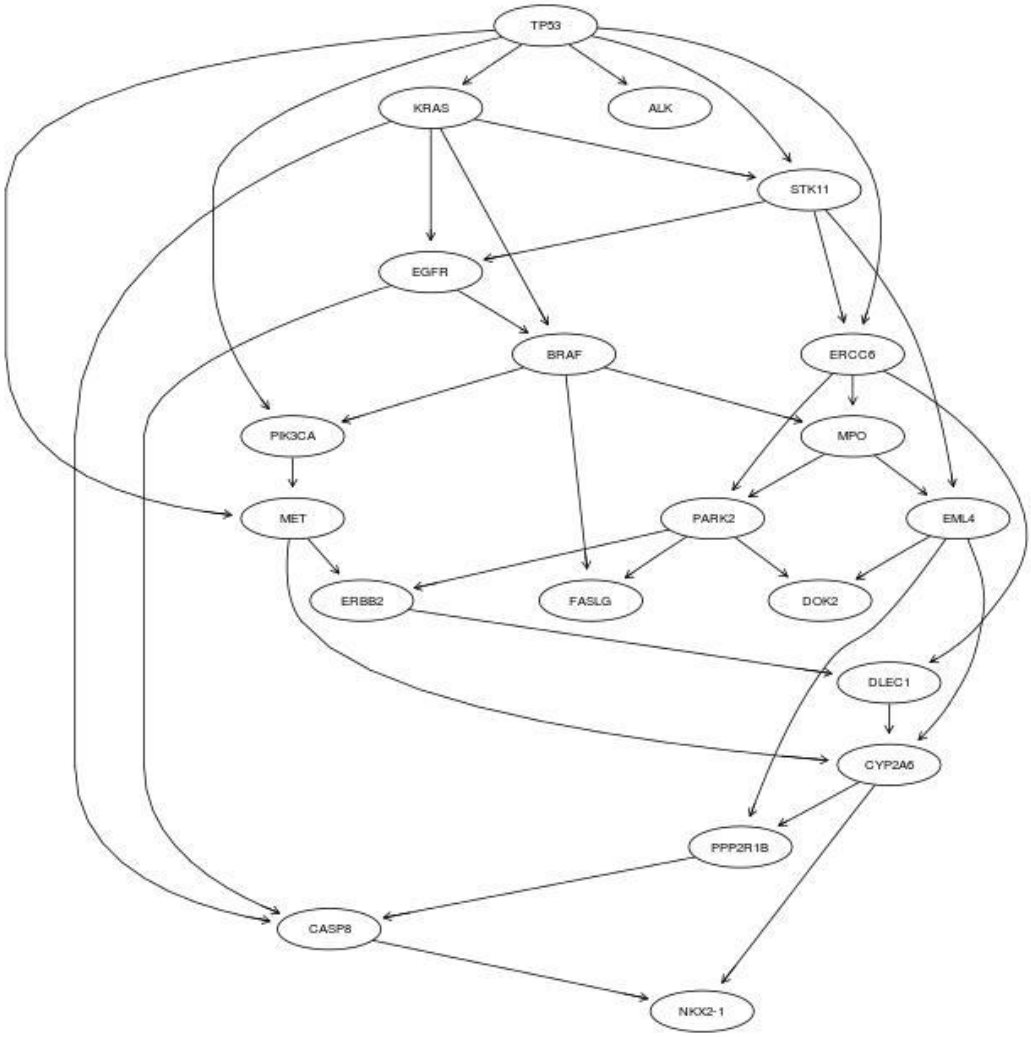
9. EKLER

Ek 1 Tabu Arama Algoritması ile Elde Edilen Sonuçlar



Tabu arama algoritması ve AIC skorlama metodu ile oluşturulan Bayesci Ağ.

Ek 1 (Devam) Tabu Arama Algoritması ile Elde Edilen Sonuçlar

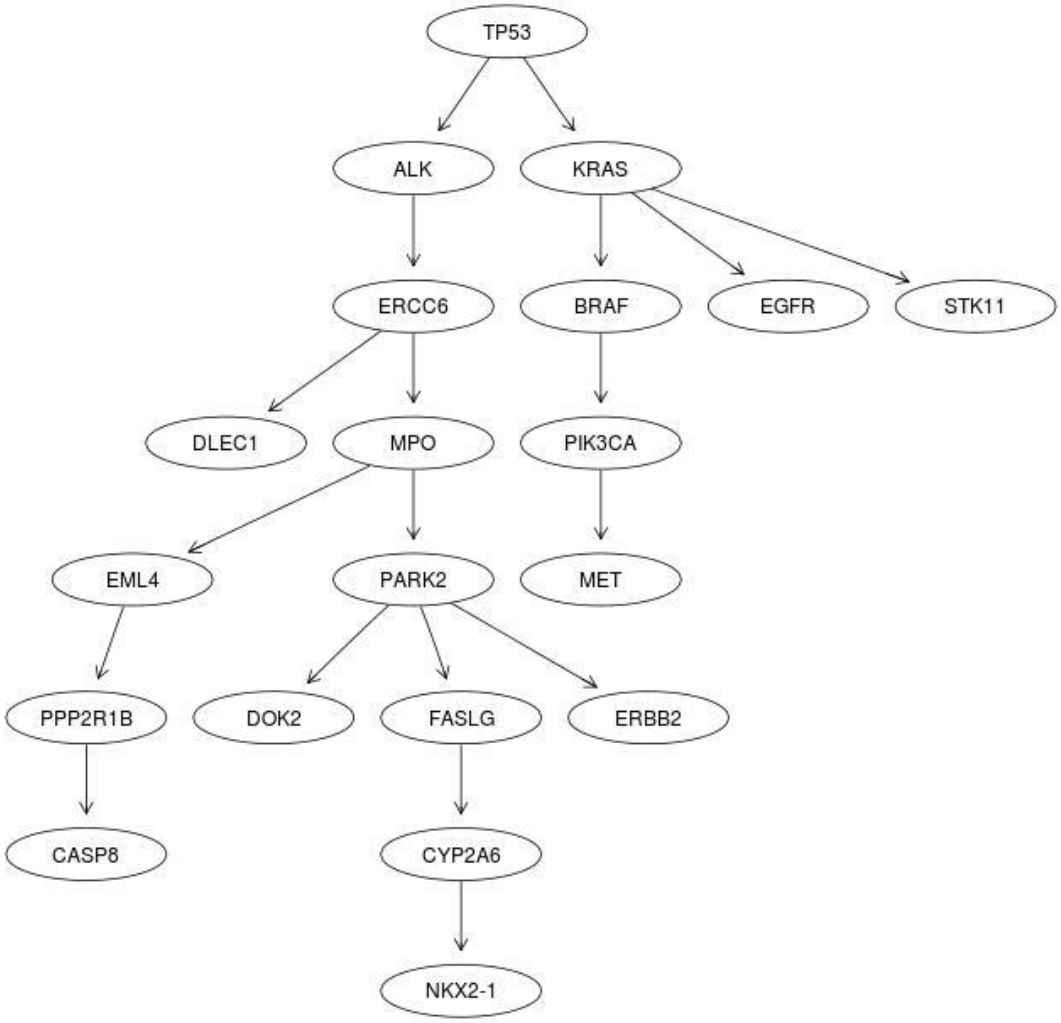


Tabu arama algoritması ve BDe skorlama metodu ile oluşturulan Bayesci Ağ.

Ek 1 (Devam) Tabu Arama Algoritması ile Elde Edilen Sonuçlar

Tabu arama algoritması ve BIC skorlama metodu ile oluşturulan Bayesci Ağ.

Ek 1 (Devam) Tabu Arama Algoritması ile Elde Edilen Sonuçlar



Tabu arama algoritması ve K2 skorlama metodu ile oluşturulan Bayesci Ağ.

10. ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Soyadı, Adı	**:** TOPRAK, Uğur
Uyruğu	**:** T.C.
Doğum tarihi ve yeri	**:** Ankara 08. 10. 1988
Medeni hali	**:** Bekar
E-Posta	**:** toprakugur@ktu.edu.tr
Yazışma adresi	**:** Karadeniz Teknik Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Merkez, Trabzon

EĞİTİMBİLGİLERİ

Derece	**Mezun Olduğu Kurumun Adı**	**Mezuniyet Yılı**
Yüksek Lisans	Karadeniz Teknik Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik ve Tıp Bilişimi Anabilim Dalı	Devam ediyor
Lisans	Karadeniz Teknik Üniversitesi Fatih Eğitim Fakültesi Bilgisayar ve Öğretim Teknolojileri Öğretmenliği Bölümü	2011
Lise	N.K.V. Beypazarı Anadolu Lisesi	2006

YABANCI DİL: İngilizce